Modèles de diffusion pour la génération d'images personnalisées et la variation guidée par référence d'effets sonores

par

Mélodie DESBOS

MÉMOIRE PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE
AVEC MÉMOIRE GÉNIE DES SYSTÈMES ET DE LA PRODUCTION
AUTOMATISÉE
M. Sc. A.

MONTRÉAL, LE "11 MAI 2026"

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CE MÉMOIRE A ÉTÉ ÉVALUÉ

PAR UN JURY COMPOSÉ DE:

Prof. Mohammadhadi Shateri, directeur de mémoire
Département de génie des systèmes, École de technologie supérieure

Prof. Eric Granger, codirecteur
Département de génie des systèmes, École de technologie supérieure

Prof. Chrisitan Desrosiers, président du jury
Département de génie logiciel et TI, École de technologie supérieure

Prof. Marco Pedersoli, membre du jury
Département de génie des systèmes, École de technologie supérieure

IL A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 13 AVRIL 2026

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

REMERCIEMENTS

Je souhaite tout d'abord exprimer ma profonde reconnaissance pour l'opportunité que j'ai eue de mener mes recherches au sein de LIVIA/ILLS à l'ETS, entourée d'esprits aussi brillants. Évoluer dans une communauté aussi bienveillante, solidaire et accueillante m'a permis de grandir, d'apprendre et de m'épanouir dans ce domaine passionnant.

J'adresse mes sincères remerciements au Professeur Shateri pour m'avoir offert l'opportunité de participer à un projet aussi exigeant que stimulant. Merci pour votre passion, votre énergie, votre encadrement, votre confiance et la liberté que vous m'avez accordée dans la conduite de mes recherches. Je remercie également chaleureusement le Professeur Granger pour sa vision, son soutien et son expérience. Merci pour votre engagement envers chacun d'entre nous à LIVIA, pour le temps que vous avez généreusement consacré malgré un emploi du temps chargé, et pour votre présence constante à travers les fuseaux horaires. À vous deux, mes directeurs, merci pour les nombreuses heures consacrées à affiner ce travail à mes côtés : en structurant ses idées, en améliorant sa rédaction et en poussant chaque figure à son plein potentiel. Merci de croire en moi et pour votre soutien face aux complexités de l'administration française.

Un merci tout particulier à la personne qui m'a accompagnée tout au long de ce parcours avec confiance, patience, bienveillance et sagesse autour de la diffusion. À mon collaborateur, merci pour toutes les nuits blanches, pour avoir porté mes idées jusqu'à ce qu'elles prennent forme, et pour ton soutien inconditionnel. Au futur Docteur Yara Bahram, merci du fond du cœur. À nos prochaines publications, et aux futures figures qui feront dire au Professeur Shateri : *« I like it too much ! »*

Au futur Docteur Oz', merci pour ta sagesse, ton accompagnement et nos nuits jazzy au LIVIA.

À ma famille, merci pour votre soutien inconditionnel et pour avoir toujours cru en moi avec autant de force. À mes frères de cœur, à mes amis, merci d'avoir été ma force et mon souffle lorsque j'en avais le plus besoin.

From passing through, to pushing through.

I find my peace in exhaustion and the harder my head meets the wall, the lighter it becomes. Satisfaction comes when consumptions ends, and yet I dance when sat and sing when spoken. I sleep tight when missions stretches towards sunrise. And, I set the bare but it is never bare.

Modèles de diffusion pour la génération d'images personnalisées et la variation guidée par référence d'effets sonores

Mélodie DESBOS

RÉSUMÉ

Les modèles de diffusion atteignent aujourd'hui des performances de pointe en génération multimodale. Néanmoins, leur adoption en pratique demeure limitée par (i) le coût élevé de l'échantillonnage itératif et (ii) la difficulté à adapter un modèle de diffusion préentraîné à de nouveaux domaines à partir de seulement quelques références. Ce mémoire étudie l'utilisation des modèles de diffusion pour l'adaptation, la génération multimodale personnalisée et l'édition dans le contexte de la génération d'images et d'effets sonores en régime de faibles données. Elle met l'accent sur l'adaptation efficace, la préservation de l'identité et la génération contrôlable de variations pour des flux de production.

Dans un premier temps, ce mémoire présente des contributions issues d'un travail collaboratif autour de l'article `Uni-DAD`. Cet article introduit un cadre d'entraînement unifié pour la distillation et l'adaptation simultanées, permettant la génération d'images en few-shot et en few-step. Ce travail confronte les pipelines classiques en deux étapes de type *adapt-then-distill* ou *distill-then-adapt*, qui impliquent des structures complexes, un risque de surapprentissage et une diversité limitée. La principale contribution porte sur la personnalisation guidée par sujet (Subject-Driven Personalization, SDP), renforcée par l'intégration d'un conditionnement textuel. Évalué sur le benchmark DreamBooth, `Uni-DAD SDP` atteint une qualité d'image compétitive par rapport aux méthodes d'adaptation de référence tout en ne nécessitant qu'une seule étape d'échantillonnage. Le résultat final est un générateur en une seule étape, capable de produire des images diversifiées et de haute qualité dans de nouveaux domaines de données, sous conditions restreintes. En définitive, cette méthode agnostique facilite le déploiement des modèles de diffusion dans des applications personnalisées en temps réel.

Dans un second temps, la continuité de ce travail contribue à la recherche appliquée sur la génération audio pour la production d'effets sonores (SFX). Plus précisément, il étudie la capacité des modèles génératifs modernes à produire des variations diversifiées à partir d'un clip de référence tout en préservant l'identité de l'événement sonore. Une analyse expérimentale évalue l'aptitude des modèles à reproduire ou éditer des caractéristiques clés du signal, telles que la structure temporelle et l'enveloppe d'énergie, sous des contraintes orientées vers la production (durée, transfert de style et indices d'alignement). Par ailleurs, cette analyse permet de structurer un état de l'art et d'établir des comparaisons équitables entre méthodes pour la variation et l'édition de SFX, afin de rapprocher les récentes avancées de la génération audio aux exigences de l'industrie. Ce travail établit également une base pour de futures recherches appliquées sur la génération efficace et contrôlable de variations dans des contextes de production.

Mots-clés: IA générative, modèles de diffusion, adaptation, distillation des connaissances, multimodalité (images et effets sonores)

Diffusion Models for Personalized Image Generation and Reference-Guided Sound-Effect Variation

Mélodie DESBOS

ABSTRACT

Diffusion models achieve state-of-the-art performance in multimodal generation. Yet, their practical adoption remains limited by (i) the high computational cost of iterative sampling and (ii) the difficulty of adapting a pretrained diffusion model to new domains from few references. This thesis investigates the use of diffusion models for adaptation, personalized multimodal generation, and editing in image and sound-effect generation under low-data regimes, with a focus on efficient adaptation, identity preservation, and controllable variation generation for production workflows.

First, this thesis presents contributions stemming from a collaborative work on the `Uni-DAD` paper. It introduces a unified training framework for simultaneous distillation and adaptation of diffusion models, enabling few-shot and few-step image generation. The work contrasts classical two-stage pipelines of *adapt-then-distill* or *distill-then-adapt*, which entail complex designs, overfitting, and limited diversity. The contribution presented here focuses on subject-driven personalization (SDP), enhanced by integrating textual conditioning. Evaluated on the DreamBooth benchmark, `Uni-DAD SDP` achieves competitive image quality compared to reference adaptation methods while requiring only a single sampling step. The end result is a few-step generator capable of adapting and producing diverse, high-quality images in novel domains under few-shot conditions. Ultimately, this checkpoint-agnostic method facilitates the deployment of diffusion models in personalized, real-time applications.

Second, this work contributes to applied research on audio generation for sound effect (SFX) production. In particular, it investigates the ability of modern generative models to produce diverse variations from a reference clip while preserving the sound event's identity. Experiments analyze the models' capacity to reproduce or edit key signal characteristics, such as temporal structure and energy curve, under production-oriented constraints (e.g., duration control, style transfer, and alignment cues). Furthermore, this analysis helps structure an overview and enables fair comparisons between existing methods for SFX variation and editing, thereby bridging recent advances in audio generation with practical industry requirements. This work also establishes a foundation for future applied research on efficient, controllable variation generation in production settings.

By bridging the efficiency of diffusion-based generation and the expressiveness of reference-guided modeling across modalities, these contributions enable user-friendly, controllable adaptation and editing of both images and SFX clips.

Keywords: Diffusion Models, Adaptation, Knowledge Distillation, Generative AI, Multimodality (Images and Sound Effects)

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

A^2SB	Audio-to-Audio Schrödinger Bridge
AC-Foley	Acoustic-Conditioned Foley
ATA	Audio-to-Audio
BBC	British Broadcasting Corporation
CFG	Classifier-Free Guidance
CLAP	Contrastive Language-Audio Pretraining
CLIP	Contrastive Language-Image Pretraining
CLIP-I	CLIP image alignment metric
CLIP-T	CLIP text alignment metric
CoT	Chain-of-Thought
CSV	Comma-Separated Values
CVPR	IEEE/CVF Conference on Computer Vision and Pattern Recognition
DAFx	International Conference on Digital Audio Effects
DDIM	Denoising Diffusion Implicit Models
DDP	Distributed Data Parallel
DDPM	Denoising Diffusion Probabilistic Models
DGM	Deep Generative Model
DiT	Diffusion Transformer
DM	Diffusion Model
DMD	Distribution Matching Distillation
DMD2	Distribution Matching Distillation v2
DINO	Self-Distillation with No Labels
ESC-50	Environmental Sound Classification dataset (50 classes)
FAD	Fréchet Audio Distance
FM	Flow Matching

GAN	Generative Adversarial Network
LDM	Latent Diffusion Model
LAION-5B	Large-scale Artificial Intelligence Open Network dataset (5B image-text pairs)
LLM	Large Language Model
LoRA	Low-Rank Adaptation
LPIPS	Learned Perceptual Image Patch Similarity
MAF	Multimodal Adaptive Fusion
MMDiT	Multimodal Diffusion Transformer
MLP	Multilayer Perceptron
MSE	Mean Squared Error
NFE	Number of Function Evaluations
ODE	Ordinary Differential Equation
PSO	Pairwise Sample Optimization
RMS	Root Mean Square
RVQ	Residual Vector Quantization
SDE	Stochastic Differential Equation
SDEdit	Stochastic Differential Editing
SDP	Subject-Driven Personalization
SDv1.5	Stable Diffusion v1.5
SDXL	Stable Diffusion XL
SFX	Sound Effects
S-MOS	Similarity Mean Opinion Score
SoTA	State of the Art
STFT	Short-Time Fourier Transform
T-Foley	Temporal Foley
TTA	Text-to-Audio
TTI	Text-to-Image

TTUR	Two-Time-Scale Update Rule
UNet	U-shaped Network architecture
Uni-DAD	Unified Distillation and Adaptation of Diffusion Models
VAE	Variational Autoencoder

LISTE DES SYMBOLES ET UNITÉS DE MESURE

$p_{\text{data}}(x)$	Unknown real-data distribution
$p_\phi(x)$	Learned model distribution parameterized by ϕ
$p(z)$	Prior distribution in latent space, typically Gaussian
ϕ	Trainable model parameters
ψ	Auxiliary trainable parameters (e.g., encoder, discriminator, or critic parameters depending on context)
x	Data sample in input space
x_t	Noisy sample at diffusion timestep t
x_0	Clean sample before noising
z	Latent representation
z_0	Clean latent representation
z_t	Noisy latent representation at timestep t
t	Diffusion or flow timestep
T	Total number of diffusion timesteps
ϵ	Gaussian noise variable
ϵ'	Independent Gaussian noise variable
ϵ_ϕ	Noise prediction network parameterized by ϕ
α_t	Timestep-dependent signal scaling coefficient
σ_t	Timestep-dependent noise scaling coefficient
σ	Noise level or standard deviation
$\mathcal{L}_{\text{LDM}}$	Latent diffusion training objective
c	Conditioning signal or prompt
c'	Instance prompt for subject-driven personalization
c^{prior}	Class-prior prompt
H_c	Multimodal conditioning embedding
e_t	Text embedding

e_v	Video embedding
e_a	Audio embedding
Y	Few-shot target set
$\lvert Y \rvert$	Number of target reference samples
$p^{\mathrm{src}}(x)$	Source-domain data distribution
$p^{\mathrm{trg}}(y)$	Target-domain data distribution
G	Student generator
θ	Parameters of the student generator
ϵ^{src}	Frozen source teacher diffusion model
ϵ^{trg}	Online target teacher diffusion model
ϵ^{fk}	Fake teacher tracking the evolving student distribution
D	Multi-head discriminator
α	Dual-domain weighting coefficient
$\mathcal{L}_{\mathrm{DB}}$	DreamBooth objective
$\mathcal{L}_{\mathrm{DM}}$	Distribution-matching objective
$\mathcal{L}_{\mathrm{DMD2}}$	Total DMD2 objective
$\mathcal{L}^{D}_{\mathrm{GAN}}$	Adversarial loss for the discriminator
$\mathcal{L}^{G}_{\mathrm{GAN}}$	Adversarial loss for the generator
λ	Generic weighting coefficient for regularization terms
λ_{GAN}	Weight of the adversarial generator loss
NFE	Number of neural function evaluations during sampling
x^{τ}	Target sample in pairwise adaptation objectives
x^{ρ}	Reference sample generated by the current distilled student
p_{ref}	Pretrained distilled reference model
β	Scaling coefficient in the PSO relative likelihood objective
$\sigma(\cdot)$	Sigmoid function
x_{ref}	Reference sample (image or audio, depending on context)

s	Source text description in audio editing
r	Reference text description in audio editing
c	Target text description or conditioning signal, depending on context
z_T^s	Noisy latent of the source audio
z_T^r	Noisy latent of the reference audio
z_T^c	Target noisy latent in the editing process
z_{t-1}^c	Intermediate denoised target latent at timestep $t-1$
$v_\phi(x_t, t, c)$	Learned velocity field in flow matching
$u_t(x_0, x_1)$	Target conditional velocity along the probability path
$\mathcal{L}_{\text{FM}}$	Flow matching objective
MAF	Multimodal Adaptive Fusion module
$\hat{\epsilon}$	Predicted noise or denoising target

INTRODUCTION

0.1 Contexte et motivations

Les avancées récentes en apprentissage profond ont conduit à des progrès remarquables en intelligence artificielle générative, permettant aux modèles de synthétiser de manière réaliste des images, des signaux audio, du texte et d'autres modalités de données complexes. Au cœur de ces avancées se trouvent les modèles génératifs profonds (DGMs), qui apprennent à approximer la distribution de probabilité sous-jacente des données du monde réel à partir de vastes ensembles d'exemples d'entraînement. Une fois entraînés, ces modèles peuvent générer de nouveaux échantillons qui demeurent statistiquement cohérents avec les données observées, tout en permettant diverses formes de conditionnement et de contrôle du processus de génération.

En plus de générer des échantillons réalistes, les DGMs modernes permettent une génération contrôlée, dans laquelle le processus de synthèse peut être guidé par des signaux de conditionnement tels que du texte, des images ou d'autres modalités.

Des méthodes populaires ont émergé pour permettre un conditionnement guidé et une génération de haute qualité : les variational auto-encoders (VAE), opérant dans un cadre encodeur–décodeur, visent à apprendre un modèle à variables latentes afin de capturer la distribution sous-jacente des données, tandis que les generative adversarial networks (GANs) entraînent de manière antagoniste un réseau discriminateur et un réseau générateur afin d'encourager ce dernier à produire des échantillons correspondant à la distribution des données réelles.

Parmi les DGMs, les diffusion models (DMs) se sont imposés comme l'une des classes de modèles génératifs les plus efficaces et les plus largement adoptées. Ils apprennent à inverser un processus de corruption directe prédéfini, transformant progressivement le bruit en échantillons provenant de la distribution cible des données à travers une séquence d'états intermédiaires (voir Fig. 1.2). Cette formulation s'est révélée particulièrement efficace pour la génération de haute

qualité et la synthèse contrôlée. De plus, les DMs sont très flexibles quant au type de données et à la représentation sur laquelle ils opèrent, incluant les espaces latents (Luo, Tan, Huang, Li & Zhao, 2023a; Sauer *et al.*, 2024a; Rombach *et al.*, 2022; Podell *et al.*, 2023; Guan *et al.*, 2024; Xing, He, Tian, Wang & Chen, 2024), les jetons (tokens) discrets (Sheffer & Adi, 2022), ainsi que les espaces d'embeddings (Girdhar *et al.*, 2023b; Wei *et al.*, 2023). Cette flexibilité rend les DMs particulièrement adaptés au conditionnement multimodal et a permis leur extension à un large éventail de modalités et de tâches, notamment la génération d'images, la génération audio, ainsi que la génération et l'édition multimodales. Bien que les DMs permettent une génération de haute fidélité et un conditionnement multimodal flexible, ils peuvent souvent être combinés à des méthodes DGM antérieures afin d'améliorer l'apprentissage des représentations et les mécanismes de guidage. En particulier, les VAE peuvent être utilisés pour encoder et décoder des images ou des signaux audio entre l'espace du signal et l'espace latent (Liu *et al.*, 2023a; Wu *et al.*, 2024), tandis qu'un entraînement adversarial avec un discriminateur de GAN peut fournir un retour discriminatif fort, complétant la génération basée sur la diffusion et renforçant le guidage lors de la synthèse (Bahram *et al.*, 2025). Par conséquent, les DMs se sont progressivement étendus de la génération unimodale vers des contextes de génération et d'édition multimodales, dans lesquels plusieurs sources d'information peuvent être intégrées afin de guider la synthèse vers un contenu, une structure ou un style souhaité. Malgré leurs performances élevées, les DMs nécessitent généralement un grand nombre d'étapes séquentielles de débruitage, ce qui entraîne un coût computationnel élevé lors de l'inférence (Rombach *et al.*, 2022).

Au-delà de la génération (in)conditionnelle, les DMs se sont également imposés comme de puissants cadres pour l'édition de contenu, où l'objectif consiste à modifier un échantillon existant tout en préservant les structures pertinentes ou les informations sémantiques. Dans les domaines de l'image et de l'audio, l'édition basée sur la diffusion peut être réalisée en corrompant partiellement un échantillon source puis en inversant le processus de diffusion vers

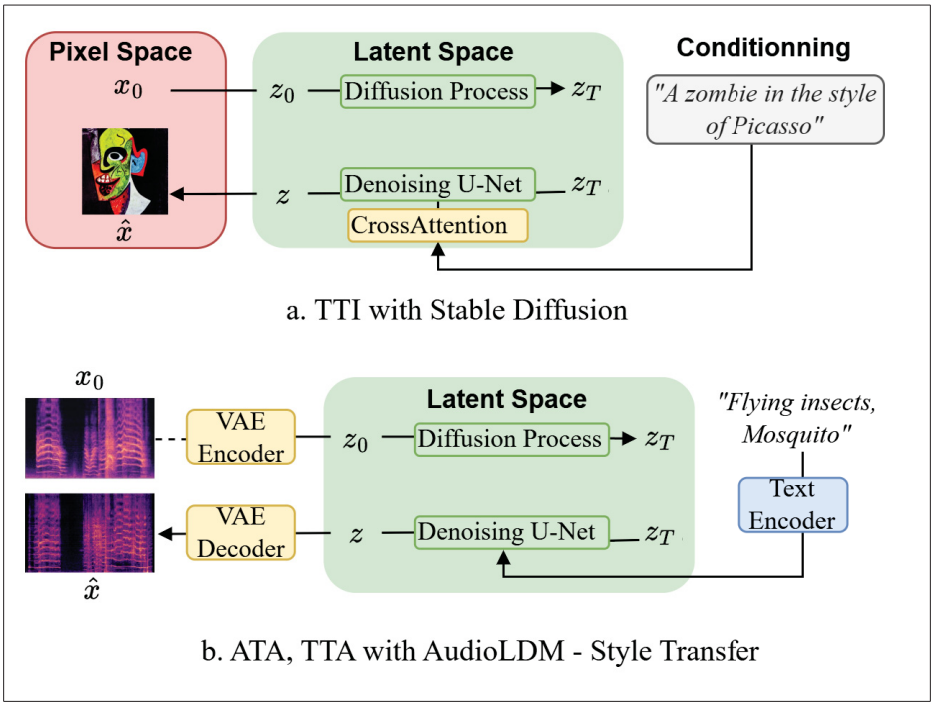

Figure 0.1 Pipelines de diffusion pour différentes applications et modalités
a) génération text-to-image (TTI) avec Stable Diffusion utilisant un réseau neuronal convolutionnel de débruitage (U-Net) et de l'attention croisée - figure adaptée de Rombach *et al.* (2022)
b) génération text-to-audio (TTA) et audio-to-audio (ATA) avec le pipeline AudioLDM utilisant un encodeur/décodeur VAE et un U-Net de débruitage pour des applications de transfert de style - figure adaptée de Liu *et al.* (2023b)

une distribution cible modifiée, comme l'ont montré SDEdit (Meng *et al.*, 2021) pour les images et AudioX (Liu *et al.*, 2025c) pour l'audio. Les DGMs ont révolutionné l'édition dans le domaine visuel en permettant un contrôle précis des modifications, notamment grâce à l'édition visuelle de bout en bout basée sur des instructions (Goodfellow *et al.*, 2014) ainsi qu'à une collaboration efficace avec l'utilisateur pour une édition de haute précision (Ling *et al.*, 2021). Dans le domaine audio, l'édition couvre un large éventail de tâches, notamment l'*inpainting*, la modification ciblée, le remplacement temporel ou spectral, le transfert de style et la génération de sons alignés sur une vidéo. Comme le décrit cette thèse, l'édition audio peut consister à réparer des segments

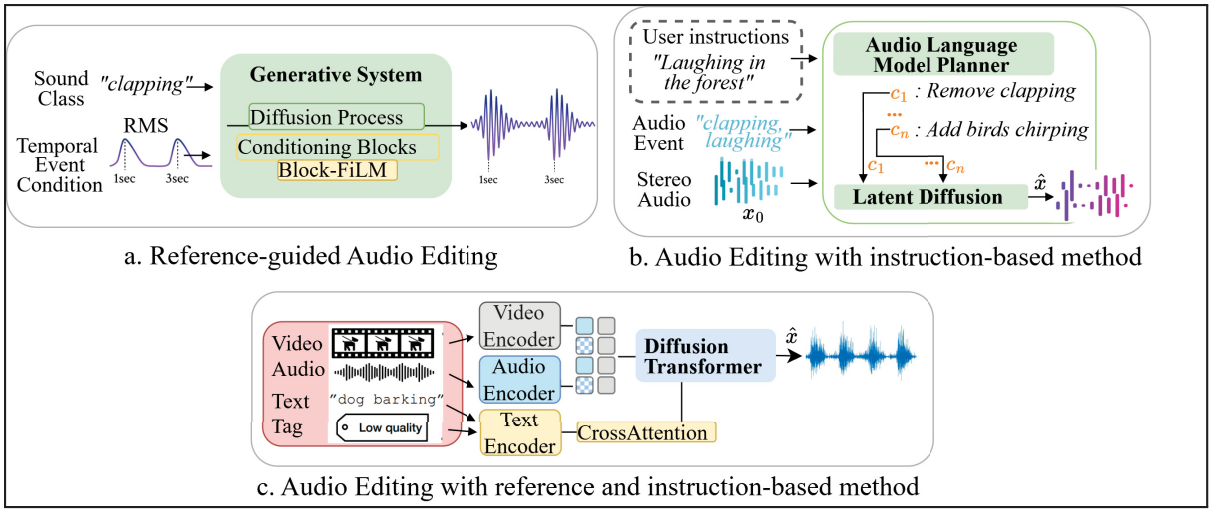

Figure 0.2 Pipelines d'édition d'audio guidés soit par des instructions explicites (p. ex., *b)* et *c)*), soit par des signaux de référence (p. ex., *a*)
a. T-Foley en *a)* adaptée de Chen *et al.* (2024b)
b. SmartDJ en *b)* adaptée de Zhang *et al.* (2025b)
c. Multi-Foley en *c* adaptée de Wu *et al.* (2024)

corrompus par *inpainting*, à modifier des régions spécifiques dans le domaine temps–fréquence, ou encore à générer plusieurs restaurations plausibles à partir d'entrées partiellement masquées.

L'une des principales forces des méthodes de diffusion modernes est que l'édition peut être guidée soit par des instructions explicites, soit par des signaux de référence. Différentes approches sont présentées à la Fig. 0.2, telles que l'édition basée sur des instructions (Zhang *et al.*, 2025b), l'édition guidée par référence (Chen *et al.*, 2024b), avec alignement d'énergie ou transfert de style dans (Liu *et al.*, 2023a) (voir Fig. 0.1), ou encore la combinaison des deux via un conditionnement multimodal (Wu *et al.*, 2024). Ces capacités motivent l'étude de la génération et de l'édition multimodales guidées par référence sous des contraintes pratiques telles que la disponibilité limitée des données, la variation contrôlée et l'efficacité de l'inférence. Relever ces défis est essentiel pour déployer des modèles basés sur la diffusion dans des applications réelles, notamment pour la génération d'images personnalisées et la génération de SFX prête pour la production.

0.2 Problématique

Malgré les progrès rapides des DGMs basés sur la diffusion dans la génération d'échantillons de haute qualité, plusieurs défis demeurent pour leur déploiement pratique dans des applications réelles. Un déploiement opérationnel requiert un modèle capable de s'adapter à un nouveau domaine ou à une nouvelle identité à partir de seulement quelques exemples, de préserver la structure globale tout en modifiant certains attributs ciblés, de générer des sorties à la fois diverses et fiables, de fonctionner avec une latence et un coût computationnel raisonnables, et de rester robuste face à des conditions d'entrée imparfaites dans des contextes réels. Pourtant, la plupart des pipelines DGMs existantes ne traitent efficacement qu'un sous-ensemble de ces exigences, plutôt que de les aborder conjointement dans un cadre unifié.

En particulier, l'efficacité computationnelle et l'adaptabilité à de nouveaux domaines sont souvent abordées indépendamment dans les approches existantes. Les méthodes de distillation appliquées aux modèles de diffusion visent à accélérer l'échantillonnage en entraînant des générateurs en peu d'étapes, tandis que les méthodes d'adaptation cherchent à transférer des modèles pré-entraînés vers de nouveaux domaines ou concepts. Cependant, ces deux objectifs sont rarement traités conjointement. En conséquence, de nombreuses pipelines existantes demeurent coûteuses sur le plan computationnel ou difficiles à adapter efficacement lorsque seul un petit nombre d'exemples cibles est disponible. En pratique, la distillation et l'adaptation sont généralement appliquées de manière séquentielle. Un modèle peut d'abord être distillé puis adapté à un nouveau domaine, ce qui peut dégrader la qualité de génération en raison de la capacité réduite et du manifold génératif restreint du modèle distillé. À l'inverse, adapter un modèle de diffusion de grande taille avant la distillation peut préserver la qualité, mais introduit un coût computationnel supplémentaire et peut réduire la capacité de généralisation au-delà du domaine adapté.

De plus, la plupart des stratégies d'adaptation reposent sur des quantités importantes de données cibles, ce qui limite leur applicabilité dans des scénarios où seuls quelques exemples

de référence sont disponibles. Ces limitations sont particulièrement problématiques dans des applications pratiques, notamment dans des contextes industriels où les pipelines prêtes pour la production doivent à la fois être efficaces et fonctionner dans des conditions de données limitées. Dans de nombreuses applications réelles, les utilisateurs ne recherchent pas une génération inconditionnelle à partir de requêtes génériques, mais plutôt une génération et une édition contrôlées, adaptées à un sujet, un domaine ou un signal de référence spécifique. Cela nécessite que les modèles de diffusion préservent les informations liées à l'identité, fonctionnent efficacement dans des régimes de données limitées et permettent une inférence efficace compatible avec des flux de production.

Ces défis apparaissent clairement dans deux contextes complémentaires. Le premier concerne la personnalisation guidée par sujet dans la génération d'images, où un modèle de diffusion pré-entraîné doit s'adapter à un nouveau concept visuel à partir de seulement quelques images de référence tout en maintenant la qualité de génération et l'efficacité de l'échantillonnage. Le second concerne la génération de *sound effects* (SFX) guidée par référence, où plusieurs variations plausibles doivent être générées à partir d'un extrait référence tout en préservant l'identité sémantique et en permettant un contrôle explicite sur la sortie.

0.3 Objectifs

L'objectif de cette thèse est d'étudier comment les DMs peuvent prendre en charge des tâches génératives efficaces et contrôlées dans des contextes de données restreintes, où seul un nombre limité d'exemples de référence est disponible. Plus précisément, ce travail examine comment des cadres basés sur la diffusion peuvent permettre la conception de pipelines pratiques de génération et d'édition combinant efficacité, contrôlabilité et préservation de l'identité. Cette thèse étudie deux cadres problématiques complémentaires. Le premier concerne la *subject-driven personalization* (SDP) pour la génération d'images, où un DM pré-entraîné doit s'adapter à un nouveau concept visuel à partir de seulement quelques exemples référence, tout en

maintenant la qualité de génération et l'efficacité de l'échantillonnage. Le second concerne la génération et l'édition pour la production de variations de SFX à partir d'un extrait référence, où plusieurs variations plausibles doivent être produites à partir d'un clip audio de référence tout en préservant l'identité sémantique et les caractéristiques structurelles de la référence donnée, sous des contraintes orientées production.

Bien que ces deux contextes diffèrent par la modalité et l'application, ils partagent un défi commun : adapter des modèles puissants basés sur la diffusion à des scénarios réels contraints, où l'efficacité et la contrôlabilité sont simultanément requis.

En conséquence, les objectifs principaux de cette thèse sont les suivants :

- Étudier des stratégies efficaces d'adaptation et de distillation pour les DMs dans des régimes de données limitées, tout en préservant la qualité et la diversité de génération.
- Étudier la SDP pour la génération d'images sous des contraintes de *few-shots* et en une seule étape.
- Étudier la génération et l'édition de variations de SFX guidées par référence sous des contraintes orientées vers la production.
- Examiner les compromis entre efficacité, contrôlabilité, préservation de l'identité et diversité des sorties dans ces deux contextes applicatifs.

0.4 Contributions

This thesis makes two main contributions to the study of efficient and controllable multimodal generation with DMs, focusing on personalized image generation and reference-guided sound-effect synthesis.

- **Chapitre 2 –** ***Dual-domain distribution matching distillation*** **(DMD) conditionnée pour la SDP :**
  - **Contribution associée à une publication CVPR 2026**[1] **avec une application de personnalisation efficace et préservant l'identité** —Un cadre de distillation DMD conditionnelle est introduit pour la SDP. L'approche aligne les fonctions de score conditionnelles tout en séparant les prompts en deux rôles distincts : un *instance prompt* c' (jeton (token) rare + nom de classe) conditionne les composantes apprises (l'étudiant G, le *fake teacher* ϵ^{fk}, et le *target teacher* ϵ^{trg}), tandis qu'un *class-prior prompt* c^{prior} (nom de classe uniquement) conditionne le *source teacher* figé ϵ^{src}. Cette conception permet de préserver la diversité au niveau de la classe tout en permettant une spécialisation vers l'identité personnalisée du sujet. De plus, une contrainte de réalisme *adversarial* conditionnée par le texte renforce le réalisme spécifique à la cible sous contrôle des prompts grâce au conditionnement textuel d'un discriminateur multi-têtes.
  - **Protocole de benchmark reproductible pour une comparaison équitable** —Le protocole complet de benchmark DreamBooth (Ruiz *et al.*, 2023) est reproduit avec des modèles de prompts standardisés et des *seed* aléatoires fixes, générant 100 échantillons par instance (25 prompts $\times$ 4 *seed*). Cela garantit une évaluation cohérente et comparable à la fois pour les instances d'objets et les sujets.
  - **Analyse au niveau des instances de la qualité, de l'identité et de la diversité** —Au-delà de l'agrégation des métriques, des évaluations détaillées par instance sont rapportées à l'aide des métriques de similarité DINO, CLIP-I et CLIP-T. Ces résultats sont complétés par des analyses de diversité et de sensibilité spécifiques à la SDP (intra- et inter-LPIPS, ainsi que des ablations sur le poids dual-domain α), explicitant le compromis entre préservation de l'identité et diversité générative dans la personnalisation *few-shots*.

[1] Y. Bahram, M. Desbos, M. Shateri, and E. Granger, "Uni-DAD : Unified Distillation and Adaptation of Diffusion Models for Few-step Few-shot Image Generation," proceedings of CVPR, 2026. https://arxiv.org/abs/2511.18281

- **Chapitre 3 – Génération de variations de SFX prête pour la production soumit à DAFx 2026**[2] **:**
  - **Sélection de baselines guidée par les exigences des contextes de production** —Une stratégie structurée de sélection des baselines associe les modèles candidats à une matrice d'exigences orientée vers la production, incluant la fidélité, le réalisme, la préservation de l'identité, la diversité, l'alignement temporel, le contrôle de l'énergie, la modification ciblée, la robustesse et l'efficacité.
  - **Benchmark d'évaluation unifié et protocole expérimental pour une comparaison équitable** —Un benchmark diagnostique complet est mis en place pour la génération de variations de SFX guidée par référence. L'évaluation combine des critères objectifs et subjectifs alignés sur les exigences de production. Les expériences sont réalisées à l'aide d'un protocole commun de fine-tuning et d'inférence sur la *soundbank few-shot* ESC-50, avec une inférence conditionnée par référence standardisée utilisant un extrait audio de référence et son étiquette de classe. Cela permet des comparaisons équitables entre différentes architectures génératives et mécanismes de conditionnement.
  - **Analyse du compromis diversité–fidélité et évaluation des capacités d'édition** —Une analyse détaillée caractérise le compromis entre la préservation de l'identité et la diversité des variations générées. La diversité est quantifiée à l'aide de distances d'embeddings entre chaque variantes générées, tandis que la fidélité est évaluée à l'aide de FAD, de scores d'alignement à la référence et d'évaluations perceptuelles humaines. Des analyses complémentaires spécifiques aux méthodes mettent également en évidence les capacités d'édition telles que le transfert de style, l'*inpainting*, l'alignement temporel et la modification ciblée, permettant d'identifier les pipelines les plus adaptés à la génération de variations de SFX prêtes pour la production.

[2] International Conference on Digital Audio Effects (DAFx)

0.5 Organisation de la thèse

Cette thèse est organisée en trois chapitres, comme suit. Le Chapitre 1 présente les notions de base nécessaires à l'ensemble de ces travaux, incluant les modèles génératifs profonds, la personnalisation basée sur la diffusion pour la génération d'images, ainsi que la génération et l'édition audio guidées par référence. Le chapitre se conclut par une synthèse des principales lacunes de la recherche qui motivent les contributions de cette thèse.

Le Chapitre 2 présente la première contribution, correspondant à un travail conjoint accepté à CVPR 2026 [3] sur la SDP. Il introduit un pipeline unifié qui traite conjointement la distillation et l'adaptation des diffusion models afin de permettre la personnalisation *few-shots* et la génération d'images en une seule étape.

Le Chapitre 3 présente la seconde contribution de cette thèse, axée sur la génération de variations de SFX prête pour la production. Ce chapitre introduit la formulation du problème, la méthodologie, le cadre expérimental, l'évaluation quantitative ainsi qu'une analyse critique de la génération de variations de SFX guidée par référence sous des contraintes orientées vers la production.

Enfin, la thèse se conclut par une synthèse des principaux résultats et par la présentation de pistes de recherche futures vers des pipelines génératives efficaces et contrôlées, avec une complexité d'entraînement réduite, une robustesse accrue et une adaptation stable pour la génération d'images, la génération de variations de SFX et la génération de scènes audio stéréo pour des applications industrielles.

---

[3] Y. Bahram, M. Desbos, M. Shateri, and E. Granger, "Uni-DAD : Unified Distillation and Adaptation of Diffusion Models for Few-step Few-shot Image Generation," proceedings of CVPR, 2026. https://arxiv.org/abs/2511.18281

CHAPITRE 1

REVUE DE LITTÉRATURE

1.1 Modèles génératifs profonds

Modèles génératifs profonds —Les *Modèles génératifs profonds* (Deep Generative Models, DGMs) sont conçus pour générer de nouveaux échantillons qui demeurent cohérents avec la distribution sous-jacente des données réelles, tout en permettant une génération contrôlée grâce à un conditionnement structuré et interprétable. Plus formellement, les DGMs visent à apprendre une approximation d'une distribution de données inconnue $p_{\text{data}}(x)$ à partir d'un ensemble fini d'observations. Comme la forme explicite de $p_{\text{data}}(x)$ n'est pas disponible, il est impossible d'échantillonner directement à partir de cette distribution. À la place, un réseau neuronal profond est utilisé pour paramétrer une distribution de modèle $p_\phi(x)$, où ϕ désigne les paramètres entraînables du réseau. L'entraînement consiste alors à optimiser ϕ de manière à minimiser la divergence entre la distribution du modèle $p_\phi(x)$ et la distribution réelle des données $p_{\text{data}}(x)$, de sorte que $p_\phi(x) \approx p_{\text{data}}(x)$. Une fois appris, ce modèle génératif peut être utilisé pour produire de nouveaux échantillons statistiquement cohérents avec les données observées.

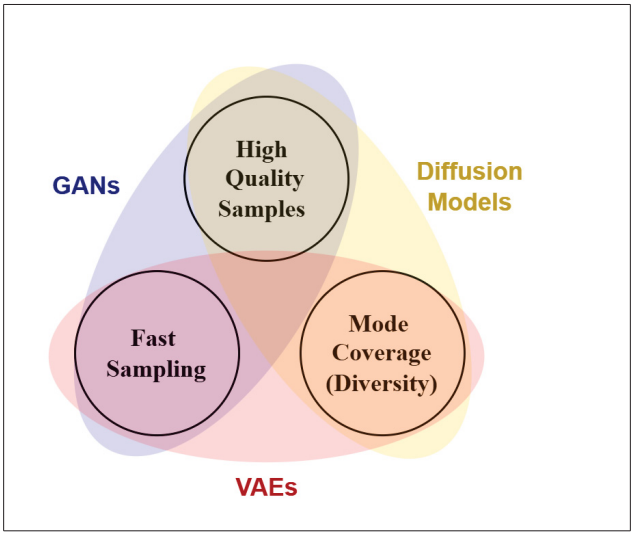

Figure 1.1 Comparaison des forces des DGM à travers les GAN, les VAE et les Diffusion Models (DM)

Generative Adversarial Networks (GAN) mettent en œuvre ce principe à travers un modèle génératif implicite défini par un générateur G_ϕ, qui projette une variable latente $z \sim p(z)$ vers un échantillon synthétique $x = G_\phi(z)$. Un discriminateur D_ψ est entraîné conjointement afin de distinguer les échantillons réels des échantillons générés, ce qui conduit à l'objectif *adversarial*

$$\min_{\phi} \max_{\psi} \; \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D_\psi(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D_\psi(G_\phi(z)))]. \tag{1.1}$$

Dans cette formulation, le générateur apprend une distribution implicite $p_\phi(x)$ qui se rapproche de $p_{\text{data}}(x)$. Les GANs peuvent produire des échantillons nets et réalistes, mais leur entraînement est souvent instable et sujet au mode collapse, ce qui limite la diversité et l'adaptation robuste (Karras, Aila, Laine & Lehtinen, 2018), voir Figure 1.1.

Des DGMs populaires tels que les **variational auto-encoders (VAE)** modélisent la distribution des données à l'aide d'une formulation de variables latentes, où un encodeur $q_\psi(z|x)$ projette une entrée vers une représentation latente et un décodeur définit la vraisemblance conditionnelle $p_\phi(x|z)$. Le modèle est entraîné en maximisant une borne inférieure variationnelle de la log-vraisemblance des données :

$$\max_{\phi,\psi} \; \mathbb{E}_{x \sim p_{\text{data}}(x)} \Big[\mathbb{E}_{z \sim q_\psi(z|x)} \log p_\phi(x|z) - KL\big(q_\psi(z|x) \| p(z)\big) \Big]. \tag{1.2}$$

Cet objectif combine un terme de reconstruction avec un terme de régularisation qui contraint le posterior approximé à se rapprocher d'une distribution a priori, généralement une gaussienne standardisée (Kingma & Welling, 2013). Les VAEs fournissent un espace latent structuré et continu qui facilite l'interpolation et la génération contrôlée. Comparés aux GANs, ils sont généralement plus stables à entraîner et moins sujets au *mode collapse*. Toutefois, les échantillons générés sont souvent sur-lissés en raison de l'approximation variationnelle et de l'expressivité limitée des échantillons latents et des décodeurs, voir Figure 1.1.

1.1.1 Modèles de diffusion

Parmi les DGMs, les modèles de diffusions (DMs) se sont imposés comme l'une des familles de modèles génératifs les plus efficaces et les plus largement adoptées pour la génération de données de haute fidélité. Ces modèles apprennent à inverser un processus de corruption directe prédéfini qui transforme progressivement les données en bruit, permettant ainsi de générer des échantillons en débruitant progressivement une initialisation aléatoire à travers une séquence d'états intermédiaires. Cette formulation s'est révélée particulièrement efficace pour la génération de haute qualité et la synthèse contrôlée sur de multiples modalités de données.

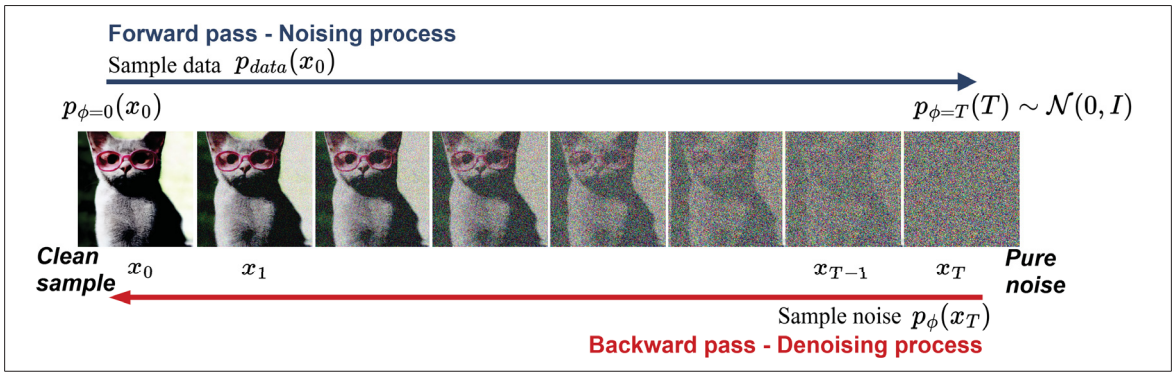

Figure 1.2 L'inversion d'une SDE en temps continu conduit à un modèle génératif basé sur les scores. Ce processus inverse nécessite d'estimer la fonction de score $\nabla_x \log p_{\phi,t}(x)$ à chaque pas de temps
Adaptée de Song *et al.* (2020d)

Un DM comporte deux étapes : un processus direct qui corrompt progressivement un échantillon propre $x_0 \sim p_{\text{data}}$ en y ajoutant du bruit, et un processus inverse qui reconstruit un échantillon propre à partir d'une variable fortement corrompue, généralement initialisée à partir d'une distribution a priori simple telle que $\mathcal{N}(0, I)$ (Song *et al.*, 2020d), voir Figure 1.2. Au-delà de leurs fortes performances génératives, les DM sont également flexibles quant à la représentation sur laquelle ils opèrent. En pratique, ils peuvent être appliqués à des espaces latents (Luo *et al.*, 2023a; Sauer *et al.*, 2024a; Rombach *et al.*, 2022; Podell *et al.*, 2023; Guan *et al.*, 2024; Xing *et al.*, 2024), à des jetons (*tokens*) discrets (Sheffer & Adi, 2022), ou à des espaces d'*embeddings* multimodaux (Girdhar *et al.*, 2023b; Wei *et al.*, 2023). Cette flexibilité a permis leur extension à

un large éventail de modalités et de tâches, notamment la génération d'images, la génération audio ainsi que la génération et l'édition multimodales.

Principales perspectives sur les DMs —Les DMs peuvent être compris à travers trois perspectives théoriques complémentaires : une perspective variationnelle, une perspective basée sur les scores et une perspective basée sur les flots. Dans tous les cas, l'idée centrale consiste à relier la distribution des données p_{data} à une distribution p_ϕ au moyen d'une séquence de distributions intermédiaires, et à apprendre la dynamique qui transforme progressivement le bruit en données. Bien que ces formulations diffèrent dans leur dérivation, elles décrivent la même évolution distributionnelle sous-jacente et peuvent être interprétées dans un cadre unifié d'équations différentielles.

Du point de vue *variationnel*, les *denoising diffusion probabilistic models* (DDPM) (Ho, Jain & Abbeel, 2020a) définissent un processus de diffusion gaussien directe fixe ainsi qu'un processus inverse de débruitage appris. Le processus direct est une chaîne de Markov fixe qui produit séquentiellement une série de variables latentes $x_1, \ldots, x_T$ en ajoutant progressivement du bruit gaussien à chaque pas de temps $t \in [1, T]$ jusqu'à ce que la distribution des données se rapproche d'une distribution gaussienne standard $\mathcal{N}(0, I)$. Formellement, le processus direct est défini comme

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\right), \qquad q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q(x_t \mid x_{t-1}), \tag{1.3}$$

ce qui induit le *kernel* de perturbation

$$q(x_t \mid x_0) = \mathcal{N}\left(x_t; \sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\, I\right), \qquad x_t = \sqrt{\bar{\alpha}_t}\, x_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \tag{1.4}$$

où $\epsilon \sim \mathcal{N}(0, I)$, $\alpha_t = 1 - \beta_t$, et $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. Lorsque t augmente, ce processus conduit progressivement les échantillons vers la distribution $\mathcal{N}(0, I)$. Le processus inverse est paramétré par un noyau conditionnel appris $p_\phi(x_{t-1} \mid x_t)$, et l'échantillonnage s'effectue récursivement

à partir de $x_T \sim \mathcal{N}(0, I)$ en tirant des échantillons à partir de ce noyau inverse. Dans la paramétrisation courante basée sur la prédiction de ϵ, la mise à jour DDPM peut s'écrire

$$x_{t-1} \leftarrow \mu_\phi(x_t, t) + \sigma_t \, \epsilon_t, \qquad \epsilon_t \sim \mathcal{N}(0, I), \tag{1.5}$$

avec

$$\mu_\phi(x_t, t) = \frac{1}{\sqrt{\alpha_t}} \left(x_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}} \, \epsilon_\phi(x_t, t) \right). \tag{1.6}$$

Des travaux ultérieurs ont introduit les *Denoising Diffusion Implicit Models* (DDIM) (Song *et al.*, 2020b), qui remplacent le processus inverse markovien par une formulation non markovienne permettant des trajectoires d'échantillonnage déterministes. Cette modification permet d'accélérer la génération tout en conservant le modèle de diffusion appris. Cette formulation DDIM est utilisée dans le pipeline de diffusion `Subject Driven Personalization (SDP) Uni-DAD` présenté au Chapitre 2.

Du point de vue *score-based*, la diffusion peut être formulée en temps continu à l'aide d'une équation différentielle stochastique (SDE) (Song *et al.*, 2020d).

$$dx(t) = f(x(t), t) \, dt + g(t) \, dw(t), \tag{1.7}$$

où f désigne le terme du *drift*, g le coefficient de diffusion, et $w(t)$ un processus de Wiener. La quantité clé dans cette formulation est la fonction de score (Song *et al.*, 2020d)

$$s(x, t) = \nabla_x \log p_t(x), \tag{1.8}$$

qui représente le gradient de la densité logarithmique de la distribution de données perturbées au temps t. Comme illustré à la Figure 1.3, une fois le processus de *forward* défini, la dynamique

inverse est déterminée par cette fonction de score, ce qui conduit à la SDE en temps inverse

$$d\bar{x}(t) = \left[f(\bar{x}(t), t) - g^2(t) \nabla_x \log p_t(\bar{x}(t)) \right] dt + g(t)\, d\bar{w}(t), \tag{1.9}$$

initialisée à partir de $\bar{x}(T) \sim p_T$.

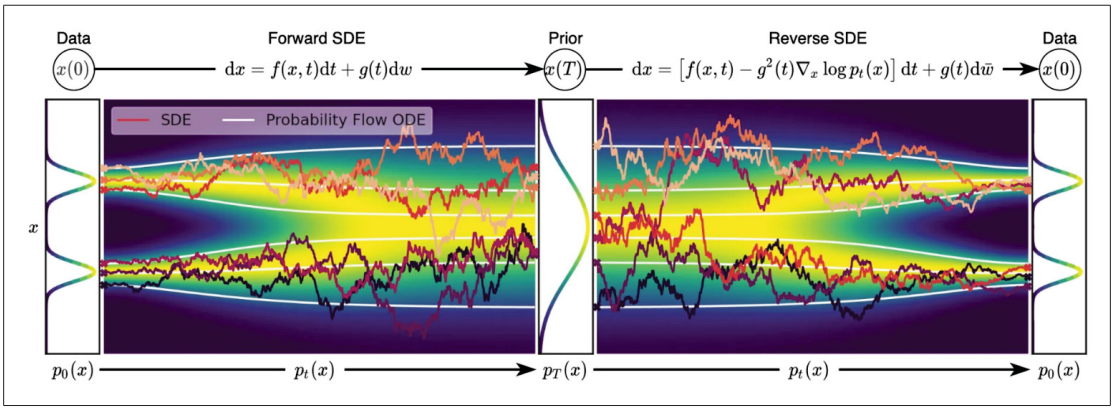

Figure 1.3 Différence des dynamiques de mouvement dans les champs de vecteurs définis par des ODE et des SDE
Tirée de Song *et al.* (2020d)

Du point de vue *flow-based*, la génération est formulée comme un transport de probabilité déterministe sous un champ de vitesse appris, voir Figure 1.3. Afin de maintenir une convention cohérente dans cette section, nous adoptons la paramétrisation temporelle JiT (Li & He, 2025)

$$z_t = tx + (1-t)\epsilon, \qquad x \sim p_{\text{data}},\ \epsilon \sim p_{\text{noise}},\ t \in [0,1], \tag{1.10}$$

où $t = 0$ correspond au bruit et $t = 1$ correspond aux données. Sous cette interpolation linéaire, la vitesse instantanée est

$$v = \frac{dz_t}{dt} = x - \epsilon. \tag{1.11}$$

Flow Matching apprend un champ de vitesse neuronal $v_\theta(z_t, t)$ en minimisant l'objectif de régression suivant

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{t,x,\epsilon} \left[\| v_\theta(z_t, t) - v \|_2^2 \right], \tag{1.12}$$

et l'échantillonnage est réalisé en résolvant l'équation différentielle ordinaire (ODE)

$$\frac{dz_t}{dt} = v_\theta(z_t, t), \tag{1.13}$$

en partant de $z_0 \sim p_{\text{noise}}$ et en intégrant jusqu'à $t = 1$. Des formulations récentes basées sur les ODE distinguent davantage l'espace de prédiction de l'espace de perte. En particulier, JiT ((Li & He, 2025)) prédit directement un échantillon débruité $x_\theta(z_t, t)$ et le convertit en champ de vitesse selon

$$v_\theta(z_t, t) = \frac{x_\theta(z_t, t) - z_t}{1 - t}, \tag{1.14}$$

tout en optimisant une perte de régression basée sur la vitesse (v-loss).

Comme les DMs bénéficient déjà de formulations efficaces d'échantillonnage et de débruitage, telles que DDIM et les méthodes de diffusion basées sur les flots, leur efficacité peut être d'avantage améliorée grâce à des méthodes de distillation.

1.1.2 Génération efficace des modèles de diffusion

Cette section passe en revue les travaux antérieurs visant à améliorer l'efficacité de l'inférence des DMs, en mettant particulièrement l'accent sur les méthodes qui réduisent le nombre d'étapes de débruitage nécessaires pour la génération.

Diffusion Distillation —Le processus itératif de débruitage des DMs nécessite généralement un grand nombre de mises à jour séquentielles, ce qui entraîne un coût computationnel élevé lors de l'inférence.

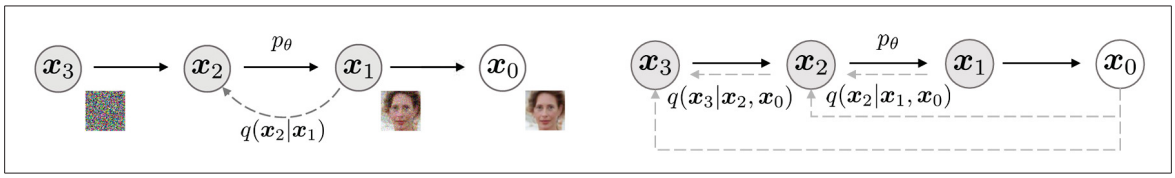

Figure 1.4 Modèles graphiques pour la diffusion avec le solveur DDPM (gauche) et le modèle d'inférence non markovien DDIM (droite)
Tirée de Song *et al.* (2020b)

Une première ligne de travaux aborde cette limitation à l'aide de solveurs améliorés qui accélèrent l'échantillonnage sans nécessiter de ré-entraînement du modèle. Des méthodes telles que les *Denoising Diffusion Implicit Models* (DDIM) (Song, Meng & Ermon, 2020a) et *DPM-Solver* (Lu *et al.*, 2022) approximent la dynamique inverse de diffusion à l'aide de schémas d'intégration numérique plus efficaces, illustrés à la Figure 1.4, ce qui permet de réduire la longueur de la trajectoire de débruitage tout en préservant la qualité de génération. En pratique, ces méthodes nécessitent généralement environ NFE ~ 10 évaluations de fonctions neuronales.

Une ligne de recherche complémentaire se concentre sur la distillation des modèles de diffusion, où un générateur "étudiant" est entraîné pour reproduire le comportement d'un modèle de diffusion "enseignant" pré-entraîné en utilisant un nombre nettement plus faible d'étapes de débruitage. Dans ce cadre, l'"étudiant" approxime directement le processus de débruitage multi-étapes de l'"enseignant", réduisant souvent l'inférence à seulement $1 \leq \text{NFE} \leq 4$ étapes.

La distillation progressive (Salimans & Ho, 2022), par exemple, réduit itérativement de moitié le nombre d'étapes d'échantillonnage en entraînant l'"étudiant" à reproduire le comportement de l'"enseignant" entre des pas de diffusion adjacents. Des approches plus récentes réduisent encore davantage le processus d'échantillonnage jusqu'à une seule étape. Les *consistency models* apprennent un mappage direct entre des entrées bruitées et des échantillons propres tout en imposant une cohérence entre différents niveaux de bruit (Song, Dhariwal, Chen & Sutskever, 2023; Luo *et al.*, 2023a). D'autres approches combinent la *score distillation* avec un entraînement *adversarial* afin d'améliorer la qualité des échantillons dans des générateurs en peu d'étapes. Parmi les exemples notables figurent *istribution Matching Distillation* (DMD) (Yin *et al.*, 2024c; Yin *et al.*, 2024a), *Adversarial Diffusion Distillation* (ADD) (Sauer, Lorenz, Blattmann & Rombach, 2024b; Sauer *et al.*, 2024a), et *Flash Diffusion* (Chadebec, Tasar, Benaroche & Aubin, 2025), qui entraînent des modèles "étudiants" capables de produire des échantillons de haute qualité avec seulement quelques étapes de débruitage.

Parmi ces méthodes de distillation progressive, DMD (Yin *et al.*, 2024c) et sa suite DMD2 (Yin *et al.*, 2024a) entraînent le modèle étudiant à correspondre à celui de l'enseignant *au*

niveau de la distribution, ce qui confère au générateur distillé davantage de flexibilité lors de l'échantillonnage. Pour permettre cette distillation, le générateur étudiant est entraîné à l'aide de la forme suivante du gradient du *distribution-matching* :

$$\nabla_\phi \mathcal{L}_{\text{DM}} \;\propto\; \mathbb{E}_{z,c,t,\epsilon}\left[\Big(s_{\text{fake}}(x_t,t,c) - s_{\text{real}}(x_t,t,c)\Big)\frac{\partial x_t}{\partial \phi}\right], \tag{1.15}$$

où $x_0 = G_\phi(z,c)$ est l'échantillon généré par l'étudiant, x_t sa version bruitée au pas de diffusion t, s_{real} est le score induit par le modèle de diffusion enseignant figé, et s_{fake} est un critique auxiliaire entraîné pour estimer le score de la distribution actuelle de l'étudiant. Contrairement à la distillation par appariement de trajectoire, cet objectif encourage l'étudiant à correspondre à l'enseignant au niveau de la distribution plutôt que le long d'un chemin de débruitage fixe. DMD2 supprime le module de régression utilisé dans la DMD et stabilise cet objectif à l'aide d'une règle de mise à jour à deux échelles temporelles, dans laquelle le critique auxiliaire est mis à jour plus fréquemment que le générateur afin de mieux suivre l'évolution de la distribution de l'étudiant.

Afin d'améliorer davantage la fidélité des échantillons, DMD2 complète de *distribution matching* par une supervision adversariale sur des images réelles. Le discriminateur est entraîné avec

$$\mathcal{L}^D_{\text{GAN}} = -\mathbb{E}_{x\sim p_{\text{data}}}\big[\log D_\psi(x_t,t,c)\big] - \mathbb{E}_{z,c}\big[\log\big(1 - D_\psi(x_t,t,c)\big)\big], \tag{1.16}$$

tandis que le générateur minimise la perte adversariale correspondante

$$\mathcal{L}^G_{\text{GAN}} = -\mathbb{E}_{z,c}\big[\log D_\psi(x_t,t,c)\big]. \tag{1.17}$$

L'objectif global d'entraînement peut alors s'écrire

$$\mathcal{L}_{\text{DMD2}} = \mathcal{L}_{\text{DM}} + \lambda_{\text{GAN}}\mathcal{L}^G_{\text{GAN}}, \tag{1.18}$$

où λ_{GAN} contrôle la contribution de la supervision adversariale. Ce terme supplémentaire expose l'étudiant non seulement au guidage de diffusion dérivé de l'enseignant, mais également au manifold des données réelles, ce qui améliore la netteté visuelle et le réalisme des échantillons. Dans sa version multi-étapes, DMD2 réduit également l'écart entre entraînement et inférence en simulant, durant l'entraînement, des états intermédiaires issus directement du modèle étudiant plutôt que de dépendre uniquement d'images réelles bruitées. Comme illustré à la Figure 1.5, ces modifications permettent aux générateurs en peu d'étapes d'atteindre une fidélité visuelle plus élevée tout en conservant les gains d'efficacité apportés par la distillation.

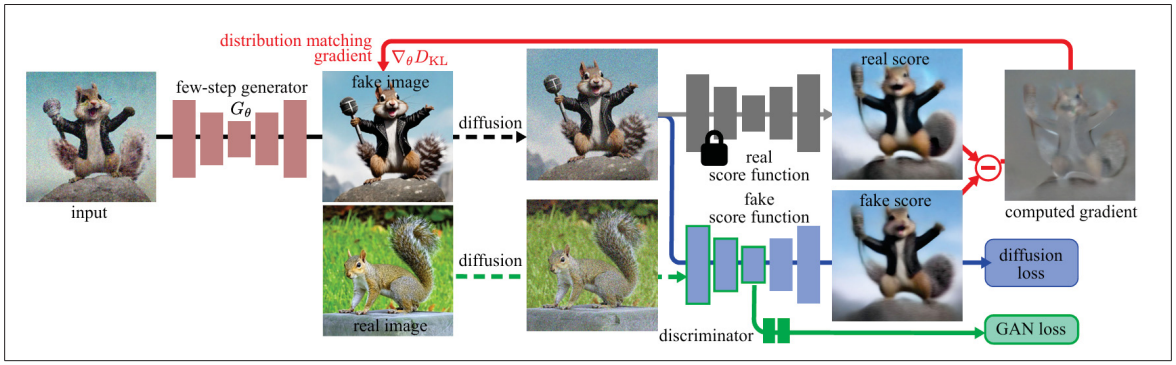

Figure 1.5 Pipeline d'entraînement de DMD2
Tirée de Yin *et al.* (2024a)

Malgré ces avancées, les DMs distillés restent généralement contraints par le manifold appris ultérieurement par le modèle enseignant, ce qui peut limiter leur flexibilité en présence de décalages de domaine. De plus, de nombreux cadres de distillation adversariale reposent sur de grands ensembles de données d'entraînement afin de maintenir la qualité de génération, ce qui peut ne pas être disponible dans des scénarios tels que la SDP où seuls quelques échantillons de référence sont fournis.

1.2 Génération d'images personnalisées avec les modèles de diffusion

Cette section passe en revue les travaux antérieurs sur la génération d'images personnalisées à l'aide des DMs. Elle examine en particulier les méthodes permettant d'adapter des DMs

préentraînés à de nouveaux concepts visuels dans des régimes de données *few-shots*. Elle analyse également les approches existantes visant à améliorer l'efficacité des pipelines de personnalisation, notamment les méthodes qui combinent adaptation et distillation afin de réduire le coût d'inférence tout en préservant la qualité de génération.

1.2.1 Adaptation des modèles de diffusion

L'adaptation désigne le processus consistant à mettre à jour un modèle pré-entraîné sur un large domaine source afin qu'il puisse représenter un domaine cible apparenté mais de plus petite taille. Bien que le *fine-tuning* soit une stratégie courante pour adapter des modèles génératifs lorsqu'un nombre relativement important d'échantillons cibles est disponible (p. ex., taille de la cible $n \sim 1000$) (Hu *et al.*, 2021), il peut facilement conduire au sur-apprentissage et à une dégradation de la diversité dans les régimes *few-shots* où seul un petit nombre d'échantillons est fourni ($n \leq 10$). Ce défi a motivé le développement de méthodes d'adaptation spécifiquement conçues pour les contextes *few-shots*, qui visent à intégrer de nouveaux concepts tout en préservant la diversité et la capacité de généralisation du modèle source.

Un benchmark représentatif dans ce cadre est DreamBooth (Ruiz *et al.*, 2023). Comme illustré à la Figure 1.6, DreamBooth réalise le *fine-tuning* d'un DM *text-to-image* (TTI) à partir d'environ ~3–5 images d'un sujet, où chaque sujet est associée à un prompt contenant un identifiant unique et le nom de classe correspondant (p. ex., "a [V] dog"). En parallèle, une *class-specififc prior preservation loss* exploite le prompt pré-appris de la classe du sujet et encourage des générations diverses à partir du seul prompt de classe (p. ex., "a dog"). L'objectif résultant combine la *subject-driven loss* avec un terme de régularisation sur les images de classe :

$$\mathcal{L}_{\text{DB}} = \mathbb{E}\left[\left\|\epsilon - \epsilon_\phi(x_t, c, t)\right\|_2^2\right] + \lambda\, \mathbb{E}\left[\left\|\epsilon' - \epsilon_\phi(x_t^{\text{pr}}, c_{\text{pr}}, t)\right\|_2^2\right], \tag{1.19}$$

où le premier terme associe l'identifiant rare au sujet cible, tandis que le second préserve la distribution au niveau de la classe afin de réduire le sur-apprentissage et la dérive sémantique, tout en maintenant la diversité des générations.

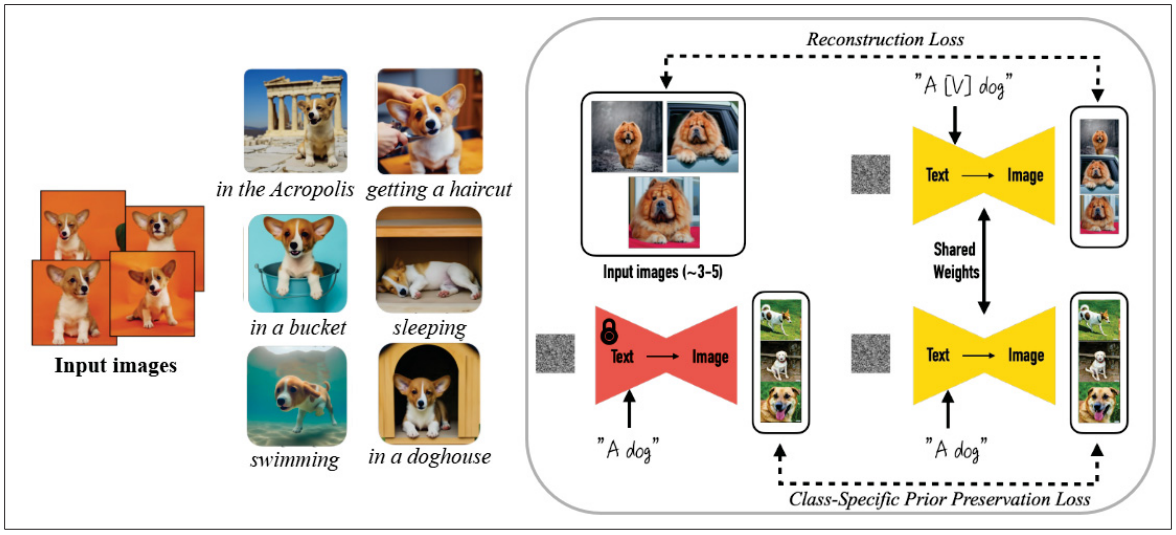

Figure 1.6 Au moment de l'inférence, avec seulement quelques images (généralement 3 à 5) d'un sujet - DreamBooth peut générer une multitude de variations du sujet en s'appuyant sur un prompt textuel - À droite, le pipeline de fine-tuning d'un DM TTI utilisant une *reconstruction loss* et une *class-specififc prior preservation loss*
Tirée de Ruiz *et al.* (2023)

1.2.2 Personnalisation efficace

Afin de développer des pipelines de personnalisation efficaces, plusieurs travaux cherchent à combiner l'adaptation de domaine avec un échantillonnage accéléré des DMs. Dans ce cadre, l'objectif n'est pas nécessairement de réduire le nombre de paramètres du modèle, mais plutôt d'obtenir un générateur capable de s'adapter à un domaine cible tout en nécessitant moins d'étapes de débruitage lors de l'inférence. Les approches existantes peuvent être globalement classées en trois stratégies : *Adapt-then-Distill*, *Distill-then-Adapt* et *Distill-and-Adapt* (Yao, Huang, Wang, Dong & Wei, 2021). La discussion suivante examine ces stratégies dans le contexte des DMs.

- Distill-then-Adapt : Dans cette stratégie, un DM pré-entraîné est d'abord distillé en un modèle étudiant en *few-stepq*, puis adapté au domaine cible. Cette approche est attrayante du point de vue de l'efficacité, puisque l'étape coûteuse de distillation peut être réalisée une seule fois avant l'adaptation en amont. Toutefois, des travaux antérieurs suggèrent que les modèles étudiants distillés possèdent souvent une capacité d'adaptation limitée. En particulier, un *fine-tuning* naïf de l'étudiant avec la perte de diffusion originale peut annuler les bénéfices de la distillation, produisant des générations floues ou pauvres en détails (Miao *et al.*, 2024).

La méthode PSO (Miao *et al.*, 2024), par exemple, entraîne l'étudiant à l'aide d'un objectif de vraisemblance relative, ce qui atténue ce problème pour des tâches de transfert de style léger dans des régimes de données plus larges ($n \sim 1000$). Plus précisément, PSO introduit une image cible x^τ provenant du domaine cible et une image de référence x^ρ échantillonnée à partir du modèle étudiant distillé courant, toutes deux conditionnées par le même prompt c, et optimise une *pairwise likelihood margin* de la forme

$$\mathcal{L}_{\text{PSO}} = -\mathbb{E}_{(x^\tau, x^\rho, c)} \left[\log \sigma \left(\beta \log \frac{p_\phi(x^\tau \mid c)}{p_{\text{ref}}(x^\tau \mid c)} - \beta \log \frac{p_\phi(x^\rho \mid c)}{p_{\text{ref}}(x^\rho \mid c)} \right) \right], \tag{1.20}$$

où p_ϕ désigne l'étudiant adapté, p_{ref} le modèle distillé pré-entraîné, $\sigma(\cdot)$ la fonction sigmoïde, et β un facteur d'échelle. De cette manière, PSO augmente la vraisemblance relative des échantillons cibles par rapport aux échantillons de référence générés par l'étudiant, ce qui permet de préserver le comportement *few-steps* hérité de la distillation tout en adaptant l'étudiant vers la distribution cible.

Plusieurs travaux visent également à produire des modèles distillés compatibles avec des méthodes d'adaptation efficaces telles que LoRA (Luo *et al.*, 2023b; Yin *et al.*, 2024a; Chadebec *et al.*, 2025). Néanmoins, l'adaptation basée sur LoRA reste souvent moins performante que le *fine-tuning* complet lorsque les décalages de distribution sont plus importants ou lorsque les données sont extrêmement limitées.

- **Adapt-then-Distill :** Une stratégie alternative consiste d'abord à adapter le DM enseignant au domaine cible, puis à distiller le modèle adapté en un générateur *few-steps*. Cette approche peut atténuer les générations sur-lissées parfois observées dans les pipelines *Distill-then-Adapt*. De plus, les objectifs adversariaux utilisés lors de la distillation peuvent aider à réduire les fuites du domaine source et les incohérences lors de l'ajustement d'un enseignant *fine-tuné* (Wang, Lin, Liu, Chen & Xu, 2024c). Toutefois, la distillation réalisée sur des données en *few-shots* est très sujette au sur-apprentissage. En outre, le modèle étudiant *few-steps* résultant n'a plus accès directement à l'ensemble des connaissances du modèle source initial et hérite plutôt des limitations de l'enseignant adapté, y compris d'éventuelles erreurs d'adaptation.

- **Distill and Adapt :** Une troisième stratégie vise à effectuer la distillation et l'adaptation conjointement au sein d'une même phase d'entraînement. Cependant, ce cadre reste encore largement sous-exploré pour les DMs. Le pipeline Codi (Mei *et al.*, 2024), par exemple, adapte conjointement un enseignant inconditionnel à des tâches conditionnées par l'image telles que l'*inpainting* et la super-résolution tout en le distillant en un étudiant *few-steps*. Son objectif consiste toutefois à fournir un contrôle basé sur l'image à l'intérieur du manifold de l'enseignant plutôt qu'une adaptation *few-shots* vers des domaines cibles situés en dehors de ce manifold.

Une ligne de travaux connexe étudie également la distillation du *classifier-free guidance* (Ho & Salimans, 2022). La *plug-and-play guidance distillation* apprend un guidage modulaire, qui par la suite peuvent être attachée à des DMs adaptés (Hsiao *et al.*, 2024), tandis que DogFit (Bahram, Shateri & Granger, 2026) intègre la *guidance distillation* dans un cadre de transfert d'apprentissage. Bien que ces approches réduisent de moitié les NFE en supprimant le coût à deux étapes du guidage, elles ne réduisent pas le *nombre d'étapes de débruitage* au même sens que la *diffusion distillation few-steps*. Plus largement, les travaux existants n'ont pas encore abordé l'adaptation et la distillation en une seule étape pour des domaines cibles *few-shots* situés en dehors du manifold du domaine source.

1.2.3 Personnalisation guidée par sujet

La génération d'images guidée par sujet vise à synthétiser des images personnalisées d'un sujet spécifique dans de nouveaux contextes tout en préservant son apparence distinctive et son identité. Plusieurs méthodes basées sur l'adaptation ont été proposées pour cette tâche. Textual Inversion (Gal *et al.*, 2022) apprend un *embedding* de jeton (tokens) dédié représentant le sujet, tandis que DreamBooth (Ruiz *et al.*, 2023) effectue le *fine-tuning* de l'ensemble de l'architecture U-Net du DM afin de capturer les caractéristiques spécifiques au sujet. Custom Diffusion (Kumari, Zhang, Zhang, Shechtman & Zhu, 2023) adapte quant à lui uniquement les couches de cross-attention, permettant une personnalisation efficace et la prise en charge de plusieurs sujets au sein d'un même modèle.

Malgré leur efficacité, ces approches restent coûteuses sur le plan computationnel, car elles nécessitent généralement un *fine-tuning* itératif et peuvent souffrir de sur-apprentissage lorsque seulement quelques images de référence sont disponibles. Afin de réduire le coût d'entraînement de la personnalisation, plusieurs stratégies d'adaptation efficaces en paramètres ont été proposées. Par exemple, SVDiff (Han *et al.*, 2023) optimise les valeurs singulières des matrices de poids afin de produire des checkpoints personnalisés compacts, tandis que LoRA (Hu *et al.*, 2021) introduit des mises à jour sur un sous-ensemble de paramètres du modèle, permettant un fine-tuning efficace en paramètres.

Par ailleurs, des méthodes sans entraînement ont récemment été explorées. ELITE (Wei *et al.*, 2023), par exemple, introduit des *mappings* d'*embeddings* globaux et locaux permettant une personnalisation en domaine ouvert sans nécessiter de *fine-tuning* explicite du modèle.

Alors que la majorité des travaux antérieurs se sont concentrés sur la réduction du coût d'entraînement de l'adaptation, relativement peu d'attention a été accordée à l'amélioration de l'*efficacité d'inférence* de la génération guidée par sujet, laquelle reste contrainte par le processus d'échantillonnage lent et multi-étapes inhérent aux DMs.

1.3 Génération et édition audio guidées par référence

1.3.1 Traitement audio pour les DGMs

Dans les DGMs et les DMs appliqués à l'audio, le choix de la représentation dépend de la structure du pipeline sous-jacent et peut varier d'un signal audio brut (*raw waveform*) à des représentations temps–fréquence telles que les *mel-spectrograms*, ou encore à des espaces latents plus compacts. Une compréhension de base du traitement du signal audio est donc nécessaire, à la fois pour préserver les caractéristiques importantes du signal et pour satisfaire les contraintes perceptuelles et liées à la production. Par ailleurs, le choix de la représentation joue un rôle central dans la conception de pipelines de génération efficaces et contrôlés, puisqu'il influence la complexité du modèle, la qualité de reconstruction et la flexibilité d'édition. Ceci est particulièrement important dans le Chapitre 3, où les méthodes comparées reposent sur différentes représentations audio, conduisant à des compromis distincts en termes de fidélité, de contrôlabilité et d'utilisabilité pratique.

Représentation audio

Les signaux audio peuvent être représentés sous plusieurs formes, chacune offrant des compromis différents en termes de résolution temporelle, de résolution fréquentielle, de pertinence perceptuelle et de complexité de modélisation. Une représentation fondamentale est la *raw waveform*, où l'amplitude du signal est exprimée en fonction du temps. Les représentations basées sur la waveform préservent l'intégralité de l'information du signal et sont directement utilisées par plusieurs modèles génératifs, notamment les approches autorégressives et basées sur la diffusion (van den Oord, Dieleman, Zen *et al.*, 2016; Mehri, Kumar, Gulrajani *et al.*, 2017; Kong, Ping, Huang *et al.*, 2021; Hai, Xu, Zhang *et al.*, 2024). Dans de nombreux cas, les waveforms sont encodées dans un espace d'*embedding* latent appris afin de faciliter une génération plus compacte et plus efficace (van den Oord, Vinyals & Kavukcuoglu, 2017; Zeghidour, Teboul, de Chaumont Quitry & Tagliasacchi, 2021; Défossez *et al.*, 2022).

Une autre approche consiste à transformer les signaux audio en représentations temps–fréquence à l'aide de la STFT, produisant des spectrogrammes qui décrivent l'évolution du contenu fréquentiel au cours du temps. Les spectrogrammes sont des représentations complexes contenant à la fois l'amplitude et la phase ; cependant, la plupart des méthodes d'apprentissage pour la génération audio opèrent principalement sur la composante d'amplitude, tandis que la phase est soit reconstruite séparément, soit approximée lors de la synthèse (Wang, Skerry-Ryan, Stanton *et al.*, 2017; Shen *et al.*, 2018). Les spectrogrammes peuvent être traités comme des représentations bidimensionnelles, ce qui les rend compatibles avec des architectures inspirées de la vision telles que les réseaux convolutionnels et les DMs.

Une variante largement utilisée est le *mel-spectrogram*, qui applique une échelle fréquentielle motivée par la perception humaine et fournit une représentation compacte et perceptuellement pertinente pour des tâches telles que le *text-to-audio* (TTA) et la génération musicale (Ristori, Bindini & Frasconi, 2025). Des approches récentes basées sur la diffusion latente et sur des transformers multimodaux montrent que ces représentations perceptuellement motivées restent centrales pour une synthèse audio de haute qualité et contrôlée (Wang, Wang, Deng *et al.*, 2025b).

Traitement du signal audio

L'audio peut être modélisé comme un signal de pression en temps continu $x(t)$, dont la représentation numérique est obtenue en échantillonnant $x(t)$ à une fréquence f_s Hz avec une période d'échantillonnage $T_s = 1/f_s$. Le signal discret résultant est

$$x[n] = x(nT_s), \qquad n \in \mathbb{Z}, \tag{1.21}$$

et est généralement appelé *waveform* audio, où n représente les instants de temps discrets. Dans cette représentation, le signal est décrit directement dans le domaine temporel par les variations de son amplitude au cours du temps. Bien que la waveform contienne toute l'information du signal, son contenu fréquentiel reste implicite et doit être révélé par une analyse spectrale.

La transformée de Fourier fournit une analyse dans le domaine fréquentiel en décomposant un signal en composantes sinusoïdales de différentes fréquences. Toutefois, pour les signaux audio non stationnaires, une transformée de Fourier globale (FT) est souvent insuffisante, car elle n'indique pas à quel moment une composante fréquentielle donnée apparaît. Afin de préserver l'information conjointe temps–fréquence, le traitement audio s'appuie sur la STFT, qui applique une transformée de Fourier sur de courts segments fenêtrés du signal, voir Figure 1.7. Pour un signal discret $x[n]$, la STFT peut s'écrire

$$X(m,k) = \sum_{n=-\infty}^{\infty} x[n]\, w[n - mR]\, e^{-j\frac{2\pi}{N}kn}, \tag{1.22}$$

où $w[\cdot]$ est une fenêtre d'analyse de longueur N, R est le pas (*hop size*) entre deux trames successives, m indexe les trames temporelles et k les indices fréquentiels. Le spectrogramme correspondant est généralement défini comme la magnitude au carré

$$S(m,k) = |X(m,k)|^2, \tag{1.23}$$

ou alternativement à l'aide d'une représentation de magnitude en échelle logarithmique.

Cependant, la segmentation du signal en segments finis (trames) introduit des discontinuités aux frontières des trames et peut provoquer un phénomène de *spectral leakage*. Afin de réduire cet effet, des fenêtres pondérées sont généralement utilisées pour lisser la transition entre les trames et atténuer les lobes secondaires, au prix d'un élargissement du lobe principal et donc d'une réduction de la résolution fréquentielle. Cela reflète le compromis classique temps–fréquence de l'analyse spectrale à court terme, où une meilleure localisation temporelle se fait au détriment de la résolution fréquentielle. La fenêtre de Hamming est couramment utilisée dans l'analyse audio et vocale car elle offre un compromis pratique entre atténuation des lobes secondaires et résolution spectrale :

$$w[n] = \begin{cases} 0.54 - 0.46\cos\left(\dfrac{2\pi n}{N-1}\right), & 0 \le n < N, \\ 0, & \text{otherwise.} \end{cases} \tag{1.24}$$

De nombreux pipelines de génération audio n'opèrent pas directement sur des spectrogrammes en fréquence linéaire, mais utilisent plutôt des représentations perceptuellement motivées telles que les *mel-spectrograms*. L'échelle mel compresse les hautes fréquences et attribue une résolution plus fine aux basses fréquences, reflétant l'observation selon laquelle la perception de la hauteur est approximativement linéaire aux basses fréquences et logarithmique aux fréquences plus élevées. Une approximation courante est

$$m(f) = 2595 \log_{10}\left(1 + \frac{f}{700}\right), \tag{1.25}$$

où f est exprimé en hertz et $m(f)$ en unités mel. En pratique, un mel-spectrogram est obtenu en calculant d'abord la magnitude de la STFT, puis en la projetant sur une banque de filtres triangulaires distribués uniformément sur l'échelle mel. Si $H_r(k)$ désigne le r-ième filtre mel, l'énergie correspondante de bande mel est

$$M(m, r) = \sum_{k=0}^{N-1} |X(m, k)|^2 H_r(k), \tag{1.26}$$

souvent suivie d'une compression logarithmique. Cette représentation réduit la dimensionnalité tout en préservant la structure spectrale pertinente.

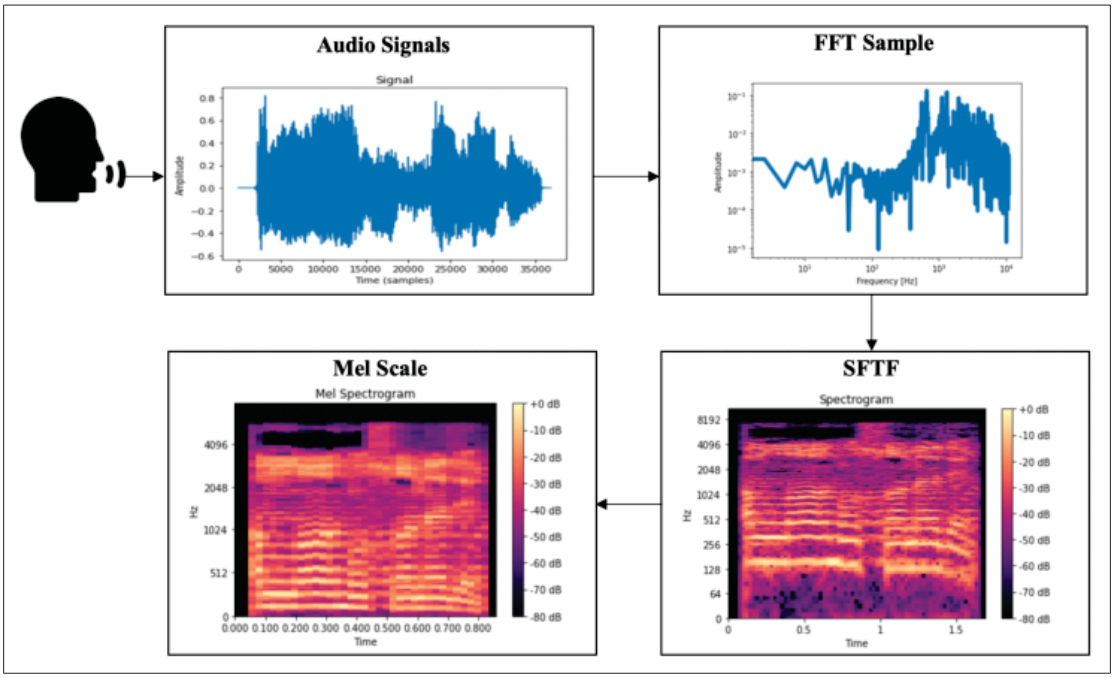

Figure 1.7 Pipeline de traitement des représentations audio, de la waveform à la représentation temps–fréquence et au mel-spectrogram
Tirée de Zhang *et al.* (2021)

Ces choix de traitement du signal sont directement pertinents pour le Chapitre 3. Les méthodes étudiées dans ce chapitre reposent sur des représentations audio hétérogènes, incluant des raw waveforms, des mel-spectrograms et des espaces latents dérivés d'encodeurs de spectrogrammes. Par conséquent, le choix de la représentation influence non seulement l'efficacité computationnelle, mais également le degré de contrôle possible lors de la génération et de l'édition, la préservation de la structure temporelle et spectrale, ainsi que la fidélité du son reconstruit. Comprendre ces représentations est donc essentiel pour analyser les compromis entre réalisme, contrôlabilité, diversité et utilisabilité pratique dans la génération de Sound Effects (SFX) orientée production.

1.3.2 Modèles génératifs audio

Autoregression—Les premières approches de génération audio reposent sur la modélisation *autoregressive*. Dans ces systèmes, la waveform audio est d'abord quantifiée en représentations

discrètes à débit binaire réduit à l'aide de codecs neuronaux tels que *EnCodec* (Défossez *et al.*, 2022) (voir Figure 1.8), *MusicGen* (Copet *et al.*, 2024), ou *VALL-E* (Wang *et al.*, 2023a). Ces codecs projettent la waveform en séquences d'indices de codes discrets. Les jetons obtenus sont ensuite davantage compressés à l'aide d'une quantification vectorielle résiduelle à multi-codebooks (RVQ), ce qui réduit la bande passante tout en préservant les informations perceptuellement pertinentes. Après cette étape de représentation, les jetons audio séquentiels ainsi que les entrées de conditionnement telles que le texte sont encodés et traités par un modèle de langage audio.

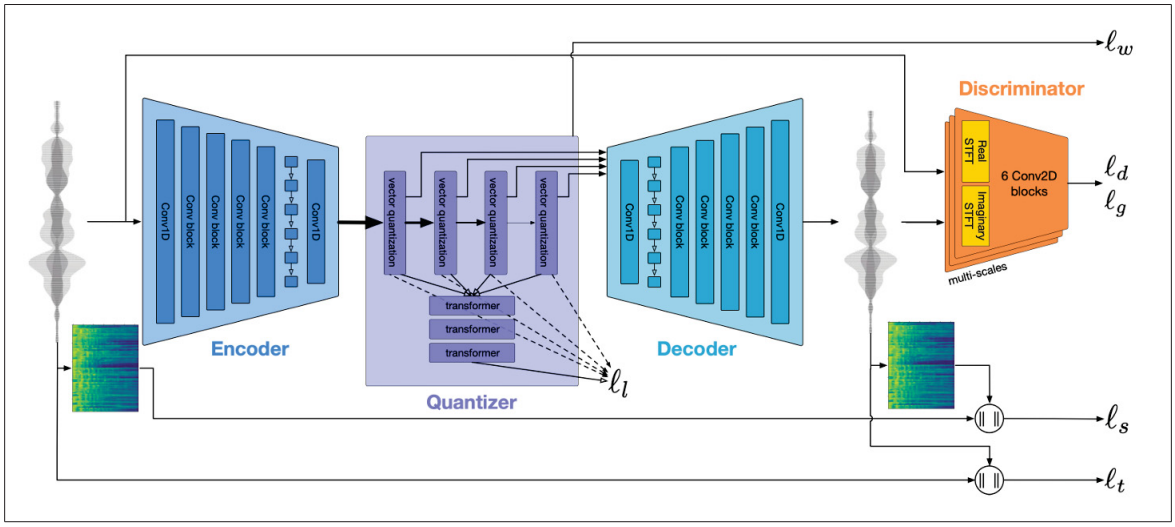

Figure 1.8 Pipeline *EnCodec* est une architecture codec encodeur–décodeur entraînée avec des pertes de reconstruction (ℓ_f et ℓ_t) ainsi que des pertes adversariales notamment ℓ_g pour le générateur et ℓ_d pour le discriminateur
Tirée de Défossez *et al.* (2022)

Le modèle utilise généralement une *causal self-attention* sur les jetons (tokens) audio et une *cross-attention* sur les jetons (tokens) de conditionnement. La séquence de sortie est générée de manière autoregressive puis décodée en waveform à l'aide du décodeur du codec neuronal, souvent entraîné avec des objectifs supplémentaires tels que des pertes adversariales et des pertes de *residual vector commitment* (Kreuk *et al.*, 2022b). Bien que la structure causale des modèles autoregressifs permette une modélisation cohérente des dépendances de longue portée dans les

séquences, la génération reste intrinsèquement séquentielle et peut donc être coûteuse sur le plan computationnel lors de l'inférence.

Score-Matching Diffusion—Dans la diffusion basée sur le *score matching*, le modèle apprend à inverser un processus de bruitage fixe en estimant soit la fonction de score, soit une cible équivalente de prédiction du bruit (Ho *et al.*, 2020a; Song *et al.*, 2020d). Plus précisément, un échantillon propre est progressivement corrompu par un processus de diffusion directe vers un bruit gaussien, tandis que la génération applique le processus inverse appris afin de reconstruire itérativement l'échantillon à partir du bruit. En génération audio, ce paradigme est généralement appliqué à des représentations de *waveform*, de spectrogrammes ou d'espaces latents, et demeure l'une des architectures dominantes pour la synthèse et l'édition audio de haute fidélité (Liu *et al.*, 2023b; Chen *et al.*, 2024b; Majumder *et al.*, 2024). Pour la diffusion latente, l'objectif d'entraînement canonique peut être écrit comme

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{z_0,\epsilon,t,c}\left[\left\|\epsilon - \epsilon_\phi(z_t, t, c)\right\|_2^2\right], \qquad z_t = \alpha_t z_0 + \sigma_t \epsilon, \tag{1.27}$$

où z_0 désigne la représentation latente propre, z_t sa version bruitée au pas de diffusion t, $\epsilon \sim \mathcal{N}(0, I)$, et c le signal de conditionnement. AudioLDM suit cette formulation de diffusion latente en entraînant le modèle de diffusion dans un espace latent continu de VAE et en le conditionnant par des embeddings audio basés sur CLAP lors de l'entraînement, tout en utilisant des embeddings textuels lors de l'échantillonnage via l'espace d'embeddings CLAP partagé. Un exemple représentatif de diffusion latente pour la génération audio est donc AudioLDM (Liu *et al.*, 2023a), qui opère dans une représentation latente basée sur VAE et utilise généralement l'échantillonnage DDIM pour l'inférence. Un autre exemple est T-Foley (Chen *et al.*, 2024b), entraîné avec une discrétisation de type DDPM d'une SDE et utilisant le *classifier-free guidance* lors de la génération.

Modèle génératifs de Flow Matching—Grâce à son échantillonnage déterministe et à ses fortes performances génératives, le *flow matching* est devenu de plus en plus populaire pour la

génération à faible latence et les applications temps réel, avec des performances remarquables dans les systèmes audio récents et les systèmes *Text-to-Speech* (Guan *et al.*, 2024; Eskimez *et al.*, 2024; Chen *et al.*, 2024a; Vyas *et al.*, 2023). Plutôt que d'apprendre un champ de score ou de bruit, le *flow matching* apprend un champ de vitesse et génère des échantillons en résolvant une ODE. Cette formulation peut également être mise en œuvre efficacement avec un nombre réduit d'étapes (Guan *et al.*, 2024) et dans des espaces latents pour améliorer la scalabilité (Guan *et al.*, 2024; Xu *et al.*, 2025).

Un exemple récent appliquant ce paradigme à la génération audio est ThinkSound (Liu *et al.*, 2025b), qui adopte une formulation de *flow matching* afin de générer des représentations audio conditionnées par des entrées multimodales telles que le texte et des indices visuels. Dans ce cadre, le modèle apprend un champ de vitesse conditionnel qui transporte les échantillons d'une distribution de bruit simple vers la représentation audio cible le long d'un chemin probabiliste continu. Le processus de génération suit ensuite cette trajectoire apprise à l'aide d'une intégration déterministe d'ODE, permettant une synthèse efficace tout en maintenant une forte qualité perceptuelle. En opérant dans une représentation audio compressée et en conditionnant le flot par des embeddings sémantiques, ThinkSound exploite la stabilité et l'efficacité du *flow matching* pour la génération de sons contrôlable dans des contextes multimodaux (Liu *et al.*, 2025b).

Rôles du Transformer—Les Transformers sont largement utilisés pour fournir un contexte global et séquentiel dans les architectures modernes de génération audio et d'images. En génération audio, une structure courante est l'architecture Diffusion Transformer (DiT), où la self-attention opère sur des jetons audio ou des patches latents et où le Transformer sert de backbone de débruitage (Tian *et al.*, 2025; Flores García, Nieto, Salamon, Pardo & Seetharaman, 2024a; Chen *et al.*, 2025; Shi *et al.*, 2025; Zhao *et al.*, 2025; Cheng *et al.*, 2024). Des architectures Transformer étroitement liées basées sur les flots prédisent un champ de vitesse plutôt qu'une cible de débruitage, ce qui les rend particulièrement adaptées à la génération efficace basée sur ODE (Kushwaha & Tian, 2024; Shi *et al.*, 2025; Xu *et al.*, 2025).

La Figure 1.9 présente un exemple représentatif : AudioX (Tian *et al.*, 2025), où des modules Transformer sont utilisés à la fois avant et pendant la génération. Des encodeurs spécialisés traitent d'abord différentes modalités, tandis que des modules Transformer temporels modélisent la structure séquentielle des caractéristiques vidéo et audio. Un module Multimodal Adaptive Fusion (MAF) unifie ensuite ces signaux en un embedding de conditionnement H_c à l'aide d'une fusion à portes (*gated fusion*), d'une cross-attention depuis des requêtes spécifiques aux modalités et d'un raffinement par self-attention. Le backbone DiT traite ensuite l'entrée latente bruitée z_t, conditionnée par H_c via cross-attention afin de générer de l'audio et de la musique de haute qualité (Tian *et al.*, 2025). Ce processus peut être résumé comme

$$c = \mathrm{MAF}(e_v, e_t, e_a), \qquad \hat{\epsilon} = \epsilon_\phi(z_t, t, c), \tag{1.28}$$

où c est la condition multimodale unifiée et $\hat{\epsilon}$ est la prédiction de bruit du DiT pour la variable latente bruitée z_t. Cette conception place le Transformer au centre à la fois de l'alignement multimodal et de la synthèse audio conditionnelle.

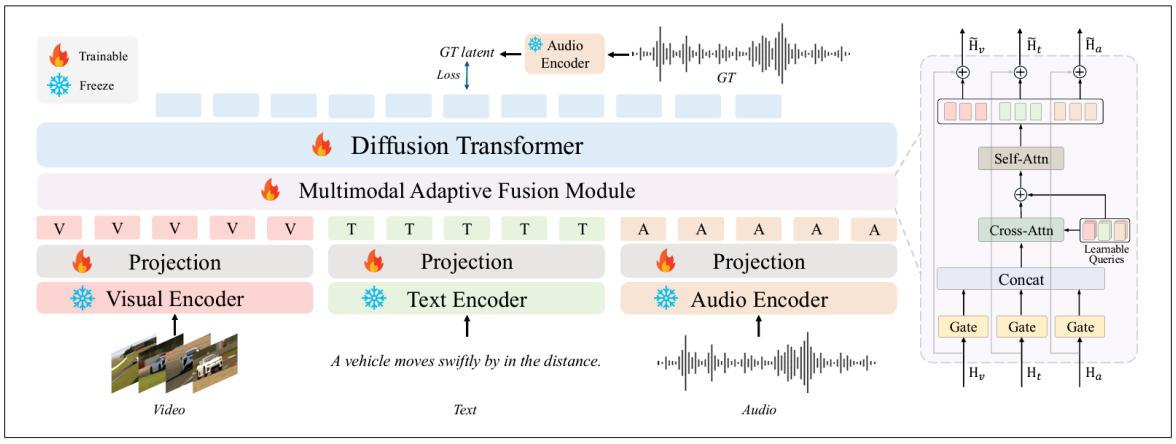

Figure 1.9 Méthode AudioX où des encodeurs spécialisés traitent plusieurs modalités et où le module Multimodal Adaptive Fusion MAF les regroupe en un *embedding* de conditionnement partagé H_c Le modèle DiT débruite ensuite une représentation latente bruitée z_t en étant conditionné par H_c via un mécanisme de cross attention afin de produire de l audio et de la musique de haute qualité Les notations z_t et H_c sont omises dans la figure pour alléger la représentation visuelle
Tirée de Tian *et al.* (2025)

1.3.3 Edition audio via diffusion

Les architectures de diffusion latente sont particulièrement efficaces pour modéliser des distributions audio complexes avec une haute qualité perceptuelle, bien que l'échantillonnage nécessite généralement plusieurs étapes de débruitage. Les cadres basés sur la diffusion peuvent également être adaptés aux tâches d'édition audio, où l'objectif est de modifier un échantillon audio existant tout en préservant des caractéristiques acoustiques pertinentes telles que la structure temporelle, le timbre ou l'identité de l'événement. Dans ce cadre, l'édition est généralement réalisée en inversant d'abord l'audio d'entrée dans la trajectoire latente d'un DM préentraîné, puis en guidant le processus inverse de débruitage vers la modification souhaitée.

Un exemple représentatif est AudioMorphix (Park *et al.*, 2025), voir Figure 1.10, une méthode d'édition audio basée sur spectrogramme et ne nécessitant pas d'entraînement (*training-free*), construite sur des modèles de diffusion latente TTA tels que AudioLDM (Liu *et al.*, 2023b) ou Tango2 (Majumder *et al.*, 2024). Plutôt que de réentraîner le backbone de diffusion, AudioMorphix réalise l'édition en appliquant d'abord une inversion DDIM afin de projeter l'audio d'entrée dans l'espace latent, puis en mettant à jour de manière itérative la représentation latente dans le domaine temps–fréquence pendant le processus de débruitage. Un mécanisme de guidage par étapes (*stepwise guidance*) permet d'orienter l'échantillon vers l'édition cible tout en conservant les caractéristiques pertinentes de l'entrée originale. En pratique, cette conception permet une édition contrôlée sans réentraînement, bien qu'elle nécessite encore un nombre relativement élevé d'étapes d'inférence, typiquement autour de 50.

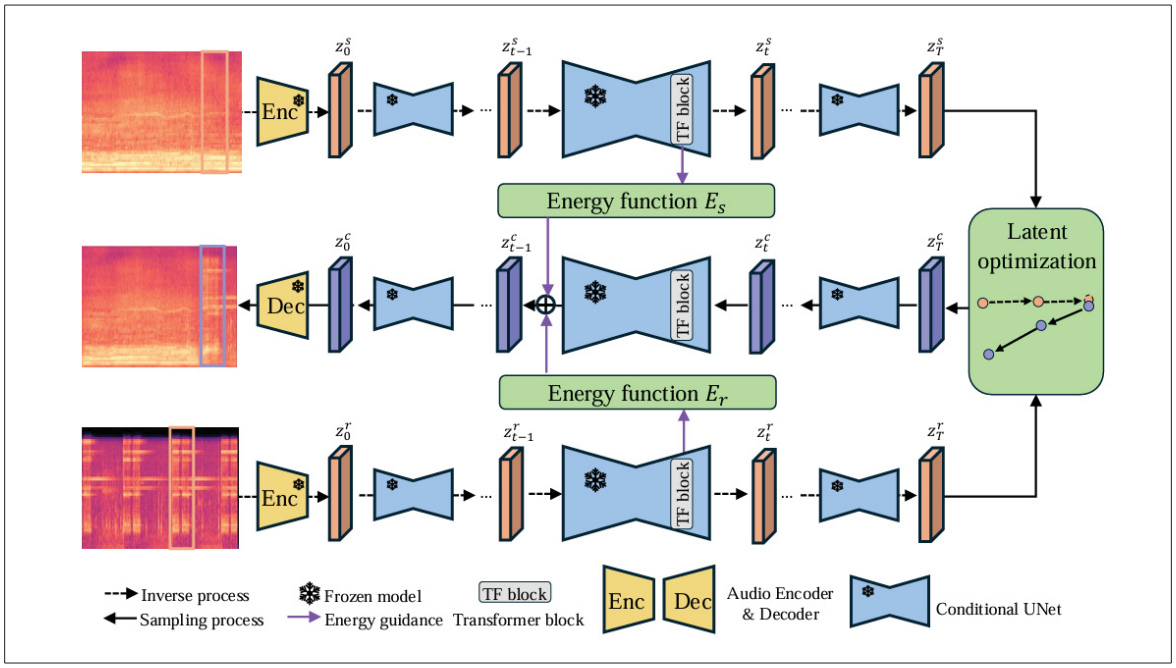

Figure 1.10 Méthode AudioMorphix sans entraînement pour l'édition audio basée sur spectrogramme
Tirée de Park *et al.* (2025)

Les modèles de *flow matching* ont également récemment été explorés pour l'édition et la génération audio. Un exemple notable est ThinkSound (Liu *et al.*, 2025b), qui utilise un raisonnement multimodal pour guider un modèle unifié de génération audio basé sur le *flow matching* pour la génération de bandes sonores et leur raffinement ciblé.

Les Transformers sont également fréquemment utilisés comme modules de conditionnement dans ces systèmes. Les invites textuelles peuvent être encodées à l'aide d'encodeurs Transformer, tandis que les pipelines multimodaux peuvent en outre modéliser des caractéristiques vidéo et le contexte temporel afin de mieux aligner l'audio généré avec le mouvement visuel (Majumder *et al.*, 2024; Tian *et al.*, 2025; Kushwaha & Tian, 2024). Plus généralement, les mécanismes de *self-attention*, de *cross-attention* et d'autres mécanismes de conditionnement restent centraux pour injecter des signaux de contrôle multimodaux dans les générateurs audio basés sur la diffusion ou les flots (Shi *et al.*, 2025; Zhao *et al.*, 2025; Flores García *et al.*, 2024a; Sheffer & Adi, 2022).

1.3.4 Génération d'effets sonores

La génération de *Sound Effects* (SFX) vise à produire des événements audio courts, sémantiquement significatifs et souvent structurés temporellement, correspondant à une action, une scène ou une référence ciblée. Comparée à la génération TTA générale, la génération de SFX met davantage l'accent sur l'identité de l'événement, la synchronisation temporelle et la contrôlabilité, puisque l'objectif est souvent non seulement de générer un audio plausible, mais également de produire un son correspondant à un timing, un mouvement ou une caractéristique acoustique spécifique. En pratique, ces exigences rapprochent la génération de SFX d'une tâche de génération contrainte ou de variation plutôt que d'une synthèse audio ouverte.

Un premier cadre important est celui de la synthèse *Foley* et de la génération *video-to-audio*, où l'objectif est de générer des SFX alignés avec des événements visuels. Ces approches sont généralement entraînées sur des ensembles de données audiovisuels soigneusement organisés, tels que des collections de sons Foley, AudioSet (**?**), ou des ensembles de données dérivés de Freesound. Par exemple, T-Foley conditionne une diffusion de waveform à la fois sur la classe sonore et sur des caractéristiques temporelles d'événement, permettant une synthèse contrôlable de sons Foley alignés temporellement (Chen *et al.*, 2024b). De même, Diff-Foley (Luo, Yan, Hu & Zhao, 2023c) étudie la synthèse audio synchronisée avec la vidéo à l'aide de DM latents, tandis que des méthodes plus récentes telles que Foley-Flow (Mo & Song, 2024) et VAFlow (Wang, Cheng, Wang, Song & Wang, 2025c) adoptent des formulations basées sur les flots afin d'améliorer la synchronisation rythmique et l'alignement multimodal. Ces travaux mettent en évidence l'importance du guidage temporel dans la génération et l'édition de SFX, en particulier lorsque le son généré doit suivre le mouvement visuel ou les limites d'événements.

Un second cadre étend la génération de SFX vers une synthèse multimodale et contrôlable. Dans ce cas, les modèles sont conditionnés par plusieurs modalités telles que le texte, la vidéo et des références audio, dépassant la génération à invite unique pour offrir un contrôle plus flexible. AudioX permet une génération *anything-to-audio* à partir de conditions textuelles, visuelles et audio via un DiT multimodal (Liu *et al.*, 2025c), tandis que MultiFoley (Wu *et al.*, 2024)

introduit des contrôles multimodaux pour la génération Foley guidée par vidéo, incluant des invites textuelles optionnelles et des embeddings audio de référence pour le transfert de style. Ces approches sont particulièrement pertinentes dans les contextes de production où plusieurs sources d'information peuvent être disponibles pour guider la sortie souhaitée.

Un troisième cadre, de plus en plus pertinent, concerne la génération de SFX guidée par référence et sensible à la source. Plutôt que de générer un effet sonore uniquement à partir d'une description textuelle ou vidéo, ces méthodes conditionnent la synthèse sur un signal audio de référence afin de préserver le timbre, le style ou l'identité acoustique. Par exemple, AC-Foley (Fang *et al.*, 2025) étudie la synthèse vidéo–audio guidée par audio de référence avec transfert acoustique, en combinant indices visuels, audio et éventuellement textuels dans un cadre conditionnel de *flow matching*. Ces approches basées sur la référence sont particulièrement pertinentes pour les workflows pratiques de SFX, où l'objectif est souvent de générer des variations d'un effet sonore existant plutôt que de synthétiser des sons entièrement nouveaux. Ainsi, de nombreux travaux récents étudient la génération de SFX dans des contextes contraints et orientés production, tels que la synthèse Foley (Wu *et al.*, 2024; Fang *et al.*, 2025; Luo *et al.*, 2023c), la génération *video-to-audio* (Liu *et al.*, 2025b; Mo & Song, 2024; Mei *et al.*, 2023), ou la génération de variations sonores guidées par référence (Chen *et al.*, 2024b; Park *et al.*, 2025; Liu *et al.*, 2025c; Zhang *et al.*, 2025b; Cífka *et al.*, 2025).

Au-delà de la génération directe, des travaux récents explorent également les pipelines d'édition comme mécanisme de génération et de manipulation sonore. Par exemple, ThinkSound (Liu *et al.*, 2025b) permet une synthèse Foley structurée et une conception sonore interactive en modifiant progressivement une scène audio générée. De même, des cadres d'édition de spectrogrammes tels que AudioMorphix (Park *et al.*, 2025) démontrent que les pipelines d'édition basés sur la diffusion peuvent être réutilisés pour une manipulation sonore ciblée, permettant des opérations telles que l'insertion, la suppression ou la transformation d'événements sonores. Ces approches basées sur l'édition suggèrent un paradigme alternatif dans lequel les effets sonores sont générés par modification itérative de représentations intermédiaires plutôt que par un unique passage de génération directe.

Enfin, des modèles de fondation open-weight tels que Stable Audio Open (Evans *et al.*, 2024), AudioLDM (Liu *et al.*, 2023b; Liu *et al.*, 2023c), et UniAudio (Yang *et al.*, 2023b) offrent des capacités de génération plus générales pouvant également servir de baselines pour les SFX. Ces modèles sont généralement entraînés sur des ensembles de données à grande échelle tels que AudioSet ou des collections Freesound et proposent des backbones préentraînés de haute qualité ainsi que des pipelines d'inférence flexibles. Cependant, ils ne sont pas spécifiquement conçus pour les workflows SFX, qui nécessitent souvent des capacités plus spécialisées telles que la génération conditionnée par référence, la variation contrôlable ou l'alignement temporel précis.

Malgré ces avancées, les systèmes actuels de génération de SFX présentent encore plusieurs limitations. Premièrement, de nombreux modèles se concentrent sur la synthèse TTA générale plutôt que sur la génération structurée d'effets sonores, ce qui rend difficile le contrôle de l'identité de l'événement et de la structure temporelle. Deuxièmement, les ensembles de données et les protocoles d'évaluation restent hétérogènes, reposant souvent sur des benchmarks audio génériques tels que ESC-50 (Piczak, 2015) ou AudioSet (Zhu, Wen & Duan, 2025), qui ne capturent pas pleinement les exigences orientées production. En particulier, les cadres d'évaluation systématiques pour la génération de variations de SFX guidées par référence restent limités dans la littérature. Troisièmement, la plupart des méthodes mettent l'accent sur la génération d'une seule instance plutôt que sur la variation systématique d'un SFX de référence, ce qui limite leur applicabilité dans les workflows professionnels de conception sonore. En pratique, les concepteurs sonores ont souvent besoin de variations contrôlables d'un clip de référence donné, qui préservent son identité acoustique tout en introduisant une diversité réaliste. Répondre à ces défis nécessite des protocoles d'évaluation et des pipelines de génération reflétant davantage les contraintes réelles de conception sonore, ce qui motive le cadre expérimental proposé dans le Chapitre 3.

1.4 Synthèse et revue des gaps dans la littérature

Ce chapitre a présenté une revue des avancées récentes en modèles génératifs pour l'image et l'audio, avec un accent particulier sur les approches basées sur la diffusion.

Dans le domaine de l'image, les DMs atteignent une forte qualité générative et permettent la SDP, tandis que les méthodes de distillation améliorent l'efficacité de l'échantillonnage. Toutefois, la personnalisation efficace demeure un défi, car l'adaptation *few-shots* et l'inférence rapide sont rarement abordées conjointement dans les travaux existants. La plupart des pipelines de personnalisation reposent sur le *fine-tuning* itératif de grands modèles de diffusion, ce qui peut entraîner du surapprentissage et une dégradation de la diversité lorsque seules quelques images de référence sont disponibles. Dans le même temps, les méthodes efficaces de diffusion distillation se concentrent principalement sur la réduction du nombre d'étapes de débruitage, tout en supposant l'accès à de grands ensembles de données d'entraînement et en restant contraintes par le manifold génératif appris par le modèle enseignant. En conséquence, les stratégies existantes telles que *Distill-then-Adapt* ou *Adapt-then-Distill* souffrent souvent d'une capacité d'adaptation limitée, d'une instabilité à l'entraînement ou d'une complexité de conception accrue. En particulier, les modèles étudiants distillés peuvent perdre en flexibilité d'adaptation en présence d'un décalage de domaine, tandis qu'une distillation effectuée sur des données *few-shots* est sujette au surapprentissage. Il subsiste donc un besoin de cadres unifiés capables de traiter conjointement la personnalisation few-shot et la génération efficace, tout en préservant l'identité, la diversité et la fidélité générative. Ce défi motive les contributions présentées au Chapitre 2, qui étudie un cadre conditionnel d'appariement de distributions à double domaine afin de réaliser adaptation et distillation dans un seul processus d'entraînement pour la SDP.

Dans le domaine audio, les modèles récents autoregressifs, basés sur la diffusion et sur le *flow matching* ont considérablement amélioré la qualité et la contrôlabilité de l'audio généré. Les modèles autoregressifs offrent une modélisation cohérente des dépendances de longue portée, les cadres basés sur la diffusion demeurent des backbones solides pour la synthèse et l'édition de haute fidélité, et les modèles de *flow matching* proposent des alternatives de plus en plus attrayantes à faible latence pour une génération efficace. En parallèle, les modules de conditionnement basés sur des Transformers et les modules de fusion multimodale ont encore élargi le champ de la génération audio contrôlable en permettant l'alignement avec des modalités textuelles, visuelles et audio. Toutefois, la génération de SFX introduit des exigences

supplémentaires, notamment l'alignement temporel, la fidélité à la référence et la variation contrôlable. En pratique, la génération de SFX n'est souvent pas une tâche de génération ouverte, mais plutôt une tâche de production contrainte dans laquelle le son généré doit correspondre à un timing, un mouvement, une identité acoustique ou un clip de référence spécifique. Bien que des méthodes récentes telles que T-Foley(Chen *et al.*, 2024b), AudioMorphix(Park *et al.*, 2025), ThinkSound(Liu *et al.*, 2025b), AudioX(Liu *et al.*, 2025c) et MultiFoley(Wu *et al.*, 2024) traitent certaines parties de ce problème à travers le guidage temporel, le conditionnement multimodal, la génération fondée sur l'édition ou le transfert à partir d'un audio de référence, les méthodes actuelles ciblent principalement la génération audio générale ou des contextes multimodaux spécifiques. En conséquence, elles ne répondent pas pleinement au besoin pratique de génération guidée par référence de variations de SFX préservant l'identité. De plus, les ensembles de données et les protocoles d'évaluation restent hétérogènes et reposent souvent sur des benchmarks audio génériques qui ne capturent pas entièrement les exigences orientées production. En particulier, les protocoles d'évaluation structurés pour la génération de variations de SFX guidées par référence demeurent limités, et la plupart des méthodes mettent l'accent sur la génération d'instances uniques plutôt que sur la variation contrôlable d'un son de référence donné. Ces limites motivent les contributions du Chapitre 3, qui se concentre sur la génération de variations de SFX prête pour la production à travers une sélection de baselines guidée par les exigences, un benchmark d'évaluation unifié, ainsi qu'une analyse détaillée du compromis diversité–fidélité et des capacités d'édition des pipelines génératifs actuels.

Dans l'ensemble, la littérature met en évidence un compromis persistant entre efficacité et contrôlabilité dans les pipelines basés sur la diffusion, à la fois pour la génération d'images personnalisées et pour l'édition de SFX. Ces limites motivent les contributions de cette thèse. Le Chapitre 2 traite le problème conjoint de l'adaptation efficace et de la distillation pour la génération d'images personnalisées guidée par sujet. Le Chapitre 3 étudie la génération de variations de SFX sous des contraintes réalistes de production et propose une analyse des pipelines de génération audio conditionnés par référence pour l'édition et la variation de SFX.

CHAPITRE 2

SUBJECT DRIVEN PERSONALIZATION AVEC UNI-DAD

Ce chapitre s'appuie sur des éléments issus d'un travail collaboratif récent sur l'article `Uni-DAD`, accepté à la conférence IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2026[1]. Plus précisément, la contribution principale présentée dans ce chapitre se concentre sur l'application de *Subject Driven Personalization* (SDP), améliorée par un conditionnement textuel et évaluée sur plusieurs ensembles de données dérivés du benchmark DreamBooth Ruiz *et al.* (2023).

Une part importante du contenu de ce chapitre est issue de l'article officiel `Uni-DAD` et a été réorganisée afin de mettre en évidence l'application spécifique étudiée ici.

2.1 Introduction

Les Diffusion Models (DMs) (Song *et al.*, 2020c) se sont imposés comme le paradigme dominant pour la modélisation générative, atteignant des performances SoTA en synthèse d'images (Dhariwal & Nichol, 2021) et en génération *text-to-image* (TTI) (Ruiz *et al.*, 2023; Rombach *et al.*, 2022; Saharia *et al.*, 2022). Ces modèles peuvent produire des images de haute qualité et diversifiées, même lorsqu'ils sont adaptés à de nouveaux domaines ou à de nouveaux sujets à partir de seulement quelques images de référence (4–6). Cela en fait une solution particulièrement adaptée à la SDP (Gal *et al.*, 2022; Ruiz *et al.*, 2023; Kumari *et al.*, 2023).

Cependant, les DMs reposent sur un processus itératif de débruitage sur de nombreux pas de temps lors de l'inférence, ce qui entraîne des temps de génération élevés au moment du test. Les modèles adaptés héritent de ce coût computationnel, ce qui limite leur utilisation pratique pour la personnalisation guidée par sujet en temps réel. Dans ce travail, nous montrons que `Uni-DAD SDP` fournit un pipeline de personnalisation efficace et de haute qualité tout en ne nécessitant qu'**une seule étape** d'échantillonnage lors de l'inférence.

---

[1] Y. Bahram, M. Desbos, M. Shateri, and E. Granger, "Uni-DAD : Unified Distillation and Adaptation of Diffusion Models for Few-step Few-shot Image Generation," Proceedings of CVPR, 2026. https://arxiv.org/abs/2511.18281.

La distillation permet d'atténuer la lenteur de l'inférence en entraînant un modèle étudiant à quelques étapes pour imiter un modèle enseignant de diffusion plus large (Salimans & Ho, 2022; Song *et al.*, 2023; Yin *et al.*, 2024c,a; Chadebec *et al.*, 2025). En définitive, la capacité à générer des images dans de nouveaux domaines en contexte few-shot tout en ne nécessitant que quelques étapes de débruitage peut faciliter le déploiement des DMs dans des applications personnalisées en temps réel.

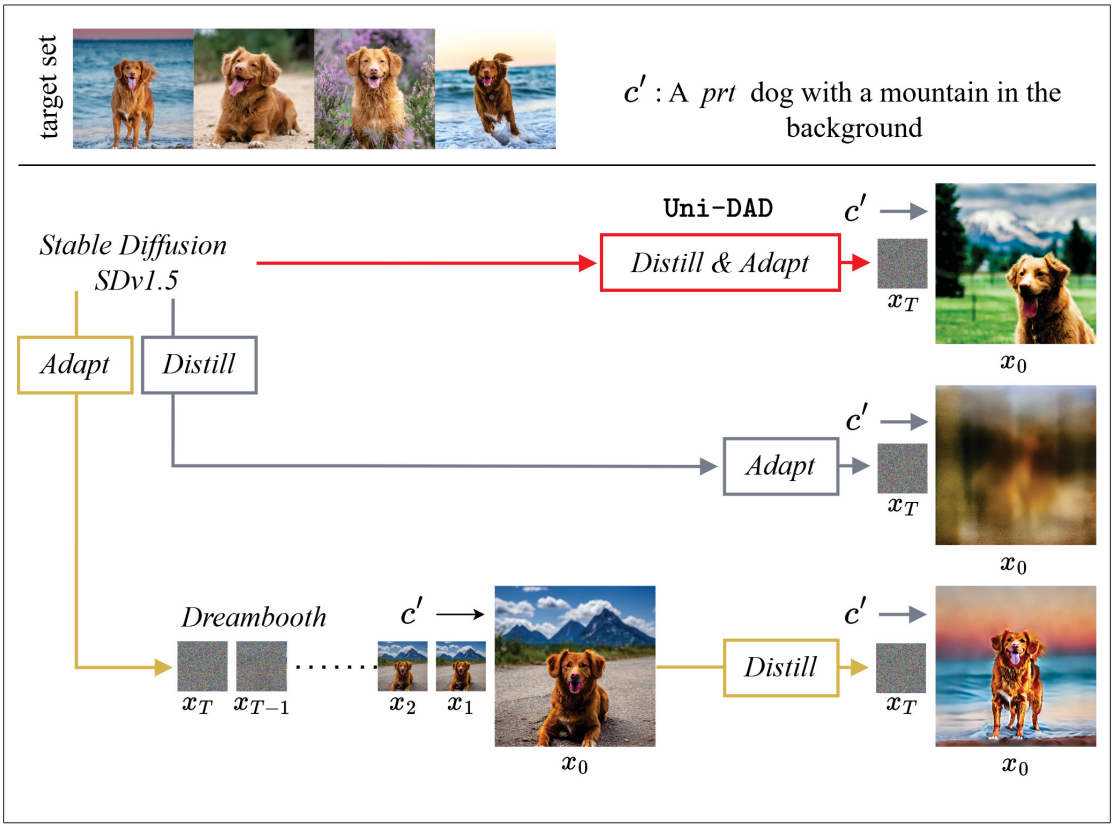

Figure 2.1 `Uni-DAD SDP` (*Distill & Adapt*) comparé aux pipelines en deux étapes (*Distill-then-Adapt* et *Adapt-then-Distill*) L'étape *Adapt* est réalisée par *fine-tuning* et l'étape *Distill* par DMD2 Yin *et al.* (2024a) - Le domaine source est représenté par 5.85B paires image–texte filtrées par CLIP Schuhmann *et al.* (2022) et le domaine cible par 5 images de l'instance dog7 - c' représente l'invite (*prompt*) fournie à chaque modèle lors de l'inférence - Les générations proviennent de l'adaptation à l'instance DreamBooth dog7 Tirée de Bahram *et al.* (2025)

Dans les travaux récents, réduire le nombre d'étapes de diffusion tout en s'adaptant à de nouveaux domaines nécessite généralement un pipeline en deux étapes (voir Figure 2.1) : *Distill-then-Adapt* ou *Adapt-then-Distill* (cf. section 1.2.2). Aucunes de ces pipelines en deux étapes ne constitue un processus véritablement *end-to-end*, et tous deux risquent de perdre une partie de l'information transférable et diversifiée issue du domaine source lors de l'entraînement.

Pour remédier à ces limitations, `Uni-DAD` introduit un pipeline d'entraînement en une seule étape qui réalise conjointement distillation et adaptation. Cette approche combine un objectif d'appariement de distributions à double domaine avec une supervision adversariale afin de préserver la diversité du domaine source tout en améliorant le réalisme du domaine cible dans un contexte d'adaptation few-shot.

`Uni-DAD SDP` atteint une qualité de génération d'images compétitive par rapport aux méthodes SoTA d'adaptation tout en ne nécessitant qu'une seule étape d'échantillonnage. La Figure 2.2 illustre comment ce générateur en une étape produit des images diversifiées et de haute qualité dans un nouveau domaine en contexte few-shot, répondant à des invites textuelles utilisateur avec un large éventail de personnalisations tout en reproduisant fidèlement l'identité du sujet. De plus, cette approche surpasse les pipelines d'entraînement classiques en deux étapes en termes de qualité et de diversité.

Figure 2.2 Aperçu des capacités de génération de `Uni-DAD Subject Driven Personalization (SDP)` à travers différentes prompts (recontextualisation, ajout d'accessoires et modification de propriétés) sur deux instances DreamBooth (cat2, teapot) - comparées au modèle DreamBooth Ruiz *et al.* (2023) - `Uni-DAD SDP` s'adapte à partir de seulement **quelques exemples** et génère en **une seule étape de diffusion** contre 100 évaluations de fonction neuronale (NFE) pour le pipeline DreamBooth Stable Diffusion - *prt* est utilisé comme *rare token* pour faciliter l'apprentissage du domaine cible
Tirée de Bahram *et al.* (2025)

La section suivante introduit le cadre `Uni-DAD SDP`.

2.2 Méthodologie

`Uni-DAD SDP` est proposé comme une pipeline en une seule étape qui compresse un DM enseignant source figé ϵ^{src}, entraîné avec un grand nombre de pas de temps ($T \sim 1000$) sur une large distribution source $p^{\text{src}}(x)$, en un générateur étudiant rapide G de paramètres θ (NFE = 1), tout en s'adaptant à une distribution cible $p^{\text{trg}}(y)$ représentée par un ensemble cible few-shot Y ($|Y| \leq 6$). La méthode s'appuie sur trois enseignants de diffusion afin de réaliser efficacement l'adaptation et la distillation. Elle combine deux signaux complémentaires pour entraîner G : (i) un objectif *dual-domain DMD* contre *source teacher* ϵ^{src} et d'un *target teacher* en ligne ϵ^{trg}, et (ii) une perte GAN multi-têtes favorisant le réalisme des échantillons cibles à plusieurs échelles de caractéristiques. Dans ce cadre, le *target teacher* ϵ^{trg} favorise l'adaptation au domaine cible, tandis que l'enseignant source figé ϵ^{src} préserve la diversité héritée du domaine source dans les prédictions adaptées. Un *fake teacher* ϵ^{fk} est maintenu afin de suivre l'évolution de la distribution du *student* et de soutenir un discriminateur multi-têtes D chargé de distinguer les générations du *student* des éléments de l'ensemble cible few-shot Y. De plus, ϵ^{trg} peut être ajusté par *fine-tuning* sur Y afin d'améliorer encore davantage l'adéquation à la distribution cible. L'entraînement alterne entre l'optimisation de trois composantes : (i) le *student* G, (ii) le *fake teacher* ϵ^{fk} conjointement avec le discriminateur GAN, et (iii) le *target teacher* ϵ^{trg} (Fig. 2.3).

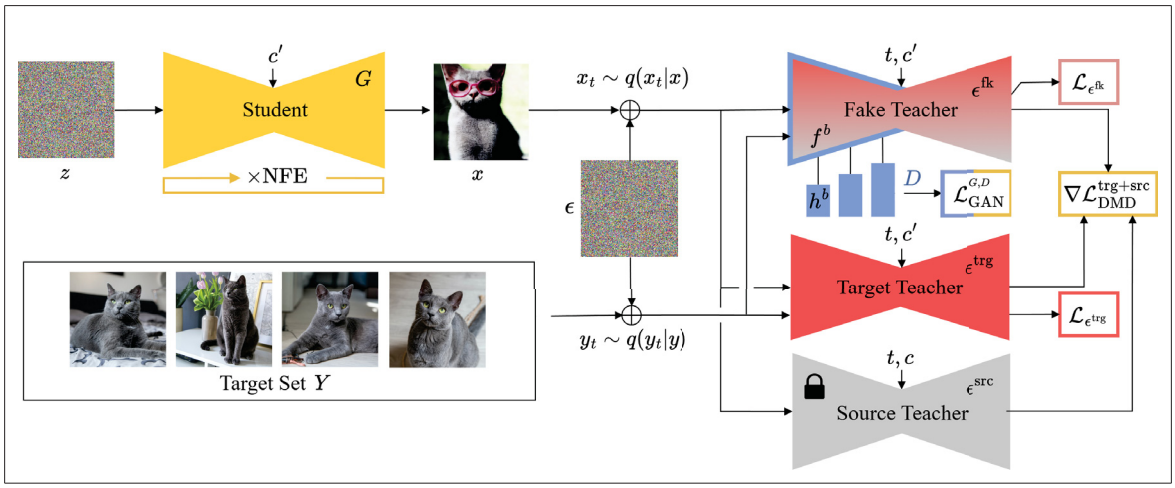

Figure 2.3 Vue d'ensemble de `Uni-DAD SDP` pour la génération d'images few-step et few-shot
Tirée et adaptée de Bahram *et al.* (2025)

DMD à double domaine— Le DMD est initialement utilisé pour aligner la distribution d'un étudiant p^{fk} sur p^{src} au sein du domaine source (Yin *et al.*, 2024c,a). Il minimise la divergence KL entre les deux distributions aux sorties courantes du *student*, en poussant le générateur étudiant vers des régions de plus forte densité de p^{src}. Le calcul explicite des densités de probabilité nécessaires pour estimer la perte $\mathcal{L}_{\text{DMD}}(\theta)$ est généralement intractable (Yin *et al.*, 2024c). Toutefois, le gradient de cette perte par rapport à θ peut être obtenu :

$$\begin{aligned} \nabla_\theta \mathcal{L}_{\text{DMD}} &= \nabla_\theta D_{\text{KL}}(p^{\text{fk}} \| p^{\text{src}}) \\ &= \mathbb{E}_z \left[\left(\nabla_x \log p^{\text{fk}}(x) - \nabla_x \log p^{\text{src}}(x) \right) \frac{dG_\theta}{d\theta} \right], \end{aligned} \tag{2.1}$$

où $x = G(z),\ z \sim \mathcal{N}(0, I)$ désigne la sortie du *student*. Sous une perturbation gaussienne, le score vérifie $s(x_t) = \nabla_{x_t} \log p(x_t) = -\frac{1}{\sigma_t}\, \epsilon(x_t)$, Song *et al.* (2020c). Par conséquent, les termes du membre de droite de l'Éq. 2.1 peuvent être approximés à l'aide de deux DM : ϵ^{src} et ϵ^{fk}, où ϵ^{src} est figé et ϵ^{fk} est entraîné conjointement pour suivre l'évolution des sorties du *student* (Sec. 2.2). En pratique, $\mathcal{L}_{\text{DMD}}$ peut être minimisée en mettant à jour $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}_{\text{DMD}}$ par descente de gradient.

Nous étendons cette formulation afin d'aligner les sorties du *student* à la fois sur p^{src} et sur p^{trg}. Les gradients des deux pertes DMD peuvent s'écrire sous forme d'estimation du bruit :

$$\begin{aligned} \nabla_\theta\, \mathcal{L}_{\text{DMD}^{\text{src}}} &\approx \mathbb{E}_{t,z} \left[\omega_t \left(\epsilon^{\text{fk}}(x_t) - \epsilon^{\mathbf{src}}(x_t) \right) \frac{dG_\theta}{d\theta} \right], \\ \nabla_\theta\, \mathcal{L}_{\text{DMD}^{\text{trg}}} &\approx \mathbb{E}_{t,z} \left[\omega_t \left(\epsilon^{\text{fk}}(x_t) - \epsilon^{\mathbf{trg}}(x_t) \right) \frac{dG_\theta}{d\theta} \right], \end{aligned} \tag{2.2}$$

où $t \sim \mathcal{U}\{0.02T, 0.98T\}$ et les pas de temps extrêmes sont exclus pour des raisons de stabilité numérique (Poole, Jain, Barron & Mildenhall, 2022). Une normalisation qui équilibre les contributions entre les pas de temps est utilisée :

$$\omega_t = \frac{\sigma_t \cdot H \cdot S}{\left\| \epsilon - \epsilon^{\text{fk}}(x_t) \right\|_1}, \tag{2.3}$$

où H désigne le nombre de canaux et S le nombre de positions spatiales (Yin *et al.*, 2024c). L'optimisation de $\mathcal{L}_{\text{DMD}}^{\text{src}}$ peut aider à préserver une information transférable diversifiée (p. ex., la pose, l'arrière-plan et l'expression faciale), compensant ainsi la rareté des données cibles. Cet objectif peut suffire pour l'adaptation en présence de faibles décalages de domaine. Cependant, des domaines cibles structurellement plus éloignés peuvent contenir des régions situées en dehors du manifold source, auquel cas $\mathcal{L}_{\text{DMD}}^{\text{src}}$ peut freiner la véritable adaptation. Un objectif DMD à double domaine permet alors de guider le *student* vers une zone commune aux deux distributions :

$$\nabla_\theta \mathcal{L}_{\text{DMD}}^{\text{trg+src}} = (1-a)\nabla_\theta \mathcal{L}_{\text{DMD}^{\text{src}}} + a\nabla_\theta \mathcal{L}_{\text{DMD}^{\text{trg}}}, \tag{2.4}$$

où $a \in [0,1]$ est un facteur de pondération contrôlant l'influence de chaque domaine. Des expériences complémentaires sont menées sur la variation de ce facteur de pondération afin d'évaluer la meilleure valeur pour équilibrer l'alignement sur la cible et la diversification des générations (Fig. 2.8).

Fake et target teachers— L'initialisation de ϵ^{fk} est faite avec les poids (checkpoints) de ϵ^{src} et une mise à jour des paramètres ϕ sur les sorties évolutives du *student* en minimisant l'objectif MSE :

$$\mathcal{L}_{\text{fk}}(\phi) = \mathbb{E}_{t,z}\left[\left\| {\epsilon^{\text{fk}}}_{\phi}(x_t) - \epsilon \right\|_2^2 \right]. \tag{2.5}$$

Lors des mises à jour de ϵ^{fk}, aucun gradient n'est propagé à travers G, et x est traité comme une entrée fixe. De manière analogue, ϵ^{trg} est initialisé avec les poids de ϵ^{src} et ses paramètres η sont mis à jour via la MSE afin de débruiter des échantillons diffusés issus de Y :

$$\mathcal{L}_{\text{trg}}(\eta) = \mathbb{E}_{t,\epsilon,y}\left[\left\| \epsilon_\eta^{\text{trg}}(y_t) - \epsilon \right\|_2^2 \right]. \tag{2.6}$$

L'entraînement et l'intégration de ϵ^{trg} ne sont pas optionnels dans `Uni-DAD SDP`, car ils facilitent l'adaptation structurelle dans les cas de forts décalages de domaine, comme certaines instances d'objets spécifiques ou de sujets réels rares (p. ex., des races de chiens peu communes).

GAN multi-têtes— Afin d'imposer une fidélité nette des sorties du *student* à Y et de stabiliser l'entraînement, nous utilisons un objectif GAN multi-têtes qui évalue le réalisme cible à plusieurs niveaux de caractéristiques. Tandis que G joue le rôle de générateur, le discriminateur D réutilise les blocs encodeur et intermédiaires de ϵ^{fk} pour l'extraction de caractéristiques : soit $f^b(\cdot)$ l'extracteur de caractéristiques au bloc $b \in \mathcal{B}$ de ϵ^{fk}, auquel on attache une tête linéaire $h^l(\cdot)$ de paramètres ψ à la sortie de chaque bloc. Cela produit un discriminateur multi-têtes D dont l'objectif est de distinguer les échantillons $y \in Y$ des images $x = G(z),\ z \sim \mathcal{N}(0, I)$. Les pertes GAN, agrégées sur les têtes par sommation, sont :

$$\mathcal{L}^G_{\text{GAN}}(\theta) = -\,\mathbb{E}_{t,z} \sum_{b \in \mathcal{B}} h^b\big(f^b(x_t)\big), \tag{2.7}$$

$$\begin{aligned}\mathcal{L}^D_{\text{GAN}}(\psi, \phi) = \mathbb{E}_{t,y} \sum_{b \in \mathcal{B}} \max\Big(0,\ 1 - h^b_\psi\big(f^b_\phi(y_t)\big)\Big) \\ + \mathbb{E}_{t,z} \sum_{b \in \mathcal{B}} \max\Big(0,\ 1 + h^b_\psi\big(f^b_\phi(x_t)\big)\Big).\end{aligned} \tag{2.8}$$

L'ajout de têtes de classification après chaque bloc encodeur permet à D d'opposer réel et *fake* à la fois aux échelles locales et globales, ce qui est particulièrement utile dans le régime few-shot $4 \leq |Y| \leq 6$, en limitant le surapprentissage et le *mode collapse*.

2.2.1 Personnalisation guidée par le sujet (*subject-driven personalization*, SDP)

Composants de `SDP`— La contribution propre à la SDP réside dans l'inclusion du conditionnement par *prompt* à chaque pas de temps. Le DMD à double domaine aligne désormais des distributions conditionnelles, guidant le *student* à préserver la sémantique générale de la classe (via $\epsilon^{\text{src}}(\cdot|c^{\text{prior}})$) tout en s'adaptant au sujet personnalisé (via $\epsilon^{\text{trg}}(\cdot|c)$). De même, le discriminateur GAN reçoit c afin d'encourager des détails réalistes spécifiques au sujet à plusieurs échelles de caractéristiques.

Conditionnement à partir de prompts de sujet— La SDP nécessite de conditionner le DM par des *prompts* textuels qui spécifient l'identité du sujet. Uni-DAD s'étend naturellement à ce cadre en incorporant l'information du *prompt* dans toutes les évaluations conditionnelles du

Algorithme 2.1 `Uni-DAD` entraînement itératif

Input : Source teacher ϵ^{src}, optional target teacher ϵ^{trg}, target set $Y = \{y\}$, weight factor a, training step $step$, update ratio $ratio$

Output : Student G (adapted and distilled)

$train\_target \leftarrow$ **False**
if $\epsilon^{trg} = \varnothing$ **then**
  $\epsilon^{\text{trg}} \leftarrow \epsilon^{\text{src}}$
  $train\_target \leftarrow$ **True**
end if

`// Prepare data`

$t \sim \mathcal{U}\{0.02T, 0.98T\}$
$(z, \epsilon) \sim \mathcal{N}(0, I)$
$y_t \leftarrow q\,(y_t \mid y),\ y \sim Y$ `// real`
$x_t \leftarrow q\,(x_t \mid x),\ x \leftarrow G\,(z)$ `// fake`

`// Student`

if *step % ratio == 0* **then**
  $\mathcal{L}_{\text{DMD}}^{\text{trg+src}} \leftarrow \text{DualDMD}(x_t, \epsilon, a)$ `// Eq 2.4`
  $\mathcal{L}_{\text{GAN}}^{G} \leftarrow \text{MhGAN}(x_t)$ `// Eq 2.7`
  $\mathcal{L}_G \leftarrow \mathcal{L}_{\text{DMD}}^{\text{trg+src}} + \lambda_{\text{GAN}}^{G}\,\mathcal{L}_{\text{GAN}}^{G}$ `// Eq 2.9`
  $G \leftarrow \text{update}(G, \mathcal{L}_G)$
end if

`// Fake teacher & discriminator`

$\mathcal{L}_{\text{fk}} \leftarrow \text{MSE}(\epsilon^{\text{fk}}(\text{stop\_grad}(x_t)),\ \epsilon)$ `// Eq 2.5`
$\mathcal{L}_{\text{GAN}}^{D} \leftarrow \text{MhGAN}(\text{stop\_grad}(x_t), y_t)$ `// Eq 2.8`
$\mathcal{L}_{\text{fk}+D} \leftarrow \mathcal{L}_{\text{fk}} + \lambda_{\text{GAN}}^{D}\,\mathcal{L}_{\text{GAN}}^{D}$ `// Eq 2.10`
$\epsilon^{\text{fk}} \leftarrow \text{update}(\epsilon^{\text{fk}}, \mathcal{L}_{\text{fk}+D})$

`// Target teacher`

if *step % ratio == 0 & train_target* **then**
  $\mathcal{L}_{\text{trg}} \leftarrow \text{MSE}(\epsilon^{\text{trg}}(y_t),\ \epsilon)$ `// Eq 2.6`
  $\epsilon^{\text{trg}} \leftarrow \text{update}(\epsilon^{\text{trg}}, \mathcal{L}_{\text{trg}})$
end if

score. Soit c' le *instance prompt*, généralement formulé comme « a [rare token] [class noun] », où le *rare token* identifie de manière unique le sujet (à partir de 4 à 6 images d'exemple), et où le *class noun* spécifie la catégorie sémantique plus générale (p. ex., « dog », « cat », « vase »). Ce *prompt* est fourni aux modèles qui sont appris, à savoir le *student* générateur G, le fake teacher ϵ^{fk} et le target teacher ϵ^{trg}. L'injection de c' garantit que ces modèles associent le *rare token* à l'identité du sujet en cours d'apprentissage.

Prompts *Class-Prior*— Afin de maintenir la généralité et de prévenir le surapprentissage, nous définissons en plus un *prior prompt* de classe c^{prior} = “a [class noun]”, qui est fourni à l’enseignant source figé ϵ^{src}. Comme ϵ^{src} n’a pas été entraîné sur le sujet spécifique, c^{prior} lui permet de produire un guidage cohérent avec la classe mais indépendant du sujet. Cette séparation des *prompts* est essentielle : ϵ^{src} continue d’agir comme un régulariseur favorisant la diversité, en stabilisant l’apprentissage de l’identité par la préservation d’une structure partagée au niveau de la classe.

2.2.2 Méthode expérimentale

En suivant le benchmark standard DreamBooth Ruiz *et al.* (2023), nous utilisons Stable Diffusion v1.5 (SDv1.5) Rombach *et al.* (2022), préentraîné sur LAION-5B Schuhmann *et al.* (2022), comme modèle source. Nous évaluons sur l’ensemble des instances du dataset DreamBooth et présentons une comparaison qualitative avec d’autres baselines sur trois instances représentatives : *cat2*, *dog6* et *vase*.

Chaque sujet est représenté par 4 à 6 images provenant du dataset DreamBooth Research (2023). Le dataset complet DreamBooth contient 30 sujets distincts (9 humains et 21 objets). Pour chaque instance cible, nous générons 100 échantillons personnalisés (25 prompts $\times$ 4 seeds), en suivant les templates de prompts de DreamBooth couvrant la recontextualisation, l’ajout d’accessoires et la modification d’attributs.

Les ensembles de prompts diffèrent pour les objets (par exemple `wolf plushie`, `vase`) et pour les sujets vivants (par exemple `cat2`, `dog6`), Fig. 2.4 et Fig. 2.5.

Comparaison des méthodes de référence— Nous comparons `Uni-DAD SDP` aux méthodes suivantes :

(i) DreamBooth *non distillé* Ruiz *et al.* (2023), utilisant un échantillonnage DDIM en 50 étapes, soit NFE = 2×50 pour l’inférence avec *classifier-free guidance* Ho & Salimans (2022) ;

```
Live Subject Prompts

prompt_list = [
'a {0} {1} in the jungle'.format(unique_token, class_token),
'a {0} {1} in the snow'.format(unique_token, class_token),
'a {0} {1} on the beach'.format(unique_token, class_token),
'a {0} {1} on a cobblestone street'.format(unique_token, class_token),
'a {0} {1} on top of pink fabric'.format(unique_token, class_token),
'a {0} {1} on top of a wooden floor'.format(unique_token, class_token),
'a {0} {1} with a city in the background'.format(unique_token, class_token),
'a {0} {1} with a mountain in the background'.format(unique_token, class_token),
'a {0} {1} with a blue house in the background'.format(unique_token, class_token),
'a {0} {1} on top of a purple rug in a forest'.format(unique_token, class_token),
'a {0} {1} wearing a red hat'.format(unique_token, class_token),
'a {0} {1} wearing a santa hat'.format(unique_token, class_token),
'a {0} {1} wearing a rainbow scarf'.format(unique_token, class_token),
'a {0} {1} wearing a black top hat and a monocle'.format(unique_token, class_token),
'a {0} {1} in a chef outfit'.format(unique_token, class_token),
'a {0} {1} in a firefighter outfit'.format(unique_token, class_token),
'a {0} {1} in a police outfit'.format(unique_token, class_token),
'a {0} {1} wearing pink glasses'.format(unique_token, class_token),
'a {0} {1} wearing a yellow shirt'.format(unique_token, class_token),
'a {0} {1} in a purple wizard outfit'.format(unique_token, class_token),
'a red {0} {1}'.format(unique_token, class_token),
'a purple {0} {1}'.format(unique_token, class_token),
'a shiny {0} {1}'.format(unique_token, class_token),
'a wet {0} {1}'.format(unique_token, class_token),
'a cube shaped {0} {1}'.format(unique_token, class_token)
]
```

Figure 2.4 Liste de prompts pour les instances de sujets vivants du benchmark DreamBooth
Tirée du code Ruiz *et al.* (2023)

(ii) PSO *distillé* Miao *et al.* (2024), qui ajuste un modèle Turbo distillé Chadebec *et al.* (2025) sur une architecture SDXL Podell *et al.* (2023) et échantillonne en 4 étapes (NFE = 4) ;

(iii) DMD2-DreamBooth *distillé* en deux étapes, où nous affinons un backbone SDv1.5 distillé avec DMD2 Yin *et al.* (2024a) à l'aide de l'objectif DreamBooth.

Comme PSO ne fournit que du code SDXL, notre comparaison utilise les résultats à résolution SDXL rapportés par les auteurs Miao *et al.* (2024). Cela introduit une différence de résolution et d'architecture (1024×1024 au lieu de 512×512 et SDXL au lieu de SDv1.5), mais constitue la comparaison la plus équitable disponible.

Métriques d'évaluation— En suivant Ruiz *et al.* (2023), la préservation de l'identité est mesurée à l'aide de la similarité DINO (ViT-S/16) Caron *et al.* (2021) et de la similarité cosinus CLIP-I

```
Object Prompts

prompt_list = [
'a {0} {1} in the jungle'.format(unique_token, class_token),
'a {0} {1} in the snow'.format(unique_token, class_token),
'a {0} {1} on the beach'.format(unique_token, class_token),
'a {0} {1} on a cobblestone street'.format(unique_token, class_token),
'a {0} {1} on top of pink fabric'.format(unique_token, class_token),
'a {0} {1} on top of a wooden floor'.format(unique_token, class_token),
'a {0} {1} with a city in the background'.format(unique_token, class_token),
'a {0} {1} with a mountain in the background'.format(unique_token, class_token),
'a {0} {1} with a blue house in the background'.format(unique_token, class_token),
'a {0} {1} on top of a purple rug in a forest'.format(unique_token, class_token),
'a {0} {1} with a wheat field in the background'.format(unique_token, class_token),
'a {0} {1} with a tree and autumn leaves in the background'.format(unique_token, class_token),
'a {0} {1} with the Eiffel Tower in the background'.format(unique_token, class_token),
'a {0} {1} floating on top of water'.format(unique_token, class_token),
'a {0} {1} floating in an ocean of milk'.format(unique_token, class_token),
'a {0} {1} on top of green grass with sunflowers around it'.format(unique_token, class_token),
'a {0} {1} on top of a mirror'.format(unique_token, class_token),
'a {0} {1} on top of the sidewalk in a crowded street'.format(unique_token, class_token),
'a {0} {1} on top of a dirt road'.format(unique_token, class_token),
'a {0} {1} on top of a white rug'.format(unique_token, class_token),
'a red {0} {1}'.format(unique_token, class_token),
'a purple {0} {1}'.format(unique_token, class_token),
'a shiny {0} {1}'.format(unique_token, class_token),
'a wet {0} {1}'.format(unique_token, class_token),
'a cube shaped {0} {1}'.format(unique_token, class_token)
]
```

Figure 2.5 Liste de prompts pour les instances d'objets du benchmark DreamBooth
Tirée du code Ruiz *et al.* (2023)

(ViT-B/32). L'alignement texte–image est quantifié à l'aide de CLIP-T (ViT-B/32) Radford *et al.* (2021).

Tous les modèles sont évalués à une résolution de 512×512, à l'exception de PSO. `Uni-DAD SDP` utilise une seule étape de débruitage (NFE=1), ce qui permet un échantillonnage significativement plus rapide que les baselines multi-étapes, PSO et DreamBooth.

2.2.3 Détails d'entraînement

Entraînement d'Uni-DAD SDP— La mise à jour d'entraînement du **student** équilibre la préservation du domaine source et l'adaptation au domaine cible en minimisant l'objectif DMD à double domaine ainsi que la perte GAN du générateur :

$$\mathcal{L}_G(\theta) = \mathcal{L}_{\text{DMD}}^{\text{trg+src}}(\theta) + \lambda_{\text{GAN}}^{G}\mathcal{L}_{\text{GAN}}^{G}(\theta), \tag{2.9}$$

La mise à jour d'entraînement du **fake teacher** combine sa perte MSE et la perte GAN du discriminateur :

$$\mathcal{L}_{\text{fk}+D}(\phi, \psi) = \mathcal{L}_{\text{fk}}(\phi) + \lambda_{\text{GAN}}^{D}\mathcal{L}_{\text{GAN}}^{D}(\psi, \phi). \tag{2.10}$$

L'entraînement de `Uni-DAD SDP` consiste à alterner la minimisation de trois pertes à chaque itération : $\mathcal{L}_{\text{G}}$, $\mathcal{L}_{\text{fk}+D}$ et $\mathcal{L}_{\text{trg}}$ (Alg. 2.1). En pratique, la minimisation de $\mathcal{L}_{\text{fk},D}$ est réalisée 5 à 10 fois Yin *et al.* (2024a) pour chaque mise à jour de $\mathcal{L}_{\text{G}}$ et $\mathcal{L}_{\text{trg}}$, afin de permettre à ϵ^{fk} de suivre l'évolution constante de la distribution de sortie de G.

Tous les modèles sont entraînés avec un taux d'apprentissage de 5×10^{-6}. Le GAN multi-têtes utilise des poids de discriminateur et de générateur de $\lambda_{\text{GAN}}^{D} = 0.01$ et $\lambda_{\text{GAN}}^{G} = 0.001$, respectivement. Le ratio de mise à jour pour G et ϵ^{trg} est fixé à 10.

Le générateur étudiant G est initialisé à partir des poids SDv1.5 pré-distillés avec DMD2 [2]. Nous fixons le facteur de pondération DMD à $a = 0.75$, correspondant à la configuration de domaine distant utilisée pour FSIG.

`Uni-DAD SDP` est entraîné pendant 5k itérations sur un seul GPU H100 (≈50 GB de mémoire utilisée), les meilleures générations apparaissant généralement entre 4k et 5k étapes. Le temps d'entraînement est d'environ 2.6 heures par sujet.

[2] Poids disponibles sur https ://github.com/tianweiy/DMD2

Tableau 2.1 Résumé des paramètres expérimentaux. DB désigne l'étape d'adaptation DreamBooth, et DMD2 l'étape de distillation.

Method	Backbone	NFE	Train Steps	Res.	CFG	LR_G	LR_D	Seeds	Batch	Hardware
DB (CVPR, 2023)	SDv1.5	2×50	800	512	–	5e-6	–	4-6	4	0.13h/RTX A6000
PSO_{SDXL} (ICLR,2025)	SDXL	4	800	512	–	2e-4	–	4-6	1	2.1h/A100 GPU
$PSO_{SDv1.5}$ (ICLR,2025)	SDv1.5	1	800	512	–	2e-4	–	4-6	1	2.1h/A100 GPU
DMD2-DB (NeurIPS, 2024)	SDv1.5	1	100+5k	512	7.5	5e-6	5e-6	4-6	4	0.01h/RTX A6000
DB-DMD2 (NeurIPS, 2024)	SDv1.5	1	5k+100	512	7.5	5e-6	5e-6	4-6	1	2h25/RTX A6000
Uni-D.	SDv1.5	1	4.5k-5k	512	7.5	5e-6	5e-6	4-6	1	1.6h/H100 80GB HBM3

Entraînement de DreamBooth— Pour la baseline DreamBooth, nous suivons l'entraînement avec préservation du prior en générant 1 000 échantillons de la forme "a [class noun]" avec SDv1.5. Nous effectuons un fine-tuning pendant 800 itérations par instance, avec un batch size de 1 et un taux d'apprentissage fixe de 5×10^{-6}. DreamBooth utilise généralement entre 400 et 1 200 étapes selon le sujet Kumari *et al.* (2023); Miao *et al.* (2024); Wei *et al.* (2023). Nous constatons que 800 étapes d'entraînement suffisent pour reproduire la qualité rapportée par les auteurs sur les différents sujets.

Entraînement de DMD2-DreamBooth— Pour construire cette baseline distillée en deux étapes, nous initialisons le pipeline DreamBooth avec des checkpoints SDv1.5 distillés par DMD2, puis ajustons le student avec l'objectif DreamBooth standard. Nous effectuons ensuite un échantillonnage en une étape conformément au modèle distillé DMD2.

Entraînement de DreamBooth-DMD2— Pour construire cette baseline distillée en deux étapes, nous initialisons le pipeline DMD2 avec des checkpoints DreamBooth SDv1.5 et ajustons le

student avec l'objectif DreamBooth standard. L'échantillonnage est ensuite réalisé en une étape selon la génération distillée DMD2.

Entraînement de PSO— PSO_{SDXL} Miao *et al.* (2024) repose sur l'architecture SDXL-Turbo et effectue l'échantillonnage en 4 étapes de débruitage. Nous utilisons la configuration officielle d'entraînement pour chaque sujet. Tous les modèles sont évalués après 800 étapes d'entraînement.

Dans un souci d'équité expérimentale, PSO est également entraîné et évalué avec une initialisation SDv1.5 en suivant une configuration d'hyperparamètres similaire à celle de PSO_{SDXL}.

2.3 Discussion des résultats

2.3.1 Évaluation qualitative

La Fig. 2.6 présente une comparaison qualitative sur trois sujets (dog6, cat2, vase). Pour l'ensemble des sujets, `Uni-DAD SDP` produit de manière cohérente des générations nettes et fidèles à l'identité, avec un fort alignement aux prompts, malgré l'utilisation d'une seule étape de débruitage (NFE = 1). En comparaison, DreamBooth, avec NFE = 2×50, présente souvent une dérive d'identité ou introduit des artefacts stylistiques indésirables. La baseline en deux étapes DMD2-DreamBooth est rapide avec NFE = 1, mais souffre d'un fort sur-lissage et d'une perte de détails. Cela confirme que l'approche *Distill-then-Adapt* sous la perte de diffusion originale dégrade la qualité de personnalisation.

PSO obtient un alignement texte–image raisonnable et préserve l'identité dans certains cas, mais produit fréquemment des textures sur-lissées ou « plastiques ». Cela se produit même si PSO utilise l'architecture plus puissante SDXL à une résolution plus élevée (1024×1024) et davantage d'étapes de débruitage (NFE = 4), ce qui rend la comparaison intrinsèquement biaisée en sa faveur. Parmi les méthodes basées sur SDv1.5, `Uni-DAD SDP` se distingue : il maintient la fidélité au prompt et la structure spécifique du sujet sous des changements difficiles de point de vue et d'arrière-plan, démontrant une forte capacité de généralisation malgré une efficacité d'échantillonnage extrême.

Figure 2.6 Comparaison qualitative de la personnalisation sur trois instances (*dog6*, *cat2*, *vase*)
Tirée de Bahram *et al.* (2025)

2.3.2 Évaluation quantitative générale

Le Tab. 2.2 présente une comparaison quantitative de `Uni-DAD SDP` avec les baselines SDP. Malgré un régime strict d'échantillonnage en une seule étape, `Uni-DAD SDP` obtient des scores de similarité d'identité et texte–image basés sur CLIP qui se rapprochent de ceux de DreamBooth multi-étapes, tout en surpassant largement son équivalent distillé en une étape, DMD2-DreamBooth. Ces résultats mettent en évidence l'efficacité de `Uni-DAD SDP` comme échantillonneur rapide et efficace pour la SDP.

Tableau 2.2 Résultats quantitatifs de DreamBooth pour la personnalisation SDP (D : DINO, CI : CLIP-I, CT : CLIP-T) - **Meilleure** et deuxième meilleure méthode distillée à NFE=1 - DMD2-DB obtient de meilleurs résultats en diversité, mais les générations présentent une dégradation marquée de la qualité et de la fidélité - DB-DMD2 montre de meilleurs résultats en similarité cosinus image-à-image (D, CI), mais les résultats qualitatifs indiquent que le modèle tend à mémoriser (voir Fig. 2.7) -`Uni-DAD SDP` apparaît comme la méthode distillée la plus stable et obtient globalement les meilleurs résultats parmi les approches distillées tout en restant compétitive par rapport à DreamBooth (DB) et PSO_{SDXL}

Method	NFE↓	D↑	CI↑	CT↑	Intra LPIPS↑	Inter LPIPS↑
DB (CVPR, 2023)	2×50	0.582	0.773	0.322	0.665 ± 0.079	0.727 ± 0.064
PSO_{SDXL} (ICLR, 2025)	4	0.496	0.696	0.300	0.417 ± 0.074	0.603 ± 0.080
$\text{PSO}_{\text{SDv1.5}}$ (ICLR, 2025)	1	0.136	0.564	0.225	0.069 ± 0.015	0.108 ± 0.023
DMD2-DB (NeurIPS, 2024)	1	0.198	0.607	0.259	$\mathbf{0.578 \pm 0.065}$	$\mathbf{0.700 \pm 0.088}$
DB-DMD2 (NeurIPS, 2024)	1	**0.567**	**0.753**	0.248	0.215 ± 0.044	0.249 ± 0.070
`Uni-DAD` (CVPR, 2026 ?)	1	0.472	0.731	**0.289**	0.507 ± 0.085	0.586 ± 0.089

Tableau 2.3 Comparaison par instance à l'aide des métriques DINO (D), CLIP-I (C-I) et CLIP-T (C-T) ↑ pour différentes méthodes de personnalisation - Le modèle Uni-DAD est entraîné entre 4k et 5k itérations - Les ID correspondent aux différentes instances - la correspondance entre les ID et les références d'instance est donnée dans le Tab. I-1

	Distilled models												Non-Distilled model		
IDs	**Uni-DAD SDP** (NFE **1**, SDv1.5)			**PSO** (NFE 4, SDXL)			**DMD2-DB** (NFE **1**, SDv1.5)			**DB-DMD2** (NFE **1**, SDv1.5)			**DB** (NFE 2×50, SDv1.5)		
	D	C-I	C-T	D	C-I	C-T	D	C-I	C-T	D	C-I	C-T	D	C-I	C-T
Live subjects															
11	0.685	0.834	0.285	0.658	0.801	0.314	0.107	0.583	0.259	0.804	0.850	0.254	0.658	0.807	0.314
12	0.723	0.850	0.266	0.609	0.795	0.310	0.124	0.578	0.255	0.758	0.737	0.203	0.609	0.795	0.310
13	0.504	0.757	0.296	0.328	0.673	0.318	0.031	0.519	0.256	0.613	0.789	0.245	0.470	0.740	0.317
14	0.517	0.695	0.275	0.619	0.757	0.305	0.071	0.522	0.255	0.674	0.781	0.237	0.619	0.757	0.305
15	0.665	0.785	0.261	0.835	0.860	0.288	0.110	0.565	0.257	0.820	0.870	0.229	0.642	0.809	0.300
16	0.593	0.806	0.291	0.590	0.809	0.308	0.084	0.549	0.256	0.674	0.836	0.236	0.590	0.809	0.308
17	0.435	0.737	0.284	0.627	0.819	0.305	0.081	0.546	0.257	0.639	0.767	0.249	0.627	0.819	0.305
7	0.686	0.818	0.259	0.665	0.811	0.300	0.708	0.826	0.298	0.764	0.841	0.243	0.665	0.811	0.300
8	0.584	0.762	0.280	0.627	0.768	0.333	0.196	0.592	0.261	0.164	0.583	0.204	0.695	0.826	0.301
Avg	0.599	0.783	0.277	0.569	0.758	0.305	0.162	0.585	0.257	0.657	0.784	0.233	0.664	0.808	0.309
Objects															
1	0.366	0.733	0.311	0.428	0.751	0.319	0.420	0.801	0.320	0.398	0.693	0.265	0.419	0.802	0.319
2	0.286	0.528	0.315	0.269	0.606	0.292	0.527	0.718	0.316	0.535	0.706	0.260	0.519	0.723	0.320
3	0.597	0.658	0.289	0.658	0.749	0.301	0.647	0.738	0.304	0.663	0.738	0.284	0.658	0.749	0.301
4	0.398	0.588	0.302	0.699	0.801	0.273	0.264	0.660	0.303	0.823	0.824	0.236	0.259	0.653	0.301
5	0.441	0.490	0.290	0.740	0.759	0.266	0.595	0.684	0.289	0.722	0.744	0.258	0.639	0.691	0.289
6	0.353	0.638	0.302	0.328	0.653	0.301	0.350	0.673	0.308	0.483	0.713	0.270	0.352	0.676	0.308
9	0.455	0.798	0.277	0.391	0.787	0.291	0.072	0.611	0.246	0.380	0.668	0.195	0.390	0.787	0.291
10	0.449	0.289	0.284	0.627	0.779	0.306	0.034	0.511	0.241	0.598	0.759	0.267	0.627	0.780	0.306
18	0.490	0.682	0.284	0.536	0.757	0.279	0.185	0.600	0.245	0.569	0.798	0.232	0.536	0.757	0.279
19	0.286	0.709	0.240	0.649	0.788	0.295	0.084	0.515	0.244	0.621	0.783	0.262	0.515	0.755	0.300
20	0.529	0.646	0.334	0.622	0.735	0.315	0.061	0.507	0.243	0.610	0.714	0.268	0.621	0.735	0.315
21	0.267	0.557	0.283	0.329	0.635	0.288	0.135	0.511	0.244	0.452	0.720	0.252	0.329	0.635	0.288
22	0.324	0.639	0.296	0.370	0.743	0.300	0.056	0.595	0.256	0.201	0.718	0.268	0.370	0.743	0.300
23	0.374	0.560	0.295	0.579	0.763	0.281	0.060	0.515	0.245	0.098	0.553	0.203	0.545	0.751	0.284
24	0.450	0.663	0.293	0.476	0.697	0.300	0.080	0.527	0.245	0.629	0.758	0.274	0.476	0.697	0.300
25	0.452	0.645	0.279	0.404	0.606	0.305	0.040	0.632	0.233	0.417	0.725	0.251	0.404	0.606	0.305
26	0.262	0.618	0.300	0.364	0.676	0.293	0.000	0.632	0.246	0.617	0.771	0.257	0.364	0.676	0.293
27	0.506	0.650	0.304	0.589	0.725	0.321	0.037	0.478	0.241	0.786	0.798	0.262	0.692	0.765	0.306
28	0.464	0.795	0.325	0.405	0.794	0.320	0.173	0.609	0.239	0.514	0.841	0.283	0.405	0.794	0.320
29	0.644	0.765	0.273	0.671	0.822	0.310	0.232	0.632	0.260	0.428	0.741	0.245	0.499	0.738	0.335
30	0.364	0.643	0.291	0.543	0.753	0.305	0.054	0.546	0.244	0.548	0.762	0.260	0.543	0.753	0.304
Avg	0.417	0.633	0.294	0.508	0.732	0.298	0.232	0.632	0.260	0.528	0.739	0.255	0.499	0.738	0.335
Avg	0.472	0.678	0.289	0.538	0.745	0.301	0.198	0.607	0.259	0.567	0.753	0.248	0.582	0.773	0.322

2.3.3 Évaluation sur la diveristé

La Fig. 2.7 et le Tab. 2.4 illustrent l'importance d'un cadre d'évaluation couvrant plusieurs aspects afin de fournir une vue d'ensemble complète des différentes méthodes. Dans le Tab. 2.4, la méthode *Distill-then-Adapt* DMD2-FT semble produire une plus grande diversité entre les prompts, avec **0.578** Intra-LPIPS et **0.700** Inter-LPIPS, comparé à *0.507* et 0.586 pour `Uni-DAD SDP`.

Cependant, comme mis en évidence en rouge dans la Fig. 2.7, les générations associées à ces scores élevés de diversité manquent de qualité et de fidélité. Cette observation est confirmée par le faible score DINO (0.198) rapporté pour DMD2-FT dans le Tab. 2.2. Par ailleurs, les résultats qualitatifs de la Fig. 2.7 suggèrent que le modèle FT-DMD2 mémorise largement l'instance cible, surapprenant l'une des images d'entraînement few-shot.

Tableau 2.4 Comparaison de la diversité à travers les prompts pour chaque instance (Intra-LPIPS) et entre les instances (Inter-LPIPS) ↑ - Chaque méthode est initialisée à partir de SDv1.5 à l'exception de PSO (SDXL) - DB désigne la méthode DreamBooth - **Meilleure** et deuxième meilleure méthode distillée à NFE=1

	Method	NFE ↓	Intra-LPIPS	Inter-LPIPS	Avg
Non-Distilled	**DreamBooth** (CVPR, 2023)	2×50	0.665 ±0.079	0.727 ±0.064	0.696
Distilled	**PSO**^{SDXL} (ICLR, 2025)	4	0.417 ±0.074	0.603 ±0.080	0.510
	DMD2-DB (NeurIPS, 2024)	1	**0.578** ±0.065	**0.700** ±0.088	**0.639**
	DB-DMD2 (NeurIPS, 2024)	1	0.215 ±0.044	0.249 ±0.070	0.232
	`Uni-DAD` (CVPR, 2026 ?)	1	0.507 ±0.085	0.586 ±0.089	0.546

Figure 2.7 Évaluation de la diversité au sein d'un même prompt (« a `prt` dog with a mountain background ») (Intra-LPIPS) et entre différents prompts (Inter-LPIPS) pour `Uni-DAD SDP` et les méthodes de référence comparatives - Toutes les méthodes sont initialisées à partir de Stable Diffusion préentraîné SDv1.5 à l'exception de PSO qui est initialisé avec SDXL-Turbo Podell *et al.* (2023) - DB désigne le modèle DreamBooth Ruiz *et al.* (2023) correspondant à l'étape d'adaptation et DMD2 Yin *et al.* (2024b) l'étape de distillation - DMD2-DB correspond à *Distill-then-Adapt*- tandis que DB-DMD2 correspond à *Adapt-then-Distill*
Tirée de Bahram *et al.* (2025)

Bien que `Uni-DAD SDP` arrive en deuxième position pour chaque prompt, il présente la plus grande stabilité et la meilleure performance globale parmi les modèles distillés. Il constitue également une baseline compétitive par rapport à DreamBooth (FT) tout en nécessitant 100 évaluations de fonction neuronale en moins, avec une adhérence plus forte à l'instance cible observable (Fig. 2.7).

De plus, `Uni-DAD SDP` surpasse PSOSDXL en diversité globale sur l'ensemble du benchmark DreamBooth Ruiz *et al.* (2023) (Tab. 2.4), tout en introduisant visuellement une fidélité et une

qualité supérieures à PSOSDXL, malgré une initialisation à partir d'un modèle préentraîné plus petit (SDv1.5 < SDXL) et l'utilisation de 4× moins d'évaluations de fonction neuronale. Globalement, le pipeline unifié en une seule étape surpasse les modèles distillés concurrents tout en restant compétitif face à des baselines plus lourdes.

2.3.4 Analyse sur la variation du poids entre le domaine source et cible

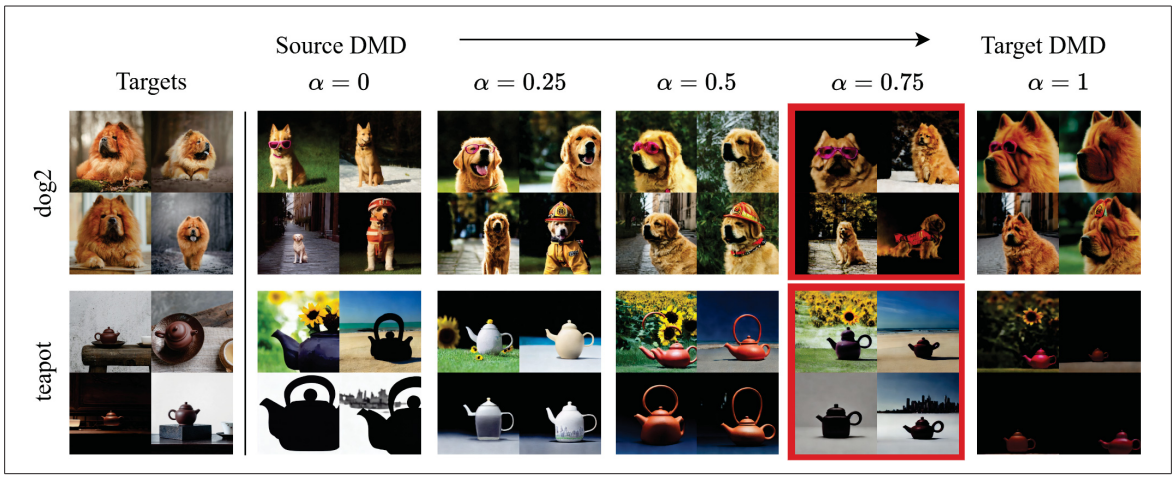

Figure 2.8 Ablation qualitative du coefficient α pour la SDP sur SDv1.5 à travers différents prompts pour un sujet (dog2) et un objet (teapot) - Avec $\alpha = 0$ correspondant à l'absence d'assistance du target teacher ϵ^{trg} dans le DMD et $\alpha = 1$ correspondant à l'absence d'assistance du source teacher ϵ^{src} lors de l'entraînement vers la distribution cible
Tirée de Bahram *et al.* (2025)

Tableau 2.5 Résultats quantitatifs pour les instances d'objets et les sujets vivants selon différentes variations du poids α où α désigne le poids attribué à l'enseignant cible dans la rétropropagation dual-DMD

	α	CLIP-I↑	CLIP-T↑	DINO↑	Intra-LPIPS↑	Inter-LPIPS↑	Avg LPIPS↑
Objects	0	0.652	**0.260**	0.304	0.509	0.595	0.552
	0.25	0.656	0.255	0.294	**0.529**	0.605	0.567
	0.5	0.661	0.256	**0.363**	0.526	**0.618**	**0.572**
	0.75	**0.674**	0.257	0.360	0.498	0.587	0.542
	1	0.657	0.258	0.330	0.466	0.545	0.506
Live subjects	0	0.771	0.247	0.380	**0.530**	**0.570**	**0.550**
	0.25	0.794	0.248	0.447	0.487	0.514	0.500
	0.5	**0.798**	**0.250**	0.485	0.455	0.494	0.475
	0.75	0.789	**0.250**	0.539	0.469	0.518	0.493
	1	0.785	0.249	**0.580**	0.472	0.497	0.484

Pour le coefficient $\boldsymbol{\alpha}$, des analyses qualitative (Fig. 2.8) et quantitative (Tab. 2.5) sont menées afin d'évaluer la sensibilité de cet hyperparamètre en fonction de la distance entre les domaines source et cible. Le paramètre α pondère l'équilibre entre l'alignement avec la cible (CLIP-I et DINO) et la diversité issue du domaine source (LPIPS), tout en conservant suffisamment d'information du domaine source pour permettre l'alignement avec les prompts (CLIP-T).

Dans ce benchmark, les instances d'objets et les sujets vivants peuvent être considérés respectivement comme des domaines *éloignés* et *plus proches* : certaines espèces de chats ou de chiens ont davantage de références dans LAION-5B Schuhmann *et al.* (2022) que des objets rares possédant des attributs distinctifs (par exemple une peluche de loup spécifique, une bougie ou une théière particulière), qui peuvent représenter un domaine plus nouveau.

Pour une conclusion complète, les analyses qualitative et quantitative doivent être considérées conjointement. Pour les instances d'objets, le Tab. 2.5 indique un fort alignement avec la cible pour des valeurs de α comprises entre 0.5 et 0.75. De manière cohérente, la Fig. 2.8 montre que la théière est visuellement la mieux alignée avec la cible lorsque $\alpha = 0.75$. Bien que le score LPIPS le plus élevé soit obtenu pour $\alpha = 0.5$, la Fig. 2.8 suggère une meilleure fidélité, diversité et un meilleur alignement aux prompts pour $\alpha = 0.75$.

Pour les sujets vivants, l'analyse mène à une conclusion similaire. La Fig. 2.8 montre que $\alpha = 1$ produit un fort alignement avec l'instance pour ce chien spécifique ; cela se reflète également dans le Tab. 2.5 par la valeur DINO la plus élevée (0.580). Cependant, les générations semblent surapprendre une pose particulière et présentent une diversité réduite, ce qui est cohérent avec les scores LPIPS plus faibles. Bien qu'un fort alignement CLIP soit observé pour $\alpha = 0.5$, la Fig. 2.8 indique que le chien généré ne correspond pas à l'espèce cible. En revanche, $\alpha = 0.75$ fournit un compromis plus équilibré entre l'alignement avec l'instance et la diversité, ainsi qu'un meilleur alignement avec les prompts, avec une valeur CLIP-T plus élevée de 0.250 et une forte diversité à l'intérieur de chaque prompt (Inter-LPIPS de 0.518).

Dans l'ensemble, ces résultats suggèrent que la méthode est raisonnablement sensible au coefficient de pondération dual-domain α, qui contrôle l'équilibre entre spécialisation sur la cible et régularisation par le domaine source. Bien que la valeur optimale varie selon les instances, l'analyse qualitative et quantitative conjointe soutient un réglage global pratique de $\alpha = 0.75$, offrant un compromis robuste entre fidélité d'identité, diversité et alignement aux prompts sans nécessiter d'ajustement spécifique par instance. De plus, dans l'article `Uni-DAD`, cette analyse est étendue à d'autres datasets, et le pipeline obtient systématiquement ses meilleures performances avec cette même configuration de α.

2.4 Analyse critique

Compromis efficacité–qualité dans `Uni-DAD SDP`— La principale force de `Uni-DAD SDP` est qu'il obtient un alignement texte compétitif tout en préservant l'identité de la cible sous la contrainte extrême d'échantillonnage NFE = 1. Notre contribution vise des pipelines de diffusion personnalisés qui permettent non seulement un échantillonnage efficace à l'inférence (1 NFE), mais qui concentrent également l'entraînement sur des générations de haute fidélité. Cependant, le régime en une seule étape rend aussi la méthode plus sensible à la stabilité de l'entraînement et à la capacité de l'étudiant à préserver les détails spécifiques à la cible tout en généralisant à une large gamme de prompts divers et créatifs. Ce compromis motive l'utilisation d'une distillation à double domaine, avec des signaux contrôlables provenant des domaines cible

et source, combinée à la contrainte GAN multi-têtes qui guide les générations de l'étudiant vers le domaine cible.

Pourquoi le routage des prompts est important pour la personnalisation— La séparation entre les *instance prompts* (c') et les *class-prior prompts* (c^{prior}) joue un rôle pratique dans l'apprentissage du domaine cible, en couplant la nouvelle distribution avec le domaine sémantique spécifique de la classe. L'utilisation de c^{prior} pour le *source teacher* figé aide à préserver une diversité cohérente avec la classe (poses, arrière-plans, attributs), tandis que l'utilisation de c' ancre l'identité du sujet dans les modèles appris grâce au *rare token* (par exemple « prt ») et à la distribution cible. En pratique, fusionner ces prompts (en fournissant c' partout) risque d'affaiblir le *source teacher* en tant que régulariseur de diversité, tandis que fournir uniquement c^{prior} partout peut affaiblir la préservation de l'identité du sujet personnalisé. Ainsi, ce routage des prompts constitue un élément clé pour la SDP et un choix de conception adopté par plusieurs baselines SDP Ruiz *et al.* (2023); Wei *et al.* (2023); Miao *et al.* (2024).

Sur-lissage et dérive d'identité comme modes d'échec des pipelines en deux étapes— Les résultats par instance dans le Tab. 2.3 et l'analyse qualitative sur trois instances dans la Fig. 2.6 reflètent des modes d'échec fréquents des pipelines en deux étapes combinant adaptation et distillation. D'une part, *Distill-then-Adapt* tend à produire des générations sur-lissées dans le régime few-shot, réduisant les indices d'identité à haute fréquence et produisant des textures floues. D'autre part, *Adapt-then-Distill* peut préserver l'apparence cible mais risque de se concentrer sur un seul exemple cible, limitant la diversité et réduisant la sensibilité aux prompts. À l'inverse, l'approche proposée en une seule étape vise à réduire ces deux effets en préservant conjointement la diversité du domaine source (via ϵ^{src}) et en imposant le réalisme du domaine cible (via ϵ^{trg} et l'objectif GAN).

Sensibilité au poids dual-domain α— Le poids dual-domain α contrôle l'équilibre qualité–diversité, particulièrement visible dans le contexte SDP. Des valeurs plus faibles de α augmentent la dépendance au *source teacher* ϵ^{src}, améliorant généralement la diversité et réduisant l'effondrement, mais pouvant freiner l'adaptation lorsque la cible est structurellement éloignée

du manifold source, notamment pour les détails à haute fréquence. Des valeurs plus élevées de α imposent un alignement plus fort avec la cible, ce qui peut améliorer la fidélité d'identité mais augmenter le surapprentissage, réduisant ainsi la diversité et la sensibilité aux prompts. Bien que les valeurs optimales puissent varier selon les instances, le pipeline montre que dans la plupart des cas un α global avec 0.75 de poids pour la cible et 0.25 pour la source permet d'adapter l'identité cible tout en conservant la diversité du domaine source. Ainsi, un α global unique est retenu afin d'éviter un ajustement spécifique à chaque cible, privilégiant une configuration équilibrée, pratique et reproductible plutôt qu'une optimisation instance par instance.

Limites des métriques pour la SDP— Les similarités basées sur CLIP (CLIP-I, CLIP-T) et la similarité DINO constituent des mesures utiles, mais elles ne capturent pas parfaitement la préservation d'identité fine ni les détails à haute fréquence (par exemple des marques subtiles, des formes spécifiques ou des textures), ni la correction d'attributs spécifiques aux prompts. Une limite importante provient du contexte few-shot : avec seulement un petit nombre d'images de référence, l'évaluation ne peut être que partiellement fiable. De plus, de nombreuses métriques d'évaluation, telles que FID, nécessitent des volumes de données plus importants pour fournir des conclusions significatives. Comme le benchmark utilisé est conçu pour un contexte few-shot, la comparaison de 100 générations avec seulement 4 à 6 images par dataset peut influencer les résultats. De même, les scores de diversité LPIPS peuvent être influencés par des variations d'arrière-plan qui ne sont pas nécessairement sémantiquement pertinentes. Par conséquent, bien que les métriques rapportées soutiennent les tendances principales, l'inspection qualitative demeure importante pour diagnostiquer les cas d'échec dans l'utilisation pratique de la SDP.

Validité et équité des baselines— Une limite de la comparaison est que certaines baselines utilisent des architectures ou des résolutions différentes (par exemple SDXL vs. SDv1.5), ce qui peut biaiser les résultats en leur faveur. Nous atténuons cet effet en signalant explicitement ces différences et en privilégiant les comparaisons utilisant la même architecture lorsque cela est possible. Une autre source de variance dans la SDP provient du choix des prompts et des seeds aléatoires ; nous réduisons cette variance en utilisant un benchmark validé avec des templates de prompts fixes et en évaluant plusieurs seeds par prompt.

Limites pratiques et pistes d'amélioration— Bien que le temps d'entraînement par sujet reste raisonnable, le maintien d'un *target teacher* en ligne augmente la complexité de l'entraînement. Une extension naturelle consiste à réduire le coût d'entraînement sans recourir à un *target teacher* supplémentaire ϵ^{trg}. Pour atteindre une adaptation optimale, une stratégie efficace pourrait consister à contrôler le champ réceptif effectif pendant l'adaptation, ce qui pourrait réduire le coût computationnel en permettant une recherche optimisée dans des dimensions plus faibles. De telles optimisations pourraient réduire le coût de calcul et permettre ainsi une personnalisation à plus haute résolution avec des architectures préentraînées plus puissantes. D'autres améliorations pourraient inclure des contraintes plus fortes de préservation de l'identité pour les *rare tokens* ainsi que des mécanismes de robustesse aux prompts afin de réduire la sensibilité aux prompts de recontextualisation qui modifient fortement la distribution d'arrière-plan.

2.5 Conclusion

Ce chapitre présente `Uni-DAD SDP`, une extension du cadre `Uni-DAD` orientée SDP qui unifie distillation et adaptation sous une contrainte d'échantillonnage extrême. En combinant un DMD à double domaine avec un objectif GAN multi-têtes, `Uni-DAD SDP` entraîne un générateur étudiant rapide qui préserve la diversité issue du domaine source tout en imposant le réalisme du domaine cible, permettant ainsi la génération d'images personnalisées avec une seule étape de débruitage à l'inférence (NFE = 1). Un composant clé spécifique à la SDP est le routage des prompts : les *instance prompts* sont utilisés pour les modèles appris (student, fake teacher et target teacher) afin d'associer le *rare token* à l'identité du sujet, tandis que les *class-prior prompts* sont utilisés pour le source teacher figé afin de conserver une diversité cohérente avec la classe et de favoriser la généralisation aux prompts.

Le chapitre met également en place une reproduction du benchmark DreamBooth ainsi qu'un pipeline d'évaluation permettant d'assurer des comparaisons équitables et reproductibles entre les baselines, incluant un rapport par instance sur l'ensemble des sujets et des métriques complémentaires capturant la préservation de l'identité (DINO, CLIP-I), l'alignement texte–image (CLIP-T) et la diversité (Intra/Inter-LPIPS). À travers les évaluations qualitatives et quantitatives,

`Uni-DAD SDP` démontre une forte fidélité d'identité et une bonne adhérence aux prompts tout en restant compétitif avec DreamBooth non distillé à plusieurs étapes et en surpassant les alternatives distillées et en deux étapes en termes de compromis global entre qualité et diversité. En particulier, les résultats mettent en évidence des modes d'échec fréquents des pipelines en deux étapes, avec un fort sur-lissage dans *Distill-then-Adapt* et un effondrement de la diversité ou une mémorisation dans *Adapt-then-Distill*. Cette comparaison globale montre que l'approche unifiée en une seule étape atténue ces problèmes en préservant conjointement l'information issue du domaine source et en imposant le réalisme du domaine cible durant l'entraînement.

Dans l'ensemble, ce chapitre démontre que l'unification de la distillation et de l'adaptation constitue une direction pratique et efficace pour une SDP rapide, permettant une génération personnalisée de haute qualité en contexte few-shot tout en réduisant significativement le coût d'inférence. Des travaux futurs pourraient viser à réduire davantage la complexité de l'entraînement (par exemple en simplifiant ou en initialisant efficacement le target teacher), améliorer la robustesse face à des décalages de domaine plus importants et renforcer la préservation de l'identité pour les *rare tokens* sous des prompts particulièrement difficiles.

CHAPITRE 3

GÉNÉRATION DE VARIATIONS D'EFFETS SONORES POUR PROTOTYPES INDUSTRIALISÉ

Ce chapitre s'intéresse à la problématique du développement d'un système capable de répondre aux exigences de l'industrie du jeu vidéo pour la génération contrôlée de variations d'effets sonores (SFX). Dans les flux de travail industriels de design sonore et les systèmes orientés production, un pipeline génératif utile doit non seulement produire un audio réaliste et de haute fidélité, mais aussi préserver l'identité perceptuelle de la référence, permettre des variations contrôlées et rester suffisamment efficace pour être déployé à grande échelle. Or, ces exigences ne sont encore que partiellement prises en compte par les méthodes actuelles de génération audio, souvent évaluées dans des cadres plus larges de génération text-to-audio (TTA) ou de synthèse inconditionnelle, plutôt que dans des scénarios de production de SFX conditionnés par une référence.

Afin de combler cet écart, ce chapitre propose une vue d'ensemble structurée ainsi qu'une analyse comparative des méthodes récentes de l'état de l'art en génération et en édition audio pouvant répondre à la variation des SFXs. En mettant en perspective les avancées récentes en audio génératif avec les contraintes concrètes de la production sonore, il établit à la fois un cadre méthodologique et développe un benchmark expérimental permettant une comparaison systématique de ces approches. L'objectif est ainsi d'identifier les méthodes les plus adaptées à des pipelines prêts pour la production, tout en mettant en évidence leurs principales limites et modes d'échec, qui constituent de nombreux défis de recherche ouverts à l'intersection de la modélisation générative audio et du design sonore industriel.

3.1 Problématique, objectifs et protocole d'évaluation pour la comparaison des méthodes de génération de variations de SFX

Dans le domaine de la génération audio, les jeux de données de SFX disponibles sont souvent limités, hétérogènes et faiblement annotés (Gemmeke *et al.*, 2017; Kim, Kim, Lee & Kim, 2019; Chen, Xie, Vedaldi & Zisserman, 2020). En conséquence, la mise en place d'un pipeline

complet de génération audio nécessite fréquemment de combiner plusieurs jeux de données, ainsi qu'un prétraitement important visant à normaliser, catégoriser et nettoyer les clips audio avant l'entraînement ou l'évaluation (Liu *et al.*, 2023a; Luo, Yan, Hu & Zhao, 2023d).

Dans les contextes de production de SFX, les clips audio sont généralement organisés en *soundbanks*, au sein desquelles chaque classe sémantique ne peut contenir qu'un nombre restreint de références réutilisables. Cette faible variabilité peut entraîner des répétitions perceptibles lorsque les mêmes sons sont mobilisés à plusieurs reprises dans des médias interactifs, tandis que l'édition manuelle et la création de variations demeurent longues et coûteuses. Afin de réduire cette charge de conception et d'atténuer le manque de diversité dans la création de sons Foley et environnementaux, plusieurs travaux récents se sont orientés vers la génération de variations à partir d'un clip audio de référence (Chung, Lee & Nam, 2024; Zhang *et al.*, 2025b; Fang *et al.*, 2025; Liang *et al.*, 2025; Liu *et al.*, 2025a), plutôt que vers la synthèse inconditionnelle (Zhu, Wen, Carbonneau & Duan, 2023b) ou la génération TTA standard (Evans *et al.*, 2024; Liu *et al.*, 2023d; Kreuk *et al.*, 2022b; Majumder *et al.*, 2024; Huang *et al.*, 2023a; Ziv *et al.*, 2024).

Par ailleurs, les cadres d'évaluation existants dans les benchmarks de génération audio se concentrent principalement sur la synthèse inconditionnelle ou conditionnée par le texte. Ils renseignent donc peu sur la capacité réelle des modèles à produire, à partir d'un son de référence, des variations à la fois contrôlées, crédibles et utiles dans un cadre de production.

Ainsi, le défi central abordé dans ce chapitre consiste à identifier un pipeline génératif capable de produire, à partir d'un ou de quelques sons de référence, des variations à la fois diversifiées et de haute qualité, tout en préservant l'identité perceptuelle de l'événement sonore d'origine (par exemple : pas, claquement de porte, rire). Un tel pipeline doit également permettre un contrôle optionnel par l'utilisateur pour un usage en production, notamment sur la durée, l'enveloppe d'énergie, les contraintes d'alignement ou encore la structure temporelle.

Comme première étape vers cet objectif, ce chapitre propose une vue d'ensemble des méthodes de l'état de l'art (SoTA) sélectionnées pour l'édition audio (plus de détails dans la sous-section 1.3.3 du Chapitre 1). Il présente en parallèle le développement d'un protocole d'adaptation par

fine-tuning, appliqué à chacune de ces méthodes sur une *soundbank* open source restreinte, spécifiquement retenue pour la génération de SFX. Ce cadre permet d'évaluer, dans une même application cible, la capacité de chaque approche à produire des variantes de SFX sous des contraintes réalistes de production. Dans cette perspective, une méthodologie d'évaluation complète a été mise en place afin d'analyser les méthodes de manière aussi équitable que possible [1], tout en tenant compte des spécificités propres à chaque modèle.

3.2 Cadre méthodologique pour la comparaison des méthodes de génération de variations de SFX

Cette section présente le cadre méthodologique adopté dans l'ensemble de ce chapitre pour analyser la génération de variations de SFX conditionnée par une référence dans une perspective de production. L'objectif n'est pas de considérer les méthodes de génération audio de manière isolée, mais d'identifier quels choix de conception sont les plus pertinents pour répondre aux contraintes concrètes de production, notamment en matière de fidélité perceptuelle, de préservation de l'identité, de contrôlabilité, de diversité et d'efficacité computationnelle.

Dans ce contexte, évaluer si un audio généré est à la fois diversifié et aligné avec l'identité de la référence revient à examiner un compromis entre deux exigences complémentaires. La première consiste à vérifier que le son généré correspond toujours au même type d'événement que la référence, en conservant ses caractéristiques essentielles, telles que la structure centrale de l'événement, l'enveloppe temporelle ou la signature timbrale. La seconde consiste à s'assurer qu'il introduit des variations perceptuellement significatives au sein de cette même identité, plutôt que de simplement reproduire la référence ou de dériver vers une autre classe sonore.

À un niveau plus fondamental, cette analyse prend en compte à la fois la représentation choisie pour modéliser l'audio et la famille de modèles génératifs utilisée pour le manipuler. Elle cherche

[1] Les méthodes évaluées dans ce chapitre reposent sur différentes représentations du signal (forme d'onde, spectrogramme ou espace latent) et utilisent des données de pré-entraînement ainsi que des paramétrisations hétérogènes, ce qui peut introduire un biais de représentation. Bien que le protocole expérimental vise à promouvoir l'équité autant que possible, une comparaison parfaitement équitable ne peut être garantie. Pour atténuer ces effets, chaque méthode est *fine-tunée*, évaluée et comparée avec une attention particulière à ces aspects.

ainsi à déterminer, d'une part, quelles représentations audio capturent le mieux les structures temporelles et spectrales fines et, d'autre part, quelles familles de modèles offrent le meilleur compromis entre réalisme, contrôlabilité, préservation de l'identité et fidélité globale.

Le cadre comparatif retenu se concentre sur la génération de sons environnementaux et non vocaux dans un contexte de génération *audio-to-audio* (ATA) conditionnée par une référence. Dans ce cadre, chaque modèle reçoit en entrée un son de référence, accompagné d'un conditionnement minimal par classe, et doit produire plusieurs variantes plausibles pouvant rester utiles dans un contexte de production.

L'objectif de cette méthodologie est donc double. Il s'agit, d'une part, de définir la terminologie et les exigences pertinentes pour la génération de variations de SFX et, d'autre part, de justifier la sélection des méthodes de référence ainsi que des critères d'évaluation retenus pour l'analyse comparative.

Cette section introduit d'abord la terminologie principale utilisée pour positionner les méthodes sélectionnées, formalise ensuite les exigences de production, puis motive enfin le choix des méthodes de référence au regard de ces exigences.

3.2.1 Terminologie en édition audio SFX

Afin de catégoriser les méthodes étudiées dans ce chapitre, certaines terminologies clés sont d'abord distinguées.

Génération audio—Dans les modèles de base les plus répandus, la **génération audio** peut être définie comme la production d'échantillons audio conditionnés par des descriptions textuelles (Kreuk *et al.*, 2022a). De manière similaire, (Liu *et al.*, 2023c) présente un cadre holistique et unifié pour la génération de parole, de musique et de SFX à travers une représentation partagée appelée *Language of Audio* (LOA) et une modélisation générative conditionnée par cette représentation. Pour la génération TTA, (Evans *et al.*, 2024) décrit explicitement un modèle open-source entraîné pour générer des échantillons audio et des SFX. De manière générale, la

génération audio est utilisée de manière interchangeable pour produire de l'audio à partir de différentes modalités : vidéo (Xu *et al.*, 2024a; Iashin & Rahtu, 2021; Wang *et al.*, 2025c), texte et vidéo (Cheng *et al.*, 2024; Shi *et al.*, 2025; Zhao *et al.*, 2025; Kushwaha & Tian, 2024), audio et texte (Park *et al.*, 2025; Xu *et al.*, 2024b; Cai, Huang, Zhang, Chen & Zhang, 2023; Zhang *et al.*, 2025b; Chen *et al.*, 2024b), ainsi que dans des contextes multimodaux plus généraux (Wu *et al.*, 2024; Yang *et al.*, 2024b; Liu *et al.*, 2025c; Fang *et al.*, 2025).

Édition audio—Une application de la génération audio est l'**édition audio**. (Cai *et al.*, 2023) propose une description concrète de l'édition audio à travers différents cas d'usage tels que l'ajout de SFX en arrière-plan, le remplacement d'un instrument ou la réparation d'un enregistrement endommagé, et formalise l'édition comme l'apprentissage d'une fonction reliant des instructions et un audio d'entrée à une sortie audio modifiée. D'autres travaux se concentrent sur des modifications localisées : Park *et al.* (2025) modifie des régions spécifiques du domaine temps–fréquence tout en conservant le reste du signal intact, en opérant sur des spectrogrammes et en utilisant un audio de référence pour cibler un contenu ou un style. L'édition audio inclut également des tâches telles que l'*inpainting* (Liu *et al.*, 2023c; Cai *et al.*, 2023; Cífka *et al.*, 2025), les opérations au niveau de la séquence (ajout, suppression, déplacement et remplacement) (Zhang *et al.*, 2025b; Park *et al.*, 2025), l'alignement temporel avec des séquences vidéo (Luo *et al.*, 2023c; Wang *et al.*, 2025c), ou encore le transfert de style ou de timbre (Fang *et al.*, 2025).

Variation audio—En complément, ce chapitre utilise le terme **audio variation**, étroitement lié à l'**audio editing**, mais avec une emphase spécifique. Ici, la **variation audio** peut être considérée comme une forme d'adaptation de domaine permettant d'apprendre l'identité sonore et les caractéristiques d'un événement (par exemple des pas, un claquement de porte ou un rire) tout en généralisant vers des versions plausibles et diversifiées (par exemple des variations de durée, d'énergie, de structure temporelle ou de contraintes d'alignement).

3.2.2 Exigences de production

L'objectif est de générer plusieurs variantes de sortie qui restent pertinentes pour la production selon les exigences résumées dans la Table 3.1 : (i) préserver l'identité de l'événement de la référence (R4), ce qui signifie que les échantillons générés doivent appartenir à la même classe perceptuelle et sémantique ; (ii) maintenir la structure temporelle lorsque cela est requis (R6), notamment les instants d'attaque et la dynamique de l'enveloppe ; (iii) introduire une variabilité contrôlée dans le timbre, la texture, l'intensité ou la coloration environnementale (R5–R7) sans provoquer de dérive sémantique ; (iv) fournir un contrôle explicite (R8) à travers des paramètres explicites (par exemple des poids de conditionnement) permettant de produire des variations prévisibles à partir d'une référence audio.

Conformément aux travaux récents en modélisation audio générative (R1), l'analyse adopte les mel-spectrograms comme représentation principale, tout en reconnaissant que la littérature plus large utilise également des spectrogrammes complexes et des espaces latents audio, chacun présentant des compromis différents en termes de fidélité, réalisme et contrôlabilité.

Tableau 3.1 Exigences relatives à la génération de variations de SFX prêtes pour la production - Chaque exigence est mise en relation avec des éléments observables et des modes d'échec diagnostiques afin de permettre une comparaison systématique des méthodes malgré l'hétérogénéité des benchmarks

ID	Requirement	Operational Def.	Associated Metrics	Failure modes
R1	Architecture & Audio Representation	Diffusion/FM backbone, self supervision, latent space or complex spectogram domain	NA (Design choice)	Text-only control, poor transient details
R2	High Fidelity	Preserving transient and spectral details, no distortions	FAD↓, MOS↑	transient smearing, noise, phasiness
R3	Realism	Real-world SFX	FAD↓, MOS↑	Synthetic texture, unatural reverb., over-regularised sound
R4	Identity Preservation	Same event class and perceptual identity class	SMOS↑, ImageBind↑,	Semantic drift, collapse to generic noise
R5	Diversity without drift	variation in texture/environment, not event type	IS↑	duplicate, diversity with identity drift
R6	Temporal Alignment	Alignment to target/ explicit timing control (onsets, envelope shape, event order)	Onset deviation (ms)↓; envelope/energy alignment (corr↑/ DTW↓)	shifted onset, stretched or compressed
R7	Energy control	Output matches target loudness/energy (scalar or curve)	energy-curve MSE↓	drift/distortion
R8	Controllability	Strength weight	IS↑, FAD↓	drifts
R9	Targeted modification	Specific segments/attributes modification	Outside/inside-target alignment	global repainting, boundary discontinuities
R10	Robustness and Stability	Stable across seeds with low failure rate	Variance↓	
R11	Efficiency	Runtime, memory, steps	Inference Time↓	Excessive latency; high VRAM; slow sampling

3.3 Comparaison des méthodes

3.3.1 Matrice des capacités au regard des exigences de production

La Table 3.2 présente la matrice de capacités des méthodes de référence sélectionnées et indique dans quelle mesure chacune d'elles répond aux exigences de production résumées dans la Table 3.1. Cette vue synthétique permet d'identifier rapidement les forces, les limites et les compromis associés aux différentes approches.

3.3.2 Justification et description des méthodes de référence sélectionnées

Au-delà de cette vue d'ensemble, cette sous-section présente plus finement les méthodes de référence retenues afin de justifier leur sélection et de préciser leur positionnement au regard des exigences de production. Il ne s'agit pas de répéter la matrice, mais d'en expliciter les principales implications pour l'application de génération de variations de SFX. Une présentation plus détaillée de chaque méthode est par ailleurs disponible dans la Section 1.3.3 du chapitre 1.

AudioLDM—AudioLDM est un système TTA basé sur une architecture de diffusion latente et un conditionnement CLAP (Liu *et al.*, 2023b). Le modèle est préentraîné sur AudioSet, AudioCaps, Freesound et des datasets BBC SFX. Le pipeline prend en charge l'édition audio via des manipulations guidées par texte (par exemple le transfert de style). Dans notre cadre expérimental, nous limitons l'entrée textuelle à l'étiquette de classe de la référence (par exemple footsteps, laughing, crow) et utilisons l'application de transfert de style pour générer des variations conditionnées par une référence. En outre, AudioLDM propose un protocole d'évaluation largement utilisé pour la modélisation audio générative à travers le benchmark AudioLDM-eval (Liu, 2023), que nous adoptons pour calculer la Fréchet Audio Distance (FAD) comme mesure objective de la qualité et du réalisme audio (R2, R3) (Table 3.2).

T-Foley—T-Foley est un modèle de diffusion sur forme d'onde guidé par des événements temporels pour la synthèse de sons Foley. Il est préentraîné sur sept classes Foley (DogBark, Footstep, GunShot, Keyboard, MovingMotorVehicle, Rain et Sneeze_Cough). À l'inférence, le

modèle est conditionné par une étiquette de classe sonore et une caractéristique d'événement temporel (par exemple une enveloppe d'énergie RMS, un signal d'attaque ou une courbe de puissance). Dans notre configuration, nous extrayons l'enveloppe RMS du clip de référence et l'utilisons comme condition d'événement, tout en utilisant l'étiquette de classe de référence comme conditionnement, puis nous générons plusieurs sorties à partir de différentes seeds de bruit. Cela fournit un contrôle du timing et de l'énergie guidé par la référence pour la génération de variations, mais ne permet pas d'éditions locales ciblées d'une région spécifique du signal, puisque la référence est utilisée uniquement comme condition d'événement et non comme contrainte ATA directe.

ThinkSound—ThinkSound est un modèle de génération audio basé sur le rectified-flow (flow matching) qui exploite un signal de raisonnement Chain-of-Thought (AudioCoT) produit par un LLM multimodal afin d'enrichir le conditionnement pour la génération audio à partir d'entrées multimodales (par exemple vidéo ou texte) (Liu *et al.*, 2025b). En incorporant cette description intermédiaire de raisonnement, le pipeline capture mieux les détails sémantiques et temporels et permet un raffinement interactif centré sur les objets pour l'édition ciblée et la génération de bandes sonores. Dans notre implémentation, nous utilisons ThinkSound dans un contexte de variation conditionnée par référence : nous fournissons un audio de référence et utilisons uniquement le texte correspondant à l'étiquette de classe afin de maintenir l'équité expérimentale, pour générer plusieurs variations plausibles au sein de la même catégorie sonore.

AudioX—AudioX est un Diffusion Transformer (DiT) pour la génération anything-to-audio intégrant diverses conditions multimodales (par exemple texte, vidéo et signaux audio). Sa conception principale inclut un module de fusion adaptative multimodale permettant d'intégrer différentes entrées, améliorant l'alignement intermodal et la qualité de génération, et prend en charge le classifier-free guidance (CFG) (Liu *et al.*, 2025c). Dans notre configuration, nous fine-tunons AudioX en utilisant l'étiquette de classe comme conditionnement textuel et générons des variations à l'inférence en initialisant l'échantillonnage à partir d'un clip de référence bruité (contrôlée).

A^{2}SB—A^{2}SB est un modèle de diffusion basé sur un Schrödinger Bridge pour la restauration audio stochastique opérant sur une représentation STFT factorisée (Kong *et al.*, 2025). Il est conçu pour des tâches d'inpainting et d'extension de bande passante, et peut générer plusieurs restaurations plausibles pour une même entrée corrompue. Dans notre cadre de variation, nous masquons un segment de longueur variable dans l'audio de référence, le remplaçons par du bruit, puis échantillonnons la région manquante conditionnée par le contexte restant (non masqué). La méthode ne comporte pas d'entrée textuelle ; nous conditionnons donc uniquement sur l'audio de référence et limitons les modifications à un segment masqué tout en conservant le reste du clip.

ImageBind—Comme composant auxiliaire, ImageBind (Girdhar *et al.*, 2023b) est utilisé comme métrique de qualité au niveau du signal, en fournissant un espace d'embedding partagé permettant de calculer des scores de similarité basés sur les embeddings : similarité cosinus pour quantifier l'alignement avec la référence (préservation de l'identité) et distances d'embedding par paires pour quantifier la diversité entre variantes. Ce benchmark fournit un composant d'évaluation indépendant du modèle, cohérent entre les méthodes de référence sélectionnées et aligné avec les exigences de production liées à la préservation de l'identité et à la variabilité contrôlée.

AudioMorphix—AudioMorphix (Park *et al.*, 2025) est une méthode d'édition audio sans entraînement basée sur des spectrogrammes, construite au-dessus d'un modèle de diffusion latente TTA préentraîné (par exemple AudioLDM ou Tango2). Elle prend en charge un large éventail d'opérations d'édition, notamment l'ajout, la suppression, le remplacement, le déplacement, l'étirement temporel et le décalage fréquentiel. Bien qu'initialement considérée, dans notre implémentation cette méthode ne satisfait pas l'exigence R11 (efficacité mémoire) : l'inférence en précision complète nécessite environ 155 GB de mémoire GPU, et la précision mixte environ 90 GB, ce qui dépasse les ressources disponibles. Nous retirons donc AudioMorphix de la sélection finale.

Tableau 3.2 Matrice des capacités des méthodes de référence sélectionnées pour l'application de génération de variations de SFX au regard des exigences principales R1–5 présentées dans le Tab. 3.1 - **NS** = non spécifié / non évalué - La matrice complète incluant des méthodes supplémentaires, est fournie en annexe dans le Tab. II-1 - Les cellules grisées correspondent aux modèles retenus

Methods	R1	R2, R3	R4	R5	R6	R7	R8	R9	R10	R11	Fit
Audio Editing											
ThinkSound $^{\gamma}$ (NeurIPS, 2025)	FM (latent); A/V/T; CoT (text)	FD↓; MOS-A↑; MOS-Q↑	**NS**	**NS**	Sync; DeSync	**NS**	Targeted editing	Targeted editing	OOD eval.	Inference time (s)↓	Medium (strong edit, text-heavy)
T-FOLEY $^{\gamma}$ (ICASSP, 2024)	Diff (wform); AR; class + time feat.	FAD↓; MOS↑	Class identity	IS	E-L1↓; MOS↑	RMS envelope cond.	Time feat. + Block-FiLM	**NS**	OOD eval.	Inference time; E-L1↓ vs. FAD↓	Medium (time/ energy, no audio-ref)
A ^{2}***SB*** $^{\gamma}$ (ArXiv, 2025)	Diffusion Schröd. Bridge; factorized STFT (3 channels); no vocoder	ViSQOL↑; MOS↑; LSD↓	SiSpec↑	Stochastic sampling (same mask)	Preserve unmasked region	**NS**	~ shape / length mask	Inpainting	**NS**	565M params	Medium (no control within the inpainted segment)
AudioMorphix $^{\gamma}$ (ArXiv, 2025)	Latent diffusion; spec / latent; A / Aref	FAD↓; KL↓; SF↑	Preserve unedited region spec.	Continuous control;	Time-stretch / shift	Pitch-shift	**NS**	Region edit; Add / Del / Rep / Move	**NS**	Training-free, high memory needs	High (audio-ref)
ImageBind* (CVPR, 2023)	Embed. space (T/A/V/D)	Acc.↑; recall↑ (aux)	Embed sim / dist (eval)	Embed sim / dist (eval)	Temp.-aligned pretrain	**NS**	**NS**	**N/A**	Cross-modal	**NS**	Auxiliary (eval / retrieval)
Audio Generation											
AudioX † (ICLR, 2026)	Latent DiT; concat cond.; V/T/A	KL↓; FD↓; FAD↓	CLAP / ImageBind↑ (align)	IS↑	CLAP / ImageBind↑ (align)	**NS**	Text control	Inpaint / complete	Masking-style train.	**NS**	Low (inpaint, text-cond)
AudioLDM † (ICML, 2023)	Latent diffusion + VAE (mel); T/A	FD↓; IS↑; KL↓; FAD↓; MOS↑	Inpainting	IS↑	Temp. order (text)	Pitch (text)	CFG; style weight; start point	Inpainting	**NS**	**NS**	Low (edit, text-cond)

* Auxiliary components. γ Primary candidates; † Secondary baselines.

La sélection finale comprend deux méthodes de référence fortes, AudioLDM (Liu *et al.*, 2023b) et AudioX (Liu *et al.*, 2025c) ; T-Foley (Chen *et al.*, 2024b) pour le contrôle conditionné par le temps ; ThinkSound (Liu *et al.*, 2025b) pour l'édition ciblée ; et A^2SB (Kong *et al.*, 2025) pour les variations basées sur l'inpainting. AudioMorphix (Park *et al.*, 2025) est exclu en raison de la contrainte d'efficacité (R11).

3.4 Cadre expérimental

3.4.1 Jeu de données et protocole d'évaluation

Parmi les méthodes SoTA d'édition audio, la littérature compare rarement les approches à l'aide d'un dataset partagé et d'un benchmark d'évaluation commun. Afin de favoriser l'équité et de permettre une analyse fidèle, cette section introduit une courte évaluation diagnostique visant à évaluer les méthodes d'édition audio pour la génération de variations à partir de seulement quelques références. Comme l'évaluation audio nécessite une combinaison de validations objectives et subjectives, ce diagnostic cible les critères requis et vise à identifier une méthode adaptée à un pipeline prêt pour la production, capable de générer des variations de haute qualité et diversifiées à partir d'une ou de quelques références tout en préservant l'identité de l'événement.

La suite présente le développement du cadre expérimental, en détaillant le traitement et le découpage du dataset sélectionné, ainsi que les protocoles de fine-tuning, d'inférence et d'évaluation.

Vue générale de l'expérimentation—Pour une analyse équitable et complète des modèles sélectionnés pour la tâche de génération de variations audio à partir d'un clip de référence, nous menons une étude en deux étapes après fine-tuning sur le benchmark. (i) **ATA génération :** à partir d'un clip audio de référence et de son étiquette de classe correspondante, chaque modèle génère des variantes, qui sont ensuite évaluées de manière complète. (ii) **Method-specific analysis :** Nous analysons chaque méthode de référence dans son cadre principal d'édition audio et mettons en évidence ses points forts, notamment le transfert de style (AudioX (Liu

et al., 2025c), AudioLDM (Liu *et al.*, 2023b)), l'inpainting (AudioX, A2SB (Cífka *et al.*, 2025)), l'alignement temporel avec une vidéo (T-Foley (Chen *et al.*, 2024b)) et l'édition ciblée avec conditionnement vidéo (ThinkSound (Liu *et al.*, 2025b)).

(A) Dataset et splits —Pour le dataset, nous utilisons ESC-50, une collection few-shot annotée pour la classification de sons environnementaux couvrant des sons humains non vocaux ainsi que des sons environnementaux intérieurs et extérieurs (Piczak, 2015). Plusieurs méthodes de référence utilisent ce dataset, notamment (Cai *et al.*, 2023; Zhang *et al.*, 2025b; Liu *et al.*, 2023c). Ce benchmark contient 2000 enregistrements audio environnementaux, chacun d'une durée de 5 secondes, organisés en 50 classes sémantiques (40 références par classe). Les classes sont regroupées en 5 grandes catégories : animaux, paysages sonores naturels et sons liés à l'eau, sons humains non vocaux, sons intérieurs et domestiques, et bruits extérieurs et urbains.

(B) Protocole du fine-tuning —Pour chaque méthode d'entraînement, nous effectuons un fine-tuning léger sur l'ensemble des classes du dataset ESC-50. Pour chacun des modèles, le fine-tuning est réalisé à partir de leur préentraînement, avec une attention particulière portée à une adaptation légère répondant aux exigences de production et respectant les configurations initiales ainsi que le cadre propre à chaque modèle. Pour chaque méthode de référence, les détails sont présentés dans la section des détails d'entraînement ci-dessous. Des pipelines de diffusion latente (AudioLDM, ThinkSound) aux pipelines de diffusion sur waveform (T-Foley) et aux modèles de rectified flow (AudioX), le processus de fine-tuning est adapté au mieux afin de répondre aux besoins d'un pipeline prêt pour la production, destiné à générer des variantes à partir d'un clip audio SFX de référence donné et de sa classe sémantique.

(C) Protocole de l'inférence —Pour l'inférence, le *fold* de test ESC-50 contient 50 classes c et, pour chaque classe, 8 clips de référence x_{ref}. Nous générons N variantes ($N \leq 10$) par référence. Cela représente un total de 4 000 clips audio générés par modèle. Pour la génération ATA, le modèle reçoit le clip de référence sous forme de *waveform* ainsi que le nom de la classe sous forme de légende. Ce cadre se concentre sur la génération guidée par référence avec une information textuelle minimale. Chaque clip prédit une durée de 5 secondes.

(D) Protocole d'évaluation et métriques —L'évaluation est réalisée sur $N = 10$ variantes générées au regard de leur classe c et de l'audio de référence. Les benchmarks suivent les implémentations fournies par les méthodes de référence considérées dans cette analyse. Pour évaluer la qualité audio, nous calculons la FAD (qualité et réalisme ; R2–R3) à l'aide du *benchmark* AudioLDM-eval (Liu, 2023). Nous calculons d'abord la FAD par classe, puis nous agrégeons les résultats sur l'ensemble des classes. Comme la FAD est plus fiable avec des tailles d'échantillon plus importantes, cette agrégation fournit une vue plus stable de la qualité de génération et capture le réalisme au niveau du signal.

Nous évaluons ensuite l'alignement d'identité (R4) et la diversité (R5) des variantes pour chaque référence. À l'aide d'ImageBind (Girdhar *et al.*, 2023b), l'alignement est mesuré en extrayant des embeddings audio et en calculant la similarité cosinus entre chaque échantillon généré et sa référence. Pour la diversité, nous calculons la distance L2 moyenne pair-à-pair entre les embeddings des variantes pour chaque référence.

Enfin, une évaluation humaine est menée afin d'évaluer la fidélité perceptuelle de l'identité par rapport à la référence. Nous réalisons une étude d'écoute S-MOS (identité de référence ; R4) sur trois classes représentatives correspondant respectivement à du bruit extérieur, à des animaux et à des vocalisations humaines non verbales : hand saw, crow et laughs. Nous recrutons 15 participants et leur demandons d'évaluer la fidélité de l'identité sur une échelle de 1 à 5 (5 étant la meilleure note) pour chaque méthode. Afin d'assurer l'équité, tous les audios générés sont standardisés à 16 kHz pour l'évaluation. Cela peut réduire certains détails de haute fréquence, par exemple pour des sons de bris de verre, et peut entraîner une légère perte de détail perceptuel, mais toutes les méthodes sont évaluées au même taux.

3.4.2 Détails d'entraînement

(A) Setup audio—Nous suivons le découpage officiel du benchmark ESC-50 (Piczak, 2015) : les folds 1 à 3 pour l'entraînement, le fold 4 pour la validation et le fold 5 pour le test, soit 8

clips par classe et par fold. L'entraînement utilise toutes les classes du benchmark, couvrant les animaux, les paysages sonores naturels et les sons liés à l'eau, les sons humains non vocaux, les sons intérieurs et domestiques, ainsi que les bruits extérieurs et urbains. La durée des clips est fixée à 5 s. Le prétraitement utilise un taux d'échantillonnage de 16 kHz, un audio mono et une normalisation Float32. Le fine-tuning en pleine précision (fp32) est coûteux sur le plan computationnel ; nous le comparons donc également à la précision mixte (fp16 ou bf16).

(B) Initialisations—Pour AudioLDM (Liu *et al.*, 2023b), nous initialisons le modèle à partir des checkpoints *audioldm-m-full* et le fine-tunons sur le benchmark ESC-50. Pour T-Foley (Chen *et al.*, 2024b), les checkpoints sont initialisés à partir du modèle préentraîné et les paramètres du pipeline sont adaptés au dataset ESC-50. Pour ThinkSound (Liu *et al.*, 2025b), nous utilisons les checkpoints d'entraînement originaux de ThinkSound. Pour AudioX (Liu *et al.*, 2025c), nous initialisons le modèle avec les checkpoints AudioX préentraînés disponibles sur HuggingFace (HKUSTAudio, 2026). Pour A^2SB (Cífka *et al.*, 2025), comme nous souhaitons tester le modèle dans une configuration de masquage comprise entre 0.3 et 1.0 s, un fine-tuning en deux découpages temporels est réalisé afin de correspondre aux checkpoints préentraînés du modèle. Le pipeline est entraîné deux fois : une première fois avec les checkpoints correspondant au découpage temporel 0.0 s–0.5 s, puis une seconde fois avec ceux du découpage 0.5 s–1.0 s.

(C) Setup du fine-tuning—Pour chaque méthode d'entraînement, nous effectuons un fine-tuning léger sur le dataset few-shot ESC-50. Pour les pipelines de diffusion latente (AudioLDM, ThinkSound), les encodeurs texte et audio sont gelés afin d'éviter une adaptation trop marquée. Seuls le backbone de diffusion (UNet ou MMDiT) et les couches de projection du conditionnement, qui projettent les embeddings de conditionnement dans les canaux du backbone, sont fine-tunés. Pour ThinkSound, les caractéristiques de conditionnement sont pré-extraites dans un dataset de type latent-directory, tandis que pour AudioLDM, les couches de conditionnement sont explicitement fine-tunées. Cette configuration de fine-tuning concentre la variation sur le timbre, le style et la distribution, plutôt que de modifier l'identité. Avant le fine-tuning, ESC-50

est transformé en dataset latent-directory afin d'extraire les caractéristiques. Nous fine-tunons ThinkSound pendant 10 époques de 150 étapes et AudioLDM pendant 200 étapes d'entraînement. Un nombre réduit d'époques est utilisé en raison du contexte few-shot et afin de limiter le risque de surapprentissage ou de dérive de distribution.

Pour les pipelines de diffusion sur *waveform*, un fine-tuning léger est également appliqué en gelant les composantes de conditionnement sémantique et en n'entraînant que les modules qui adaptent la distribution et le timbre. En suivant ces principes, T-Foley est fine-tuné en gelant le générateur et en entraînant les embeddings de classe et de MLP, ainsi que les couches MLP de conditionnement FiLM, qui permettent un conditionnement flexible des cartes de caractéristiques MLP. Nous réduisons le fine-tuning de 500 à 25 époques, avec 250 étapes par époque, pour un total de 6 250 étapes d'entraînement. T-Foley (Chen *et al.*, 2024b) fonctionne à un taux d'échantillonnage de 22 kHz, et ThinkSound à 44 kHz. Comme nous fine-tunons à partir de modèles préentraînés, nous ne modifions pas le taux d'échantillonnage à 16 kHz, car cela pourrait introduire un décalage de distribution.

Pour AudioX (Liu *et al.*, 2025c), nous fine-tunons pendant 20 époques avec 200 étapes par époque, soit un total de 4 000 étapes d'entraînement, en conservant la configuration initiale du pipeline. Le fine-tuning d'AudioX est réalisé via le pipeline d'entraînement Stability *stable-audio-tools* (HKUSTAudio, 2026). Comme AudioX utilise une fenêtre fixe de 11 s, chaque clip ESC-50 de 5 s est complété par des zéros jusqu'à 11 s et entraîné avec une perte à masque de padding. Dans ce cadre, l'objectif de diffusion n'est calculé que sur la portion réelle (non paddée). Le conditionnement pendant l'entraînement n'utilise que le nom de classe de chaque clip audio, tandis que les modalités optionnelles audio et vidéo sont fournies sous forme d'entrées vides afin de satisfaire l'interface du modèle.

Pour A^2SB (Cífka *et al.*, 2025), dans le cadre du *fine-tuning* sur ESC-50 tout en respectant les exigences du pipeline, les waveforms sont converties en caractéristiques STFT avec leur configuration par défaut : n_fft=2048, hop_length=512, et un taux d'échantillonnage de 44.1 kHz. Ensuite, une corruption par inpainting est appliquée sous la forme d'un segment temporel bruité

aléatoire afin d'apprendre au modèle à reconstruire une STFT propre à partir d'une référence corrompue.

(D) Protocole d'implémentation pour la génération ATA—Pour T-Foley (Chen *et al.*, 2024b), nous créons des variations à partir d'un clip de référence en extrayant une condition d'événement temporel (enveloppe RMS ou signal d'attaque) et en conditionnant un DM pour générer de nouveaux échantillons de la même classe suivant ce profil temporel. La racine de la moyenne quadratique (RMS) de la waveform est utilisée comme caractéristique d'enveloppe d'amplitude au niveau des trames. Cette méthode est particulièrement adaptée aux classes qui ne possèdent pas un unique début et une unique fin distinctifs, mais présentent plutôt des motifs temporels avec des réalisations variables (par exemple *rain*, *sneeze* (Chen *et al.*, 2024b).

Pour ThinkSound (Liu *et al.*, 2025b), le pipeline prend le tenseur latent correspondant au clip audio de référence, lui ajoute du bruit, puis démarre à partir de cette référence bruitée plutôt qu'à partir d'un bruit pur. Le niveau de bruit est contrôlé par σ dans $z_0 = z_{\text{ref}} + \sigma \cdot \epsilon$, où $\epsilon \sim \mathcal{N}(0, I)$, et nous fixons $\sigma = 0.9$. Une valeur plus élevée de σ peut accroître la variation mais aussi entraîner une dérive par rapport à la référence, tandis qu'une valeur plus faible encourage le modèle à rester plus proche du même domaine et de la même identité. Pour chaque variante, ϵ est échantillonné aléatoirement selon la distribution gaussienne. Le latent bruité est ensuite débruité en utilisant le nom de la classe comme conditionnement afin de préserver l'identité. Avec cette méthode, les variations conservent le timbre, la texture et le style d'arrière-plan de la référence. Le CFG est appliqué pendant la génération et contrôle l'intensité avec laquelle le modèle suit le conditionnement textuel. Pour la ATA generation, nous fixons $cfg\_scale = 5$, car des valeurs plus faibles peuvent induire une dérive perceptuelle de l'identité.

Pour AudioX (Liu *et al.*, 2025c), nous suivons une procédure de génération similaire. À l'évaluation, nous réalisons une variation ATA guidée par référence en initialisant l'échantillonnage à partir du clip de référence (`init_audio`, contrôlé par `init_noise_level`), puis en recadrant les sorties générées à 5 s pour le calcul des métriques. Nous augmentons le bruit ajouté à la référence à $\sigma = 1.5$ afin d'encourager la diversité. Le sampler utilisé est DPMPP 3M SDE.

Pour A^2SB (Cífka *et al.*, 2025), conformément à l'objectif de cette méthode de référence, nous visons un inpainting stochastique à l'intérieur d'un clip réel. Le pipeline prend donc une référence ESC-50 de 5 s, masque un segment (0.3–1.0 s), puis échantillonne $K = 10$ restaurations différentes en modifiant la seed (même entrée corrompue et même masque). Pour le choix de la durée de masquage, l'intervalle 0.3–1 s est retenu car l'entraînement initial est réalisé sur des fractions d'inpainting allant jusqu'à ~ 0.54 d'un segment de $\sim$2.96 s, de sorte que des lacunes de 0.3 s à 1.0 s à l'intérieur d'un clip de 5 s restent raisonnables. Le bruit σ ajouté à la référence est augmenté de 0.3 à 0.9.

(E) Setup d'évaluation —Pour l'évaluation de la FAD, de la diversité et de l'alignement, toutes les méthodes sont évaluées sur des clips audio de 5 secondes (correspondant à la durée initiale des références ESC-50), à l'exception de T-Foley et A^2SB, initialement entraînés sur 4 secondes. Les clips de référence sont donc recadrés pour correspondre à cette durée et permettre une évaluation équitable. Par ailleurs, une évaluation humaine est menée sur la fidélité perceptuelle de l'identité à l'aide du Similarity-Mean Opinion Score (S-MOS, (Loizou, 2011)) entre la référence et les variations, à partir des notes attribuées par les participants. Pour chaque méthode, 30 ensembles par modèle sont échantillonnés à partir de trois classes représentatives (laughing, crow, hand_saw), en sélectionnant 2 clips de référence par classe et 5 variations par référence (3×2×5). Chaque ensemble présente un audio de référence et une variation, et les participants évaluent la similarité d'identité sur une échelle de 1 à 5 (5 : plus similaire). Afin d'éviter tout biais, les noms des méthodes sont supprimés, et tous les essais sont fusionnés dans une liste aléatoire unique ; un fichier-clé privé permet de relier chaque identifiant anonymisé à son origine pour l'analyse. L'étude est déployée sous la forme d'une interface web légère hors ligne (HTML), qui lit la référence avec ses variations et enregistre les notes dans un fichier CSV. Les scores sont agrégés en calculant d'abord, pour chaque participant, la note moyenne par méthode, puis en moyennant sur l'ensemble des participants pour obtenir l'évaluation S-MOS, avec tous les audios standardisés à 16 kHz afin d'assurer une lecture cohérente et une comparaison équitable.

Tableau 3.3 Comparaison quantitative globale des méthodes de référence pour la génération *audio-to-audio* (ATA), après *fine-tuning* sur l'ensemble complet du benchmark ESC-50 Piczak (2015) - Toutes les méthodes sont évaluées à partir d'un clip de référence et du nom de classe associé - Les métriques de diversité et d'alignement sont calculées dans l'espace du benchmark ImageBind Girdhar *et al.* (2023b) - La diversité désigne la distance pairwise ℓ_2 entre les variantes générées pour une même référence tandis que l'alignement mesure la similarité cosinus entre chaque variante et sa référence dans l'espace d'embedding *audio* de ImageBind. Les **meilleurs résultats** et les deuxièmes meilleurs résultats sont indiqués - Comme **A^{2}SB** est une méthode d'inpainting les résultats présentés ici ne concernent que des modifications appliquées entre 0.3,s et 1,s sur une référence de 5 secondes - Une analyse complémentaire est fournie dans les études d'ablation afin de proposer une évaluation plus adaptée au cadre de génération de variations de SFX

Methods	**FAD↓**	**S-MOS↑**	**Diversity↑**	**Alignment↑**
AudioLDM (ICML, 2023)	20.09	2.22 ± 0.45	**0.85**	0.39
T-FOLEY* (ICASSP, 2024)	24.53	1.89 ±0.36	0.77	0.22
ThinkSound (NeurIPS, 2025)	16.51	2.57 ± 0.56	0.73	0.51
A^{2}SB* (ArXiv, 2025)	**1.59**	**4.81 ± 0.10**	0.135	**0.963**
AudioX (ICLR, 2026)	9.34	3.37 ± 0.56	0.67	0.59

* Durée des clips générés : 4 secondes. L'évaluation est réalisée sur des clips de 4 secondes, avec des références également tronquées à 4 secondes.

3.5 Résultats et discussion

La Table. 3.3 compare l'ensemble des méthodes de référence sur la tâche de génération ATA avec ESC-50. Globalement, AudioX offre le meilleur compromis entre fidélité, préservation de l'identité, diversité et alignement, tandis qu'AudioLDM et T-Foley révèlent des compromis entre diversité et qualité perceptuelle. A^{2}SB obtient les meilleurs scores quantitatifs, mais ces résultats doivent être interprétés avec prudence, car les variations évaluées correspondent uniquement à de courtes régions *inpainted* de 0.3 à 1 s au sein de clips de référence largement préservés.

Concernant AudioLDM (Liu *et al.*, 2023b), la Table. 3.3 montre le score de diversité le plus élevé (**0.85**) entre les variantes issues d'une même référence, mais au prix d'un alignement plus faible (0.39), d'une FAD relativement élevée (20.09) et d'un S-MOS faible (2.22). Ces résultats sont cohérents avec les observations qualitatives de la Figure 3.1, où les échantillons générés présentent une qualité audio dégradée, des changements temporels abrupts et des textures

bruitées, produisant ainsi des sorties perceptuellement déformées et fragiles. Cette tendance est également visible dans les courbes d'énergie de la Figure II-2 : bien que le motif dynamique global soit partiellement préservé, les variations générées présentent une distribution d'énergie plus plate que les références.

Comme T-Foley n'est préentraîné que sur sept classes (*DogBark*, *Footstep*, *GunShot*, *Keyboard*, *MovingMotorVehicle*, *Rain*, *Sneeze_Cough*), son *fine-tuning* sur ESC-50 reste limité par le manifold de préentraînement. Pour les classes dont les distributions sont éloignées de ces classes sources, la qualité de génération et la fidélité à la classe se dégradent sensiblement. Par exemple, lors de la génération de SFX de bébé qui pleure, des motifs timbraux rappelant un aboiement peuvent apparaître dans la sortie. Pour atténuer ce problème, nous augmentons l'échelle de conditionnement à 10 (au lieu de 0.3 dans la configuration originale) afin de mieux imposer le conditionnement par classe. Comme le montre la Table. 3.4, T-Foley fonctionne nettement mieux sur les classes recouvrant son manifold de préentraînement, atteignant une FAD de **13.51**, alors que les performances chutent à **25.75** sur les 45 classes restantes. Cette limitation se reflète également dans le S-MOS global de seulement **1.49** dans la Table. 3.3, indiquant une faible préservation perceptuelle de l'identité après adaptation. Dans la Figure 3.1, les générations de T-Foley en dehors du manifold de préentraînement présentent des textures bruitées, un contenu à haute fréquence atténué et une plage spectrale réduite par rapport aux autres méthodes. En même temps, la Figure II-2 montre que T-Foley suit plus fidèlement le profil énergétique de la référence que les autres méthodes de référence, ce qui est attendu étant donné son conditionnement explicite par l'énergie. Dans l'ensemble, ces résultats suggèrent que T-Foley préserve bien les dynamiques temporelles-énergétiques grossières, mais peine à maintenir l'alignement de classe et la fidélité perceptuelle au-delà de sa distribution d'entraînement initiale.

Tableau 3.4 Comparaison quantitative pour la génération avec T-Foley, préentraîné sur 7 classes (*DogBark*, *Footstep*, *GunShot*, *Keyboard*, *MovingMotorVehicle*, *Rain*, *Sneeze_Cough*), puis *fine-tuné* sur les classes de ESC-50 - Lorsque les classes générées suivent des distributions trop éloignées de celles vues au préentraînement (c'est-à-dire hors de la variété apprise), la fidélité à la classe peut se dégrader - À l'inverse, les classes restant dans le support du préentraînement et cohérentes avec celui-ci (par exemple *coughing*, *dog*, *footsteps*, *keyboard_typing* et *rain*) obtiennent de meilleurs résultats

T-Foley* (ICASSP, 2024)	***Dans l'espace de pré-entraînement*** **(5 classes)**	***Hors pré-traitement*** **(45 classes)**
FAD Overall↓	**13.51**	25.75

∗ Durée des clips générés : 4 secondes. L'évaluation est effectuée sur des clips de 4 secondes, à partir de références elles aussi tronquées à 4 secondes.

ThinkSound (Liu *et al.*, 2025b) et AudioX offrent le meilleur compromis entre diversité et alignement dans la Table. 3.3, avec des scores de (0.73, 0.51) pour ThinkSound et (0.67, 0.59) pour AudioX. ThinkSound améliore l'alignement par rapport à AudioLDM et T-Foley tout en maintenant une diversité relativement élevée, mais sa qualité perceptuelle reste limitée, avec une FAD de **16.51** et un S-MOS de **2.57**. Qualitativement, les spectrogrammes générés préservent la structure temporelle globale au niveau de l'événement, mais présentent un léger étalement spectral et des attaques plus floues, comme l'illustre l'exemple du corbeau dans la Figure 3.1. Globalement, ThinkSound produit des variations diversifiées qui restent structurellement liées à la référence, mais avec une fidélité perceptuelle de l'identité seulement modérée.

Du côté de A^2SB, il est important de noter que cette méthode réalise de l'inpainting plutôt qu'une véritable génération complète de variations ATA. Par conséquent, les bons résultats quantitatifs rapportés dans la Table. 3.3 reflètent principalement de courtes modifications locales au sein de clips de référence autrement préservés, et ne doivent pas être interprétés comme directement comparables à ceux des autres méthodes de référence. Dans les régions encadrées en vert de la Figure 3.1, la méthode régénère généralement des textures restant proches de la référence tout en respectant la temporalité de l'événement. Par exemple, l'exemple du corbeau montre une structure temporelle préservée avec une concentration d'énergie plus marquée dans les bandes médianes, tandis que l'exemple du rire contient davantage d'événements isolés en

fréquence. Comme le montre la Figure II-2, les variations générées peuvent introduire des pics locaux d'énergie marqués, mais l'évolution énergétique globale reste proche de la référence. Nous analysons plus en détail cette méthode dans les ablations ci-dessous. Dans la Table. 3.7, l'augmentation de la durée du masque de 0.3–1.0 s à 0.3–2.0 s dégrade la fidélité (FAD passant de 1.54 à 3.92) et l'alignement (de 0.959 à 0.916), tout en augmentant légèrement la diversité. Cela suggère que A^2SB gère bien les segments masqués courts, mais devient moins stable lorsqu'il doit régénérer des régions manquantes plus longues, où peuvent apparaître des prédictions vides ou de fortes discontinuités temporelles.

AudioX, quant à lui, met en évidence le meilleur équilibre global entre les métriques, avec une FAD faible de **9.34**, un fort alignement ImageBind de **0.59**, et le meilleur S-MOS non fondé sur l'inpainting, soit **3.37**, tout en conservant un bon score de diversité de **0.67**. Dans la Figure 3.1, les échantillons générés préservent la structure temporelle globale de la référence, avec des bouffées d'énergie survenant à des intervalles similaires, tout en introduisant des variations perceptibles, comme une activité haute fréquence supplémentaire dans l'exemple du corbeau. La Figure II-2 confirme également cette observation, en montrant qu'AudioX conserve la structure énergétique globale tout en permettant des variations modérées d'intensité entre les échantillons générés.

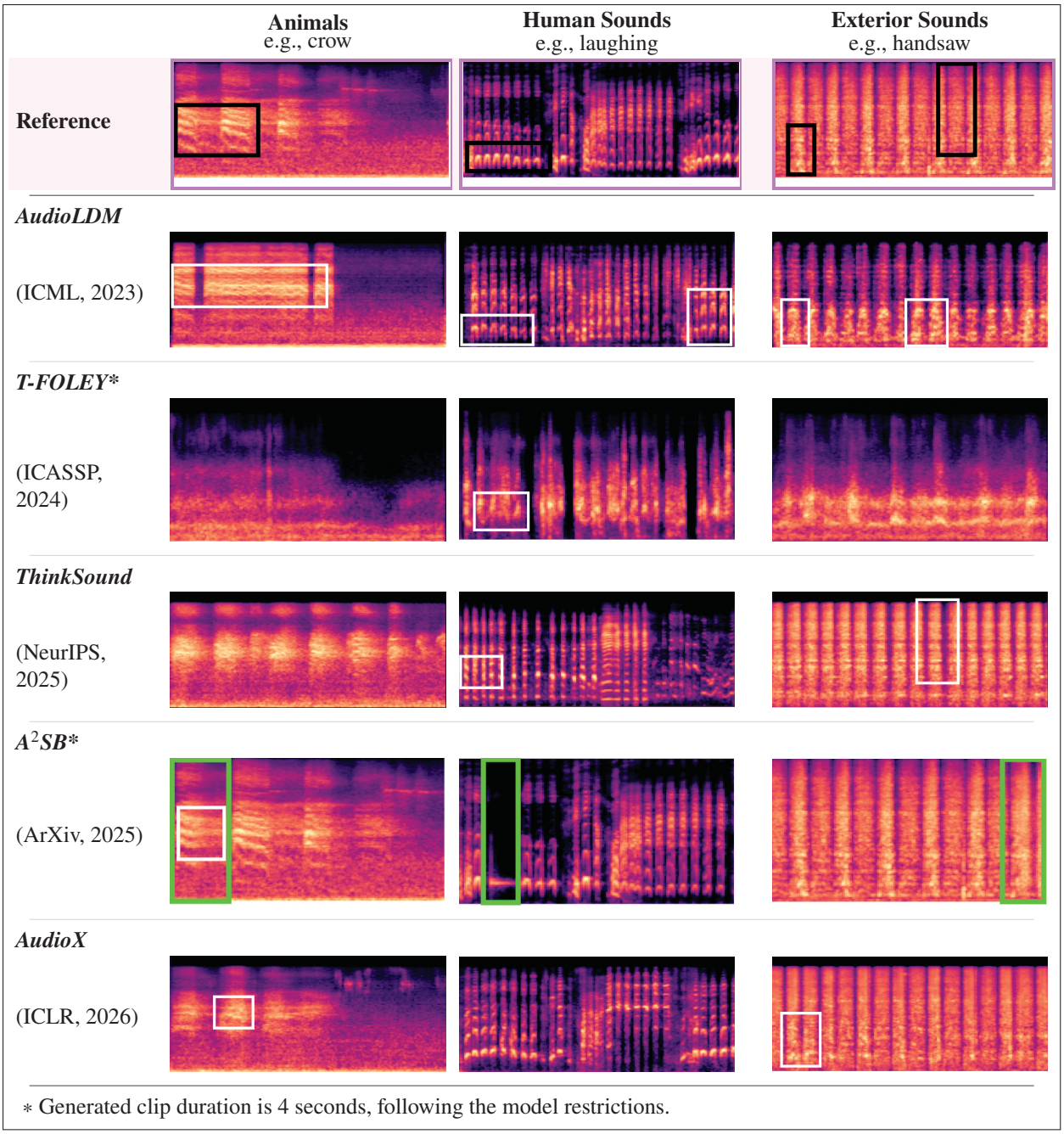

Figure 3.1 Visualisation de l'évolution fréquentielle et de l'amplitude des clips pour chaque modèle dans la tâche de variation ATA, à partir de représentations en mel-spectrogrammes - Pour la tâche d'inpainting de A^2SB, les sections masquées puis reconstruites sont encadrées en vert, car l'évaluation sur l'ensemble complet consiste à masquer et régénérer uniquement le segment compris entre 0.3,s et 1,s - Des expériences complémentaires avec des durées de masquage plus étendues sont également présentées afin de mieux se rapprocher du cadre de génération de variations sonores - La boîte noire met en évidence des régions à texture marquée dans les mel-spectrogrammes de référence - Les boîtes blanches montrent les zones où les motifs texturaux sont préservés de manière similaire à la référence - La figure illustre également dans quelle mesure les méthodes conservent une structure proche de celle du signal de référence

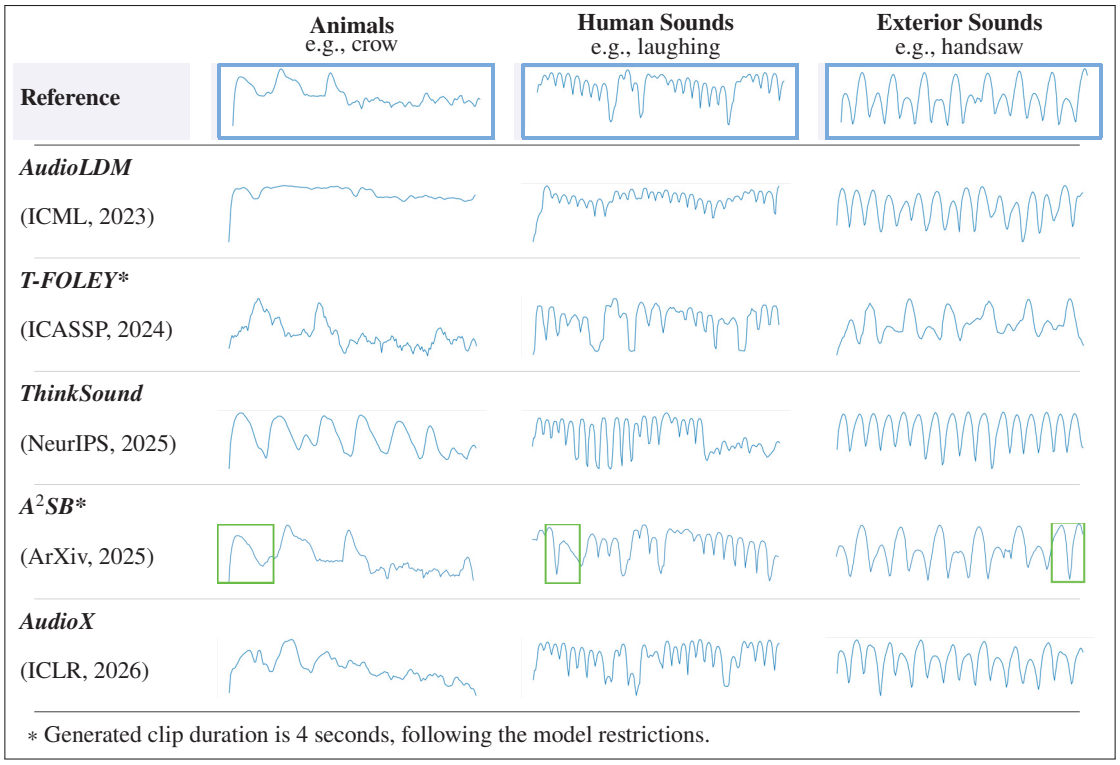

Figure 3.2 Visualisation des variations d'énergie des clips audio pour chaque modèle dans la tâche d'édition ATA - Chaque panneau représente l'évolution de l'énergie (dB) en fonction du temps (s) - La majeure partie de l'énergie des clips se situe entre [-80, -5] dB - Tous les clips ont une durée de 5 secondes, à l'exception de T-Foley et de A^2SB, qui produisent des clips de 4 secondes - Cette figure permet d'observer qualitativement dans quelle mesure chaque méthode reproduit ou s'éloigne du profil énergétique du signal de référence - Pour la tâche d'*inpainting* de A^2SB, les sections masquées puis reconstruites sont indiquées sur la courbe d'énergie, car l'évaluation sur l'ensemble complet consiste à masquer et régénérer uniquement le segment compris entre 0.3,s et 1,s - Des expériences complémentaires avec des masques plus étendues sont également présentées afin de mieux se rapprocher du cadre de génération de variations sonores

Coût de Calcul —La Table. 3.5 rapporte le coût computationnel des pipelines évalués à l'inférence pour une génération à grande échelle de variantes sur l'ensemble du benchmark ESC-50, correspondant à 4000 générations au total (10 variantes pour chacune des 8 références à travers 50 classes). Évaluer le coût à cette échelle fournit une vue pratique de l'efficacité de chaque méthode et, par conséquent, de son adéquation à un pipeline prêt pour la production où de nombreuses variations doivent être générées de manière fiable et dans des contraintes raisonnables de temps et de ressources.

Parmi les méthodes de référence ATA, les méthodes diffèrent principalement par la complexité de leur pipeline de génération et par le nombre d'étapes d'échantillonnage requises lors de l'inférence. Les méthodes basées sur un échantillonnage par diffusion latente, telles qu'AudioLDM et AudioX, nécessitent plusieurs itérations de débruitage pour générer chaque échantillon, ce qui augmente le temps d'inférence et le coût computationnel par rapport à des pipelines de génération plus directs.

Le pipeline A^2SB présente le coût computationnel global le plus élevé, car il repose sur un échantillonnage itératif par diffusion pour reconstruire les régions audio masquées. Contrairement aux autres méthodes, qui génèrent directement des variations complètes, A^2SB effectue une reconstruction conditionnelle par des étapes répétées de débruitage, ce qui entraîne un coût d'inférence sensiblement plus élevé. Bien que ce comportement soit acceptable pour des tâches d'édition localisée ou d'inpainting court, il est moins favorable à un déploiement à grande échelle dans un pipeline de production où le débit constitue une contrainte centrale.

À l'inverse, les méthodes reposant sur des mécanismes de génération plus directs nécessitent moins d'opérations itératives et offrent donc une génération plus rapide. Cette différence devient particulièrement importante lorsqu'il s'agit de générer plusieurs variantes candidates par référence, de traiter un grand nombre de classes, ou d'intégrer le modèle dans un cadre de production interactif ou contraint en ressources.

Dans l'ensemble, ces résultats mettent en évidence un compromis pratique entre qualité de génération et efficacité. Bien que les pipelines basés sur la diffusion tels que A^2SB puissent

Tableau 3.5 Résumé des paramètres expérimentaux utilisés à l'inférence pour la génération de $N = 10$ variantes par référence sur 50 classes, avec 8 références par classe, soit un total de 4000 générations par modèle, suivant la soundbank ESC-50 Piczak (2015) - L'évaluation est menée après standardisation de l'ensemble des sorties à une fréquence d'échantillonnage de 16,kHz et à une durée de clip audio de 4 ou 5 secondes - Pour les modèles fonctionnant à des fréquences d'échantillonnage plus élevées, un post-traitement est appliqué avant la sauvegarde

Method	Backbone	NFE	Sample Rate	Inference (ESC-50)	CFG	Steps	Batch	Hardware
ThinkSound (ICLR, 2025)	MM-DiT	100	16 kHz	~0.66 h	5	24	2	A100-SXM4-40GB
T-Foley (ICLR, 2025)	Block FiLM	–	22 kHz	~28.8 h	–	250	16	RTX-A6000-48GB
A^{2}SB (ArXiv, 2025)	Schrödinger bridge	50	16 kHz	~33 h	–	300	1	A100-SXM4-40GB
AudioX (NeurIPS, 2024)	Rectified Flow	–	48 kHz	~10 h	5	250	1	A100-SXM4-40GB
AudioLDM (CVPR, 2023)	DDIM	200	16 kHz	~0.75 h	2.5	200	1	RTX-A6000-48GB

atteindre de bonnes performances de reconstruction pour l'inpainting localisé, leur coût computationnel les rend moins adaptés à un usage de production à grande échelle que les autres approches évaluées.

Étude d'ablation du paramètre d'initialisation du bruit α —Dans la Table 3.6, l'évaluation porte sur l'effet du paramètre d'initialisation du bruit α sur la qualité des variations générées pour trois méthodes introduisant une contrôlabilité via l'initialisation du bruit, également utilisée dans les applications de transfert de style pour contrôler l'intensité du transfert. À cet égard, AudioX, ThinkSound et A^{2}SB sont sélectionnés. Comme le montre la Table 3.6, α contrôle la quantité de perturbation appliquée au latent de référence avant la génération, régulant ainsi

Tableau 3.6 Ablation sur le niveau de bruit initialisé dans l'espace latent pour la génération de variations à partir d'une référence donnée avec **A^{2}SB** (Cífka *et al.*, 2025) - L'évaluation est menée sur trois classes de ESC-50 (*crow*, *laughing* et *handsaw*) - AudioX produit des clips audio silencieux pour α0.6 - Les meilleurs résultats globaux de chaque méthode sont mis en évidence en couleur

α	**Method**	**FAD↓**	**Diversity↑**	**Alignment↑**
0.0*	AudioX	null	null	null
	ThinkSound	3.17	0.105	0.852
	A^{2}SB	2.41	0.109	0.943
0.3	AudioX	null	null	null
	ThinkSound	3.06	0.323	0.825
	A^{2}SB	1.55	0.177	0.948
0.6	AudioX	4.57	0.394	0.774
	ThinkSound	4.93	0.494	0.767
	A^{2}SB	2.02	0.285	0.922
0.9	AudioX	3.97	0.513	0.734
	ThinkSound	11.09	0.678	0.666
	A^{2}SB	3.03	0.274	0.889
1.5	AudioX	5.04	0.598	0.674
	ThinkSound	14.03	0.764	0.586
	A^{2}SB	2.83	0.175	0.899
2.0	AudioX	6.64	0.659	0.643
	ThinkSound	13.93	0.764	0.575
	A^{2}SB	2.83	0.175	0.899

* Deterministic.

le compromis entre fidélité à la référence et intensité de variation. Des valeurs plus faibles de α produisent des sorties qui restent plus proches du signal de référence, avec de meilleurs scores d'alignement mais une diversité limitée. À mesure que α augmente, les échantillons générés présentent une plus grande diversité tout en s'éloignant progressivement de l'énergie de référence et de la structure spectrale. Ainsi, des niveaux de bruit modérés à élevés (par exemple $\alpha = 0.6 - 0.9$) offrent le meilleur compromis entre fidélité et variation, en atteignant un **meilleur alignement** tout en maintenant une diversité raisonnable. Des niveaux de bruit plus élevés conduisent à une déviation structurelle et à un alignement réduit avec la référence. Ces résultats soulignent l'importance de contrôler soigneusement le niveau de bruit initial afin d'obtenir des

Tableau 3.7 Étude d'ablation de la durée de masquage pour la génération par inpainting de variations à partir d'une référence donnée avec **A^{2}SB** (Cífka *et al.*, 2025) - L'évaluation est menée sur trois classes de ESC-50 (*crow*, *laughing* et *handsaw*)

Mask durations	**FAD↓**	**Diversity↑**	**Alignment↑**
0.3–1.0	1.54	0.162	0.959
0.3–2.0	3.92	0.230	0.916

variations de SFX perceptuellement significatives.

Afin d'analyser plus en détail le comportement de la méthode basée sur l'inpainting, nous évaluons l'impact de la durée du masque sur la qualité de reconstruction.

Études d'ablation sur la tâche d'*inpainting* —La Table. 3.7 évalue la robustesse du pipeline d'inpainting A^{2}SB lorsque la durée masquée augmente. Deux régimes de masquage sont considérés : des segments masqués courts entre 0.3 et 1.0 seconde, et des segments plus longs entre 0.3 et 2.0 secondes à l'intérieur de l'audio de référence. Il est important de noter que les métriques rapportées doivent être interprétées avec prudence, car les sorties générées sont constituées en grande partie de l'audio de référence préservé, seule une courte région masquée étant régénérée. En conséquence, les scores de fidélité et d'alignement reflètent partiellement les portions inchangées du signal et tendent donc à paraître plus élevés que dans le cadre d'une génération ATA complète.

Lorsque la région masquée reste courte (0.3–1.0 s), la méthode atteint une forte fidélité de reconstruction, avec une FAD faible de 1.54 et un alignement ImageBind élevé de 0.959. Cela indique que le modèle est capable de régénérer des structures audio locales restant cohérentes avec le contexte de référence environnant.

Cependant, lorsque la durée du masque augmente à 0.3–2.0 s, la qualité de reconstruction se dégrade de manière notable. La FAD augmente de 1.54 à 3.92, tandis que l'alignement diminue de 0.959 à 0.916. Cela suggère que la méthode devient moins fiable lorsque de plus grandes

portions du signal doivent être synthétisées, car moins d'informations contextuelles issues de la référence restent disponibles pour guider la reconstruction.

Dans le même temps, la diversité augmente légèrement lorsque des segments masqués plus longs sont utilisés. Ce comportement est attendu, puisqu'une région manquante plus large offre davantage de liberté au processus génératif, permettant plusieurs reconstructions plausibles du segment audio manquant.

Études d'ablation sur le transfert de style —En suivant la configuration de transfert de style audio d'AudioLDM (Liu *et al.*, 2023b), adaptée au benchmark ESC-50, nous évaluons trois exemples de transfert de style : *toilet flush* vers *children singing*, *sheep* vers *narration/monologue*, et *coughing* vers *ambient music*. La Figure 3.3 montre l'audio de référence ainsi que les échantillons générés obtenus avec différents niveaux de bruit d'initialisation σ (intensité du transfert).

Pour de petites valeurs de σ, AudioX préserve plus fidèlement le timbre source et la structure spectrale, en conservant les motifs harmoniques et temporels d'origine tout en n'introduisant que de faibles variations stylistiques. À mesure que σ augmente, la transformation devient plus marquée et les échantillons générés s'éloignent progressivement de la référence. À l'inverse, AudioLDM suit plus agressivement la condition textuelle cible lorsque σ augmente, produisant des sorties dont les motifs spectro-temporels ressemblent davantage au style audio cible (par exemple une modulation proche de la parole pour *narration* ou des structures harmoniques pour *singing*). Cependant, cette transformation plus forte tend également à réduire la préservation de l'identité audio originale.

Dans l'ensemble, ces résultats illustrent un compromis entre préservation de la source et alignement stylistique : AudioX fournit des transformations plus progressives et plus douces, tandis qu'AudioLDM atteint une plus forte adhérence au style cible guidé par le texte lorsque l'intensité du transfert est élevée.

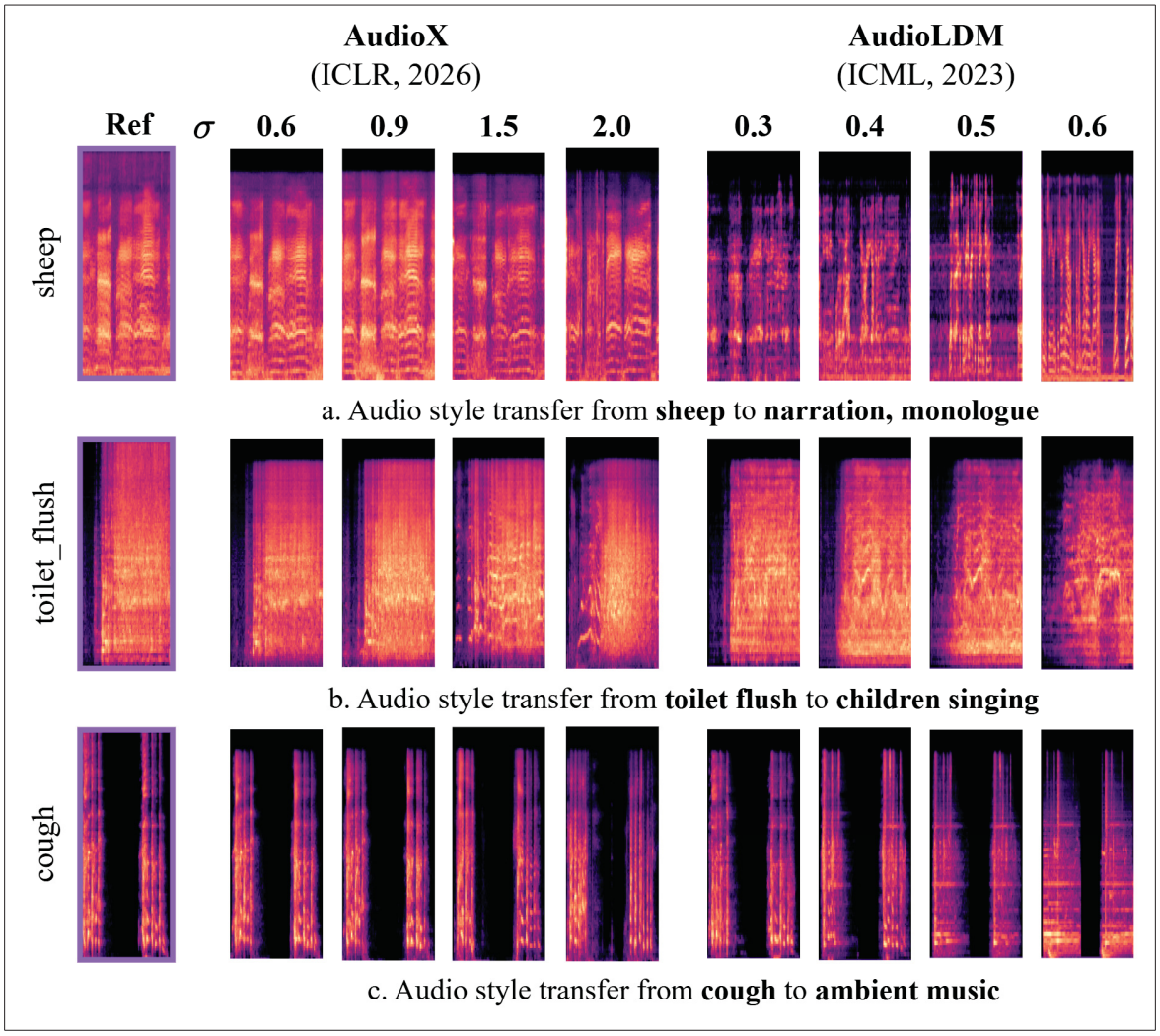

Figure 3.3 Ablations sur la tâche de transfert de style audio pour AudioLDM (Liu *et al.*, 2023b) et AudioX (Liu *et al.*, 2025c) sur ESC-50 (Piczak, 2015). De gauche à droite : l'audio de référence (par exemple *sheep*, *toilet_flush*, *cough*) et quatre échantillons générés conditionnés par le prompt textuel cible (par exemple "narration, monologue", "children singing", "ambient music") avec différents niveaux de bruit d'initialisation σ (intensité du transfert) - Pour des valeurs plus faibles de σ (à gauche), les échantillons générés restent plus proches de la référence, tandis que des valeurs plus élevées produisent des sorties plus fortement alignées avec la condition textuelle - Chaque méthode utilise sa propre échelle de bruit, cohérente avec son pipeline

3.6 Analyse critique

L'objectif de ce chapitre n'est pas seulement de comparer des méthodes de référence récentes pour la génération audio conditionnée par une référence, mais également d'évaluer leur pertinence pour un pipeline prêt pour la production dédié à la génération de variations de SFX. Dans ce cadre, une méthode utile en pratique doit satisfaire plusieurs exigences : préserver l'identité de l'événement de la référence (R4), générer un audio perceptuellement plausible (R2–R4), produire des variations à la fois diversifiées (R5) et contrôlées (R8), respecter des contraintes structurelles optionnelles telles que le timing (R6) ou l'énergie (R7), et rester suffisamment efficace pour un déploiement à grande échelle (R10, R11). Les résultats montrent que, dans cette implémentation de la génération audio de variations guidée par référence, chaque méthode reflète un compromis différent entre fidélité, contrôlabilité, diversité et efficacité, et peut ainsi représenter soit un point fort, soit un mode d'échec pour de futures recherches sur les pipelines industriels.

Le cas de A^2SB —Une première observation importante est qu'une forte performance quantitative n'implique pas toujours une adéquation à l'application visée. En particulier, les très bons scores obtenus par A^2SB doivent être interprétés avec prudence, puisque la méthode réalise un inpainting local à l'intérieur de clips de référence largement préservés plutôt qu'une véritable génération complète de variations ATA. En conséquence, ses scores de fidélité et d'alignement sont en partie soutenus par des portions inchangées du signal. Cela rend A^2SB particulièrement pertinent pour la réparation localisée ou les éditions courtes sous contrainte, mais moins représentatif du problème complet de génération de variations étudié dans ce chapitre. Son coût computationnel plus élevé limite également sa scalabilité dans des contextes de production où de nombreuses variantes candidates doivent être générées.

AudioX la méthode analysée optimale—Parmi les méthodes de référence de génération complète, AudioX présente le comportement global le plus équilibré. Il atteint le meilleur compromis entre qualité perceptuelle, alignement, diversité et préservation de l'identité évaluée par des jugements humains, tout en préservant qualitativement la structure temporelle de la

référence et en introduisant des variations significatives. Cela fait d'AudioX la méthode de référence la plus adaptée à l'objectif central de cette analyse : générer plusieurs variations plausibles d'une référence donnée tout en maintenant son identité perceptuelle. De plus, l'ablation sur le transfert de style montre qu'AudioX préserve plus efficacement le timbre source pour de faibles intensités de transfert, ce qui est souhaitable dans les flux de production où des éditions subtiles et progressives sont souvent préférées à des transformations brusques.

En termes de diversité —À l'inverse, **AudioLDM** se distingue par sa diversité et par sa capacité à suivre plus fortement les transformations guidées par le texte lorsque l'intensité du transfert augmente. Cependant, cela se fait au prix d'un alignement plus faible, d'une préservation de l'identité jugée plus faible par les participants, et d'une qualité perceptuelle dégradée. L'analyse qualitative montre également des textures plus bruitées et des profils d'énergie plus plats. Ainsi, AudioLDM semble davantage adapté à une génération exploratoire ou fortement stylisée, où une grande variation est préférée à une préservation stricte de l'identité. Il est moins approprié lorsque l'exigence principale est de rester étroitement ancré à un son de référence donné.

ThinkSound occupe une position intermédiaire. Comparé à AudioLDM et T-Foley, il offre un meilleur compromis entre diversité et alignement, et il préserve raisonnablement bien l'organisation temporelle globale de la référence. Toutefois, sa qualité perceptuelle et sa préservation de l'identité restent plus limitées que celles d'AudioX. En pratique, ThinkSound peut donc être pertinent lorsque l'on souhaite une cohérence structurelle modérée accompagnée de variation, mais il ne constitue pas la réponse la plus forte à la contrainte de production consistant à préserver avec réalisme l'identité d'une référence spécifique.

Conditionnement temporel et limitations du modèle T-Foley —T-Foley est la méthode de référence qui traite le plus explicitement le conditionnement temporel et énergétique. Les échantillons qu'il génère suivent plus fidèlement le profil énergétique de la référence que les autres méthodes de référence, ce qui constitue une propriété utile pour les tâches nécessitant un contrôle du timing ou de l'enveloppe. Cependant, son préentraînement sur seulement sept classes source limite fortement son adaptabilité à des distributions de SFX plus larges. Cela se reflète à

la fois dans l'analyse du manifold et dans l'évaluation perceptuelle, où des fuites de classe et une fidélité réduite apparaissent en dehors du domaine de préentraînement. En conséquence, T-Foley doit plutôt être considéré comme une méthode spécialisée pour la génération sous contrainte temporelle dans une distribution de classes restreinte, plutôt que comme une solution générale pour la génération de variations prêtes pour la production sur des *soundbanks* diversifiés.

En termes de production —Les résultats suggèrent que la méthode la plus pertinente dépend de la priorité de l'application. Si l'objectif est de générer des variations complètes à partir d'une référence avec le meilleur compromis global entre préservation de l'identité, réalisme et diversité, **AudioX** constitue le choix le plus approprié. Si la tâche requiert plutôt une édition localisée ou la réparation d'un court segment audio tout en préservant le signal environnant, **A^{2}SB** est l'option la plus pertinente, bien que ses résultats ne doivent pas être comparés directement à ceux des méthodes de génération complète et que son coût d'inférence demeure élevé. Si l'objectif principal est un contrôle explicite du timing ou de l'énergie, **T-Foley** fournit le mécanisme le plus direct, mais uniquement dans un manifold limité. Enfin, si une stylisation plus marquée ou une transformation guidée par le texte est préférée à une stricte préservation de la source, **AudioLDM** offre le comportement de transfert de style le plus agressif, tandis que **ThinkSound** fournit un compromis plus modéré.

Dans l'ensemble, cette analyse met en évidence que la variation audio prête pour la production ne peut pas être réduite à une seule métrique. Un pipeline adapté doit équilibrer conjointement le réalisme, la préservation de l'identité, la contrôlabilité, la diversité et le coût d'inférence. Parmi les méthodes de référence évaluées, AudioX offre le compromis le plus solide à usage général pour la variation de SFX conditionnée par référence, tandis que A^{2}SB, T-Foley et AudioLDM doivent plutôt être compris comme des solutions spécifiques à certaines tâches, préférables uniquement lorsque l'édition locale, le contrôle de l'enveloppe temporelle ou une forte stylisation constitue l'exigence dominante. Ces observations indiquent également que les recherches futures devraient viser à combiner ces forces complémentaires au sein d'un cadre unifié : préservation robuste de l'identité de référence, contrôlabilité explicite, capacité d'édition localisée et inférence efficace pour la production.

3.7 Conclusion

Ce chapitre a étudié la génération de variations de SFX conditionnée par une référence afin d'évaluer l'adéquation des pipelines récents de génération audio pour des flux de production prêts à l'emploi. Contrairement à des contextes génératifs plus généraux, cette tâche exige de préserver l'identité perceptuelle d'un son de référence, de produire des variations à la fois diversifiées et plausibles, de respecter les propriétés structurelles du signal et de rester computationnellement efficace.

Synthèse comparative des performances des méthodes de référence —À travers une évaluation quantitative, une analyse perceptuelle humaine et une analyse qualitative de spectrogrammes, les expériences montrent que les méthodes de référence actuelles présentent des compromis clairs entre ces différentes exigences. **AudioX** fournit la performance globale la plus équilibrée pour la génération complète de variations conditionnées par une référence, en combinant une forte fidélité perceptuelle, une bonne préservation de l'identité et une diversité satisfaisante avec un coût computationnel raisonnable. À l'inverse, **AudioLDM** offre une forte diversité et des transformations guidées par le texte plus agressives, mais peine à préserver l'identité perceptuelle. **ThinkSound** fournit un compromis intermédiaire entre diversité et alignement, mais n'atteint pas la qualité perceptuelle d'AudioX. **T-Foley** montre de fortes capacités de conditionnement temporel et énergétique, bien que son manifold de préentraînement limité restreigne sa généralisation à des distributions de SFX plus larges. Enfin, **A^{2}SB** obtient de bonnes performances de reconstruction pour des segments masqués courts, mais est conçu pour l'*inpainting* localisé plutôt que pour la génération complète de variations, de sorte que ses résultats doivent être interprétés en conséquence.

Limites des méthodes actuelles et perspectives de recherche —Globalement, aucune méthode de référence ne fournit simultanément une forte préservation de l'identité, une contrôlabilité explicite, une capacité d'édition localisée et une inférence efficace. Cela suggère que les travaux futurs devraient développer des pipelines unifiés combinant ces forces complémentaires, par

exemple à travers une génération ancrée sur la référence, des modules de contrôle explicites, des opérateurs d'édition localisée et des stratégies de génération en peu d'étapes plus efficaces.

Limites des protocoles d'évaluation et perspectives pour les futurs benchmarks —Ce chapitre met également en évidence que les comparaisons équitables restent difficiles en raison de biais de représentation, de pré-entraînements hétérogènes et de différences dans les taux d'échantillonnage ou les procédures d'inférence. Les futurs benchmarks devraient donc adopter des interfaces d'évaluation plus standardisées, une séparation plus claire des tâches et des protocoles plus alignés sur la perception, incluant des tests d'écoute structurés et des analyses de robustesse à travers les classes, la difficulté des références et différents régimes de contrôle.

Plus largement, les expériences montrent que les métriques fixes de benchmark ne reflètent pas toujours l'utilité perceptuelle des variations générées, en particulier pour les modèles produisant des distributions audio plus larges ou exploratoires. Avancer vers une génération audio prête pour la production nécessitera donc à la fois des pipelines génératifs plus robustes et des protocoles d'évaluation mieux alignés avec les besoins réels du design sonore.

3.8 Travaux futurs

Plusieurs pistes peuvent renforcer la génération de variations de SFX guidée par référence dans un cadre prêt pour la production, tout en améliorant la contrôlabilité des modèles et l'équité des comparaisons entre pipelines hétérogènes. Une première direction concerne le choix des représentations audio et la définition de signaux de contrôle pertinents pour les usages de production.

Choix des représentations audio —Les méthodes étudiées dans ce chapitre reposent sur des espaces de représentation différents, tels que la forme d'onde, les représentations temps–fréquence ou des latents appris. Or, ce choix influence directement la fidélité des transitoires, la localité des éditions, l'efficacité de l'échantillonnage et la robustesse globale. Les travaux futurs pourraient ainsi comparer ces représentations dans un cadre unifié de génération conditionnée par référence. Cela inclut notamment des paramétrisations spectrales à plus haute résolution pour

les événements riches en transitoires, des espaces latents audio appris, par exemple fondés sur des codecs (Défossez *et al.*, 2022), pour réduire le coût computationnel à l'inférence, ainsi que des approches hybrides permettant de préserver une structure temporelle fine tout en assurant une génération stable.

Conditionnement explicite fondé sur la structure du signal —Au-delà des simples étiquettes de classe, il apparaît également pertinent d'explorer des formes de conditionnement plus explicites et mieux alignées avec les pratiques de production. De tels signaux peuvent être dérivés de la référence, par exemple sous la forme d'enveloppes d'énergie, de pistes d'attaque, de contraintes de durée ou de descripteurs de réverbération. Ces informations s'accordent naturellement avec les pratiques du *sound design* et permettraient d'analyser plus systématiquement la contrôlabilité, en particulier le compromis entre intensité de la variation et préservation de l'identité.

Vers des pipelines unifiés —Une deuxième direction consiste à développer des pipelines capables de réunir les forces complémentaires observées parmi les méthodes de référence. Les résultats de ce chapitre montrent en effet que la préservation de l'identité, l'édition localisée, le contrôle temporel explicite et l'efficacité à l'inférence sont rarement combinés dans une même approche. Il paraît donc pertinent de concevoir des systèmes pensés pour répondre simultanément à ces différentes exigences.

Architectures modulaires et contrôle structuré —Dans cette perspective, des architectures modulaires pourraient constituer une voie prometteuse. Un ancrage sur la référence permettrait de préserver la cohérence d'identité et de timbre, tandis que des modules dédiés imposeraient des contraintes temporelles et énergétiques. En parallèle, des opérateurs d'édition localisée permettraient d'intervenir sur des régions temps–fréquence ou sur des segments masqués, afin de mieux répondre aux besoins concrets de production.

Équilibre entre fidélité, diversité et efficacité —Les travaux futurs gagneraient aussi à introduire des objectifs d'entraînement équilibrant explicitement la fidélité à la référence et la diversité des variations générées, plutôt que de laisser cette dernière émerger comme simple effet de l'échantillonnage stochastique. Par ailleurs, l'efficacité d'inférence reste un enjeu central pour

les usages en production. Dans cette optique, la génération en peu d'étapes, la distillation adaptée à l'audio et les stratégies d'échantillonnage adaptatif apparaissent comme des directions prometteuses, en particulier dans les contextes d'itération interactive où rapidité de retour et prévisibilité du contrôle sont essentielles.

Robustesse et stabilité en conditions réelles —Une troisième direction consiste à élargir les applications visées et à évaluer la robustesse des modèles dans des conditions plus réalistes. Au-delà des performances mesurées sur benchmark, l'usage pratique dans des *soundbanks* requiert une cohérence sous des conditions d'enregistrement variées, une robustesse au bruit de fond et aux environnements réverbérants, ainsi qu'une compatibilité avec des contraintes aval telles qu'une durée fixe, des cibles de *loudness*, des signaux sans discontinuité et un débit élevé en traitement par lots. Les futurs travaux pourraient donc évaluer explicitement les modèles sous ces contraintes et intégrer la génération guidée par référence dans des chaînes d'outils de *sound design* permettant un raffinement itératif, des contrôles de variation bornés et des contraintes compatibles avec le mixage. L'extension de ce cadre aux Foley alignés sur la vidéo et aux contextes d'édition interactive apparaît également naturelle, à condition que les objectifs d'alignement et les protocoles d'évaluation reflètent réellement les besoins temporels des flux de travail aval.

Évaluation équitable et alignée sur la perception —Enfin, une évaluation à la fois équitable et perceptuellement pertinente demeure un défi ouvert. Les biais de représentation, l'hétérogénéité des corpus de pré-entraînement et les différences entre procédures natives d'échantillonnage compliquent encore les comparaisons directes. En s'appuyant sur la matrice de capacités et les protocoles partagés introduits dans ce chapitre, les futurs benchmarks pourraient améliorer la comparabilité en standardisant le prétraitement, la gestion de la durée et du taux d'échantillonnage, tout en rapportant les coûts de calcul et de mémoire en parallèle des métriques de qualité. Sur le plan quantitatif, l'évaluation peut continuer à mobiliser des métriques au niveau du signal, comme la FAD, ainsi que des mesures fondées sur des *embeddings* audio partagés, par exemple ImageBind, pour quantifier l'alignement à la référence et la diversité. Néanmoins, la validité perceptuelle devrait être renforcée par des protocoles d'écoute structurés et par des évaluations

d'attributs fondées sur un vocabulaire audio partagé, tel que le réalisme, la fidélité de l'identité, la netteté des transitoires, la brillance spectrale, la texture ou encore les caractéristiques de réverbération. Combinées à des ablations contrôlées sur l'intensité du conditionnement, ces évaluations permettraient de mieux distinguer les améliorations réellement utiles en production des artefacts liés à une représentation donnée ou à une interface d'évaluation particulière.

Ensemble, ces directions pourraient faire évoluer la génération de variations de SFX guidée par référence d'un simple exercice de comparaison sur benchmark vers des pipelines plus fiables, plus contrôlables et plus efficaces, mieux adaptés aux exigences concrètes de la production.

CONCLUSION ET RECOMMANDATIONS

Les modèles de diffusion sont devenus un paradigme dominant pour la modélisation générative à haute fidélité à travers différentes modalités. Cependant, leur déploiement pratique reste contraint par deux défis récurrents : le coût computationnel élevé de l'échantillonnage itératif et la difficulté d'adapter des modèles préentraînés puissants à de nouveaux domaines à partir de seulement quelques références. Ces limitations sont particulièrement saillantes dans deux contextes d'application étudiés dans cette thèse : la *Subject-Driven Personalization (SDP)* pour la génération d'images et la génération et édition de variations de *Sound Effects (SFX)* guidées par référence dans des contextes de production. Dans ces deux cadres, une même tension fondamentale apparaît entre efficacité, contrôlabilité, préservation de l'identité et diversité des sorties.

Les Chapitres 1 et 2 établissent le contexte conceptuel et méthodologique, en mettant en évidence que les approches existantes traitent souvent l'accélération (distillation) et l'adaptation séparément, ce qui conduit à des pipelines en deux étapes qui soit sacrifient la qualité sous un échantillonnage rapide, soit nécessitent une adaptation coûteuse avant l'accélération. S'appuyant sur cette lacune, le Chapitre 2 introduit `Uni-DAD SDP`, qui unifie distillation et adaptation pour la personnalisation few-shot sous une contrainte d'échantillonnage extrême. La méthode étend le *dual-domain Distribution Matching (DMD)* au cadre conditionnel via un mécanisme de *prompt routing* : les *instance prompts* (token rare + nom de classe) conditionnent les composantes apprises afin d'associer l'identité du sujet, tandis que les *class-prior prompts* conditionnent le *source teacher* gelé afin de préserver la diversité au niveau de la classe et d'améliorer la généralisation des prompts. Une contrainte adversariale de réalisme conditionnée par le texte renforce en outre le réalisme spécifique au sujet sous contrôle du prompt. Évalué sur le benchmark DreamBooth avec un protocole reproductible et un reporting par instance, `Uni-DAD SDP` démontre qu'une génération personnalisée de haute qualité peut rester compétitive avec des baselines multi-étapes tout en ne nécessitant qu'une seule étape de débruitage à l'inférence

(NFE = 1). L'analyse met également en évidence les modes d'échec des stratégies en deux étapes, notamment le sur-lissage dans les pipelines *distill-then-adapt* et l'effondrement de diversité ou la mémorisation dans les configurations *adapt-then-distill*, soutenant la conclusion plus générale selon laquelle traiter conjointement adaptation et distillation constitue une direction pratique pour la personnalisation rapide.

En complément du cadre image, le Chapitre 3 étudie la génération de variations de SFX guidée par référence dans une perspective de production, où l'évaluation doit refléter des exigences allant au-delà de la simple qualité audio. Une matrice d'exigences orientée production est introduite afin de formaliser les contraintes pertinentes pour le design sonore, notamment la préservation de l'identité, la fidélité et le réalisme, la variation contrôlable, l'alignement temporel et le contrôle énergétique, la modification ciblée, la robustesse et l'efficacité. Une matrice de capacités et un protocole expérimental partagé sur la soundbank ESC-50 permettent une sélection structurée des baselines et des comparaisons plus équitables entre familles de modèles et représentations hétérogènes. Le benchmark résultant combine une évaluation de qualité au niveau du signal via la FAD (à l'aide d'AudioLDM-eval), des mesures d'alignement et de diversité basées sur des embeddings (via ImageBind), et une validation perceptuelle à travers une étude S-MOS. Les expériences révèlent des compromis distincts entre les baselines : AudioX fournit la performance la plus équilibrée pour la génération complète de variations conditionnées par référence, avec une forte fidélité et une bonne préservation de l'identité tout en maintenant une diversité significative ; AudioLDM permet des transformations agressives guidées par le texte mais présente une préservation perceptuelle de l'identité plus faible dans les contextes conditionnés par référence ; ThinkSound offre un compromis intermédiaire entre alignement et diversité mais reste limité en netteté perceptuelle ; et T-Foley préserve les dynamiques temporelles-énergétiques grossières mais généralise difficilement au-delà de son manifold de préentraînement. À l'inverse, A^2SB obtient de bons résultats pour l'inpainting localisé, bien que ses sorties ne soient pas directement comparables à celles de la génération complète de

variations, soulignant l'importance d'adapter les protocoles d'évaluation à la définition de la tâche. Au-delà de la comparaison des modèles, ce chapitre met également en évidence que l'évaluation équitable reste difficile lorsque les méthodes diffèrent en termes de représentation, de données de préentraînement, de procédures d'échantillonnage et de formats audio natifs, et que les métriques standard peuvent sous-représenter l'utilité perceptuelle ou la variabilité créative dans des contextes de production.

Dans l'ensemble, les contributions de cette thèse répondent aux problématiques initiales de deux manières complémentaires. Premièrement, l'étude sur la personnalisation d'images démontre que distillation et adaptation peuvent être unifiées afin de réduire le coût d'inférence sans compromettre la qualité de la personnalisation few-shot, réduisant ainsi l'écart entre les pipelines de diffusion de recherche et les contraintes de personnalisation en temps réel. Deuxièmement, l'étude sur les SFX propose une méthodologie structurée pour sélectionner et évaluer les systèmes de génération audio guidés par référence dans des contextes de production, en clarifiant quelles capacités sont actuellement satisfaites, lesquelles restent partielles et où la conception des modèles et des benchmarks doit évoluer pour permettre un déploiement fiable.

Plusieurs directions de recherche découlent naturellement de ces résultats. Du côté de la génération d'images, les travaux futurs pourraient réduire davantage la complexité d'entraînement en simplifiant la dépendance aux enseignants, améliorer la robustesse face à des changements de domaine sévères et renforcer les contraintes d'identité sous des prompts difficiles tout en conservant une génération en une seule étape. Du côté de l'audio, les progrès sont attendus à travers des choix de représentation et des interfaces de contrôle explicites mieux alignés avec les pratiques du design sonore (par exemple des descripteurs d'enveloppe, d'attaque, de durée et de réverbération), ainsi que des pipelines modulaires combinant ancrage sur la référence, mécanismes de variation contrôlable et édition localisée dans une interface unifiée. À travers les deux modalités, une priorité récurrente concerne l'évaluation : des benchmarks plus alignés

sur la perception, une standardisation du reporting des coûts de calcul et de mémoire, et des protocoles humains fondés sur des attributs perceptuels définis dans un vocabulaire audio partagé pourraient améliorer la comparabilité et mieux refléter la valeur réelle en production. Relier ces directions contribuerait à faire évoluer la génération basée sur les modèles de diffusion, d'une synthèse de haute qualité vers des systèmes d'édition multimodale efficaces, contrôlables et orientés utilisateur, adaptés à un déploiement pratique.

ANNEXE I

APPENDIX CHAPTER 2

Tableau-A I-1 Identification des instances utilisées pour la table 2.3 depuis le benchmark de Dreambooth Ruiz *et al.* (2023)

ID	Instance	ID	Instance
1	backpack	16	dog7
2	backpack_dog	17	dog8
3	bear_plushie	18	duck_toy
4	berry_bowl	19	fancy_boot
5	can	20	grey_sloth_plushie
6	candle	21	monster_toy
7	cat	22	pink_sunglasses
8	cat2	23	poop_emoji
9	clock	24	rc_car
10	colorful_sneaker	25	red_cartoon
11	dog	26	robot_toy
12	dog2	27	shiny_sneaker
13	dog3	28	teapot
14	dog5	29	vase
15	dog6	30	wolf_plushie

ANNEXE II

APPENDIX CHAPTER 3

Tableau-A II-1 Matrice complète des capacités des méthodes de référence pour l'application de génération de variations de SFX, au regard des exigences principales R1-5 du Tab. 3.1 - **NS** = non spécifié / non évalué

Methods	R1	R2, R3	R4	R5	R6	R7	R8	R9	R10	R11	Fit
Audio Editing											
ThinkSound$^{\gamma}$ (NeurIPS, 2025)	FM (latent); A/V/T; CoT (text)	FD↓; $MOS-A$ ↑; $MOS-Q$ ↑	**NS**	**NS**	Sync; DeSync	**NS**	Targeted editing	Targeted editing	OOD eval.	Inference time (s)↓	Medium (strong edit, text-heavy)
AudioEditor (ICASSP, 2025)	Diff. TTA; mel-spec; VAE-latent; A/T	FAD↓; FD ↓; $CLAP$ ↑; IS ↑	Preserve unedited; Null-text inv.; EOT-supp.	**NS**	**NS**	**NS**	Add Del Rep (text regions)	Add Del Rep (text regions)	**NS**	Training-free	Low (edit, text-cond, code unclear)
AudioMorphix$^{\gamma}$ (ArxiV, 2025)	Latent diffusion; spec/ latent; A/ Aref	FAD↓; KL ↓; SF ↑	Preserve unedited; region spec.	Continuous control	Time-stretch/ shift	Pitch-shift	**NS**	Region edit; Add/ Del/ Rep/ Move	**NS**	Training-free, High memory needs	High (audio-ref)
T-FOLEY$^{\gamma}$ (ICASSP, 2024)	Diff (wform); AR; class + time feat.	FAD↓; MOS ↑	Class identity	IS	E-L1↓; MOS ↑	RMS envelope cond.	Time feat. + Block-FiLM	**NS**	OOD eval.	Inference time; E-L1↓ $vsFAD$ ↓	Medium (time/energy, no audio-ref)
A^{2}SB$^{\gamma}$ (ArxiV, 2025)	Diffusion Schrödinger Bridge; factorized STFT (3 channels); no vocoder	ViSQOL↑ MOS ↑; LSD ↓	SiSpec↑	stochastic sampling (same mask)	preserve unmask region	**NS**	~ shape length mask	Inpainting	**NS**	565M params	Medium (no controllability within the inpainted segment)
Solo-Audio* (ICASSP, 2025)	Diff-TSE; A/T	ViSQOL↑; FD ↓; KL ↓; $CLAP$ ↑	CLAP-audio↑ (aux)	**NS**	**NS**	**NS**	**NS**	Extract/ mask target	OOD eval.	**NS**	Auxiliary (preprocess)
ImageBind* (CVPR, 2023)	Embedding space (T/A/V/D)	Acc.↑; $recall$ ↑ (aux)	Embed sim/ dist (eval)	Embed sim/ dist (eval)	Temp.-aligned pretrain	**NS**	**NS**	**N/A**	Cross-modal	**NS**	Auxiliary (eval/ retrieval)
Audio Generation											
AudioX† (ICLR, 2026)	Latent DiT; concat cond.; V/T/A	KL↓; FD ↓; FAD ↓	CLAP/ ImageBind $(align)$	IS↑	CLAP/ ImageBind↑ $(align)$	**NS**	Text control	Inpaint/ complete	Masking-style train.	**NS**	Low (inpaint, text-cond)
AudioLDM (ICML, 2023)	Latent diffusion + VAE (mel); T/A	FD↓; IS ↑; KL ↓; FAD ↓; MOS ↑	Inpainting	IS↑	Temp. order (text)	Pitch (text)	CFG; style weight; start point	Inpainting	**NS**	**NS**	Low (edit, text-cond)
AudioGen† (ICLR, 2023)	AR-T; disc. tokens; T/A	FAD↓; MOS ↑	Continu. (audio prefix)	CFG sweep (FAD/ KL)	Continu.	**NS**	Guidance scale γ	**NS**	**NS**	Multi-stream	Low (no full-clip variation; limited control)
EDMSound† (SD, 2023)	EDM diffusion; complex spec; class-cond.	FAD↓	Class identity	IS	**NS**	**NS**	CFG; class labels	**NS**	Copy detection	10-50 steps	Low (fidelity/ speed; no audio-ref)

* Composants auxiliaires. γ Candidats principaux ; † méthodes de référence secondaires.

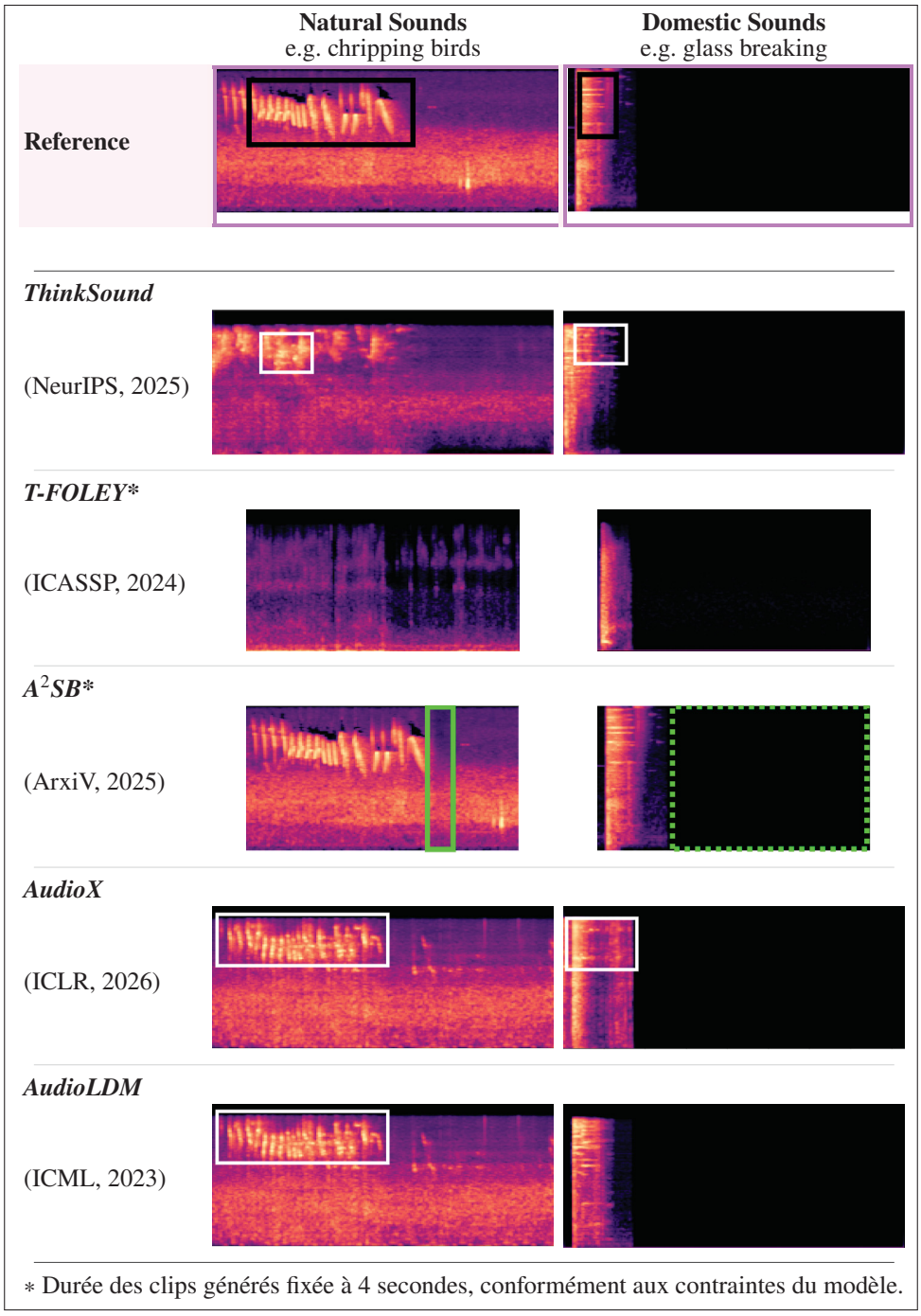

Figure-A II-1 Visualisations complémentaires de l'évolution fréquentielle et de l'amplitude des clips pour chaque modèle dans la tâche de variation A2A, à partir de mel-spectrogrammes - Pour la tâche d'inpainting de A^2SB, les sections masquées puis reconstruites sont encadrées en vert, car l'évaluation sur l'ensemble complet ne sélectionne que le segment compris entre 0.3,s et 1,s pour le masquer puis le régénérer - Les lignes pointillées indiquent que les variations peuvent s'être produites dans des zones de bruit blanc (silence) - Des expériences complémentaires avec des durées de masquage plus larges sont également menées afin de mieux se rapprocher de la tâche de génération de variations

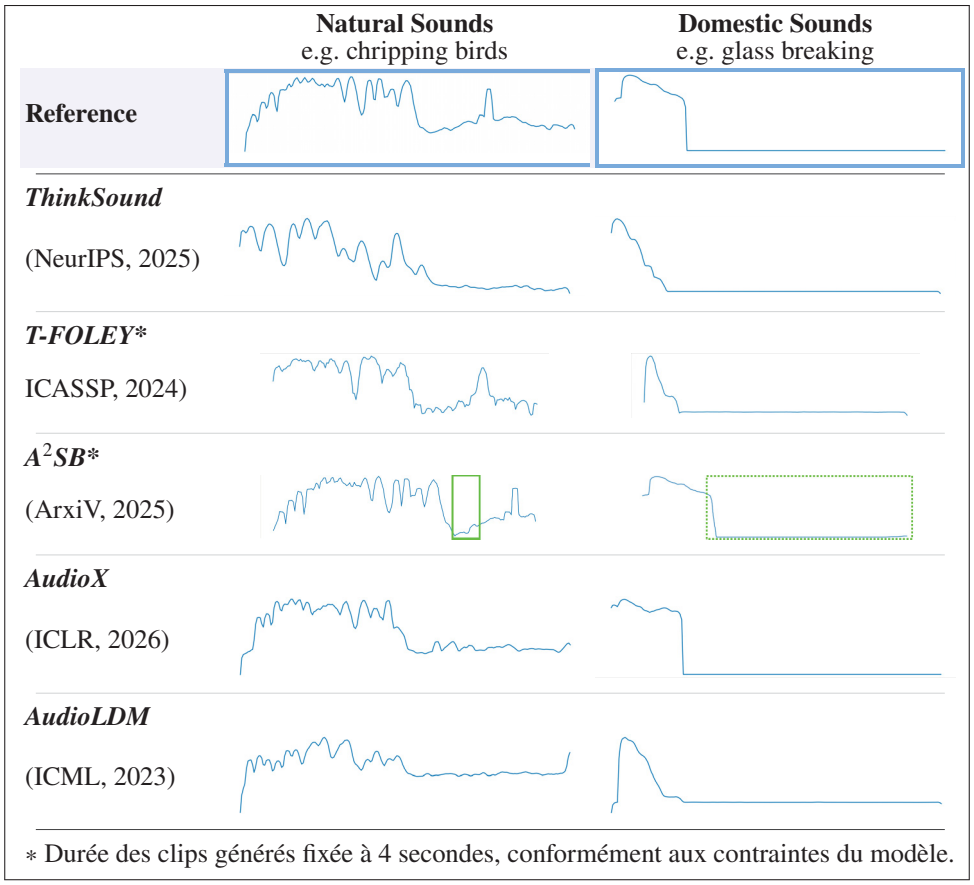

Figure-A II-2 Visualisation des variations d'énergie des clips pour chaque modèle dans la tâche d'édition A2A - Chaque panneau représente l'évolution de l'énergie (dB) en fonction du temps (s) - La majeure partie de l'énergie des clips se situe entre [-5, -80] dB - Tous les clips durent 5 secondes, à l'exception de T-Foley et de A^2SB, qui produisent des générations de 4 secondes - Cette figure permet d'observer dans quelle mesure chaque méthode reproduit ou s'éloigne du profil énergétique du signal de référence - Pour la tâche d'inpainting de A^2SB, les sections masquées puis reconstruites sont indiquées sur la courbe d'énergie, car l'évaluation sur l'ensemble complet ne sélectionne que le segment compris entre 0.3,s et 1,s pour le masquer puis le régénérer - Des expériences complémentaires avec des durées de masquage plus étendues sont également menées afin de mieux se rapprocher de la tâche de génération de variations

BIBLIOGRAPHIE

Agostinelli, A., Denk, T. I., Borsos, Z., Engel, J., Verzetti, M., Caillon, A., Huang, Q., Jansen, A., Roberts, A., Tagliasacchi, M., Sharifi, M., Zeghidour, N. & Frank, C. (2023). MusicLM : Generating Music From Text. *arXiv preprint arXiv :2301.11325*. Repéré à https://arxiv.org/abs/2301.11325.

Arjovsky, M., Chintala, S. & Bottou, L. (2017). Wasserstein generative adversarial networks. *International conference on machine learning*, pp. 214–223.

Bahram, Y., Desbos, M., Shateri, M. & Granger, E. (2025). Uni-DAD : Unified Distillation and Adaptation of Diffusion Models for Few-step Few-shot Image Generation. *arXiv preprint arXiv :2511.18281*. doi : 10.48550/arXiv.2511.18281.

Bahram, Y., Shateri, M. & Granger, E. (2026). DogFit : Domain-guided Fine-tuning for Efficient Transfer Learning of Diffusion Models. *Proceedings of the AAAI Conference on Artificial Intelligence*.

Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V. & Babenko, A. (2021). Label-efficient semantic segmentation with diffusion models. *arXiv preprint arXiv :2112.03126*.

Cai, X., Huang, Q., Zhang, Y., Chen, H. & Zhang, W. (2023). AUDIT : Audio Editing by Following Instructions with Latent Diffusion Models.

Cao, Y. & Gong, S. (2024). Few-shot image generation by conditional relaxing diffusion inversion. *European Conference on Computer Vision*, pp. 20–37.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P. & Joulin, A. (2021). Emerging Properties in Self-Supervised Vision Transformers. *arXiv preprint arXiv :2104.14294*. Repéré à https://arxiv.org/abs/2104.14294. Submitted 29 Apr 2021 ; Revised 24 May 2021.

Cartwright, M. & Pardo, B. (2015). VocalSketch : Vocally imitating audio concepts. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pp. 43–46.

Chadebec, C., Tasar, O., Benaroche, E. & Aubin, B. (2025). Flash diffusion : Accelerating any conditional diffusion model for few steps image generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(15), 15686–15695.

Chae, Y. & Lee, K. (2025). MGE-LDM : Joint Latent Diffusion for Simultaneous Music Generation and Source Extraction. *NeurIPS 2025 (Poster)*. Repéré à https://neurips.cc/virtual/2025/poster/120257.

Chen, H., Xie, W., Vedaldi, A. & Zisserman, A. (2020). VGG-Sound : A Large-Scale Audio-Visual Dataset. *arXiv preprint arXiv :2004.14368*.

Chen, K., Wu, Y., Liu, H., Nezhurina, M., Berg-Kirkpatrick, T. & Dubnov, S. (2023). MusicLDM : Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies. Repéré à https://arxiv.org/abs/2308.01546.

Chen, Y., Niu, Z., Ma, Z., Deng, K., Wang, C., Zhao, J., Yu, K. & Chen, X. (2024a). F5-TTS : A Fairytaler that Fakes Fluent and Faithful Speech with Flow Matching. Repéré à https://arxiv.org/abs/2410.06885.

Chen, Z., Seetharaman, P., Russell, B., Nieto, O., Bourgin, D., Owens, A. & Salamon, J. (2025). Video-Guided Foley Sound Generation with Multimodal Controls. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18770–18781. doi : 10.1109/CVPR52734.2025.01749.

Chen, Z., Yu, Q., Niu, H., Liu, H. & Wang, W. (2024b). T-FOLEY : A Controllable Waveform-Domain Diffusion Model for Temporal Event Guided Foley Sound Synthesis.

Cheng, H. K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A. & Mitsufuji, Y. [Accepted to CVPR 2025]. (2024). MMAudio : Taming Multimodal Joint Training for High-Quality Video-to-Audio Synthesis. Repéré à https://arxiv.org/abs/2412.15322.

Choi, K., Im, J., Heller, L. M., McFee, B., Imoto, K., Okamoto, Y., Lagrange, M. & Takamichi, S. (2023). Foley sound synthesis at the DCASE 2023 challenge. *Proceedings of the 8th Detection and Classification of Acoustic Scenes and Events 2023 Workshop (DCASE2023)*, pp. 16–20.

Choi, Y., Uh, Y., Yoo, J. & Ha, J.-W. (2020). Stargan v2 : Diverse image synthesis for multiple domains. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8188–8197.

Chung, Y., Lee, J. & Nam, J. (2024). T-FOLEY : A Controllable Waveform-Domain Diffusion Model for Temporal-Event-Guided Foley Sound Synthesis. Repéré à https://arxiv.org/abs/2401.09294.

Cífka, O., Kumar, A., Dupont, E., Mas, N., Rybakov, O., Gross, J., Essid, S. & Afouras, T. (2025). A2SB : Audio-to-Audio Schrödinger Bridges.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y. & Défossez, A. [Published at NeurIPS 2023 ; arXiv v3 (Jan 2024)]. (2024). Simple and Controllable Music Generation. Repéré à https://arxiv.org/abs/2306.05284.

Défossez, A., Zeghidour, N., Usunier, N. & Bottou, L. (2022). High Fidelity Neural Audio Compression. *Advances in Neural Information Processing Systems (NeurIPS)*.

Deja, K., Kuzina, A., Trzcinski, T. & Tomczak, J. (2022). On analyzing generative and denoising capabilities of diffusion-based deep generative models. *Advances in Neural Information Processing Systems*, 35, 26218–26229.

Dhariwal, P. & Nichol, A. (2021). Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34, 8780–8794.

Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A. & Sutskever, I. (2020). Jukebox : A Generative Model for Music. Repéré à https://arxiv.org/abs/2005.00341.

Dong, H.-W., Liu, X., Pons, J., Bhattacharya, G., Pascual, S., Serrà, J., Berg-Kirkpatrick, T. & McAuley, J. (2023). Clipsonic : Text-to-audio synthesis with unlabeled videos and pretrained language-vision models. *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5.

Elsevier. (n.d.). Acoustic Modeling. Repéré le 2026-01-30 à https://www.sciencedirect.com/topics/physics-and-astronomy/acoustic-modeling.

Eskimez, S. E., Wang, X., Thakker, M., Li, C., Tsai, C.-H., Xiao, Z., Yang, H., Zhu, Z., Tang, M., Tan, X., Liu, Y., Zhao, S. & Kanda, N. (2024). E2 TTS : Embarrassingly Easy Fully Non-Autoregressive Zero-Shot TTS. *IEEE Spoken Language Technology Workshop, SLT 2024, Macao, December 2–5, 2024*, pp. 682–689. doi : 10.1109/SLT61566.2024.10832320.

Evans, Z., Parker, J. D., Carr, C., Zukowski, Z., Taylor, J. & Pons, J. (2024). Stable Audio Open.

Fang, P., He, Y., Xing, Y., Chen, Q., Lim, S.-N. & Yang, H. [OpenReview paper page (verify venue/year if needed)]. (2025). AC-FOLEY : Reference-Audio-Guided Video-to-Audio Synthesis with Acoustic Transfer.

Flores García, H., Nieto, O., Salamon, J., Pardo, B. & Seetharaman, P. (2024a). Sketch2Sound : Controllable Audio Generation via Time-Varying Signals and Sonic Imitations. Repéré à https://arxiv.org/abs/2412.08550.

Flores García, H., Nieto, O., Salamon, J., Pardo, B. & Seetharaman, P. (2024b). Sketch2Sound : Controllable Audio Generation via Time-Varying Signals and Sonic Imitations. *arXiv preprint arXiv :2412.08550*. doi : 10.48550/arXiv.2412.08550.

Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G. & Cohen-Or, D. (2022). An Image is Worth One Word : Personalizing Text-to-Image Generation using Textual Inversion. *arXiv preprint arXiv :2208.01618*.

Gandikota, R. & Bau, D. (2025). Distilling diversity and control in diffusion models. *arXiv preprint arXiv :2503.10637*.

Gemmeke, J. F., Ellis, D. P. W., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M. & Ritter, M. (2017). AudioSet : An Ontology and Human-Labeled Dataset for Audio Events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A. & Misra, I. (2023a). ImageBind : One Embedding Space To Bind Them All. *arXiv preprint arXiv :2305.05665*. doi : 10.48550/arXiv.2305.05665.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K., Joulin, A. & Misra, I. (2023b). ImageBind : One Embedding Space To Bind Them All. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Grassucci, E., Galadini, G., Cicchetti, G., Uncini, A., Antonacci, F. & Comminiello, D. (2025). Training-Free Multimodal Guidance for Video to Audio Generation. *arXiv preprint arXiv :2509.24550*. doi : 10.48550/arXiv.2509.24550.

Grötschla, F., Solak, A., Lanzendörfer, L. A. & Wattenhofer, R. (2025). Benchmarking Music Generation Models and Metrics via Human Preference Studies. *arXiv preprint arXiv :2506.19085*. Repéré à https://arxiv.org/abs/2506.19085.

Guan, W., Wang, K., Zhou, W., Wang, Y., Deng, F., Wang, H., Li, L., Hong, Q. & Qin, Y. (2024). LAFMA : A Latent Flow Matching Model for Text-to-Audio Generation. *Proc. Interspeech 2024*, pp. 4813–4817. doi : 10.21437/Interspeech.2024-1848.

Hai, J., Xu, Y., Zhang, H. et al. (2024). EzAudio : Enhancing Text-to-Audio Generation with Efficient Diffusion Transformer. *arXiv preprint arXiv :2409.10819*.

Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D. & Yang, F. (2023). SVDiff : Compact Parameter Space for Diffusion Fine-Tuning. *arXiv preprint arXiv :2303.11305*.

Hawthorne, C., Simon, I., Roberts, A., Zeghidour, N., Gardner, J., Manilow, E. & Engel, J. (2022). Multi-instrument Music Synthesis with Spectrogram Diffusion. Repéré à https://arxiv.org/abs/2206.05408.

HKUSTAudio. [Hugging Face. Last updated February 10, 2026. Retrieved February 25, 2026]. (2026). AudioX (HKUSTAudio/AudioX) [Model repository]. Repéré à https://huggingface.co/HKUSTAudio/AudioX/tree/main.

Ho, J. & Salimans, T. (2022). Classifier-free diffusion guidance. *arXiv preprint arXiv :2207.12598*.

Ho, J., Jain, A. & Abbeel, P. (2020a). Denoising Diffusion Probabilistic Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33, 6840–6851. Repéré à https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Ho, J., Jain, A. & Abbeel, P. (2020b). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33, 6840–6851.

Hsiao, Y.-T., Khodadadeh, S., Duarte, K., Lin, W.-A., Qu, H., Kwon, M. & Kalarot, R. (2024). Plug-and-play diffusion distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13743–13752.

Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L. & Chen, W. (2021). LoRA : Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv :2106.09685*.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. et al. (2022). Lora : Low-rank adaptation of large language models. *ICLR*, 1(2), 3.

Huang, J., Ren, Y., Huang, R., Yang, D., Ye, Z., Zhang, C., Liu, J., Yin, X., Ma, Z. & Zhao, Z. [arXiv preprint]. (2023a). Make-An-Audio 2 : Temporal-Enhanced Text-to-Audio Generation. Repéré à https://arxiv.org/abs/2305.18474.

Huang, Q., Park, D. S., Wang, T., Denk, T. I., Ly, A., Chen, N., Zhang, Z., Zhang, Z., Yu, J., Frank, C., Engel, J., Le, Q. V., Chan, W., Chen, Z. & Han, W. (2023b). Noise2Music : Text-conditioned Music Generation with Diffusion Models. Repéré à https://arxiv.org/abs/2302.03917.

Huang, R., Li, M., Yang, D., Shi, J., Chang, X., Ye, Z., Wu, Y., Hong, Z., Huang, J., Liu, J., Ren, Y., Zhao, Z. & Watanabe, S. (2023c). AudioGPT : Understanding and Generating Speech, Music, Sound, and Talking Head. Repéré à https://arxiv.org/abs/2304.12995.

Huo, J., Hou, J., Wang, X., Ye, Z., Liu, X. & Yuan, J. (2024). AudioEditor : A Training-Free Diffusion-Based Audio Editing System.

Iashin, V. & Rahtu, E. (2021). Taming Visually Guided Sound Generation. *arXiv preprint arXiv :2110.08791*. doi : 10.48550/arXiv.2110.08791.

Jeong, Y., Kim, Y., Chun, S. & Lee, J. (2024). Read, Watch and Scream ! Sound Generation from Text and Video. *arXiv preprint arXiv :2407.05551*. doi : 10.48550/arXiv.2407.05551.

Jia, Y., Chen, Y., Zhao, J., Zhao, S., Zeng, W., Chen, Y. & Qin, Y. (2024). AudioEditor : A Training-Free Diffusion-Based Audio Editing Framework. *arXiv preprint arXiv :2409.12466*. doi : 10.48550/arXiv.2409.12466.

Karras, T., Aila, T., Laine, S. & Lehtinen, J. (2018). Progressive Growing of GANs for Improved Quality, Stability, and Variation. *International Conference on Learning Representations (ICLR)*.

Karras, T., Laine, S. & Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410.

Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J. & Aila, T. (2020a). Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33, 12104–12114.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. & Aila, T. (2020b). Analyzing and improving the image quality of stylegan. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119.

Kim, C. D., Kim, B., Lee, H. & Kim, G. (2019). AudioCaps : Generating Captions for Audios in the Wild. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, pp. 119–132.

Kim, J., Hong, Y. & Ye, J. C. (2025). FlowAlign : Trajectory-Regularized, Inversion-Free Flow-based Image Editing.

Kingma, D. P. & Welling, M. (2013). Auto-Encoding Variational Bayes. *arXiv preprint arXiv :1312.6114*.

Kong, Z., Ping, W., Huang, J. et al. (2021). DiffWave : A Versatile Diffusion Model for Audio Synthesis. *International Conference on Learning Representations (ICLR)*.

Kong, Z., Shih, K. J., Nie, W., Vahdat, A., Lee, S.-g., Santos, J. F., Jukic, A., Valle, R. & Catanzaro, B. (2025). A2SB : Audio-to-Audio Schrodinger Bridges. *arXiv preprint arXiv :2501.11311*.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y. & Adi, Y. (2022a). AudioGen : Textually Guided Audio Generation.

Kreuk, F., Synnaeve, G., Polyak, A., Singer, U., Défossez, A., Copet, J., Parikh, D., Taigman, Y. & Adi, Y. (2022b). AudioGen : Textually Guided Audio Generation. Repéré à https://arxiv.org/abs/2209.15352.

Kumari, N., Zhang, B., Zhang, R., Shechtman, E. & Zhu, J.-Y. (2023). Multi-Concept Customization of Text-to-Image Diffusion. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19200–19210.

Kushwaha, S. S. & Tian, Y. (2024). VinTAGe : Joint Video and Text Conditioning for Holistic Audio Generation. Repéré à https://arxiv.org/abs/2412.10768.

Lai, C.-H., Song, Y., Kim, D., Mitsufuji, Y. & Ermon, S. (2025). The Principles of Diffusion Models. Repéré à https://arxiv.org/abs/2510.21890.

Lee, W.-J., Hsieh, F.-C., Chen, X., Tsai, F.-D. & Yang, Y.-H. (2026). Training-Efficient Text-to-Music Generation with State-Space Modeling. *arXiv preprint arXiv :2601.14786*. Repéré à https://arxiv.org/abs/2601.14786.

Leng, X. [Accessed : 2025-10-20]. (2022). LAION Aesthetics v2 6.25+ Dataset. Repéré à https://huggingface.co/datasets/xingjianleng/laion_aesthetics_v2_6.25plus.

Levy, M., Di Giorgi, B., Weers, F., Katharopoulos, A. & Nickson, T. (2023). Controllable Music Production with Diffusion Models and Guidance Gradients. Repéré à https://arxiv.org/abs/2311.00613.

Li, S., Zhang, Y., Tang, F., Ma, C., Dong, W. & Xu, C. (2024). Music Style Transfer with Time-Varying Inversion of Diffusion Models. Repéré à https://arxiv.org/abs/2402.13763.

Li, T. & He, K. (2025). Back to Basics : Let Denoising Generative Models Denoise. Repéré à https://arxiv.org/abs/2511.13720.

Liang, J., Chen, Y., Yuan, Y., Jia, D., Zhuang, X., Chen, Z., Wang, Y. & Wang, Y. (2025). AudioMorphix : Training-free audio editing with diffusion probabilistic models. *arXiv preprint arXiv :2505.16076*. doi : 10.48550/arXiv.2505.16076.

Lim, J. H. & Ye, J. C. (2017). Geometric gan. *arXiv preprint arXiv :1705.02894*.

Ling, H., Kreis, K., Li, D., Kim, S. W., Torralba, A. & Fidler, S. (2021). EditGAN : High-Precision Semantic Image Editing. *arXiv preprint arXiv :2111.03186*. doi : 10.48550/arXiv.2111.03186.

Lipman, Y., Chen, R. T. Q., Ben-Hamu, H., Nickel, M. & Le, M. (2023). Flow Matching for Generative Modeling. *arXiv preprint arXiv :2210.02747*.

Liu, H. (2023). audioldm_eval : Audio Generation Evaluation. Repéré le 2026-02-26 à https://github.com/haoheliu/audioldm_eval.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W. & Plumbley, M. D. (2023a). AudioLDM : Text-to-Audio Generation with Latent Diffusion Models. *arXiv preprint arXiv :2301.12503*. doi : 10.48550/arXiv.2301.12503.

Liu, H., Chen, Z., Yuan, Y., Mei, X., Liu, X., Mandic, D., Wang, W. & Plumbley, M. D. (2023b). AudioLDM : Text-to-Audio Generation with Latent Diffusion Models.

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y. & Plumbley, M. D. (2023c). AudioLDM 2 : Learning Holistic Audio Generation with Self-supervised Pretraining.

Liu, H., Yuan, Y., Liu, X., Mei, X., Kong, Q., Tian, Q., Wang, Y., Wang, W., Wang, Y. & Plumbley, M. D. (2023d). AudioLDM 2 : Learning Holistic Audio Generation with Self-supervised Pretraining. *arXiv preprint arXiv :2308.05734*. Repéré à https://arxiv.org/abs/2308.05734.

Liu, H., Luo, K., Wang, J., Wang, W., Chen, Q., Zhao, Z. & Xue, W. (2025a). ThinkSound : Chain-of-Thought Reasoning in Multimodal Large Language Models for Audio Generation and Editing. *arXiv preprint arXiv :2506.21448*. doi : 10.48550/arXiv.2506.21448.

Liu, H., Luo, K., Wang, J., Wang, W., Chen, Q., Zhao, Z. & Xue, W. (2025b). ThinkSound : Chain-of-Thought Reasoning in Multimodal Large Language Models for Audio Generation and Editing.

Liu, S., Hussain, A. S., Wu, Q., Sun, C. & Shan, Y. (2023e). M^2UGen : Multi-modal Music Understanding and Generation with the Power of Large Language Models. Repéré à https://arxiv.org/abs/2311.11255.

Liu, S., Pan, X., Yuan, T., Wang, C., Ren, Z., Yang, D., Zhang, W., Yin, G., Wu, X. et al. (2025c). AudioX : Diffusion Transformer for Anything-to-Audio Generation.

Liu, X., Li, H., Ye, Z., Wu, Y., Liu, W., Chen, B. & Liu, X. (2025d). Music Style Transfer with Time-Varying Inversion of Diffusion Models.

Liu, Z., Ding, S., Zhang, Z., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D. & Wang, J. (2025e). SongGen : A Single Stage Auto-regressive Transformer for Text-to-Song Generation. *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 267(Proceedings of Machine Learning Research), 38351–38364. Repéré à https://proceedings.mlr.press/v267/liu25m.html.

Loizou, P. C. (2011). Speech Quality Assessment. Dans Lin, W., Tao, D., Kacprzyk, J., Li, Z., Izquierdo, E. & Wang, H. (Éds.), *Multimedia Analysis, Processing and Communications* (pp. 623–654). Berlin, Heidelberg : Springer Berlin Heidelberg. doi : 10.1007/978-3-642-19551-8_23.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C. & Zhu, J. (2022). Dpm-solver : A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35, 5775–5787.

Luo, S., Tan, Y., Huang, L., Li, J. & Zhao, H. (2023a). Latent consistency models : Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv :2310.04378*.

Luo, S., Tan, Y., Patil, S., Gu, D., von Platen, P., Passos, A., Huang, L., Li, J. & Zhao, H. (2023b). Lcm-lora : A universal stable-diffusion acceleration module. *arXiv preprint arXiv :2311.05556*.

Luo, S., Yan, C., Hu, C. & Zhao, H. (2023c). Diff-Foley : Synchronized Video-to-Audio Synthesis with Latent Diffusion Models.

Luo, S., Yan, C., Hu, C. & Zhao, H. (2023d). Diff-Foley : Synchronized Video-to-Audio Synthesis with Latent Diffusion Models. Repéré à https://arxiv.org/abs/2306.17203.

Ma, J., Liang, J., Chen, C. & Lu, H. (2024a). Subject-Diffusion : Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. *ACM SIGGRAPH Conference Papers*, pp. 1–12.

Ma, N., Goldstein, M., Albergo, M. S., Boffi, N. M., Vanden-Eijnden, E. & Xie, S. (2024b). SiT : Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers. Repéré à https://arxiv.org/abs/2401.08740.

Majumder, N., Hung, C.-Y., Ghosal, D., Hsu, W.-N., Mihalcea, R. & Poria, S. [Accepted at ACM MM 2024]. (2024). Tango 2 : Aligning Diffusion-based Text-to-Audio Generations through Direct Preference Optimization. Repéré à https://arxiv.org/abs/2404.09956.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z. & Paul Smolley, S. (2017). Least squares generative adversarial networks. *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802.

Mehri, S., Kumar, K., Gulrajani, I. et al. (2017). SampleRNN : An Unconditional End-to-End Neural Audio Generation Model. *International Conference on Learning Representations (ICLR)*.

Mei, K., Delbracio, M., Talebi, H., Tu, Z., Patel, V. M. & Milanfar, P. (2024). Codi : Conditional diffusion distillation for higher-fidelity and faster image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9048–9058.

Mei, X., Nagaraja, V., Le Lan, G., Ni, Z., Chang, E., Shi, Y. & Chandra, V. (2023). FoleyGen : Visually-Guided Audio Generation. *arXiv preprint arXiv :2309.10537*. doi : 10.48550/arXiv.2309.10537.

Melechovsky, J., Guo, Z., Ghosal, D., Majumder, N., Herremans, D. & Poria, S. (2023). Mustango : Toward Controllable Text-to-Music Generation. *arXiv preprint arXiv :2311.08355*. Repéré à https://arxiv.org/abs/2311.08355.

Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y. & Ermon, S. (2021). Sdedit : Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv :2108.01073*.

Meng, C., Rombach, R., Gao, R., Kingma, D., Ermon, S., Ho, J. & Salimans, T. (2023). On distillation of guided diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14297–14306.

Miao, Z., Yang, Z., Lin, K., Wang, Z., Liu, Z., Wang, L. & Qiu, Q. (2024). Tuning timestep-distilled diffusion model using pairwise sample optimization. *arXiv preprint arXiv :2410.03190*.

Mittal, G., Engel, J., Hawthorne, C. & Simon, I. (2021a). Symbolic Music Generation with Diffusion Models.

Mittal, G., Engel, J. H., Hawthorne, C. & Simon, I. (2021b). Symbolic Music Generation with Diffusion Models. *Proceedings of the 22nd International Society for Music Information Retrieval Conference (ISMIR 2021)*, pp. 468–475. Repéré à https://archives.ismir.net/ismir2021/paper/000058.pdf.

Mo, S. & Song, Y. (2024). Foley-Flow : Coordinated Video-to-Audio Generation with Masked Audio-Visual Alignment and Dynamic Conditional Flows.

Mo, S. & Song, Y. (2025, June). Foley-Flow : Coordinated Video-to-Audio Generation with Masked Audio-Visual Alignment and Dynamic Conditional Flows. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28912–28921.

Novack, Z., McAuley, J., Berg-Kirkpatrick, T. & Bryan, N. J. (2024). DITTO : Diffusion Inference-Time T-Optimization for Music Generation. Repéré à https://arxiv.org/abs/2401.12179.

Ojha, U., Li, Y., Lu, J., Efros, A. A., Lee, Y. J., Shechtman, E. & Zhang, R. (2021). Few-shot image generation via cross-domain correspondence. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10743–10752.

Park, S., Hong, S., Kim, H. & Yang, E. (2025). AudioMorphix : Training-free Audio Editing with Diffusion Probabilistic Models.

Piczak, K. J. (2015). ESC : Dataset for Environmental Sound Classification. *Proceedings of the 23rd Annual ACM Conference on Multimedia*. doi : 10.1145/2733373.2806390.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J. & Rombach, R. (2023). Sdxl : Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv :2307.01952*.

Poole, B., Jain, A., Barron, J. T. & Mildenhall, B. (2022). Dreamfusion : Text-to-3d using 2d diffusion. *arXiv preprint arXiv :2209.14988*.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021). Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv :2103.00020*. Repéré à https://arxiv.org/abs/2103.00020. Version v1.

Ram, S., Neiman, T., Feng, Q., Stuart, A. M., Tran, S. & Chilimbi, T. A. (2025, February). DreamBlend : Advancing Personalized Fine-tuning of Text-to-Image Diffusion Models. *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 3614–3623. Repéré à https://www.amazon.science/publications/dreamblend-advancing-personalized-fine-tuning-of-text-to-image-diffusion-models.

Research, G. [Archived : May 9, 2024 ; CC-BY-4.0 license]. (2023). dreambooth : Fine-Tuning Text-to-Image Diffusion Models for Subject-Driven Generation (GitHub Repository). Repéré à https://github.com/google/dreambooth.

Ristori, E., Bindini, L. & Frasconi, P. (2025). MARS : Audio Generation via Multi-Channel Autoregression on Spectrograms. *arXiv preprint arXiv :2509.26007*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695.

Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M. & Aberman, K. (2023). DreamBooth : Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. *arXiv preprint arXiv :2208.12242*.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., Salimans, T., Ho, J., Fleet, D. J. & Norouzi, M. [arXiv :2205.11487]. (2022). Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. Repéré à https://arxiv.org/abs/2205.11487.

Salimans, T. & Ho, J. (2022). Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv :2202.00512*.

Sauer, A., Boesel, F., Dockhorn, T., Blattmann, A., Esser, P. & Rombach, R. (2024a). Fast high-resolution image synthesis with latent adversarial diffusion distillation. *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–11.

Sauer, A., Lorenz, D., Blattmann, A. & Rombach, R. (2024b). Adversarial diffusion distillation. *European Conference on Computer Vision*, pp. 87–103.

Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M. et al. (2022). Laion-5b : An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35, 25278–25294.

Sheffer, R. & Adi, Y. [Accepted at ICASSP 2023]. (2022). I Hear Your True Colors : Image Guided Audio Generation. Repéré à https://arxiv.org/abs/2211.03089.

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerrv-Ryan, R. et al. (2018). Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4779–4783.

Shi, B., Tjandra, A., Hoffman, J., Wang, H., Wu, Y.-C., Gao, L., Richter, J., Le, M., Vyas, A., Chen, S., Feichtenhofer, C., Dollár, P., Hsu, W.-N. & Lee, A. (2025). SAM Audio : Segment Anything in Audio. Repéré à https://arxiv.org/abs/2512.18099.

Smith, J. O. (2011). *Spectral Audio Signal Processing*. W3K Publishing. Repéré à https://ccrma.stanford.edu/~jos/sasp/sasp.html.

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N. & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International conference on machine learning*, pp. 2256–2265.

Song, J., Meng, C. & Ermon, S. (2020a). Denoising diffusion implicit models. *arXiv preprint arXiv :2010.02502*.

Song, J., Meng, C. & Ermon, S. (2020b). Denoising Diffusion Implicit Models. Repéré à https://arxiv.org/abs/2010.02502.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S. & Poole, B. (2020c). Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv :2011.13456*.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S. & Poole, B. (2020d). Score-Based Generative Modeling through Stochastic Differential Equations. *arXiv preprint arXiv :2011.13456*. Repéré à https://arxiv.org/abs/2011.13456.

Song, Y., Dhariwal, P., Chen, M. & Sutskever, I. (2023). Consistency Models. *International Conference on Machine Learning*, pp. 32211–32252.

Tang, W. & Zhao, H. (2024). Score-based Diffusion Models via Stochastic Differential Equations – a Technical Tutorial. *arXiv preprint arXiv :2402.07487*.

Tian, Y., Chen, Q., Wang, W., Zhao, Z., Xue, W. et al. (2024). Parallel Multimodal Large Diffusion Language Models for Thinking-Aware Editing and Generation.

Tian, Z., Jin, Y., Liu, Z., Yuan, R., Tan, X., Chen, Q., Xue, W. & Guo, Y. (2025). AudioX : Diffusion Transformer for Anything-to-Audio Generation. Repéré à https://arxiv.org/abs/2503.10522.

Turland, S. et al. (2025). Similarity-Guided Diffusion for Long-Gap Music Inpainting. *arXiv preprint arXiv :2509.16342*. Repéré à https://arxiv.org/abs/2509.16342.

van den Oord, A., Dieleman, S., Zen, H. et al. (2016). WaveNet : A Generative Model for Raw Audio. *Advances in Neural Information Processing Systems (NeurIPS)*.

van den Oord, A., Vinyals, O. & Kavukcuoglu, K. (2017). Neural Discrete Representation Learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Vyas, A., Shi, B., Le, M., Tjandra, A., Wu, Y.-C., Guo, B., Zhang, J., Zhang, X., Adkins, R., Ngan, W., Wang, J., Cruz, I., Akula, B., Akinyemi, A., Ellis, B., Moritz, R., Yungster, Y., Rakotoarison, A., Tan, L., Summers, C., Wood, C., Lane, J., Williamson, M. & Hsu, W.-N. (2023). Audiobox : Unified Audio Generation with Natural Language Prompts. Repéré à https://arxiv.org/abs/2312.15821.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., He, L., Zhao, S. & Wei, F. (2023a). Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers. Repéré à https://arxiv.org/abs/2301.02111.

Wang, H., Hai, J., Lu, Y.-J., Thakkar, K., Elhilali, M. & Dehak, N. (2024a). SoloAudio : Target Sound Extraction with Language-oriented Audio Diffusion Transformer.

Wang, H., Hai, J., Lu, Y.-J., Thakkar, K., Elhilali, M. & Dehak, N. (2024b). SoloAudio : Target Sound Extraction with Language-oriented Audio Diffusion Transformer. *arXiv preprint arXiv :2409.08425*. doi : 10.48550/arXiv.2409.08425.

Wang, J., Xu, C., Yu, C., Shang, L., Hu, Z., Wang, S. & Bo, L. (2025a). Synchronized Video-to-Audio Generation via Mel Quantization-Continuum Decomposition. Accepted to CVPR 2025.

Wang, L., Wang, J., Deng, F. et al. (2025b). AudioGen-Omni : A Unified Multimodal Diffusion Transformer for Video-Synchronized Audio, Speech, and Song Generation. *arXiv preprint arXiv :2508.00733*.

Wang, X., Cheng, X., Wang, Y., Song, R. & Wang, Y. [CVF Open Access (verify exact venue/year if you need the official proceedings entry)]. (2025c). VAFlow : Video-to-Audio Generation with Cross-Modality Flow Matching.

Wang, X., Cheng, X., Wang, Y., Song, R. & Wang, Y. (2025d, October). VAFlow : Video-to-Audio Generation with Cross-Modality Flow Matching. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11777–11786. Repéré à https://openaccess.thecvf.com/content/ICCV2025/html/Wang_VAFlow_Video-to-Audio_Generation_with_Cross-Modality_Flow_Matching_ICCV_2025_paper.html.

Wang, X., Lin, B., Liu, D., Chen, Y.-C. & Xu, C. (2024c). Bridging data gaps in diffusion models with adversarial noise-based transfer learning. *Forty-first International Conference on Machine Learning*.

Wang, Y., Skerry-Ryan, R., Stanton, D. et al. (2017). Tacotron : Towards End-to-End Speech Synthesis. *Interspeech*.

Wang, Z., Lu, C., Wang, Y., Bao, F., Li, C., Su, H. & Zhu, J. (2023b). Prolificdreamer : High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36, 8406–8441.

Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L. & Zuo, W. (2023). ELITE : Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation. *arXiv preprint arXiv :2302.13848*.

Wu, J. Z., Ge, Y., Wang, X., Lei, S. W., Gu, Y., Shi, Y., Hsu, W., Shan, Y., Qie, X. & Shou, M. Z. (2023a). Tune-a-video : One-shot tuning of image diffusion models for text-to-video generation. *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7623–7633.

Wu, S.-L., Donahue, C., Watanabe, S. & Bryan, N. J. [Submitted to IEEE/ACM TASLP]. (2023b). Music ControlNet : Multiple Time-varying Controls for Music Generation. Repéré à https://arxiv.org/abs/2311.07069.

Wu, Y., Chen, Z., Wang, C., Zhang, Y. & Wang, W. [ArXiv id not pinned here ; add if you want fully-resolved bib entry]. (2024). Multi-Foley : Video-Guided Foley Sound Generation with Multimodal Controls.

Xie, Z., Xu, X., Wu, Z. & Wu, M. (2024a). PicoAudio : Enabling Precise Timestamp and Frequency Controllability of Audio Events in Text-to-audio Generation.

Xie, Z., Xu, X., Wu, Z. & Wu, M. (2024b). PicoAudio : Enabling Precise Timestamp and Frequency Controllability of Audio Events in Text-to-audio Generation. *arXiv preprint arXiv :2407.02869*. doi : 10.48550/arXiv.2407.02869.

Xing, Y., He, Y., Tian, Z., Wang, X. & Chen, Q. (2024). Seeing and Hearing : Open-domain Visual-Audio Generation with Diffusion Latent Aligners. *arXiv preprint arXiv :2402.17723*. doi : 10.48550/arXiv.2402.17723.

Xu, M., Li, C., Tu, X., Ren, Y., Chen, R., Gu, Y., Liang, W. & Yu, D. (2024a). Video-to-Audio Generation with Hidden Alignment. *arXiv preprint arXiv :2407.07464*. doi : 10.48550/arXiv.2407.07464.

Xu, M., Li, C., Zhang, D., Su, D., Liang, W. & Yu, D. (2024b). Prompt-guided Precise Audio Editing with Diffusion Models. *Proceedings of the International Conference on Machine Learning (ICML)*.

Xu, X., Mei, J., Zheng, Z., Tao, Y., Xie, Z., Zhang, Y., Liu, H., Wu, Y., Yan, M., Wu, W., Zhang, C. & Wu, M. (2025). UniFlow-Audio : Unified Flow Matching for Audio Generation from Omni-Modalities. Repéré à https://arxiv.org/abs/2509.24391.

Yang, C., Shen, Y., Zhang, Z., Xu, Y., Zhu, J., Wu, Z. & Zhou, B. (2023a). One-shot generative domain adaptation. *Proceedings of the ieee/cvf international conference on computer vision*, pp. 7733–7742.

Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Chang, X., Shi, J., Zhao, S., Bian, J., Zhao, Z., Wu, X. & Meng, H. (2023b). UniAudio : An Audio Foundation Model Toward Universal Audio Generation.

Yang, D., Tian, J., Tan, X., Huang, R., Liu, S., Chang, X., Shi, J., Zhao, S., Bian, J., Zhao, Z., Wu, X. & Meng, H. [arXiv preprint (submitted 2023, revised 2024)]. (2024a). UniAudio : An Audio Foundation Model Toward Universal Audio Generation. Repéré à https://arxiv.org/abs/2310.00704.

Yang, Q., Mao, B., Wang, Z., Nie, X., Gao, P., Guo, Y., Zhen, C., Yan, P. & Xiang, S. (2024b). Draw an Audio : Leveraging Multi-Instruction for Video-to-Audio Synthesis. *arXiv preprint arXiv :2409.06135*. doi : 10.48550/arXiv.2409.06135.

Yao, Y., Li, P., Chen, B. & Wang, A. (2025). JEN-1 Composer : A Unified Framework for High-Fidelity Multi-Track Music Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*. doi : 10.1609/aaai.v39i13.33584.

Yao, Y., Huang, S., Wang, W., Dong, L. & Wei, F. (2021). Adapt-and-Distill : Developing Small, Fast and Effective Pretrained Language Models for Domains. *Findings of the Association for Computational Linguistics : ACL-IJCNLP 2021*, pp. 460–470.

Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F. & Freeman, B. (2024a). Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37, 47455–47487.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T. & Park, T. (2024b). One-step Diffusion with Distribution Matching Distillation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. Repéré à https://openaccess.thecvf.com/content/CVPR2024/papers/Yin_One-step_Diffusion_with_Distribution_Matching_Distillation_CVPR_2024_paper.pdf.

Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W. T. & Park, T. (2024c). One-step diffusion with distribution matching distillation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623.

Yin, T., Zhang, Q., Zhang, R., Freeman, W. T., Durand, F., Shechtman, E. & Huang, X. (2025). From slow bidirectional to fast autoregressive video diffusion models. *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22963–22974.

Yuan, R. et al. (2024a). ChatMusician : Understanding and Generating Music Intrinsically with LLM. Repéré à https://arxiv.org/abs/2402.16153.

Yuan, Y., Jia, D., Zhuang, X., Chen, Y., Liu, Z., Chen, Z., Wang, Y., Wang, Y., Liu, X., Plumbley, M. D. et al. (2024b). Improving Audio Generation with Visual Enhanced Caption. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Yuan, Y., Liu, H., Liu, X., Huang, Q., Plumbley, M. D. & Wang, W. (2024c). Retrieval-Augmented Text-to-Audio Generation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 581–585.

Yuan, Y., Liu, X., Liu, H., Plumbley, M. D. & Wang, W. (2024d). FlowSep : Language-Queried Sound Separation with Rectified Flow Matching. Repéré à https://arxiv.org/abs/2409.07614.

Zeghidour, N., Teboul, O., de Chaumont Quitry, F. & Tagliasacchi, M. (2021). SoundStream : An End-to-End Neural Audio Codec. *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhang, K., Pham, T. X., Lee, S., Niu, A., Senocak, A. & Chung, J. S. (2025a). Model-Guided Dual-Role Alignment for High-Fidelity Open-Domain Video-to-Audio Generation. *Advances in Neural Information Processing Systems (NeurIPS)*. Repéré à https://neurips.cc/virtual/2025/poster/120356.

Zhang, W., Jiang, Z., Chen, Z., Xiao, N. & Ou, Y. (2021). NUMA-Aware DGEMM Based on 64-Bit ARMv8 Multicore Processors Architecture. *Electronics*, 10(16), 1984. doi : 10.3390/electronics10161984.

Zhang, Y., Maezawa, A., Xia, G., Yamamoto, K. & Dixon, S. [arXiv preprint (submitted 2023, revised 2024)]. (2024). Loop Copilot : Conducting AI Ensembles for Music Generation and Iterative Editing. Repéré à https://arxiv.org/abs/2310.12404.

Zhang, Y., Sharon, R., Dyer, G. F., Allard, E., Yona, A., Chen, Z., Verbin, D. & Cohen-Or, D. (2023). Loop Copilot : Generating Loopable Music from Text and Audio Prompts.

Zhang, Y., Wu, Y., Wang, C., Chen, Z., Lin, S. & Wang, W. (2025b). SmartDJ : Declarative Audio Editing with Audio Language Model.

Zhao, L., Feng, L., Ge, D., Chen, R., Yi, F., Zhang, C., Zhang, X.-L. & Li, X. (2025). UniForm : A Unified Multi-Task Diffusion Transformer for Audio-Video Generation. Repéré à https://arxiv.org/abs/2502.03897.

Zhao, Y., Chandrasegaran, K., Abdollahzadeh, M. & Cheung, N.-M. M. (2022). Few-shot image generation via adaptation-aware kernel modulation. *Advances in Neural Information Processing Systems*, 35, 19427–19440.

Zhao, Y., Du, C., Abdollahzadeh, M., Pang, T., Lin, M., Yan, S. & Cheung, N.-M. (2023). Exploring incompatible knowledge transfer in few-shot image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7380–7391.

Zheng, H., Nie, W., Vahdat, A. & Anandkumar, A. (2023a). Fast Training of Diffusion Models with Masked Transformers. Repéré à https://arxiv.org/abs/2306.09305.

Zheng, H., Nie, W., Vahdat, A. & Anandkumar, A. (2024). Fast Training of Diffusion Models with Masked Transformers. *Transactions on Machine Learning Research*. Repéré à https://openreview.net/forum?id=vTBjBtGioE. Also available as arXiv :2306.09305.

Zheng, K., Lu, C., Chen, J. & Zhu, J. (2023b). Improved Techniques for Maximum Likelihood Estimation for Diffusion ODEs. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 202(Proceedings of Machine Learning Research), 42363–42389. Repéré à https://proceedings.mlr.press/v202/zheng23c.html.

Zheng, Y. & Yang, Y. [CVPR Workshop paper (verify exact workshop/venue fields if needed)]. (2024). Flow-Optimizer : A Transformer-based Conditional Flow Model for Image Interpolation.

Zhou, L., Lou, A., Khanna, S. & Ermon, S. (2023). Denoising Diffusion Bridge Models. Repéré à https://arxiv.org/abs/2309.16948.

Zhou, L., Lou, A., Khanna, S. & Ermon, S. (2024). Denoising Diffusion Bridge Models. *The Twelfth International Conference on Learning Representations (ICLR)*. Repéré à https://arxiv.org/abs/2309.16948.

Zhu, G., Darefsky, J., Jiang, F., Selitskiy, A. & Duan, Z. (2022a). Music Source Separation with Generative Flow. *IEEE Signal Processing Letters*, 29, 2288–2292. doi : 10.1109/LSP.2022.3219355.

Zhu, G., Wen, Y., Carbonneau, M.-A. & Duan, Z. (2023a). EDMSound : Spectrogram Based Diffusion Models for Efficient and High-Quality Audio Synthesis.

Zhu, G., Wen, Y., Carbonneau, M.-A. & Duan, Z. (2023b). EDMSound : Spectrogram Based Diffusion Models for Efficient and High-Quality Audio Synthesis. Repéré à https://arxiv.org/abs/2311.08667.

Zhu, G., Wen, Y. & Duan, Z. (2025). Audio Generation Through Score-Based Generative Modeling : Design Principles and Implementation. Repéré à https://arxiv.org/abs/2506.08457.

Zhu, J., Ma, H., Chen, J. & Yuan, J. (2022b). Few-shot image generation with diffusion models. *arXiv preprint arXiv :2211.03264*.

Ziv, A., Gat, I., Le Lan, G., Remez, T., Kreuk, F., Défossez, A., Copet, J., Synnaeve, G. & Adi, Y. [arXiv preprint]. (2024). Masked Audio Generation using a Single Non-Autoregressive Transformer. Repéré à https://arxiv.org/abs/2401.04577.