A PRODUCTION-ORIENTED FRAMEWORK FOR EVALUATION OF SFX GENERATION

Mélodie Desbos[*]*, Yara Bahram, Eric Granger, Mohammadhadi Shateri*

LIVIA, Dept. of Systems Engineering, ETS Montréal, Canada
melodie.desbos@livia.etsmtl.ca

ABSTRACT

Industrial sound design requires audio generation systems that not only produce realistic audio, but also preserve the perceptual identity of a reference, support controllable variation, and remain efficient for practical workflows. Existing evaluations are usually tied to text-to-audio (TTA), unconditional, or task-specific settings, limiting assessment for reference-guided sound effects (SFX) variation. To address this gap, we present a production-oriented evaluation framework for structured comparison of heterogeneous audio generation and editing methods. Our framework identifies nine production requirements and explicitly accounts for differences in model capabilities, enabling comparison under a common production objective. A two-stage protocol is introduced: (1) a reference-guided audio-to-audio (ATA) variation task, in which all methods are evaluated under the same ESC-50 SFX adaptation setup, and (2) capability-specific analyses of native operations such as SFX morphing, temporal and energy alignment, inpainting, and targeted editing. This framework combines objective metrics (including FAD, ImageBind-based reference alignment, and diversity across generated variants), together with a human study of perceptual identity preservation and transient diagnosis. Our study reveals complementary strengths and trade-offs across baselines for different production needs. Among the full-generation baselines evaluated under a shared ATA setting, AudioX provides the strongest overall trade-off between reference alignment and diversity while still supporting SFX morphing. Other baselines remain most suitable for specific editing operations. Our framework establishes a structured evaluation and decision protocol for reference-guided SFX variation and provides a practical basis for designing future unified industrial audio generation pipelines. Audio demos and further details can be found on the accompanying web page[1].

1. INTRODUCTION

1.1. Production Motivation

In SFX production, sound designers often rely on few reusable recordings, since manually designing variations is costly and time-consuming. This low variability can lead to perceptible repetition in interactive media, motivating systems that can produce useful variations from a reference audio clip [1, 2, 3, 4, 5] or from standard TTA generation [6, 7, 8], rather than synthesizing audio unconditionally [9]. In practice, an effective sound design system must do more than produce plausible audio as it is subject to multiple constraints. Unlike generic TTA generation, this setting requires jointly balancing realism, identity preservation, controllability, and usability.

1.2. Gaps in Current Evaluations

Recent audio generation progress has made SFX generation increasingly accessible. Diffusion, flow-matching, and multimodal pipelines now support style transfer [10, 11] (or SFX morphing [12]), temporal conditioning [1], targeted editing [5], and inpainting [13]. However, these methods are typically evaluated only within their original task settings, such as TTA generation [10, 6], video-to-audio (VTA) synthesis [5, 14], foley alignment [1, 15], or localized restoration [13]. Therefore, their reported performance offers limited insight into controlled and production-useful reference-guided variation. In addition, differences in task formulation, architecture, and evaluation criteria make methods difficult to compare and complicate the selection of the most suitable approach for specific SFX generation needs.

1.3. Proposal of Evaluation Framework

This work aims to address this gap by proposing a production-oriented evaluation framework for reference-guided SFX variation. The proposed framework defines practical production requirements, selects complementary state-of-the-art baselines accordingly, and evaluates them through a two-stage protocol: (1) All methods are compared under a shared reference-conditioned ATA variation task using ESC-50, an open-source few-shot soundbank, to reflect a realistic production usage [16]. (2) Since the evaluated baselines differ in conditioning, training, and editing capabilities, our goal is not to produce a single ranking but to characterize their suitability for creative SFX variation. The proposed framework further provides a capability-specific analyses to assess strengths and limitations of native editing operations such as reference-guided SFX morphing, temporal alignment, inpainting, and targeted editing. This design enables a structured and fair comparison of heterogeneous methods under a shared production objective, while creating a decision protocol to emphasize distinctive strengths of each approach.

1.4. Contributions

(i) We introduce a production-oriented evaluation framework for reference-guided SFX variation grounded in practical requirements. **(ii)** A two-stage protocol is proposed that combines a shared reference-conditioned ATA variation task with capability-specific analyses. This enables a decision-oriented analysis and more structured comparison across heterogeneous baselines. **(iii)** Comparison of recent audio generation and editing methods through objective, perceptual, qualitative, and capability-specific evaluations. It reveals clear trade-offs and complementary strengths of different methods for reference-guided SFX variation.

[1] https://melodiedesbos.github.io/sfx-eval-framework

2. RELATED WORK

2.1. Audio Generation for SFX

In audio generation, diffusion-based models remain a dominant paradigm for high-fidelity synthesis and are increasingly used for unconditional or text-conditioned synthesis, while operating on waveform, spectrogram, or latent representations [10, 1, 17]. AudioLDM adopts a latent-diffusion formulation in a VAE-based latent space and relies on CLAP embeddings for audio-text conditioning [10]. T-Foley [1] is guided by sound-class and temporal-event information for Foley generation. Recently, flow-matching models have also been explored for controllable audio generation, conditioning the flow on multimodal semantic embeddings [5]. Another common denoising architecture is the Diffusion Transformer (DiT), in which self-attention operates on multimodal features together with audio tokens or latent audio representations, while the Transformer serves as the denoising network [11, 18, 15, 19, 20, 21]. Earlier SFX-oriented works also studied controllable SFX generation, through synthesis of environmental sounds from onomatopoeic words [22], or from acoustic features with explicit control over pitch, loudness, timbre, and transients [23]. These works motivate SFX generation as controllable synthesis, but lack inputs over recent generative strengths.

2.2. Reference-Guided Audio Variation and Editing

Beyond unconditional or TTA synthesis, audio editing aims to modify existing samples while preserving timbre, temporal structure, or event identity. Recent generative audio frameworks support editing driven either by explicit instructions [5, 24], reference signals [13, 1], or both [25, 26, 2, 3, 27, 28, 15]. In practice, these pipelines cover complementary objectives, including full reference-guided variation [2, 28], temporal control [1], localized modification such as inpainting [13] or region-specific editing [5]. These developments are particularly relevant to SFX generation, where the objective is often not only to synthesize plausible audio, but also to preserve event identity and temporal structure while supporting controllable variation [1, 15]. In addition, a few methods are explicitly designed to adapt to new domains under restricted-data settings, including retrieval-augmented approaches for zero-shot and few-shot TTA generation [29] and customized generation from a few reference audio samples [30]. Yet these methods remain primarily generation-oriented and are not necessarily suited to editing or reference-guided SFX workflows.

2.3. Evaluation Gaps for Production-oriented SFX Variation

SFX generation focuses on producing short, semantically meaningful, and often temporally structured events tied to an action, scene, or reference. [1, 15]. Compared with general TTA generation, it places stronger emphasis on event identity, temporal synchronization, and controllability. Despite recent progress, such methods are rarely evaluated under a shared SFX-variation objective or a common benchmark, and their evaluation protocols remain tied to each method's original task: AudioLDM mainly on TTA generation [31], with style transfer, high resolution, and inpainting shown separately; T-Foley on temporal-event control, combining quality, diversity, and temporal-adherence metrics with MOS ratings; ThinkSound on VTA generation, object-focused generation, and editing, mainly on VGGSound [32] and MovieGen Audio Bench [33]; AudioX on a large protocol spanning TTA [34, 35], VTA, joint text-and-video, and instruction-driven benchmarks; and A^2SB on localized restoration. Thus, evaluations reflect different task definitions, datasets, and control assumptions, making a direct cross-method comparison inherently difficult. This limitation is important for production-oriented SFX variation, where the objective is not only to generate plausible audio but also to preserve event identity, support controllable variation, and remain feasible under practical constraints such as limited data and efficient inference. Such limitations motivate a shared evaluation framework suitable for practical SFX deployment (this being the focus of our work).

3. PRODUCTION-ORIENTED EVALUATION FRAMEWORK AND EXPERIMENTAL SET-UP

In production settings, a reference sound often requires diverse variations under constraints such as identity preservation, controllability, and temporal alignment. Existing baselines address complementary capabilities, but are rarely compared under a shared objective. Figure 1 summarizes the proposed evaluation framework to standardize comparison through a common ESC-50 [16] reference-guided ATA setting complemented by method-specific analyses. Each method is then evaluated using objective, perceptual, and qualitative criteria to characterize its relevance to different SFX production needs.

3.1. Production Requirements

Production-ready SFX variation requires more than fidelity: a useful method should preserve identity, support controlled variation, while remaining efficient for iterative deployment. In this work, the evaluation is structured around nine production requirements, each associated with an operational definition, specific evaluation signals, and typical failure modes; see Table 1. These requirements guide the baseline selection and the evaluation protocol.

3.2. Heterogeneous Baselines

We evaluate five representative baselines covering complementary SFX variation capabilities. AudioLDM [10] and AudioX [11] for SFX morphing; T-Foley [1] for time-conditioned control; ThinkSound [5] for targeted editing; and A^2SB [13] for inpainted variations. A detailed mapping between the production requirements and each baseline can be found on our web page[1]. Below, we provide a brief description mapping of their conditioning signals, editing scope and production strengths.

We first select **AudioLDM** [10], a method which transfers the reference style to generate reference-conditioned variations. Its primary control mechanism is text conditioning, while its editing scope covers the entire generated sample. Thus, it serves as a general-purpose baseline for full-sample variation, quality and realism (R1), diversity (R3), and flexible user control through text or reference-guided editing (R6). Then **T-Foley** [1], as it uses a protocol centered on temporal-event control through sound-class and RMS envelope information for Foley generation. This makes T-Foley particularly relevant for temporal alignment (R4) and energy control (R5), while supporting identity preservation (R2). However, it does not support targeted local edits of a specific waveform region (R7), since the reference is used only as an event condition rather than as a direct ATA edition. **ThinkSound** [5] is chosen as the third baseline, to evaluate its control mechanism

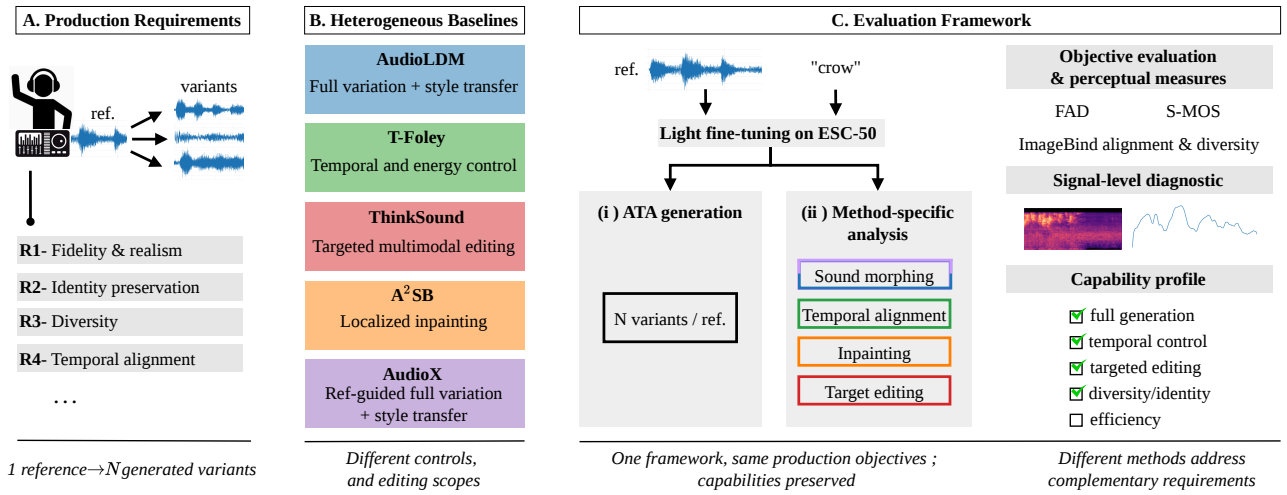


Figure 1: Overview of the proposed production-oriented evaluation framework for reference-guided SFX variation. A. In production settings, a reference sound often requires multiple diverse variations (Sec. 3.1); B. Existing baselines address complementary capabilities [10, 1, 5, 13, 11], including full variation generation (Sec. 3.2); C. We propose an evaluation framework to assess models' abilities at generating variations of SFX in production settings (Sec. 3) and report a full capability profile for each baseline on our web page [1].

Table 1: Production requirements for reference-guided SFX variation, with evaluation signals and typical failure modes.

Requirement	Assessment	Typical failure modes
R1 Fidelity and realism	FAD↓, MOS↑, spectrogram inspection, transient diagnosis	Transient smearing, noise, phasiness, synthetic texture, unnatural reverberation
R2 Identity preservation	S-MOS↑, ImageBind alignment↑	Semantic drift, generic textures, unrelated events
R3 Diversity without drift	ImageBind diversity↑, pairwise variation analysis	Near-duplicate outputs, limited variation, identity drift
R4 Temporal alignment	Energy-curve comparison, onset timing error, FWHM ratio	Shifted onsets, stretched/compressed events, incorrect ordering
R5 Energy control	Energy-curve & envelope comparison, pre-onset Δ	Loudness drift, distorted dynamics, poor envelope matching
R6 Controllability	Comparison across control settings	Weak control response, uncontrolled drift, over-transformation
R7 Targeted modification	Local qualitative and editing analysis	Global rewriting, boundary discontinuities, unintended off-target changes
R8 Robustness and stability	Cross-condition comparison, failure analysis	Unstable behavior, out-of-domain failure, inconsistent quality
R9 Efficiency	Inference time and compute-cost comparison	Excessive latency, slow sampling, impractical compute cost

that combines semantic instructions with reference conditioning, enabling targeted modifications while preserving overall plausibility. ThinkSound supports a localized editing scope, making it suitable for studying controllable semantic variation and localized content modification. This allows evaluation of targeted modification (R7), controllability (R6), and semantic editing while maintaining plausible variation (R3). Further, **AudioX** [11] is selected as it combines text, video, and audio signals through multimodal fusion, enabling audio generation under multiple conditioning signals. AudioX spans TTA, VTA, joint text-and-video conditioning, and instruction-driven conditioning. This is particularly relevant as a recent generation model with strong identity preservation (R2), diversity without drift (R3), and controllable reference-guided generation (R6), while also providing a useful reference point for fidelity and realism (R1). Lastly, **A^{2}SB** [13] masks a segment and reconstructs it from the surrounding unmasked context. Unlike the other baselines, control is provided directly through the preserved waveform context rather than text or temporal conditioning. Its editing scope is strictly local, affecting only the masked region while leaving the remainder of the signal unchanged. This makes A^{2}SB relevant for evaluating targeted modification (R7), the preservation of an unedited context as part of identity preservation (R2), and localized variation with limited drift (R3).

Overall, the selected baselines cover four distinct editing scopes: **(i)** full-sample generation (AudioLDM, AudioX), **(ii)** temporally constrained generation (T-Foley), **(iii)** semantic reference-guided editing (ThinkSound), and **(iv)** localized waveform inpainting (A^{2}SB). This diversity enables a broader assessment of production-oriented SFX variation beyond fidelity alone.

3.3. Evaluation Framework

We evaluate production-oriented SFX variation through a two-stage protocol after lightweight adaptation on the few-shot soundbank ESC-50 [16]: (i) **ATA generation:** Given a reference audio clip and its corresponding class label, each model generates variants, which are then evaluated comprehensively. (ii) **Method-specific analysis:** Selected methods are further analyzed in their primary audio editing setting to highlight their strengths, including SFX morphing, inpainting, temporal alignment, and targeted editing. Experiments rely on **ESC-50 dataset**, with 2,000 five-second recordings across 50 classes. We use the official split, with folds 1-3 for training, fold 4 for validation, and fold 5 for testing,

yielding 8 clips per class per fold. This low-data setting approximates practical adaptation of pretrained models to limited SFX soundbanks. Although relatively small in scale, the ESC-50 setting reflects practical deployment scenarios in which pretrained models are fine-tuned on limited subsets of target classes rather than trained from scratch on large-scale datasets.

Our **fine-tuning protocol** is performed on top of each model's initial pretraining, with particular attention to preserving each model's original setup and framework. All baselines are initialized from their released pretrained checkpoints. Fine-tuning is deliberately lightweight to adapt each method to ESC-50 while preserving its native behavior. Encoders are kept frozen when applicable, and only task-relevant generation or conditioning layers are adapted; method-specific details are reported in our web page[1].

For the **ATA generation protocol**, we generate N = 10 variants, for each of the 400 ESC-50 test references, yielding 4,000 outputs per model. Each method receives the waveform reference and class name as minimal conditioning. ATA generation is implemented according to each baseline's native design: T-Foley uses reference-derived temporal features and RMS-envelope conditioning; ThinkSound and AudioX sample from a noisy reference latent or waveform, respectively, with class-name conditioning; and A^2SB is evaluated in its native inpainting regime by masking a segment of the reference and sampling multiple restorations from the same corrupted input. Table 2 reports inference time and number of generated variants per hour for each method, serving as a practical efficiency indicator (R9). Results in Tab. 2 are diagnostic rather than directly comparable, since methods use different backbones, batch sizes, and sampling steps.

Because the selected baselines differ in their conditioning signals and editing scope, we complement the shared ATA framework with a **method-specific analysis** on each method. These analyses are diagnostic and are not intended for direct cross-method comparison. For A^2SB, we evaluate localized inpainting by comparing short masks of 0.3-1.0,s with longer masks of 0.3-2.0,s, using metrics computed exclusively on the inpainted regions. For T-Foley, we assess the effect of its restricted pretraining manifold (only seven classes, e.g., *coughing*, *dog*, *footsteps*, *keyboard_typing*, and *rain*) by separating ESC-50 classes that overlap with its native training domain from the remaining unseen classes. For AudioLDM and AudioX, we evaluate SFX morphing using reference-to-target transformations under different noise levels σ, which control the strength of the transformation. For ThinkSound, we evaluate object-centric editing on masked regions using attenuation, enhancement, and reverberation instructions. These native analyses preserve each model's intended control mechanism while complementing the requirement-level interpretation of the shared framework.

Table 2: Practical inference cost under each method's setting for generating $N = 10$ variants per reference over 50 ESC-50 classes with 8 references per class (4 000 generations per model) [16].

Method	Backbone	Hardware	Steps	Batch	Inference	Variants/ h
ThinkSound (NeurIPS'25)	MMDiT	A100 40GB	24	2	0.66 h	∼6061
T-Foley (ICASSP'24)	Wave. Diff. + FiLM	RTX A6000	250	16	28.8 h	∼139
A^2SB (ArXiv'25)	Attn. UNet	A100 40GB	300	1	33 h	∼121
AudioX (ICLR'26)	DiT	A100 40GB	250	1	10 h	400
AudioLDM (ICML'23)	UNet (LDM)	RTX A6000	200	1	0.75 h	∼5333

3.4. Evaluation Metrics and Listening Study

Objective evaluation is performed on cropped 4 s clips for all methods. Distributional audio quality and realism (R1) are measured with Fréchet Audio Distance (FAD) using the AudioLDM-eval implementation [36], computed per class and pooled across classes.

For the **reference alignment and diversity**, we use ImageBind audio embeddings [37] for identity preservation (R2) and diversity (R3), since the shared audio-to-audio (ATA) protocol is reference-conditioned rather than text-driven: an audio-text metric such as CLAP would mainly reflect agreement with the class label rather than preservation of the specific reference event. Let $\hat{z}_r^{\text{ref}}$ and $\hat{z}_{r,v}^{\text{gen}}$ be the ℓ_2-normalised embeddings of reference r and its variant v. Alignment is the reference-variant cosine similarity $A_{r,v} = \text{cos\_sim}(\hat{z}_r^{ref}, \hat{z}_{r,v}^{gen}) = \frac{\langle \hat{z}_r^{ref}, \hat{z}_{r,v}^{gen} \rangle}{\|\hat{z}_r^{ref}\|_2 \, \|\hat{z}_{r,v}^{gen}\|_2}$ and diversity D_r is the mean pairwise cosine distance $D_r = \frac{2}{V_r(V_r-1)} \sum_{1 \le i < j \le V_r} \left(d_{\cos}\left(\hat{z}_{r,i}^{\text{gen}}, \hat{z}_{r,j}^{\text{gen}}\right)\right)$, where $d_{\cos}(\cdot,\cdot) = 1 - \text{cos\_sim}(\cdot,\cdot)$ and V_r=10 denotes the variants of a reference, following [38]. High diversity must be read jointly with FAD and alignment, as a large spread may also indicate semantic drift. Reguarding the **transient diagnostic**, we assess temporal behaviour (R4-R5) from a normalised onset-strength envelope $\tilde{e}_t$, built by summing positive log-mel increments over mel bins and rescaling to $[0, 1]$, so the comparison reflects transient shape rather than loudness. Peaks are detected with a 40 ms minimum spacing and prominence ρ=0.10, yielding three measures: (i) *FWHM ratio*, the generated-to-reference ratio of median peak full-width at half-maximum ($\to$ 1 ideal; $>$1 smeared, $<$1 sharpened); (ii) *pre-onset* Δ, the difference in normalised pre-transient energy (40 ms before vs. 80 ms after the peak; $\to$ 0 ideal, positive values indicate pre-echo); and (iii) *onset error*, the median timing gap between matched reference and generated onsets (tolerance $\delta_{\max}$=150 ms). All detection parameters are fixed across methods. Finally, the **listening study** evaluates perceptual identity (R1-R2), the Similarity Mean Opinion Score (S-MOS) is rated by 15 participants on a 1-5 identity scale over reference-variation pairs drawn from the 4,000 outputs per method. Participants used headphones in a quiet environment, could replay each pair freely, and rated 100 randomized trials per method with the prompt: *rate the identity fidelity (1–5) as the similarity to the reference event/source, excluding loudness.* Scores are reported with 95% CI.

4. EXPERIMENTAL RESULTS

This section presents results of the proposed production-oriented evaluation framework for reference-guided SFX variation. We report the shared ATA evaluation, Pareto-inspired trade-offs, and qualitative analyses, followed by method-specific results on native capabilities not fully captured by the shared framework.

4.1. ATA Generation

Table 3 reports the shared reference-conditioned ATA comparison along with transient-level diagnostics, further discussed in Sec. 4.3. We evaluate each method's ability to generate variants from a reference SFX across reference preservation, quality, diversity, and perceptual identity under common production constraints.

AudioX emerges as the strongest full-generation baseline, providing the most balanced trade-off across audio quality, identity preservation, and human evaluation. It remains consistently competitive across the main criteria, with a low FAD of 9.34, the

Table 3: Overall quantitative and transient-level diagnosis comparison of baselines for ATA on SFX variation. FAD, S-MOS, diversity, and ImageBind-audio alignment summarize global generation quality, perceptual quality, variation diversity, and reference consistency. Transient diagnostics measure onset-local behavior using log-mel onset-strength envelopes: FWHM Ratio measures onset broadening or sharpening, Pre-onset Δ measures extra energy before the transient, and Onset Error evaluates temporal alignment. S-MOS is reported as the mean with 95% confidence intervals over 15 raters. **Best** and second best on full generation methods.

	Overall Quantitative Results				Transient Level Diagnostics		
Methods	**FAD↓**	**S-MOS↑**	**Div.↑**	**Align.↑**	**FWHM Ratio → 1**	**Pre-onset Δ→ 0**	**Onset Err. ms↓**
AudioX (ICLR, 2026)	**9.34**	**3.37** [3.08, 3.66]	0.27	**0.59**	0.985	**-0.004**	**43.52**
AudioLDM (ICML, 2023)	20.09	2.22 [1.97, 2.48]	**0.43**	0.39	0.969	-0.005	60.72
ThinkSound (NeurIPS, 2025)	16.51	2.57 [2.25, 2.89]	0.32	0.51	0.978	-0.011	53.01
T-Foley (ICASSP, 2024)	24.53	1.89 [1.62, 2.16]	0.32	0.22	**1.011**	0.008	63.53
A^{2}SB* (ArXiv, NVIDIA, 2025)	4.43	4.81 [4.75, 4.87]	0.28	0.49	0.992	0.005	54.79

* All reported objective metrics are computed on cropped 4 s clips for all methods, except for A^{2}SB where only inpainted regions are evaluated.

strongest alignment score of 0.59, and the highest S-MOS among full-generation methods (3.37) [3.08, 3.66]. Its diversity score is more restrained than AudioLDM (0.27), suggesting that AudioX favors reference consistency over large variation, but without collapsing to over-fitting the reference.

In contrast, AudioLDM reaches the highest diversity among full-generation baselines (0.43), but with weaker alignment (0.39), lower S-MOS (2.22 [1.97, 2.48]), and higher FAD (20.09), suggesting identity drift. ThinkSound presents an intermediate profile, with diversity (0.32), alignment (0.51), and S-MOS (2.57 [2.25, 2.89]), indicating that controlled variation does not fully translate into stronger perceptual fidelity in this ATA setting.

Although T-Foley explicitly conditions on temporal structure and energy, it performs poorly in the shared ATA setting, with the highest full-generation FAD (24.53), the lowest S-MOS (1.89 [1.62, 2.16]), and weak alignment (0.22), despite diversity comparable to ThinkSound (0.32). This suggests that temporal-event conditioning alone is insufficient to preserve the broader perceptual identity of the reference. A^{2}SB achieves the strongest S-MOS and FAD values, but should be interpreted separately because its variations are restricted to short inpainted regions of 0.3-1.0 s while most of the reference remains preserved. On inpainted regions, it obtains FAD (4.43), S-MOS (4.81 [4.75, 4.87]), diversity (0.28), and alignment (0.49), confirming its strength for localized repair rather than full ATA generation. Among the full-generation baselines evaluated under the shared ATA setting, results indicate that AudioX provides the strongest production-oriented compromise when identity preservation and perceptual fidelity are prioritized, while AudioLDM and ThinkSound expose different diversity-identity trade-offs that are further analyzed in Sec. 4.4.

4.2. Pareto-inspired Analysis in Production Settings

Figure 2 summarizes the diversity-alignment trade-off AudioLDM provides the highest diversity (R3: 0.43), but weaker identity alignment (R2: 0.39), while AudioX reaches the strongest alignment (R2: 0.59) with lower diversity (R3: 0.27). ThinkSound lies between these two regimes, with R3: 0.32 and R2: 0.51, for a moderate balance between variation and reference consistency. T-Foley does not occupy a favorable Pareto position, since it has diversity comparable to ThinkSound (R3: 0.32), but much weaker alignment (R2: 0.22). The position of A^{2}SB should again be interpreted separately, since its generation is restricted to a short inpainted segment rather than full-clip generation. Overall, the analysis confirms that higher diversity alone is insufficient when accompanied by weaker identity preservation [2].

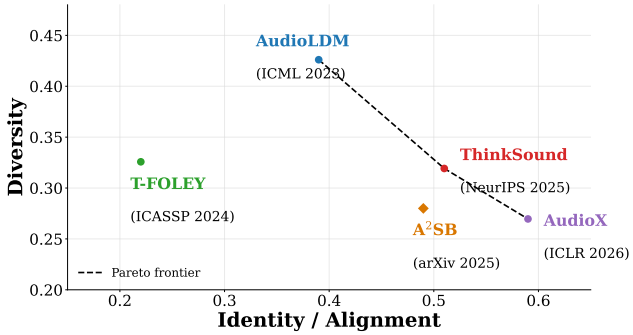


Figure 2: Diversity–identity alignment (R3-R2) trade-off across reference-guided SFX variation methods. Each point represents a model positioned according to its ability to preserve reference identity and generate diverse variations. Higher diversity and alignment are better.

4.3. Further Qualitative Analysis

Figure 3 provides a visual analysis of the spectro-temporal structure and frequency behavior of the baseline generations. While the quantitative results capture overall trends in quality, alignment, and diversity, the spectrogram and energy-curve visualizations reveal additional differences in local texture preservation, temporal organization, and failure modes across methods. In terms of spectral texture, AudioLDM shows the best-preserved fine local structure, as highlighted by the white boxes, whereas T-Foley generations exhibit noisier textures, attenuated high-frequency content, and a reduced spectral range. ThinkSound and AudioX remain closer to the reference at the event level, although ThinkSound shows more local spectral smoothing. The transient diagnostics support this observation: ThinkSound has an FWHM ratio close to 1 but a moderate onset error (53.01 ms), while AudioX achieves the lowest onset error among full-generation methods (43.52 ms), supporting stronger temporal organization.

These differences also appear in the temporal-energy curves. T-Foley sometimes follows the coarse reference energy profile, which is expected from its explicit energy conditioning, but this does not translate into better onset-level alignment, since it has the largest onset error among the full-generation baselines (63.53 ms). This suggests that envelope following does not necessarily pre-

[2] Further Pareto-inspired analyses are provided in our web page[1]

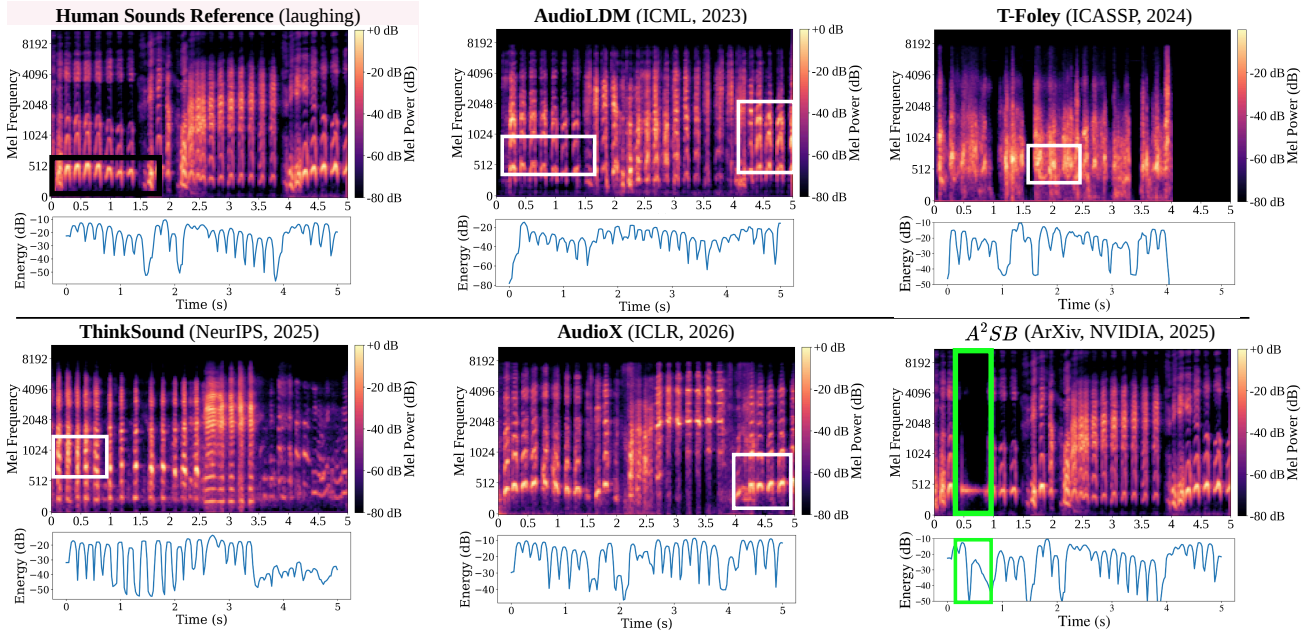


Figure 3: Qualitative comparison for the ATA variation task on the human-sound example *laughing*. For A²SB inpainting, the masked and regenerated sections between 0.3 s and 1 s are outlined. Similar texture to reference is outlined in white. Top: mel-spectrogram. Bottom: energy curve.

serve sound texture. A²SB preserves the global energy evolution well because most of the reference remains unchanged, while local transient diagnostics on the inpainted region remain close to the reference, with a FWHM ratio of 0.992 and a pre-onset difference of 0.005. However, its inpainted regions may still introduce localized energy changes. Overall, the qualitative analysis complements the quantitative comparison by revealing method-specific trade-offs in local texture preservation, temporal organization, and energy behavior that are not fully captured by scalar metrics alone.

4.4. Method-specific Analysis

Table 4 reports compact capability-specific diagnostics under each method's native setting defined in Sec 3.3. These results complement the shared ATA benchmark by showing where each method is most suitable within its intended editing scope. Below, the key observations from each method are examined in greater detail.

Table 4 shows A²SB remains strong for local repair on **inpainting task**, but degrades as mask duration increases. Specifically, FAD rises from 4.74 to 7.57 and alignment stabilizes from 0.35 to 0.31, while diversity decreases from 0.39 to 0.11. This suggests that longer masked regions are harder to reconstruct and may reduce both consistency and variation, likely because less local context remains available to guide the reconstruction. The transient diagnostic in our web page[1] further shows that both mask regimes remain temporally stable at the onset level, but a longer mask has a lower FWHM ratio (0.89), suggesting narrow or sharp reconstructed transients. On the **temporal-energy control and a restricted pretraining manifold** T-Foley generates noisy spectral texture, visible in Figure 3. This observation supports the quantitative results: although temporal-event conditioning can impose coarse structure, it does not guarantee identity preservation, and resulting variations show degraded fidelity and semantic consistency.

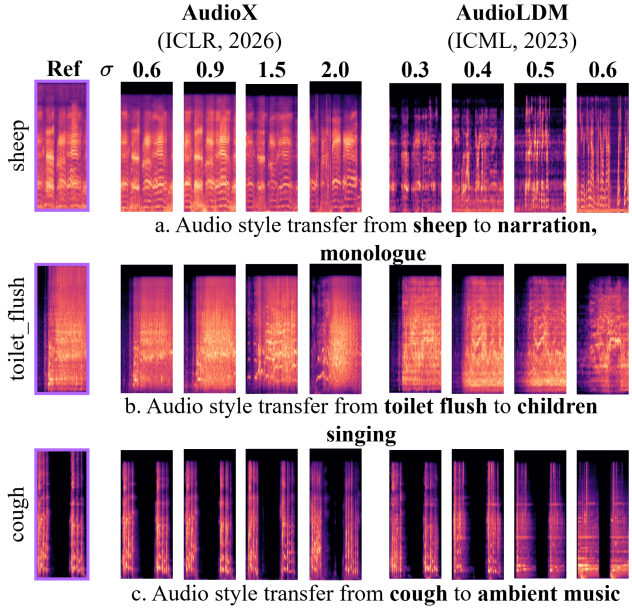
This behavior is explained with T-Foley's restricted pretraining manifold; Table 4 shows substantially better within-manifold performance, with FAD improving from 25.75 to 13.51, S-MOS from 1.76 ± 0.62 to 3.05 ± 0.78, and alignment from 0.21 to 0.34, while diversity remains stable ($0.32 \rightarrow 0.34$). Together, these results suggest that T-Foley's weaker ATA performance is driven more by fidelity, identity preservation, and domain-transfer limitations than by onset timing alone.

Table 4: Capability-specific diagnostics under each method's native setting. Results are diagnostic and not intended for direct cross-method comparison.

Methods	FAD↓	S-MOS↑	Div.↑	Align.↑
A²SB Inpainting (evaluated on 3 classes)				
Mask 0.3–1.0 s	4.74	–	0.39	0.35
Mask 0.3–2.0 s	7.57	–	0.11	0.31
SFX morphing (evaluated on 3 classes)				
AudioX ↓ σ / ↑ σ	4.57 / 6.64	–	0.10 / 0.26	0.77 / 0.64
AudioLDM ↓ σ / ↑ σ	18.97 / 24.09	–	0.34 / 0.56	0.40 / 0.15
T-Foley Restricted pretraining manifold				
Seen classes (5 classes)	13.51	3.05 ± 0.78	0.34	0.34
Unseen classes (45 classes)	25.75	1.76 ± 0.62	0.32	0.21
ThinkSound Object-centric region editing (evaluated on 5 classes)				
Attenuation Mask 1.0–4.0 s	9.06	–	0.19	0.67
Enhancement Mask 1.0–4.0 s	9.30	–	0.21	0.64
Reverberation Mask 1.0–4.0 s	8.97	–	0.22	0.64

On the **SFX morphing task**, AudioLDM and AudioX methods follow the AudioLDM audio style transfer setup [10] adapted to ESC-50. Three examples are evaluated: toilet flush to children singing, sheep to narration/monologue, and coughing to ambient music. For smaller σ values, the outputs remain closer to the ref-

erence, whereas larger σ values increase adherence to the target text condition. Figure 4 shows the reference audio together with generated samples obtained at different initialization noise levels σ, which control the transfer strength. At low σ, AudioX preserves the source timbre and spectral structure more closely, consistent with its higher alignment (0.77 vs. 0.40) and lower diversity (0.10 vs. 0.34). As σ increases, both methods trade alignment for diversity, with sharper results for AudioLDM (alignment 0.40 $\rightarrow$ 0.15; while diversity increases 0.34 $\rightarrow$ 0.56). AudioX changes more progressively (alignment 0.77 $\rightarrow$ 0.64; diversity 0.10 $\rightarrow$ 0.26). The transient diagnostic in our web page[1] follows the same tendency, with onset error increasing at higher σ for both methods. This yields a clear trade-off: AudioX provides smoother and more identity-preserving transformations, whereas AudioLDM produces stronger transformations at the cost of weaker preservation of the original audio identity.



a. Audio style transfer from **sheep** to **narration, monologue**

b. Audio style transfer from **toilet flush** to **children singing**

c. Audio style transfer from **cough** to **ambient music**

Figure 4: Ablations on **SFX Morphing task** for **AudioLDM** [10] and **AudioX** [11] on ESC-50 [16]. From left to right: the reference audio (e.g., sheep, toilet_flush, cough) conditioned on the target prompt with different noise levels σ (transfer strength).

On **object-centric editing**, ThinkSound strives to refine or modify sounds through user-specified regions and semantic instructions. Since the shared ATA protocol only provides restricted class-level conditioning, we further evaluate it in its native object-centric editing setting. Table 4 shows that the edited regions remain close to the reference, with alignment (0.64-0.67) and moderate diversity (0.19-0.22).

Figure 5 visually supports these results through light but directionally consistent fluctuations. Attenuation slightly reduces the target transient regions, enhancement increases transient strength, and reverberation introduces more diffuse repeated patterns. However, the editing scope remains limited in this setting. Since the prompts are deliberately simple and the evaluated references often contain a single dominant sound event, the model has limited semantic or acoustic structure to selectively modify. Further, the transient diagnostic in our web page[1] shows reasonable local timing preservation, with onset errors between 38.3 and 43.8 ms, although FWHM ratios above 1.1 suggest some transient broadening. As a result, ThinkSound produces plausible local changes, but the edits remain relatively subtle and do not always correspond to strong, fine-grained acoustic transformations.

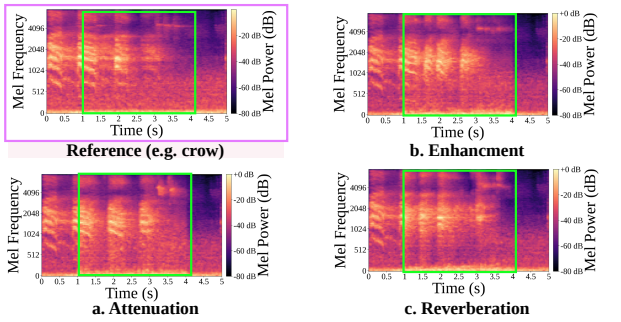


Figure 5: Ablation on **ThinkSound** with object-centric local editing over masked regions of the reference *crow*[16]) and 3 instructions: a. Attenuation, *"reduce the {class_name} volume."*; b. Enhancement, *"make {class_name} sound stronger and sharper."*; and c. Reverberation, *"make {class_name} event more distant and reverberant."*. Masked and edited sections [1 s; 3 s] are outlined.

5. DISCUSSION AND CONCLUSION

Production-ready SFX variation requires evaluating heterogeneous generation and editing methods under shared production requirements. By combining a common reference-guided ATA task with capability-specific analyses, our framework exposes key trade-offs without reducing performance to a single score. AudioX provides the strongest overall balance among full-generation methods, while AudioLDM favors stronger stylization and variation, T-Foley offers temporal and energy control, A^{2}SB supports localized inpainting, and ThinkSound enables semantically guided editing. Thus, no method dominates across all production needs.

These results show that generic audio-quality benchmarks alone are insufficient for evaluating SFX variation. Our requirement-driven protocol enables structured comparison while preserving each method's native strengths. Limitations include heterogeneous pretrained configurations and the restricted scope of ESC-50. Future work should unify realism, identity preservation, explicit control, localized editing, and efficient deployment within a single production-oriented generative pipeline.

6. ACKNOWLEDGMENTS

This work was supported by the Natural Sciences and Engineering Research Council of Canada, with additional computational resources provided by the Digital Research Alliance of Canada. We thank Dr. Hugo Seuté for his valuable support, the La Forge R&D department and the Alice sound team at Ubisoft for contributing to the problem formulation and insights in defining the production requirements.

7. REFERENCES

[1] Y. Chung, J. Lee, and J. Nam, "T-FOLEY: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis," in *ICASSP*, 2024.

[2] Z. Lan, Y. Hao, and M. Zhao, "SmartDJ: Declarative audio editing with audio language model," in *ICLR*, 2026.

[3] P. Fang, Y. He, Y. Xing, Q. Chen, S.-N. Lim, and H. Yang, "Ac-foley: Reference-audio-guided video-to-audio synthesis with acoustic transfer," in *ICLR*, 2026.

[4] J. Liang, Y. Chen, Y. Yuan, D. Jia, X. Zhuang, Z. Chen, Y. Wang, and Y. Wang, "Audiomorphix: Training-free audio editing with diffusion probabilistic models," *arXiv*, 2025.

[5] H. Liu, K. Luo, J. Wang, W. Wang, Q. Chen, Z. Zhao, and W. Xue, "Thinksound: Chain-of-thought reasoning in multimodal large language models for audio generation and editing," in *NeurIPS*, 2025.

[6] Z. Evans, J. D. Parker, C. Carr, Z. Zukowski, J. Taylor, and J. Pons, "Stable audio open," *ICASSP*, 2025.

[7] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *ACM TASLP*, 2024.

[8] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "AudioGen: Textually guided audio generation," in *ICLR*, 2023.

[9] G. Zhu, Y. Wen, M.-A. Carbonneau, and Z. Duan, "Edmsound: Spectrogram based diffusion models for efficient and high-quality audio synthesis," *NeurIPS workshop Machine Learning for audio*, 2023.

[10] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "AudioLDM: Text-to-audio generation with latent diffusion models," in *ICML*, 2023.

[11] Z. Tian, Z. Liu, Y. Jin, R. Yuan, L. Xue, X. Tan, Q. Chen, W. Xue, and Y. Guo, "AudioX: Diffusion transformer for anything-to-audio generation," in *ICLR*, 2026.

[12] X. Niu, J. Zhang, and C. P. Martin, "SoundMorpher: Perceptually-uniform sound morphing with diffusion model," 2024.

[13] Z. Kong, K. J. Shih, W. Nie, A. Vahdat, S. gil Lee, J. F. Santos, A. Jukic, R. Valle, and B. Catanzaro, "A2sb: Audio-to-audio schrodinger bridges," *arXiv*, 2025.

[14] S. Luo, C. Yan, C. Hu, and H. Zhao, "Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models," in *NeurIPS*, 2023.

[15] Z. Chen, P. Seetharaman, B. Russell, O. Nieto, D. Bourgin, A. Owens, and J. Salamon, "Video-guided foley sound generation with multimodal controls," in *CVPR*, 2025.

[16] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *ACM MM*, 2015.

[17] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, "Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization," in *ACM ICM*, 2024.

[18] H. F. García, O. Nieto, J. Salamon, B. Pardo, and P. Seetharaman, "Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations," in *ICASSP*, 2025.

[19] L. Zhao *et al.*, "Uniform: A unified multi-task diffusion transformer for audio-video generation," *arXiv*, 2025.

[20] B. Shi, A. Tjandra, J. Hoffman, H. Wang, Y.-C. Wu, L. Gao, J. Richter, M. Le, A. Vyas, S. Chen, C. Feichtenhofer, P. Dollár, W.-N. Hsu, and A. Lee, "Sam audio: Segment anything in audio," *arXiv*, 2025.

[21] H. K. Cheng, M. Ishii, A. Hayakawa, T. Shibuya, A. Schwing, and Y. Mitsufuji, "MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis," in *CVPR*, 2025.

[22] Y. Okamoto, K. Imoto, S. Takamichi, R. Yamanishi, T. Fukumori, and Y. Yamashita, "Onoma-to-wave: Environmental sound synthesis from onomatopoeic words," *APSIPA Transactions*, 2021.

[23] Y. Liu, C. Jin, and D. Gunawan, "Ddsp-sfx: Acoustically-guided sound effects generation with differentiable digital signal processing," *DAFx*, 2023.

[24] Y. Zhang, A. Maezawa, G. Xia, K. Yamamoto, and S. Dixon, "Loop copilot: Conducting ai ensembles for music generation and iterative editing," 2023.

[25] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, Z. Zhao, X. Wu, and H. Meng, "Uniaudio: An audio foundation model toward universal audio generation," 2024.

[26] L. Zhao, L. Feng, D. Ge, R. Chen, F. Yi, C. Zhang, X.-L. Zhang, and X. Li, "Prompt-guided precise audio editing with diffusion models," in *PMLR*. PMLR, 2024.

[27] Y. Wang, Z. Ju, X. Tan, L. He, Z. Wu, J. Bian, and S. Zhao, "AUDIT: Audio editing by following instructions with latent diffusion models," in *NeurIPS*, 2023.

[28] Y. Jia, Y. Chen, J. Zhao, S. Zhao, W. Zeng, Y. Chen, and Y. Qin, "Audioeditor: A training-free diffusion-based audio editing framework," *arXiv preprint arXiv:2409.12466*, 2024.

[29] M. Yang, B. Shi, M. Le, W.-N. Hsu, and A. Tjandra, "Audiobox tta-rag: Retrieval-augmented generation for zero-shot and few-shot text-to-audio," *Arxiv*, 2025.

[30] Y. Yuan, X. Liu, H. Liu, X. Kang, Z. Chen, Y. Wang, M. D. Plumbley, and W. Wang, "Dreamaudio: Few-shot customized text-to-audio generation via natural language supervision," *arXiv*, 2025.

[31] C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in *NAACL-HLT*, 2019, pp. 119–132.

[32] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "VGGSound: A large-scale audio-visual dataset," in *ICASSP*, 2020.

[33] T. M. G. team @Meta, "Movie gen: A cast of media foundation models," *CVPR*, 2024.

[34] Z. Xie, X. Xu, Z. Wu, and M. Wu, "Audiotime: A temporally-aligned audio-text benchmark dataset," *ICASSP*, 2025.

[35] H. Wang, C. Liu, J. Chen, H. Liu, Y. Jia, S. Zhao, J. Zhou, H. Sun, H. Bu, and Y. Qin, "Tta-bench: A comprehensive benchmark for evaluating text-to-audio models," *AAAI*, 2026.

[36] H. Liu, Y. Zhang, R. Mira, Y. Zang, and I. E. Ashimine. (2023) audioldm_eval: Audio generation evaluation. GitHub repository.

[37] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One embedding space to bind them all," in *CVPR*, 2023.

[38] Y. Xing, Y. He, Z. Tian, X. Wang, and Q. Chen, "Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners," in *CVPR*, 2024.