WHERE TO PERTURB: REFERENCE-STRUCTURED CORRUPTION FOR ONE-SHOT SFX VARIATION

Anonymous authors
Paper under double-blind review

ABSTRACT

One-shot SFX variation requires generating diverse yet recognizable alternatives from a single reference. Limited variability can lead to perceptual repetition in interactive media (e.g. video games), while manually designing variations is costly. Yet reference-guided flow models typically condition on the reference while applying a largely uniform corruption level across time, despite its heterogeneous acoustic structure. While recent asynchronous flow methods demonstrate that corruption can vary across a sequence, they do not specifically exploit the internal structure of an audio reference to allocate stochasticity for identity-preserving variation. We ask whether the internal structure of an audio latent can directly determine where and how strongly stochasticity should be introduced. We propose a Reference-Structured Corruption (RSC) that transforms a single reference into multiple stochastic trajectories without auxiliary conditioning or per-reference optimization. The method partitions the latent into content-aligned regions from latent geometry and activity, then assigns activity-conditioned regional flow times, yielding continuous control between preservation and regeneration. On Woosh, our method increases within-reference diversity from 1.29 to 1.41 WCIS while maintaining identity drift close to the level observed in real production SFX variations, and improves FAD from 4.82 to 4.41. On Stable Audio 3, it similarly increases WCIS from 1.69 to 1.79 while preserving generation quality (6.39 vs. 6.41 FAD). These results show that the learned latent representation can serve not only as the generation space, but also as a signal for structuring the stochastic process itself. An anonymized, frozen demo page is available online[1]Code is included in the supplementary material and will be released publicly.

1 INTRODUCTION

In audio modeling, generating variations of an audio sample requires preserving the perceived identity of the reference, which is conveyed by its spectral and temporal characteristics, including spectral envelope, brightness and noise structure, onset dynamics, and spectrotemporal modulations and texture. An identity-preserving variation space should maintain these identity-defining characteristics while allowing other dimensions, such as subtle timing, intensity, spectral detail, or local texture, to vary and produce perceptually diverse audio samples. This highlights the difficulty of framing an identity-preserving variation space, and raises the central question of this work: *How can the generative strengths and flexibility of a pre-trained flow-matching model be leveraged to adapt its generation process to the complex and heterogeneous structure of a single reference sound, while producing realistic and diverse variations without compromising its own identity?* A variation of a reference identity requires preserving reference-defined acoustic identity, broad temporal-event structure Chung et al. (2024); Xie et al. (2024), and timbral profile McAdams et al. (1995); García et al. (2025), while allowing changes in micro-timing, excitation intensity, spectral detail, and stochastic texture. Depending on the application, this identity can involve timbre, spectral envelope, temporal structure Hermes (2023), transient characteristics, energy evolution, or harmonic and noise componentsGordon & Grey (1978). The problem then requires a balance between reference fidelity, variation diversity, generation quality, and practical inference cost Desbos et al. (2026).

[1]https://anonymous.4open.science/r/variation-audio-demo/index.html

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

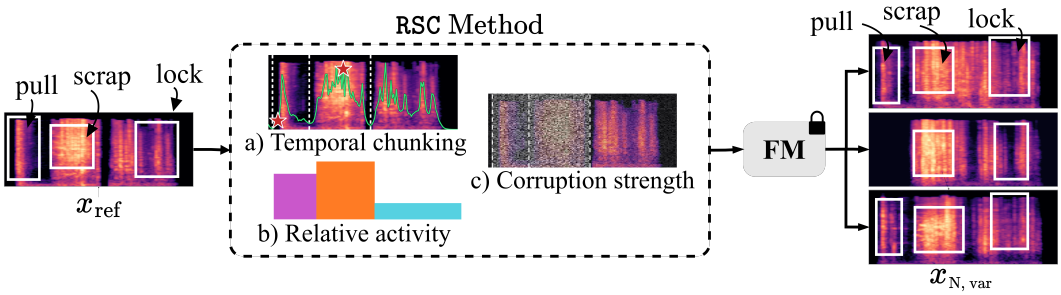


Figure 1: RSC method (Reference-Structured Corruption). A 3-stage training-free corruption scheme to generate variations given one reference (e.g., a two-stage metallic reloading sound).

Diffusion and Flow Models have demonstrated remarkable generalization capabilities in adapting and generating across a large range of audio domains Evans et al. (2026); Tian et al. (2026); Hadjeres et al. (2026) including settings with a limited number of target samples Yao et al. (2026); Fang et al. (2026); Yuan et al. (2026a;b). Prior analysis of self-supervised speech models shows that information is not uniformly distributed across temporal frames, as frame-level representations within a segment carry different amounts of identifying information for modeling generalization Pasad et al. (2024). This raises the question of whether a non-uniform representation of a sequence should necessarily receive uniform corruption. Existing works typically model complex audio structure by adapting the conditioning rather than the corruption process itself, through acoustic control signals García et al. (2025), an adversarial refinement stage Yao et al. (2026), and dedicated losses Luo et al. (2025); Hadjeres et al. (2026). Yet the structure already encoded within the reference latent provides a natural guide on *where* and *how strongly* stochastic corruption should be applied. Notably, recent vision methods have explored spatially or temporally varying corruption schedules informed by local visual or paired audio information Hu et al. (2026); Schusterbauer et al. (2026); Liu et al. (2026), motivating us to investigate whether the same principle can be extended to reference-structured corruption (RSC) in the audio latent space.

In this work, we propose a training-free method which leverages the internal structure of a reference audio latent to directly guide its own asynchronous flow corruption. Rather than relying on binary preservation and regeneration, as in inpainting-based methods Kong et al. (2025), we partition the latent into activity- and geometry-aligned temporal regions to assign each region an independent corruption strength. Inspired by diffusion forcing sampling Chen et al. (2024) and asynchronous flows Hu et al. (2026); Schusterbauer et al. (2026); Liu et al. (2026), this region-wise formulation provides continuous control over how much reference information is preserved throughout the sequence, exposing an identity-diversity frontier without additional training or per-reference optimization. Sound effects (SFX) are used as the main application and experimental validation domain of this research, while the underlying methodological questions remain relevant to controlled audio generation more broadly. Its production requirements provide a demanding setting for studying representation fidelity, sparse-reference conditioning, multimodal control, and low-latency generation.

Contributions. We formulate a **single-reference audio variation** as a stochastic process structured by the internal representation of a reference latent, rather than relying solely on external conditioning. Moreover, this work introduces a **training-free latent segmentation mechanism** and independently sampled **regional flow-levels** for a continuous variation mechanism without per-reference optimization. Additionally, we propose an improved **validation protocol for SFX variations application**, designed to characterize the balance between reference alignment and diversity.

2 Related Work

Reference-guided audio generation. One-shot generative modeling aims to capture the structure of a previously unseen sample and generate relevant alternative realizations from a single observation Rezende et al. (2016). In audio, SpecSinGAN Barahona Rios & Collins (2022) learns the internal distribution of a single SFX spectrogram to synthesize variations of the same sound, while WaveFlow Andreu & Villanueva Aylagas (2022) conditions waveform generation on mel spectro-

grams to produce controllable SFX variations. Recent generative methods increasingly separate semantic sound information from finer reference-specific acoustic characteristics. Simi-SFX Liu & Jin (2024) models continuous timbral similarity for SFX, DreamAudio Yuan et al. (2026b) uses few reference recordings to capture customized acoustic characteristics beyond semantic conditioning, and AC-Foley Fang et al. (2026) uses reference audio to resolve fine-grained acoustic differences not precisely conveyed by text. AnySynth Jing et al. (2026) similarly preserves the acoustic identity of an unseen reference instrument while changing its musical structure. More broadly, reference information has been incorporated through embeddings and multimodal objectives Hadjeres et al. (2026), as well as temporal control signals such as RMS envelopes Chung et al. (2024). These approaches primarily improve the information supplied to or learned by the generative model through conditioning, control signals, or training objectives. In contrast, we study whether the internal structure of the reference representation can directly organize the stochastic generation process itself, without additional training or a learned control representation.

Heterogeneous and asynchronous corruption Diffusion Forcing demonstrates that sequence elements do not need to share a single noise level, and training a diffusion or flow model with independent per-token noise levels can bridge autoregressive prediction with full-sequence diffusion Chen et al. (2024). More recently, asynchronous flow and diffusion schemes have extended this principle by varying corruption schedules across spatial or temporal regions, in some cases using local visual or paired audio information to inform these schedules Hu et al. (2026); Schusterbauer et al. (2026); Liu et al. (2026). Our method builds on these heterogeneous flow levels but leverages it on a different purpose. Rather than extending the generation horizon or imposing an autoregressive ordering, we extract content-aligned temporal regions directly from the reference audio latent and dependently shape sampled flow levels to each selected region. This allows control on how strongly regeneration happens within a fixed-duration sequence.

Content-dependant segmentation and representation geometry Evidence from self-supervised speech representations suggests that individual temporal frames carry different amounts of identity-relevant information Pasad et al. (2024). Such representational structure can itself provide a signal for segmentation: Pasad et al. Pasad et al. (2024) derive training-free word boundaries from changes between neighboring frame representations, while ZeroSyl Visser et al. (2026) extracts syllable boundaries from peaks in the L_2 norm of frozen WavLM representations. Beyond speech, adaptive segmentation has been explored through learned content-dependent chunking, as in H-Net Hwang et al. (2026), and data- and timestep-dependent dynamic chunking in diffusion transformers, as in DC-DiT Haridas et al. (2026). Related change-point approaches identify structural transitions from representation changes, including neighboring temporal representations in TS-CP^2 Deldari et al. (2021) and temporal self-similarity in UBoCo Kang et al. (2022). Our objective differs from segmentation or boundary detection as an end task: we use the geometry and activity of a pretrained generative latent as training-free signals to partition a reference into regions over which independent stochastic corruption can act.

3 Proposed Method

3.1 Flow-Matching Background and Reference Corruption

Latent representation and flow path. Given a reference audio waveform x_{ref}, we aim to generate diverse variations that preserve its perceived sound identity. Let E and D denote a pretrained audio encoder and decoder, and let $\mathbf{z}_{\mathrm{ref}} = E(x_{\mathrm{ref}}) \in \mathbb{R}^{C\times T}$ be the reference latent, with C channels and T temporal frames. We index frames by $t \in \{0, \dots, T-1\}$ and reserve τ for flow time. For a generic clean latent $\mathbf{z}_0$ drawn from the encoded data distribution, we adopt the linear flow-matching path used by Woosh Hadjeres et al. (2026):

$$\mathbf{z}_\tau = (1-\tau)\mathbf{z}_0 + \tau\boldsymbol{\epsilon}, \qquad \tau \in [0,1], \tag{1}$$

where $\boldsymbol{\epsilon} \in \mathbb{R}^{C\times T}$ has independent standard Gaussian entries and is independent of $\mathbf{z}_0$. Thus, $\tau = 0$ corresponds to clean data and $\tau = 1$ to isotropic Gaussian noise.

Flow matching and reference-guided variation. The conditional velocity along the Eq. 1 is $\mathbf{u}_\tau = \partial\mathbf{z}_\tau/\partial\tau = \boldsymbol{\epsilon} - \mathbf{z}_0$. A velocity field $\mathbf{v}_\theta$ is trained using

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\mathbf{z}_0,\boldsymbol{\epsilon},\tau}\left[\|\mathbf{v}_\theta(\mathbf{z}_\tau,\tau) - (\boldsymbol{\epsilon}-\mathbf{z}_0)\|_F^2\right], \tag{2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, and the expectation is over training latents, Gaussian noise, and flow times sampled according to the backbone's training distribution. We suppress any existing backbone conditioning inputs for clarity. Generation integrates the ODE $\mathrm{d}\mathbf{z}_\tau/\mathrm{d}\tau = \mathbf{v}_\theta(\mathbf{z}_\tau, \tau)$ from $\tau = 1$ to $\tau = 0$. For reference-guided variation, we instead choose a starting level $\tau^\star \in (0, 1)$, initialize $\mathbf{z}_{\tau^\star} = (1 - \tau^\star)\mathbf{z}_{\text{ref}} + \tau^\star \boldsymbol{\epsilon}$, and integrate the same ODE from $\tau^\star$ to zero. The starting level controls the relative contributions of the reference and Gaussian noise in the initialization.

From global to heterogeneous flow levels. A global corruption level applies the same interpolation strength to every temporal frame Evans et al. (2026). To vary corruption across the reference, we introduce a temporal field $\boldsymbol{\tau} = (\tau(0), \ldots, \tau(T-1)) \in [0, 1]^T$ and initialize each frame as

$$\mathbf{z}_{\boldsymbol{\tau},t} = \big(1 - \tau(t)\big)\mathbf{z}_{\text{ref},t} + \tau(t)\boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_C). \tag{3}$$

The all-one field recovers the isotropic Gaussian endpoint, while heterogeneous levels retain different amounts of reference information across frames. We assume a velocity model $\mathbf{v}_\theta(\mathbf{z}, \boldsymbol{\tau})$ conditioned on frame-wise flow times and trained to denoise heterogeneously corrupted latents. In our experiments, this capability is obtained by adapting the pretrained backbones, as described in the experimental setup. After adaptation, the model remains fixed across references. The next subsection describes how the reference latent determines the temporal regions and their initial corruption levels, without additional training or per-reference optimization.

3.2 Proposed Reference-Structured Corruption (RSC) Method

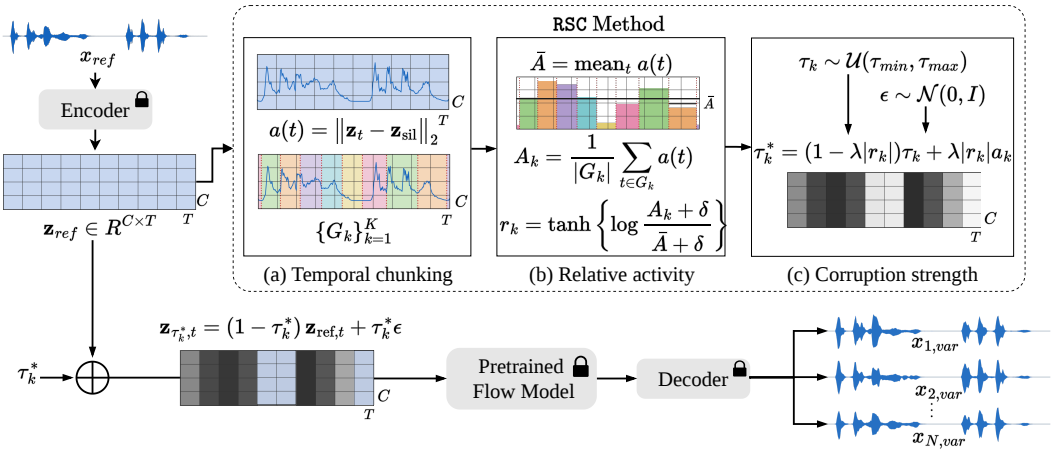


Figure 2: **RSC for one-shot SFX variation**. Given a reference latent, our method **(a)** partitions the sequence into activity-aligned temporal regions, **(b)** measures each region's activity relative to the full reference, and **(c)** uses this signal to modulate independently sampled regional flow times. The resulting RSC is denoised by the pretrained flow model to produce variations.

Figure 2 illustrates the proposed RSC scheme. Given the reference latent representation $\mathbf{z}$, we first localize its temporal structure, then use relative latent activity to determine where variation should be allocated, and finally control how much corruption each region receives through a magnitude-conditioned flow-time field.

Where regions are formed: Training-free Activity-Aligned Temporal Chunking. As shown in panel (a) of Figure 2, the boundary rule operates entirely in the latent domain: only $\mathbf{z}$ is available, so no waveform-domain feature is used. First, we denote the **activity of the latent** $\mathbf{z}$. Digital silence does not encode to the origin; thus, to create a representative silence latent $\mathbf{z}_{\text{sil}}$, we fix a reference $\mathbf{z}_{\text{sil}} \in \mathbb{R}^C$, obtained once by encoding a silent buffer and selecting a central frame, away from STFT center padding and convolutional edge padding, to capture the steady-state response rather than an edge artifact. We define the activity by

$$a(t) = \big\|\mathbf{z}_t - \mathbf{z}_{\text{sil}}\big\|_2, \qquad t = 0, \ldots, T-1, \tag{4}$$

with the Euclidean distance in the C-dimensional latent space; it is silence-relative latent activity. Because $\mathbf{z}_{\text{sil}}$ is fixed, $a(t)$ is comparable across frames. We show in Figure 7 that the latent activity curve closely follows the temporality and amplitude of the reference waveform's RMS curve,

while preserving more details at the same window size of 10 ms. Then, a **structural anchor detection** identifies locations of important transitions b_k within the latent sequence $\mathbf{z}$. Inspired by representation-based change-point detection for time series Deldari et al. (2021), which detects transition points using the cosine similarity of consecutive temporal window pairs, we directly operate on a pretrained generative latent space and compare the mean latent representations μ_t of two contrastive windows W, with

$$\mu_t^- = \frac{1}{N}\sum_{i=t-N}^{t-1}\Delta\mathbf{z}_i, \quad \mu_t^+ = \frac{1}{N}\sum_{i=t}^{t+N-1}\Delta\mathbf{z}_i, \tag{5}$$

where $\Delta\mathbf{z}_i = \mathbf{z}_i - \mathbf{z}_{\text{sil}}$ and N is the number of temporal frames within W. Near the sequence temporal edges, each window is shortened to exclude padded frames, while requiring at least two latent frames on each side of t. Following the broader observation that temporal boundaries manifest as local breaks in representation self-similarity for video frames Kang et al. (2022), we replace the temporal self-similarity matrix with the mean latent representation of each temporal window and compute the contrast c_t as a function of the cosine similarity between μ_t^- and μ_t^+ as follows

$$c_t = 1 - \frac{\langle \mu_t^-, \mu_t^+ \rangle}{\|\mu_t^-\|_2 \|\mu_t^+\|_2 + \varepsilon}, \tag{6}$$

with ε, set by default to a small value, e.g. 10^{-8}, avoiding division by zero, and a near-zero guard with $c_t = 0$ when both windows have no reliable latent direction, $\max\left(\|\mu_t^-\|_2, \|\mu_t^+\|_2\right) \le \rho s_{\text{ref}}$, with $\rho = 10^{-3}$ and s_{ref} the reference window-norm scale. Figure 3 highlights, with a red star, how we select the prominent local maxima as anchors b_i with $\mathcal{B} = \{b_1, \ldots, b_M\} = \mathcal{P}(c_t; p, d)$, where $\mathcal{P}$ selects local maxima of c_t with prominence greater than or equal to the peak prominence threshold p and each boundary respects a minimum separation d between selected peaks. We set $(W, p, d) = (5, 0.30, 7)$ as default. In this SFX application, $M = 0$ is allowed, as some sounds may not contain strong contrasts and an abrupt disruption could harm the sound structure.

Finally, as a boundary placed at a high contrast may cause a strong audible disruption in sound rendering, we refine the anchor set $\mathcal{B} = \{b_1, \ldots, b_M\}$ towards nearby minima of the latent activity. More precisely, for each anchor we search within its associated interval $\mathcal{V}_i$ and select

$$b_i^* = \arg\min_{t \in \mathcal{V}_i} a(t). \tag{7}$$

See Figure 3 for the visualization of this two-stage boundary scheme applied to the latent activity. We further compare the sme method on the RMS of the waveform x_{ref} and the spectral flux $X_{\text{ref}}(f, t)$; further results comparing both representations are observable in Table 4. To accommodate transient SFX whose attack can happen near a sequence edge, we additionally search the intervals between the sequence edges and the first or last anchors. The selected internal cuts b_k^* are subject to a minimum separation defined as a fraction γ of the active sequence and bounded by $[m_{\min}, m_{\max}]$. We set this range between 7 and 50 frames (0.07 s–0.5 s) for flexibility over structure details in short impulse and long sequence sounds, and γ to 15% of the active signal. After refinement, the selected cuts partition the latent into an adaptive set of content-aligned chunks $\{G_k\}_{k=1}^K$, as shown in the bottom-left panel of Figure 2. Recent dynamic sequence models have shown the benefit of replacing fixed partitions with content- and context-dependent segmentations Hwang et al. (2026). In contrast to a learned chunking mechanism, our partition is obtained directly from the **geometry of a pretrained latent sequence** and refined by its **latent activity**, requiring no additional training.

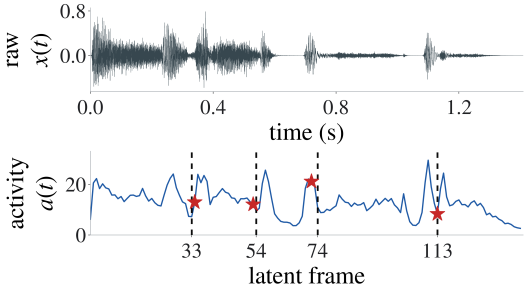


Figure 3: Stage **(a)** Two-stage activity-aligned temporal chunking.

Where to allocate variation: Bounded relative activity. Panel (b) of Figure 2 compares the activity of each detected chunk with the activity of the full reference. For each chunk G_k, we summarize its activity as

$$A_k = \frac{1}{|G_k|} \sum_{t \in G_k} a(t), \tag{8}$$

where $a(t)$ is the silence-relative latent activity defined in Equation 4. This chunk-level summary indicates which regions should receive relatively more or less stochastic freedom: low-activity regions are encouraged towards $\tau_{\min}$, while high-activity regions are encouraged towards $\tau_{\max}$, allowing more active acoustic regions greater freedom to vary. To measure the activity of a chunk relative to the full reference, we define the mean reference activity $\bar{A} = \frac{1}{T} \sum_{t=1}^{T} a(t)$ and compute the bounded relative activity $r_k = \tanh\left(\log \frac{A_k + \delta}{\bar{A} + \delta}\right)$, with $r_k \in (-1, 1)$ and δ a small numerical constant. Hence, $r_k > 0$ denotes a region that is more active than the reference average, $r_k < 0$ denotes a less active region, and $|r_k|$ measures the strength of this deviation. Bounding the activity contrast prevents arbitrarily large modulation of the regional flow time.

How much: Magnitude-dependent regional flow field. As shown in panel (c) of Figure 2, we draw K stochastic regional flow times $\tau_k \sim \mathcal{U}(\tau_{\min}, \tau_{\max})$, preserving stochastic variation across generations. The relative activity r_k then determines how strongly each sampled flow time is moved towards one of the limits of the pretrained flow-time range. We define the activity-dependent target as $a_k = \tau_{\max}$ if $r_k \geq 0$ and $a_k = \tau_{\min}$ otherwise, and obtain the magnitude-conditioned flow time

$$\tau_k^* = (1 - \lambda |r_k|)\, \tau_k + \lambda |r_k| a_k, \qquad \lambda \in [0, 1]. \tag{9}$$

Thus, regions whose activity is close to the reference average retain their initially sampled flow time, whereas increasingly active or inactive regions are progressively moved towards $\tau_{\max}$ or $\tau_{\min}$, respectively. The parameter λ controls the strength of reference conditioning. The resulting chunk-wise constant temporal flow field is $\tau(t) = \tau_k^*$ for $t \in G_k$. Since τ_k^* is a convex interpolation between $\tau_k \in [\tau_{\min}, \tau_{\max}]$ and one of the two interval boundaries, it remains within the pretrained flow-time range by construction. We use $\tau(t)$ as the per-frame flow schedule, in Equation 3, while retaining the original isotropic Gaussian endpoint.

4 Results and Discussion

4.1 Experimental Setup

Asynchronous flow adaptation. Our corruption requires a frame-wise flow-time field $\boldsymbol{\tau} \in [\tau_{\min}, \tau_{\max}]^T$ rather than a single global flow time. We adapt Woosh Hadjeres et al. (2026) and Stable Audio 3 (SAO3) Evans et al. (2026) by evaluating their existing timestep embedding independently at each temporal position and propagating the resulting embeddings through their original adaptive-normalization pathway. When $\tau_1 = \cdots = \tau_T$, this reduces to the original scalar-time computation. Woosh is fine-tuned end-to-end, whereas for SAO3 only the timestep-to-modulation pathway is fine-tuned. During training, contiguous temporal chunks receive sampled flow levels tau_k independently. Architecture and optimization details are provided in Appendix A.4.

Datasets and Evaluation Metrics. We evaluate on 500 SFX references sampled from our internal dataset, generating 10 variations per reference ($5,000$ outputs). Generation quality is measured on PANNs Wavegram-Logmel model Kong et al. (2020b) with precision/recall from Sony CSLGrachten & Nistal (2025) and FAD_{5K} on VGGish embeddingsKong et al. (2020b) with the 500 references and $4.5k$ sampled from an internal corpus. To measure within-reference diversity, we adapt WCIS Benny et al. (2021) by replacing the class label with the reference clip as a condition; the marginal is pooled over a reference's own variations instead of over a class: $p_r(y) = \frac{1}{N_{\text{ref}}} \sum_{j=1}^{N_{\text{ref}}} p(y \mid x_{\text{ref},j})$, where the classifier $p(y \mid x_{\text{ref}})$ is obtained from the 527-class PANNs Wavegram-Logmel model Kong et al. (2020a;b); higher WCIS indicates greater diversity among variations of the same reference. Following the calibration study in Appendix A.2, identity preservation is primarily measured by event-distribution $e_{\text{JSD}} = JSD(p_{ref} || p_{var})$, with p the distributions of referenced and generated data in PANNs space and CLAP cosine similarity between each reference to its variations for the $\mathcal{S}_{id}$ score and pairwise cosine distance between variations of a reference for $\mathcal{S}_{div}$ Desbos et al. (2026). Additional acoustic and temporal diagnostics are reported in the Appendix A.2.

4.2 Empirical Results

To validate our method, we first calibrate the preservation-variation regime by relying on different perspectives of variation: real production and ESC-50 Piczak (2015) samples. This allows us to frame the evaluation, supported by a human study, on our RSC applied to Woosh Hadjeres et al. (2026), and Stable Audio 3 Evans et al. (2026). We analyze controllability through activity, λ, and flow-time ablations, and finally we assess representation robustness.

Table 1: Corruption schemes on Woosh Hadjeres et al. (2026) and SAO3 Evans et al. (2026) at $[\tau_{\min}, \tau_{\max}] = [0.70, 0.85]$. DF denotes asynchronous Diffusion Forcing. Closest to ground truth in **bold**, second closest underlined.

		Quality$_{\text{PANNs}}$		FAD ↓	$\mathcal{S}_{div}$		$\mathcal{S}_{\text{id}}$	
		P% ↑	R% ↑	VGGish	WCIS$_{\text{PANNs}}$ ↑	CLAP$_{\text{LAION}}$	e_{JSD}	CLAP$_{\text{LAION}}$
GT	**Production SFX variation**	**1.00**	**0.98**	**NA**	**1.23**	**0.25**	**0.26**	**0.78**
	Broader semantic class variation	0.99	0.69	NA	1.47	0.30	0.29	0.69
Woosh	Random+ gated silence	0.85	0.76	4.82	1.29	**0.23**	**0.13**	0.71
	0.5 **s clear context**							
	ODE (dopri5, a,r_tol=0.001)	**0.92**	0.88	4.44	1.17	0.14	0.05	0.83
	SDE (schedule=σ, churn=2, steps=300)	**0.92**	**0.93**	**4.38**	**1.19**	0.15	0.05	**0.81**
	Woosh Hadjeres et al. (2026)	**0.89**	**0.89**	4.65	1.46	0.34	0.23	0.64
	Woosh+DF	0.85	0.77	4.82	**1.29**	**0.23**	0.19	**0.71**
	Woosh+DF + Ours	**0.89**	0.88	**4.41**	1.41	0.31	**0.24**	0.67
	SAO3 Evans et al. (2026)	0.70	0.26	6.83	**1.45**	**0.24**	0.59	0.29
	SAO3+DF	**0.84**	**0.30**	**6.39**	1.69	0.42	**0.37**	**0.48**
	SAO3+DF + Ours	0.83	0.29	6.41	1.79	0.44	0.40	0.46

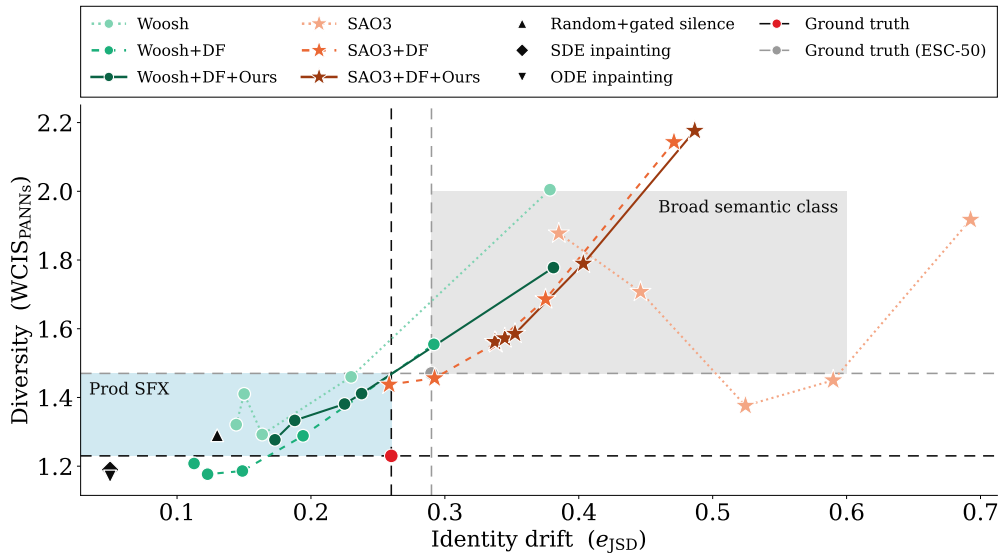


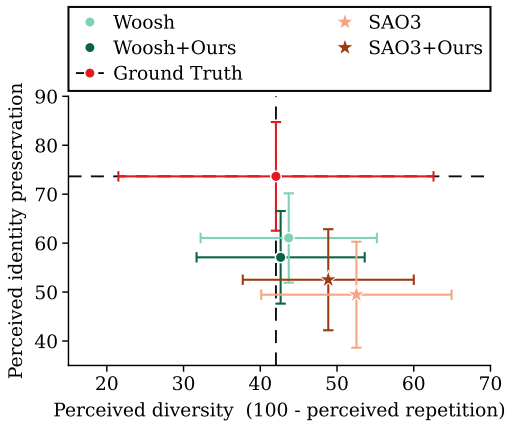
Figure 4: Preservation-variation frontier across flow-time bands (Appendix Table 5). Production SFX and within-class ESC-50 Piczak (2015) provide empirical reference regimes. DF denotes Diffusion Forcing fine-tuning for asynchronous flow levels.

Ground truth calibration for SFX preservation-variation regime. Our evaluation is calibrated using two real-data regimes. Our internal paired dataset contains 20 specific SFX identities (e.g., aggressive lion growling, Neanderthal snorts, etc.), with 3-4 production-made variations per identity. In addition, ESC-50 Piczak (2015) is used, where samples are paired within 50 broader sound classes (e.g., airplane, crow, etc.). Further details on this calibration and the metric selection are provided in Appendix A.2. Production SFX variations occupy a specific-identity regime with WCIS$_{\text{PANNs}} = 1.23$ and $e_{\text{JSD}} = 0.26$, while broader within-class ESC-50 variations reach 1.47 and 0.29, respectively. These empirical reference regimes were used rather than optimization targets, as increasing diversity is meaningful only while retaining the reference identity. Figure 4 shows how

RSC consistently shifts models on asynchronous flow level (DF: Diffusion forcing fine-tuning) towards higher diversity at matched flow-time ranges on both backbones. Resulting trajectories move through the empirically calibrated region from production SFX identity variations toward broader semantic variation, demonstrating direct control over the preservation-variation trade-off.

Impact of RSC. We demonstrate the relevance of aligning the activity of the representation for structured corruption at a fixed flow field range $[\tau_{\min}, \tau_{\max}] = [0.70, 0.85]$, adding RSC to Woosh+DF increases WCIS from 1.29 to 1.41 and CLAP diversity from 0.23 to 0.31, while moving e_{JSD} from 0.19 to 0.24. Quality improves simultaneously (P/R: $0.85/0.77 \rightarrow 0.89/0.88$; FAD: $4.82 \rightarrow 4.41$). On SAO3, the same allocation shifts WCIS from 1.69 to 1.79 (e_{JSD} : $0.37 \rightarrow 0.40$) with nearly unchanged FAD ($6.39 \rightarrow 6.41$), showing that the effect is not specific to Woosh.

Perceptual Evaluation. The objective frontier is complemented with a human evaluation of perceived diversity and reference-identity preservation. 15 participants rated five variations for each of seven diverse references on a 0-100 scale, comparing our method with the baselines. Figure 5 shows distinct perceptual preservation-variation operating points relative to production SFX, complementing Figure 4.



Method	Diversity (↑)	Identity (↑)
Ground Truth	42.04 [21.5, 65.6]	73.64 [62.5, 84.7]
Woosh+Ours	**42.6** [31.7, 53.6]	<u>57.1</u> [47.6, 66.6]
Woosh	<u>43.7</u> [32.2, 55.2]	**61.0** [51.9, 70.2]
SAO3+Ours	48.9 [37.7, 60.0]	52.5 [51.9, 70.2]
SAO3	52.5 [40.1, 64.9]	49.5 [38.6, 60.3]

Mean human ratings with 95% confidence intervals. Diversity is computed as 100 − repetition.

Figure 5: **Human evaluation of the diversity-identity trade-off.** Left: mean perceived diversity and identity preservation with 95% confidence intervals. Right: corresponding numerical results.

4.3 Ablation Studies

Activity-conditioned Allocation. Table 6 isolates how reference structure allocates a fixed corruption budget. Starting from content-aligned chunks with random flow times (1.35/0.20 WCIS/e_{JSD}), activity ranking reaches 1.37/0.22, while magnitude-conditioned allocation reaches 1.41/0.24. Thus, reference activity provides a direct signal for allocating stochasticity across the non-uniform temporal structure of SFX. Rather than increasing corruption uniformly over the sequence, the stochasticity is allocated according to the local structure of the reference, allowing the corruption process itself to depend on the non-uniform complex structure of SFX.

Table 2: Ablation on λ. Best per column in **bold**.

		Quality$_{\text{PANNs}}$		FAD ↓	$\mathcal{S}_{div}$		$\mathcal{S}_{\text{id}}$	
	λ	P% ↑	R% ↑	VGGish	WCIS$_{\text{PANNs}}$	CLAP$_{\text{LAION}}$	e$_{JSD}$	CLAP$_{\text{LAION}}$
GT	**Production SF**	**1.00**	**0.98**	**NA**	**1.23**	**0.25**	**0.26**	**0.78**
GT	Broader semantic class	0.99	0.69	NA	1.47	0.30	0.29	0.69
	$\lambda = 0.25$	**0.90**	<u>0.88</u>	4.38	**1.36**	**0.29**	0.21	**0.70**
	$\lambda = 0.50$	<u>0.89</u>	**0.89**	4.39	<u>1.38</u>	**0.29**	0.22	<u>0.69</u>
	$\lambda = 0.75$	<u>0.89</u>	<u>0.88</u>	4.40	1.39	<u>0.30</u>	<u>0.23</u>	0.68
	$\lambda = 1.00$	<u>0.89</u>	<u>0.88</u>	4.41	1.41	0.31	**0.24**	0.67

Controllability on the Preservation-variation Frontier. In Table 2, we demonstrate that increasing λ strengthens the activity-conditioned corruption and systematically shifts the operating

point: WCIS increases monotonically $1.36 \rightarrow 1.38 \rightarrow 1.39 \rightarrow 1.41$, while e_{JSD} increases $0.21 \rightarrow 0.22 \rightarrow 0.23 \rightarrow 0.24$, approaching the production-SFX regime (0.26). Quality remains stable (P/R $= 0.90 - 0.89/0.88 - 0.89$; $\text{FAD}_{5K} = 4.38$-4.41), showing λ as a direct control over the identity-diversity trade-off.

Chunk-independant Flow Fields $\tau(i)$. The flow-time range provides a second control over variation. Increasing $\mathcal{U}(0.60, 0.80) \rightarrow \mathcal{U}(0.70, 0.85) \rightarrow \mathcal{U}(0.80, 0.95)$ moves WCIS $1.28 \rightarrow 1.41 \rightarrow 1.78$ and e_{JSD} $0.17 \rightarrow 0.24 \rightarrow 0.38$. Thus, stronger corruption increases variation, but eventually moves generation beyond the reference-preserving regime and degrades quality.

Table 3: Fluctuation of timesteps range at sampling time. Best per column in **bold**.

		$\text{Quality}_{\text{PANNs}}$		FAD ↓	$\mathcal{S}_{div}$		$\mathcal{S}_{\text{id}}$	
	Flow times t	P% ↑	R% ↑	VGGish	$\text{WCIS}_{\text{PANNs}}$	$\text{CLAP}_{\text{LAION}}$	e_{JSD}	$\text{CLAP}_{\text{LAION}}$
GT	**Production SFX**	**1.00**	**0.98**	**NA**	**1.23**	**0.25**	**0.26**	**0.78**
	Broader semantic class	0.99	0.69	NA	1.47	0.30	0.29	0.69
	$\mathcal{U}(0, 1)$	**0.90**	<u>0.86</u>	4.74	1.38	0.29	<u>0.22</u>	0.67
	$\mathcal{U}(0.3, 0.9)$	**0.90**	<u>0.86</u>	4.60	<u>1.33</u>	**0.26**	0.10	<u>0.72</u>
	$\mathcal{U}(0.6, 0.8)$	<u>0.89</u>	0.85	<u>4.44</u>	**1.28**	<u>0.23</u>	0.17	**0.74**
	$\mathcal{U}(0.7, 0.85)$	<u>0.89</u>	**0.88**	**4.41**	1.41	0.31	**0.24**	0.67
	$\mathcal{U}(0.8, 0.95)$	<u>0.89</u>	0.78	5.15	1.78	0.49	0.38	0.48

Latent Representation Relevance. We extend our latent reference-structured method and prove its representation-agnostic behaviour through waveform-derived representations: RMS from $x(t)$ and spectral flux from $X(f, t)$. For each representation, contrast determines candidate boundaries and activity refines the final cuts at the same 10 ms resolution. All arms share the same references, random seed, noise field, flow-time sampling $\tau_k \sim \mathcal{U}[0.70, 0.85]$, sampler, and generation setup, isolating the representation used to structure corruption. Figures 7 compare the resulting regions, while Table 4 shows that RMS and spectral flux both reach $\text{WCIS}_{\text{PANNs}}/e_{\text{JSD}} = 1.36/0.21$, compared with $1.41/0.24$ for the latent representation, at comparable FAD ($4.38/4.38/4.41$). Thus, `RSC` is not restricted to latent-space structure; we retain the latent formulation as it directly derives both boundaries and activity from the same representation used for generation.

Table 4: Ablation on `RSC` method applied with the activity information taken from the latent activity itself, the spetral flux and the RMS directly from the waveform. Best per column in **bold**.

	$\text{Quality}_{\text{PANNs}}$		FAD ↓	$\mathcal{S}_{div}$		$\mathcal{S}_{\text{id}}$	
Data Representation	P% ↑	R% ↑	VGGish	$\text{WCIS}_{\text{PANNs}}$	$\text{CLAP}_{\text{LAION}}$	e_{JSD}	$\text{CLAP}_{\text{LAION}}$
Ground truth	1.00	0.98	NA	1.23	0.25	0.26	0.78
RMS $x(t)$	**0.90**	<u>0.89</u>	**4.38**	**1.36**	**0.28**	<u>0.21</u>	**0.70**
Spectral flux $X(f, t)$	**0.90**	**0.90**	**4.38**	**1.36**	**0.28**	<u>0.21</u>	**0.70**
Latent $z(t)$	<u>0.89</u>	0.88	<u>4.41</u>	<u>1.41</u>	<u>0.31</u>	**0.24**	<u>0.67</u>

5 Conclusion

We introduced reference-structured corruption (`RSC`) for one-shot SFX variation, where the internal structure of a single reference determines where and how strongly stochasticity is introduced. By combining activity-aligned temporal chunking, bounded relative activity, and region-wise flow levels, our method turns a pretrained flow model into a controllable variation mechanism without per-reference optimization. Across Woosh and Stable Audio 3, reference-structured corruption moves generation along the preservation-variation frontier while maintaining generation quality. Our analysis further shows that the latent representation can serve not only as the generation space, but also as a signal for organizing the stochastic process itself. This suggests a broader role for learned representations: not only as spaces in which generation occurs, but as signals that can directly structure the stochastic generation process.

Supplementary Material: It provides additional ablations and identity diagnostics, evaluation calibration and metric analysis, human-evaluation details, implementation and computational costs, algorithms, and qualitative analysis of the proposed chunking scheme.

AI USE STATEMENT

In this work, we used generative AI tools for software-development assistance, including drafting, debugging, and refactoring research and evaluation code. Additionally, we used LLMs for writing refinement, including improvements to grammar, clarity, organization, and presentation of author-written text. We reviewed all AI-assisted work. AI-assisted code was manually inspected and tested by the authors, and AI-assisted text was reviewed and revised by the authors for technical accuracy and consistency with the experiments and claims presented in the paper. We take responsibility for the final content of this work, including text, claims, code, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide the methodological and experimental details required to reproduce the proposed reference-structured corruption framework. Section A.4 describes the backbone adaptation and training setup, while Section A.6 provides pseudocode for asynchronous flow adaptation and the proposed reference-structured corruption procedure. The supplementary material also reports the evaluation calibration (Section A.2), computational cost (Section A.5), human-evaluation protocol (Section A.3), and additional ablations and qualitative analyses. We additionally provide an anonymized, frozen demo page with representative audio examples from the evaluated methods. Code will be released with the submission materials.

REFERENCES

S. Andreu and M. Villanueva Aylagas. Neural synthesis of sound effects using flow-based deep generative models. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 18, pp. 2–9, 2022. doi: 10.1609/aiide.v18i1.21941.

Adrian Barahona Rios and Tom Collins. Specsingan: Sound effect variation synthesis using single-image gans. In *Proceedings of the Sound and Music Computing Conference*, 2022. This is an author-produced version of the published paper. Uploaded in accordance with the publisher's self-archiving policy. Further copying may not be permitted; contact the publisher for details.

Yaniv Benny, Tomer Galanti, Sagie Benaim, and Lior Wolf. Evaluation metrics for conditional image generation. *International Journal of Computer Vision*, 129(5):1712–1731, 2021. doi: 10.1007/s11263-020-01424-w.

Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 24081–24125, 2024.

Yoonjin Chung, Junwon Lee, and Juhan Nam. T-FOLEY: A controllable waveform-domain diffusion model for temporal-event-guided foley sound synthesis. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. doi: 10.1109/ICASSP48485.2024.10447380.

Shohreh Deldari, Daniel V. Smith, Hao Xue, and Flora D. Salim. Time series change point detection with self-supervised contrastive predictive coding. In *Proceedings of the Web Conference 2021*, WWW '21, pp. 3124–3135. ACM, April 2021. doi: 10.1145/3442381.3449903. URL http://dx.doi.org/10.1145/3442381.3449903.

Mélodie Desbos, Yara Bahram, Eric Granger, and Mohammadhadi Shateri. A production-oriented framework for evaluation of sfx generation. In *Proceedings of the 29th International Conference on Digital Audio Effects (DAFx)*, Cambridge, MA, USA, 2026.

Zach Evans, Julian D. Parker, Matthew Rice, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio 3, 2026. URL https://arxiv.org/abs/2605.17991.

Pengjun Fang, Yingqing He, Yazhou Xing, Qifeng Chen, Ser-Nam Lim, and Harry Yang. AC-Foley: Reference-audio-guided video-to-audio synthesis with acoustic transfer. In *International Conference on Learning Representations (ICLR)*, 2026.

Hugo Flores García, Oriol Nieto, Justin Salamon, Bryan Pardo, and Prem Seetharaman. Sketch2sound: Controllable audio generation via time-varying signals and sonic imitations. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

John W. Gordon and John M. Grey. Perception of spectral modifications on orchestral instrument tones. *Computer Music Journal*, 2(1):24–31, 1978. doi: 10.2307/3680135.

Maarten Grachten and Javier Nistal. Accompaniment prompt adherence: A measure for evaluating music accompaniment systems. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, Hyderabad, India, 2025. IEEE.

Gaëtan Hadjeres, Marc Ferras, Khaled Koutini, Benno Weck, Alexandre Bittar, Thomas Hummel, Zineb Lahrichi, Hakim Missoum, Joan Serrà, and Yuki Mitsufuji. Woosh: A sound effects foundation model, 2026. URL https://arxiv.org/abs/2604.01929.

Akash Haridas, Utkarsh Saxena, Parsa Ashrafi Fashi, Mehdi Rezagholizadeh, Vikram Appia, and Emad Barsoum. Dc-dit: Adaptive compute and elastic inference for visual generation via dynamic chunking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Dik J. Hermes. *The Perceptual Structure of Sound*, volume 11 of *Current Research in Systematic Musicology*. Springer, Cham, Switzerland, 2023. ISBN 978-3-031-25565-6. doi: 10.1007/978-3-031-25566-3. URL https://doi.org/10.1007/978-3-031-25566-3.

Zijing Hu, Yunze Tong, Fengda Zhang, Junkun Yuan, Jun Xiao, and Kun Kuang. Asynchronous denoising diffusion models for aligning text-to-image generation. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 27693–27717, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2ecc80084c96cc25b11b0ab995c25f47-Paper-Conference.pdf.

Sukjun Hwang, Brandon Wang, and Albert Gu. Dynamic chunking for end-to-end hierarchical sequence modeling. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 149273–149313, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f1aa53cac69ef6980ca4a911ffcf278b-Paper-Conference.pdf.

Chong Jing, Junan Zhang, Jing Yang, Yulun Wu, Fan Fan, and Zhizheng Wu. AnySynth: Zero-shot instrument cloning via in-context learning and asymmetric hierarchical guidance, 2026.

Hyolim Kang, Jinwoo Kim, Taehyun Kim, and Seon Joo Kim. Uboco: Unsupervised boundary contrastive learning for generic event boundary detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20073–20082, June 2022.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. PANNs: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2880–2894, 2020a.

Qiuqiang Kong et al. PANNs: Audioset tagging cnn. https://github.com/qiuqiangkong/audioset_tagging_cnn, 2020b. GitHub repository.

Zhifeng Kong, Kevin J Shih, Weili Nie, Arash Vahdat, Sang gil Lee, Joao Felipe Santos, Ante Jukic, Rafael Valle, and Bryan Catanzaro. A2sb: Audio-to-audio schrodinger bridges. *arXiv*, 2025.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. AudioGen: Textually guided audio generation. In *International Conference on Learning Representations*, 2023.

Sang-gil Lee, Zhifeng Kong, Arushi Goel, Sungwon Kim, Rafael Valle, and Bryan Catanzaro. ETTA: Elucidating the design space of text-to-audio models. In *Proceedings of the International Conference on Machine Learning*, 2025.

Yunyi Liu and Craig Jin. Simi-SFX: A similarity-based conditioning method for controllable sound effect synthesis, 2024.

Yunyi Liu, Craig Jin, and David Gunawan. DDSP-SFX: Acoustically-guided sound effects generation with differentiable digital signal processing. In *Proceedings of the 27th International Conference on Digital Audio Effects (DAFx24)*, pp. 216–221, Guildford, United Kingdom, 2024.

Z. Liu, J. Lu, R. Lu, C. Liang, and S. Wang. Consisttalk: Intensity controllable temporally consistent talking head generation with diffusion noise search. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 7449–7457, 2026. doi: 10.1609/aaai.v40i9.37684.

Tianze Luo, Xingchen Miao, and Wenbo Duan. Wavefm: A high-fidelity and efficient vocoder based on flow matching. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 2187–2198, 2025. doi: 10.18653/v1/2025.naacl-long.110.

Stephen McAdams, Suzanne Winsberg, Sabine Donnadieu, Geert De Soete, and Jochen Krimphoff. Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, 58(3):177–192, 1995. doi: 10.1007/BF00419633.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12: 372–391, 2024. doi: 10.1162/tacl_a_00656. URL https://aclanthology.org/2024.tacl-1.21/.

K. J. Piczak. ESC: Dataset for environmental sound classification. In *ACM MM*, 2015. doi: 10.1145/2733373.2806390.

Danilo Rezende, Shakir, Ivo Danihelka, Karol Gregor, and Daan Wierstra. One-shot generalization in deep generative models. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1521–1529, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/rezende16.html.

Johannes Schusterbauer, Ming Gui, Yusong Li, Pingchuan Ma, Felix Krause, and Björn Ommer. Denoising, fast and slow: Difficulty-aware adaptive sampling for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43260–43270, June 2026.

Haokun Tian, Stefan Lattner, and Charalampos Saitis. Assessing the alignment of audio representations with timbre similarity ratings. In *Proceedings of the 26th International Society for Music Information Retrieval Conference*, pp. 710–718, 2025.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. AudioX: Diffusion transformer for anything-to-audio generation. In *ICLR*, 2026.

Nicol Visser, Simon Malan, Danel Slabbert, and Herman Kamper. ZeroSyl: Simple Zero-Resource Syllable Tokenization for Spoken Language Modeling. In *Interspeech 2026*, pp. 4062–4066, 2026. doi: 10.21437/Interspeech.2026-315.

Zeyu Xie, Xuenan Xu, Zhizheng Wu, and Mengyue Wu. Picoaudio: Enabling precise timestamp and frequency controllability of audio events in text-to-audio generation, 2024. URL https://arxiv.org/abs/2407.02869.

Zengwei Yao, Wei Kang, Han Zhu, Liyong Guo, Lingxuan Ye, Fangjun Kuang, Weiji Zhuang, Zhaoqing Li, Zhifeng Han, Long Lin, and Daniel Povey. Flow2gan: Hybrid flow matching and gan with multi-resolution network for few-step high-fidelity audio generation. In *International Conference on Learning Representations (ICLR)*, 2026.

Yi Yuan, Xubo Liu, Haohe Liu, Xiyuan Kang, Zhuo Chen, Yuxuan Wang, Mark D. Plumbley, and Wenwu Wang. Dreamaudio: Customized text-to-audio generation with diffusion models. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2421–2435, 2026a. doi: 10.1109/TASLPRO.2026.3678172.

Yi Yuan, Xubo Liu, Haohe Liu, Xiyuan Kang, Zhuo Chen, Yuxuan Wang, Mark D. Plumbley, and Wenwu Wang. DreamAudio: Customized text-to-audio generation with diffusion models. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2421–2435, 2026b. doi: 10.1109/TASLPRO.2026.3678172.

A APPENDIX

Supplementary Material

A.1 Additional Results and Ablations 13
A.1.1 Flow-Time Range Analysis 13
A.1.2 Corruption-Allocation Ablations 14
A.1.3 Additional Identity Diagnostics 14
A.2 Evaluation Calibration and Metrics 14
A.2.1 Ground-Truth Calibration Protocol 14
A.2.2 Metric Analysis 15
A.3 Human Evaluation Protocol 16
A.4 Implementation Details 17
A.4.1 Asynchronous Flow Adaptation 17
A.4.2 Backbone and Training Setup 17
A.5 Computational Cost 17
A.6 Algorithms 17
A.7 Representation and Chunking Analysis 18

A.1 ADDITIONAL RESULTS AND ABLATIONS

This section complements the main quantitative analysis with the full flow-time sweeps, corruption-allocation ablations, and additional reference-alignment diagnostics.

A.1.1 FLOW-TIME RANGE ANALYSIS

Table 5: Fluctuation of timesteps range at sampling time. Lowest identity drift highlight in **bold**.

Woosh			**SAO**		
Flow times t	$\text{WRIS}_{\text{PANNs}}$	e_{JSD}	Flow times t	$\text{WRIS}_{\text{PANNs}}$	e_{JSD}
Woosh ours			SAO+ DF + ours		
$\mathcal{U}(0,1)$	1.38	0.22	$\mathcal{U}(0,1)$	1.57	**0.34**
$\mathcal{U}(0.3,0.9)$	1.33	0.19	$\mathcal{U}(0.3,0.9)$	1.56	**0.34**
$\mathcal{U}(0.6,0.8)$	1.28	0.17	$\mathcal{U}(0.6,0.8)$	1.58	0.35
$\mathcal{U}(0.7,0.85)$	1.41	**0.24**	$\mathcal{U}(0.7,0.85)$	1.79	0.40
$\mathcal{U}(0.8,0.95)$	1.78	0.38	$\mathcal{U}(0.8,0.95)$	2.18	0.49
Woosh+ DF			SAO+ DF		
$\mathcal{U}(0,1)$	1.21	0.11	$\mathcal{U}(0,1)$	1.44	**0.26**
$\mathcal{U}(0.3,0.9)$	1.18	0.12	$\mathcal{U}(0.3,0.9)$	1.46	0.29
$\mathcal{U}(0.6,0.8)$	1.19	0.15	$\mathcal{U}(0.6,0.8)$	1.56	0.34
$\mathcal{U}(0.7,0.85)$	1.29	0.19	$\mathcal{U}(0.7,0.85)$	1.69	0.37
$\mathcal{U}(0.8,0.95)$	1.55	**0.29**	$\mathcal{U}(0.8,0.95)$	2.14	0.47
Woosh baseline			SAO baseline		
$\mathcal{U}(0,1)$	1.41	0.15	$\mathcal{U}(0,1)$	1.88	**0.38**
$\mathcal{U}(0.3,0.9)$	1.32	0.14	$\mathcal{U}(0.3,0.9)$	1.71	0.45
$\mathcal{U}(0.6,0.8)$	1.29	0.16	$\mathcal{U}(0.6,0.8)$	1.38	0.52
$\mathcal{U}(0.7,0.85)$	1.46	**0.23**	$\mathcal{U}(0.7,0.85)$	1.45	0.59
$\mathcal{U}(0.8,0.95)$	2.00	0.38	$\mathcal{U}(0.8,0.95)$	1.92	0.69

The complete flow-time sweeps complement Figure 4 in the main paper. Across both backbones, increasing the corruption range generally moves generation toward higher variation and greater identity drift. The effect is most clearly observed over $\mathcal{U}(0.60,0.80)$, $\mathcal{U}(0.70,0.85)$, and $\mathcal{U}(0.80,0.95)$, which progressively move the operating point along the preservation–variation frontier.

A.1.2 CORRUPTION-ALLOCATION ABLATIONS

We progressively isolate the contribution of reference structure to the regional corruption field. Random chunking assigns independently sampled flow times without using the reference structure. Content-aligned chunking replaces these random boundaries with the detected regions $\{G_k\}_{k=1}^{K}$. Activity ranking then assigns larger sampled flow times to more active regions, while the final magnitude-conditioned scheme continuously modulates each sampled τ_k using $|r_k|$. Table 6 reports these intermediate variants.

Table 6: Ablation of temporal structure, activity conditioning, and regional flow-time allocation.

	Quality	FAD ↓	$\mathcal{S}_{div}$		$\mathcal{S}_{\text{id}}$	
Corruption strength	P% / R% ↑	VGGish	$\text{WCIS}_{\text{PANNs}}$	$\text{CLAP}_{W/L}$	e_{JSD}	$\text{CLAP}_{W/L}$
Ground truth	1.00/ 0.98	8.45	1.23	0.21/ 0.25	0.26	0.78/ 0.78
Global τ	0.89/ 0.89	4.65	1.46	0.35/ 0.34	0.23	0.66/ 0.64
Random chunk + Random $\tau_k \sim \mathcal{U}(0.70, 0.85)$	0.85/ 0.77	4.82	**1.29**	0.32/ 0.23	0.19	0.70/ 0.71
G_k + Random $\tau_k \sim \mathcal{U}(0.70, 0.85)$	0.90/ 0.88	4.37	1.35	**0.31**/ 0.28	0.20	**0.71**/ 0.70
G_k +Activity rank $\tau_k \sim \mathcal{U}(0.70, 0.85)$	0.89/ 0.86	4.43	1.37	0.32/ 0.29	0.22	0.69/ 0.69
G_k + $\tau_k \sim \mathcal{U}(0.75, 0.85)$	0.89/ 0.83	4.43	1.42	0.35/ 0.32	0.24	0.67/ 0.66
G_k + $\tau_k \sim \mathcal{U}(0.80, 0.85)$	0.89/ 0.83	4.51	1.48	0.38/ 0.35	**0.27**	0.63/ 0.63
G_k+ $\lambda = 1.00$, $\tau_k \sim \mathcal{U}(0.70, 0.85)$	0.89/ 0.88	4.41	1.41	0.34/ 0.31	0.24	0.67/ 0.67

A.1.3 ADDITIONAL IDENTITY DIAGNOSTICS

In addition to the primary event-distribution and embedding metrics used in the main paper, we report complementary acoustic diagnostics for semantic event agreement, timbre, brightness, and loudness. These measures are used as diagnostics rather than as a single optimization criterion, since reference preservation depends jointly on several acoustic attributes.

Table 7: Activity-aligned latent chunking vs. random chunking. Best per column in **bold**.

	e_{KL} ↓	e_{timb} ↓	Δ_{bright} ↓	Δ_{loud} ↓
Ground truth	1.09	0.05	7.46	13.77
Global	**0.807**	0.038	4.515	4.367
Random+ gated silence	0.739	0.036	4.224	3.647
0.5 **s clear context**				
ODE (dopri5, a,r_tol=0.001)	0.346	0.015	2.215	1.922
SDE (schedule=σ, churn=2, steps=300)	0.382	0.016	2.452	2.312
Ours	0.805	**0.039**	**4.568**	**4.546**

A.2 GROUND-TRUTH EVALUATION CALIBRATION WITH METRIC ANALYSIS

A.2.1 GROUND-TRUTH CALIBRATION PROTOCOL

We calibrate the preservation–variation evaluation using two real-data regimes. The internal paired SFX set contains 20 specific sound identities with multiple production-made variations per identity, whereas ESC-50 Piczak (2015) provides 50 broader semantic sound classes.

For each corpus, we construct two pairing regimes. **AA** compares samples belonging to the same identity or class and therefore represents natural within-identity variation. **AB** compares samples belonging to different identities or classes and provides a contrasting regime in which substantially larger semantic and acoustic differences are expected. Generative systems are interpreted relative to the **AA** regime, rather than toward the maximum separation represented by **AB**. For **AA**, each internal SFX identity contains 3–4 production variations, while ESC-50 uses 12 samples per class in the calibration experiment. The **AB** analysis contains 3,300 inter-class pairs.

A.2.2 METRIC ANALYSIS

We characterize reference identity through complementary semantic, timbral, transient, texture, and structural dimensions:

$$\mathbf{S}_{\mathrm{id}}(x_{\mathrm{gen}}, x_{\mathrm{ref}}) = [S_{\mathrm{event}}, S_{\mathrm{timbre}}, S_{\mathrm{transient}}, S_{\mathrm{texture}}, S_{\mathrm{structure}}]\,. \tag{10}$$

Following prior work elucidating the design space of audio Lee et al. (2025), S_{event} measures preservation of the semantic event or sound class. We use the paired KL divergence Kreuk et al. (2023) between the reference and each variation in the PANNs Wavegram-Logmel space Kong et al. (2020a;b), $D_{\mathrm{KL}}(p_{\mathrm{ref}} \| p_{\mathrm{var}})$, where p denotes the corresponding event distributions. We additionally report the symmetric event-distribution distance $e_{\mathrm{JSD}} = \mathrm{JSD}(p_{\mathrm{ref}} \| p_{\mathrm{var}})$.

For S_{timbre}, we rely on Mel-Frequency Cepstral Coefficient representations Tian et al. (2025). Although originally studied in musical timbre, this representation provides a timbre-oriented space whose agreement with perceptual dissimilarity has been evaluated in prior work. We use it as a complementary measure of reference-specific acoustic similarity.

$S_{\mathrm{transient}}$ measures localized onset and transient preservation, which is particularly important for impulsive SFX Liu et al. (2024). We follow the transient diagnostics of the production-oriented SFX evaluation framework Desbos et al. (2026), comparing each variation with its reference through the peak-width ratio Δ_{FWHM}, pre-onset error, and overall onset-envelope error. These diagnostics are especially relevant for sounds whose attack carries a substantial part of their perceptual identity.

Following Sketch2Sound García et al. (2025), we use loudness and brightness as complementary diagnostics of $S_{\mathrm{structure}}$ and S_{texture}. $S_{\mathrm{structure}}$ evaluates whether the variation preserves the reference loudness shape and level. We report loud_l1_dbfs as

$$\text{loud_l1_dbfs} = \frac{1}{|G|} \sum_{t \in G} |L_{\mathrm{ref}}(t) - L_{\mathrm{var}}(t)|\,, \tag{11}$$

where G contains the reference frames satisfying $L(t) \geq -40\,\mathrm{dBFS}$. This measure can therefore also reflect shape or alignment discontinuities. For S_{texture}, we measure brightness through the frame-wise spectral centroid, providing a diagnostic of finer spectrotemporal and noise characteristics associated with the sound's timbral profile Hermes (2023).

Table 8 reports the resulting benchmark. We additionally compare the AA and AB metric distributions in Figure 6 using kernel density estimates, allowing us to inspect which measures most clearly distinguish within-identity from between-identity comparisons.

Table 8: **Ground truth validation for expected variations.** Every metric is scored on two corpora with the same protocol: pairs within one identity class (AA) against pairs across identity classes (AB). Arrows mark the direction of greater identity preservation, so AA is expected to lead AB on every identity column. Diversity is the exception: it is a distance over a reference's variation set, so AB is expected to lead, and a generative arm is read *against* AA rather than towards the maximum. Best per column in **bold**.

		Quality			Diversity			Alignment	
Corpus	Arm	P%	R%	FAD_{VGGish}	$\mathrm{WRIS}_{\mathrm{PANNs}}$	$CLAP_W$	$CLAP_{LAION}$	$CLAP_W$	$CLAP_{LAION}$
ESC-50	AA	0.99	0.69	3.80	1.47	0.42	0.30	**0.57**	**0.69**
	AB	0.99	0.73	4.36	4.33	**0.73**	**0.81**	0.25	0.17
Internal DS	AA	1.00	0.98	8.45	1.23	0.21	0.25	**0.78**	**0.78**
	AB	1.00	0.76	7.93	2.62	**0.68**	**0.84**	0.33	0.15

		Event		Timbre / texture		Structure		Transient (alignment-dependent)			
Corpus	Arm	$e_{\mathrm{KL}} \downarrow$ nats	$e_{\mathrm{JSD}} \downarrow$ bits	$e_{\mathrm{timb}} \downarrow$ $1-\cos$	$\Delta_{\mathrm{bright}} \downarrow$ st	$\Delta_{\mathrm{loud}} \downarrow$ dB	$r_{\mathrm{env}} \uparrow$ r	$R_{\mathrm{FWHM}} \to 1$ ratio	$\Delta t_{\mathrm{ons}} \downarrow$ ms	$\Delta t_{\mathrm{pw}} \downarrow$ ms	$m_{\mathrm{ons}} \uparrow$ frac
ESC-50	AA	**1.37**	**0.29**	**0.17**	**9.79**	**19.54**	**0.06**	1.03	57.62	59.99	**0.71**
	AB	7.33	0.90	0.30	13.33	25.18	0.03	1.03	**56.12**	**57.94**	0.70
Internal DS	AA	**1.09**	**0.26**	**0.05**	**7.46**	**13.77**	**0.38**	**0.99**	**46.68**	**49.13**	**0.70**
	AB	4.14	0.69	0.29	13.23	29.03	0.04	1.03	48.45	50.86	0.49

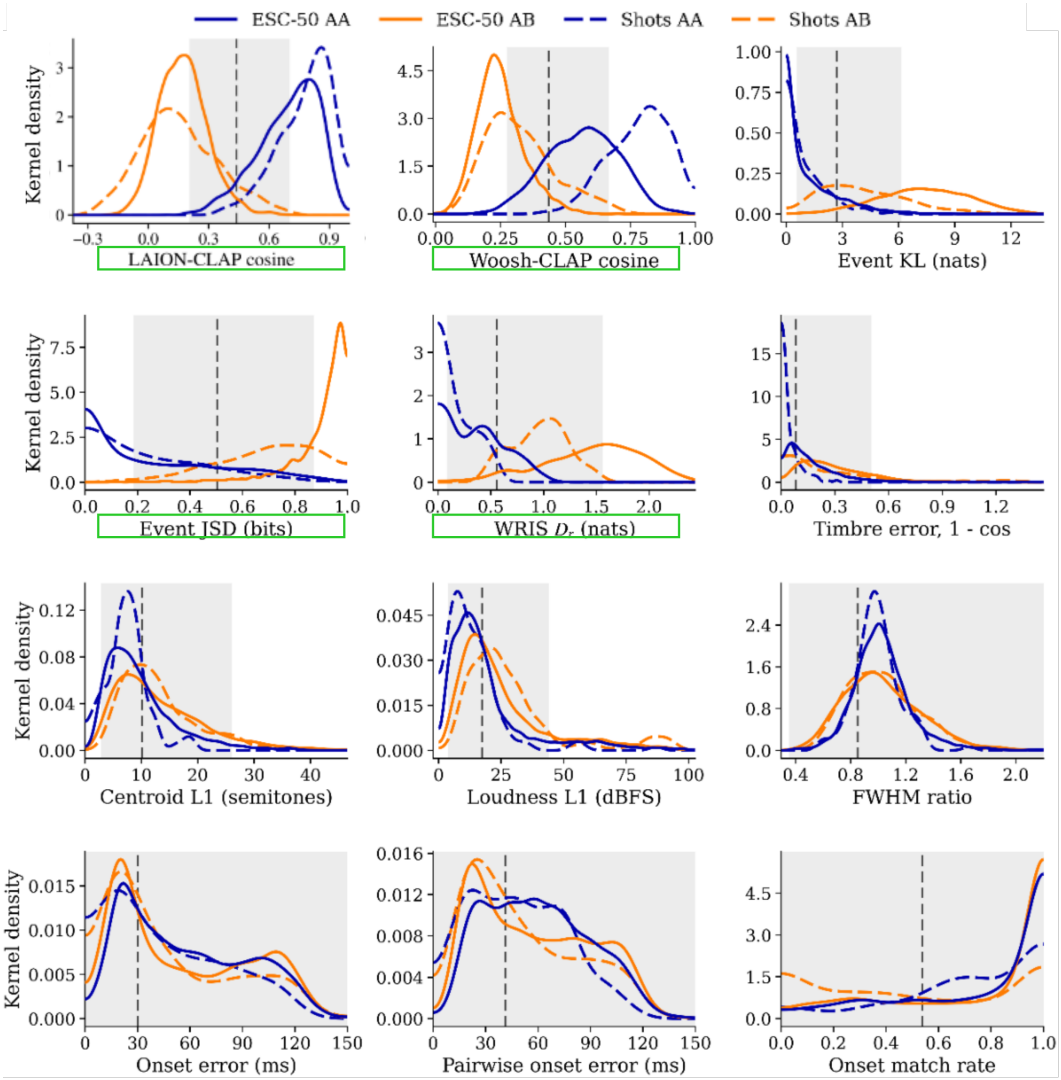


Figure 6: Kernel density estimates of the evaluated identity and diversity metrics under within-identity (**AA**) and between-identity (**AB**) comparisons. We evaluate the internal paired SFX set with 20 specific identities and ESC-50 with 50 broader semantic classes. Outlined metrics correspond to those retained for the main evaluation.

From Table 8 and Figure 6, the AA and AB regimes are clearly separated by the event-distribution and embedding metrics retained for the main evaluation. For the internal production SFX set, within-identity variation yields $\mathrm{WCIS}_{\mathrm{PANNs}} = 1.23$ and $e_{\mathrm{JSD}} = 0.26$, whereas between-identity comparisons increase to 2.62 and 0.69, respectively. ESC-50 exhibits a broader within-class regime, with $\mathrm{WCIS}_{\mathrm{PANNs}} = 1.47$ and $e_{\mathrm{JSD}} = 0.29$, and again increases substantially under **AB** pairing.

These observations motivate using $\mathrm{WCIS}_{\mathrm{PANNs}}$ to characterize within-reference variation and e_{JSD} as the primary event-level identity-drift measure in the main paper. The remaining acoustic, timbral, structural, and transient measures are retained as complementary diagnostics rather than combined into a single score.

A.3 Human Evaluation Protocol

We conduct a perceptual study to complement the objective preservation–variation analysis. Fifteen participants evaluate five variations for each of seven reference sounds sampled from our internal dataset. Scores are reported on a 0–100 scale, where higher diversity scores indicate greater per-

ceived variation and higher identity scores indicate stronger preservation of the reference identity. For each question, the four evaluated methods are presented through their concatenated variations, with individual samples separated by a short, volume-normalized *beep*. Participants are instructed to wear headphones during the evaluation.

A.4 Implementation Details

A.4.1 Asynchronous Flow Adaptation

Reference-structured corruption requires the velocity model to accept a frame-wise flow-time field rather than a single scalar flow time. We therefore adapt each backbone once using contiguous regions assigned independent flow levels. After this adaptation, the model is frozen, and the proposed reference-structured corruption requires no further training or per-reference optimization.

For Woosh, all parameters of the flow model are optimized during this adaptation. We adapt Stable Audio 3 (medium) by fine-tuning only its timestep and global-conditioning embedders (19.3M parameters, 1.3% of the DiT; all other weights frozen) with Diffusion Forcing ($K_{\mathrm{DF}} = 5$ chunks per clip, minimum 3 latent frames, levels $\sim \mathcal{U}[0, 1]$). We train on the training split of the Soundly SFX library (137 shards, ≈ 1.4M clips) with 5 s random crops, using AdamW (lr 10^{-4}, weight decay 0.01, 500 linear warm-up steps followed by cosine decay), batch size 4, FP16 mixed precision, gradient clipping at 1.0, and an EMA with decay 0.999, for 50K steps (3.35 h on a single NVIDIA H100 NVL).

A.4.2 Backbone and Training Setup

The Woosh-Flow diffusion transformer contains approximately 337.2M trainable parameters. Woosh-AE remains frozen and is used to encode waveforms into latent sequences and decode generated latents back into audio. Asynchronous adaptation is performed on an NVIDIA Quadro RTX 8000 with 48 GB of memory using a batch size of 8. The training set contains 50,000 audio clips of 5.01 s each. The Woosh latent has $C = 128$ channels and $T = 501$ temporal frames at 100 Hz.

For SAO3, only the timestep and global-conditioning embedders are optimized during asynchronous adaptation; the remaining pretrained parameters are kept fixed.

A.5 Computational Cost

We report the inference cost of RSC on both backbones in Table 9. For a 5 s generated variation on a single NVIDIA H100 NVL, we measure the number of function evaluations (NFE), steady-state latency, real-time factor (RTF), and peak VRAM usage. This comparison isolates the practical sampling cost introduced by the regional flow-time formulation relative to each backbone baseline.

Table 9: Inference cost for one 5 s variation on a single NVIDIA H100 NVL. Latency reports steady-state generation time. RTF denotes real-time factor.

Method	NFE	Latency (s)	RTF	VRAM (GB)
Woosh	76	1.082 ± 0.095	0.216	3.53
Woosh + Ours	48	0.712 ± 0.030	0.142	3.54
Stable Audio 3	10	0.708 ± 0.074	0.142	9.16
Stable Audio 3 + Ours	48	3.043 ± 0.034	0.609	9.16

A.6 Algorithms

$\mathbf{z}_{\boldsymbol{\tau},t}$

Diffusion Forcing Fine-Tuning Implementation. For Woosh, all parameters of $\mathbf{v}_\theta$ are optimized; for SAO3, only the timestep-to-modulation pathway is updated. When $\tau(0) = \cdots = \tau(T-1)$, the formulation recovers the original scalar-time conditioning.

$\mathbf{z}_{\boldsymbol{\tau},t}$

Algorithm 1: Asynchronous Diffusion-Forcing Adaptation

Input : Training set $\mathcal{D}_{\text{train}}$; pretrained encoder E and velocity field $\mathbf{v}_\theta$; number of random regions K_{DF}
Output: Adapted velocity field $\mathbf{v}_\theta$

$\mathbf{v}_\theta(\mathbf{z}, \tau) \to \mathbf{v}_\theta(\mathbf{z}, \boldsymbol{\tau}), \quad \boldsymbol{\tau} = (\tau(0), \ldots, \tau(T-1))$
while *not converged* **do**
  $x \sim \mathcal{D}_{\text{train}}$
  $\mathbf{z}_0 \leftarrow E(x)$
  Randomly partition $\{0, \ldots, T-1\}$ into K_{DF} contiguous regions $\{R_k\}$
  foreach R_k **do**
    $\tau_k \sim \mathcal{U}(0, 1)$
    $\tau(t) \leftarrow \tau_k, \qquad t \in R_k$
  $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  $\mathbf{z}_{\boldsymbol{\tau},t} \leftarrow \big(1 - \tau(t)\big)\mathbf{z}_{0,t} + \tau(t)\boldsymbol{\epsilon}_t$ (3)
  $\mathcal{L}_{\text{FM}} \leftarrow \|\mathbf{v}_\theta(\mathbf{z}_{\boldsymbol{\tau}}, \boldsymbol{\tau}) - (\boldsymbol{\epsilon} - \mathbf{z}_0)\|_2^2$ (2)
  $\mathbf{v}_\theta \leftarrow \text{update}\,(\mathbf{v}_\theta, \mathcal{L}_{\text{FM}})$
Freeze $\mathbf{v}_\theta$

A.7 Representation and Chunking Analysis

Figure 7 provides qualitative examples of the two-stage chunking procedure across the Woosh and Stable Audio 3 latent representations, waveform RMS, and spectral flux. The first stage detects structural transitions from neighboring-window contrast, while the second stage moves candidate boundaries toward local activity minima. These examples complement the quantitative representation ablation in Table 4.

Algorithm 2: Reference-Structured Corruption for One-Shot SFX Variation

Input : Reference waveform x_{ref}; adapted and frozen flow model $(E, D, \mathbf{v}_\theta)$; silence latent $\mathbf{z}_{\mathrm{sil}}$; flow-time band $[\tau_{\min}, \tau_{\max}]$; conditioning strength λ; number of variations V

Output : V variations $\{X_{\mathrm{var}.N}\}_{N=1}^{V}$ of x_{ref}

$\mathbf{z}_{\mathrm{ref}} \leftarrow E(x_{\mathrm{ref}})$

$\Delta\mathbf{z}_t \leftarrow \mathbf{z}_{\mathrm{ref},t} - \mathbf{z}_{\mathrm{sil}}$

$a(t) \leftarrow \|\mathbf{z}_{\mathrm{ref},t} - \mathbf{z}_{\mathrm{sil}}\|_2$ (4)

// **Where: Activity-Aligned Temporal Chunking**

for $t = 0, \ldots, T-1$ **do**

$$\mu_t^- \leftarrow \frac{1}{N} \sum_{i=t-N}^{t-1} \Delta\mathbf{z}_i \quad (5)$$

$$\mu_t^+ \leftarrow \frac{1}{N} \sum_{i=t}^{t+N-1} \Delta\mathbf{z}_i$$

$$c_t \leftarrow 1 - \frac{\langle \mu_t^-, \mu_t^+ \rangle}{\|\mu_t^-\|_2 \|\mu_t^+\|_2 + \varepsilon} \quad (6)$$

if $\max\left(\|\mu_t^-\|_2, \|\mu_t^+\|_2\right) \le \rho s_{\mathrm{ref}}$ **then**

$c_t \leftarrow 0$

$\mathcal{B} = \{b_1, \ldots, b_M\} \leftarrow \mathcal{P}(c_t; p, d)$

foreach $b_k \in \mathcal{B}$ **do**

$$b_k^* \leftarrow \arg\min_{t \in \mathcal{V}_k} a(t) \quad (7)$$

$m_{\mathrm{eff}} \leftarrow \max\left(m_{\min}, \min\left(m_{\max}, \lfloor \gamma T_{\mathrm{act}} \rfloor\right)\right)$

Retain valid b_k^* with minimum separation m_{eff}

$\{G_k\}_{k=1}^{K} \leftarrow \mathrm{Partition}\left(\{b_k^*\}, T\right)$

// **Where to Allocate: Bounded Relative Activity**

foreach G_k **do**

$$A_k \leftarrow \frac{1}{|G_k|} \sum_{t \in G_k} a(t) \quad (8)$$

$$\bar{A} \leftarrow \frac{1}{T} \sum_{t=1}^{T} a(t) \quad (3.2)$$

$$r_k \leftarrow \tanh\left(\log \frac{A_k + \delta}{\bar{A} + \delta}\right)$$

// **How Much: Magnitude-Dependent Regional Flow Field**

for $N = 1, \ldots, V$ **do**

$\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$

foreach G_k **do**

$\tau_k \sim \mathcal{U}(\tau_{\min}, \tau_{\max})$

$$a_k \leftarrow \begin{cases} \tau_{\max}, & r_k \ge 0, \\ \tau_{\min}, & r_k < 0 \end{cases}$$

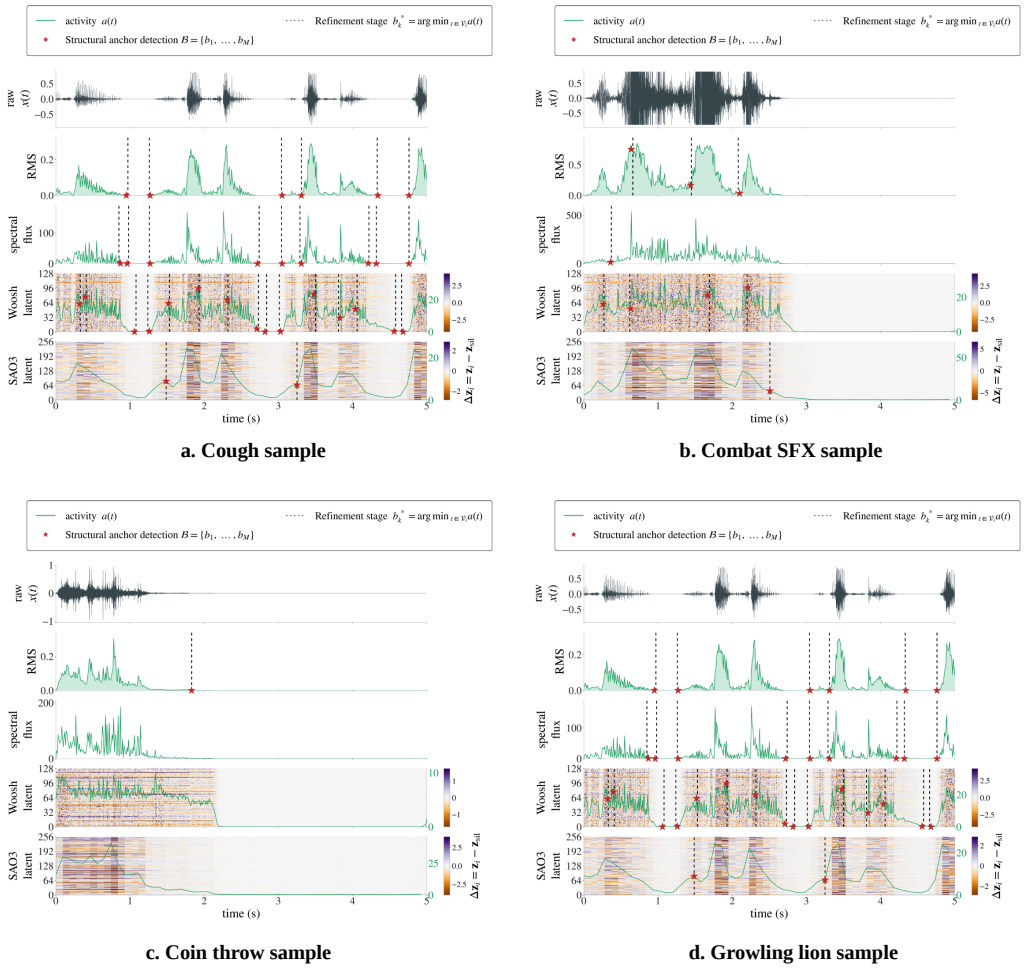
$$\tau_k^* \leftarrow (1 - \lambda|r_k|)\,\tau_k + \lambda|r_k| a_k \quad (9)$$

$\tau(t) \leftarrow \tau_k^*, \quad t \in G_k$

$$\mathbf{z}_{\boldsymbol{\tau},t} \leftarrow \big(1 - \tau(t)\big)\mathbf{z}_{\mathrm{ref},t} + \tau(t)\boldsymbol{\epsilon}_t \quad (3)$$

$\mathbf{z} \leftarrow \mathrm{ODEsolve}\left(\mathbf{v}_\theta, \mathbf{z}_{\boldsymbol{\tau}}, \boldsymbol{\tau} \to \mathbf{0}\right)$

$X_{\mathrm{var}.N} \leftarrow D(\mathbf{z})$



a. Cough sample

b. Combat SFX sample

c. Coin throw sample

d. Growling lion sample

Figure 7: Comparison of our two-stage chunking scheme across the Woosh and Stable Audio 3 latent spaces, waveform-derived representation (RMS envelope), and spectral flux of the Woosh encoder on different examples.