SOUNDWEAVER: COMPOSITIONAL WARM-STARTING FOR TEXT-TO-AUDIO DIFFUSION SERVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-audio (T2A) diffusion models generate high-quality audio but require tens of neural function evaluations (NFEs), resulting in substantial inference latency. Existing acceleration methods primarily optimize each generation trajectory independently, even when reusable acoustic structure exists in prior examples. We introduce SoundWeaver, the first training-free framework for compositional cross-example acoustic reuse. Rather than generating from pure noise, SoundWeaver constructs a prompt-conditioned acoustic prior from segments distributed across cached audios and warm-starts diffusion from an intermediate noise level. We introduce an *Acoustic Composition Graph* (ACG), which uses neural-codec representations and segment-level identification to identify candidate cross-cache transitions and structured path search to jointly optimize prompt relevance and acoustic continuity. We further establish a warm-start error bound showing how prior quality and diffusion SNR jointly govern the admissible skip depth, motivating an adaptive contextual-bandit *Skip Gater* that selects the warm-start level for each request. Across T2A diffusion backbones, SoundWeaver achieves 1.48–$3.74\times$ generation speedup with only a 1K-clip cache while improving generation quality.

1 INTRODUCTION

Diffusion and flow-matching models have become the foundation of text-to-audio (T2A) generation, such as for music composition and sound-effect synthesis (Liu et al., 2024a; 2023). These models iteratively transform Gaussian noise into high-fidelity waveforms or spectrograms over tens of sampling steps, conditioning on natural language prompts. Despite their quality advantages, inference remains inherently computationally intensive: generating a single clip can take multiple seconds on modern GPUs (Li et al., 2025). At production scale, where services process millions of requests daily, this translates into high user-perceived latency and substantial infrastructure cost.

Most existing acceleration methods reduce the cost of an individual generation trajectory. Training-based approaches distill multi-step diffusion into few-step generators, but require additional training for each target model, which implies substantial computational cost and data resources (Salimans & Ho, 2022; Song et al., 2023). Training-free approaches instead employ improved numerical solvers (Liu et al., 2022; Lu et al., 2022), adaptive timestep schedules (Chen et al., 2024), or early-exit strategies (Moon et al., 2024). Concurrently, systems work improves execution through multi-GPU parallelism and hardware–software co-design (Fang et al., 2024; Qiao et al., 2026; Agarwal et al., 2024; Xia et al., 2025).

We investigate a different source of efficiency: *can acoustic structure from prior examples be reused instead of repeatedly reconstructed from noise?* Real-world audio contains substantial semantic recurrence—such as rainfall, footsteps, engines, speech, or combinations of common acoustic events. Our analysis shows that even a modest 1K-audio cache frequently contains semantically related content (Fig. 1). At the same time, diffusion warm-starting provides a mechanism for exploiting such structure: rather than initializing from pure noise, a related audio prior can be noised to an intermediate timestep and denoised only from that point onward, skipping part of the reverse trajectory (Agarwal et al., 2024; Xia et al., 2025). We further show theoretically that a prior closer to the prompt-conditioned target distribution can tolerate a less-noised warm start, allowing more of the reverse process to be skipped (§3.1).

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.

The central difficulty is that, for audio, the useful unit of reuse is rarely an entire audio. Semantic relevance is temporally localized: a cached clip may contain only part of the requested event sequence, mix relevant and irrelevant intervals, or differ from the requested duration. Our analysis on real-world datasets shows that progressively partitioning audios into shorter segments exposes substantially stronger prompt–audio matches than whole-clip retrieval (§1). This suggests moving beyond the single-reference assumption used by prior warm-start systems for image generation (Xia et al., 2025; Agarwal et al., 2024). Instead, a useful acoustic prior may need to combine complementary structure distributed across different times and different audios.

We introduce SoundWeaver, a training-free framework for *compositional acoustic reuse*. Rather than retrieving one cached output, SoundWeaver constructs a prompt-conditioned acoustic prior from reusable segments across audios. To make such composition feasible, we introduce an *Acoustic Composition Graph* (ACG). Its nodes represent cached audio segments, while candidate cross-audio transitions are indexed using coarse-to-fine neural-codec prefixes and scored for acoustic continuity. Given a prompt and target duration, SoundWeaver assigns segment-level semantic scores and solves a dynamic program that jointly optimizes prompt relevance and transition compatibility. A small set of high-scoring trajectories is then assembled from the original waveforms and globally reranked against the complete prompt, yielding the final acoustic prior.

A strong prior alone does not determine how aggressively it should be reused. Preserving more of the prior allows more denoising steps to be skipped, but also leaves less opportunity for the generative model to revise mismatched content. SoundWeaver introduces an adaptive *Skip Gater* that selects the warm-start timestep for each request. Guided by our warm-start error analysis, the gater operates over log-SNR-spaced skip actions and uses a contextual bandit conditioned on the prompt, composed prior, denoising budget, and serving state. Together, ACG and the Skip Gater address the two sides of warm-start acceleration: constructing more reusable acoustic structure and determining how much of it can be preserved.

Contributions. Overall, we make three key contributions:

- To our knowledge, SoundWeaver is the first T2A inference framework to accelerate generation through compositional cross-example acoustic reuse. We formalize why a better acoustic prior permits a less-noised warm start and hence greater denoising savings (§3.1).
- We introduce the *Acoustic Composition Graph (ACG)*, which composes prompt-relevant segments across caches while accounting for acoustic transition compatibility and target duration. A structured dynamic program efficiently identifies high-utility compositional priors (§3.2).
- We develop an adaptive *Skip Gater* that selects the amount of reuse for each request using prompt–prior context and serving conditions, enabling context-dependent decisions (§3.3).

Our evaluations across AudioLDM2 and Stable Audio Open show that SoundWeaver achieves 1.48–$3.74\times$ generation speedup with only $\sim$1K cached audios while preserving or improving perceptual quality, without retraining the model (§4).

2 Related Work

Efficient audio diffusion. Faster samplers reduce sampling steps without modifying the model, via deterministic sampling (Song et al., 2021) or higher-order ODE solvers (Lu et al., 2025). Distillation trains a student to match the teacher in fewer steps, by progressively halving the step count (Salimans & Ho, 2022) or learning a consistency mapping to the trajectory endpoint (Song et al., 2023). Caching methods reuse intermediate features or model outputs across adjacent timesteps to skip computation (Ma et al., 2024; Liu et al., 2025a). Audio-specific methods include consistency distillation for one- or few-step T2A generation (Bai et al., 2024; Liu et al., 2024b; Saito et al., 2025), step and layer distillation for music generation (Novack et al., 2025b), adversarial post-training (Novack et al., 2025a), and rectified- and mean-flow T2A models (Liu et al., 2025b; Li et al., 2025). However, these methods largely adapt designs originally developed for image generation, rather than exploiting properties specific to audio. SoundWeaver is instead an audio-centric system, built around the temporally localized and recurring structure of sound.

RAG for diffusion models. RAG has been explored in T2A primarily to improve generation quality and generalization. Re-AudioLDM (Yuan et al., 2024) retrieves relevant text–audio pairs to

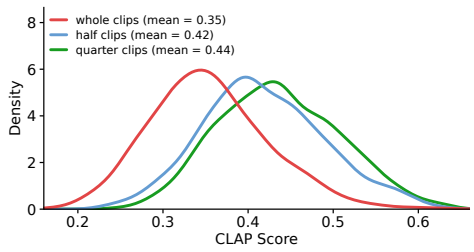


Figure 1: Whole and segmented audio often have many semantically similar neighbors.

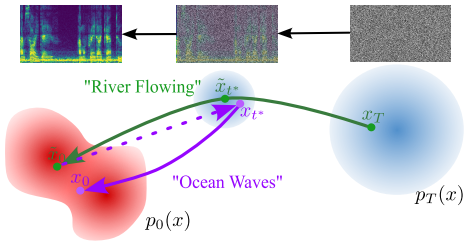


Figure 2: An intuitive geometric illustration of SoundWeaver's warm-starting.

condition generation of rare and unseen sound events, while Audiobox TTA RAG (Yang et al., 2025) conditions on retrieved audio for zero- and few-shot generation. Recent work, including NIRVANA (Agarwal et al., 2024) and MoDM (Xia et al., 2025), shows that generation requests can be accelerated by warm-starting from a similar previous request, motivated by core image editing algorithms Meng et al. (2022). These methods, however, reuse a single retrieved sample for image generation. SoundWeaver instead extends to multi-sample compositional reuse for T2A generation.

3 SoundWeaver Design

3.1 Motivation and Overview

Cross-example reuse is promising only if prior audio contains structure relevant to new requests. In practice, such similarity is prevalent: as shown in Fig. 1, many prompts have semantically aligned examples even in a modest (1K) audio cache. However, the useful structure is often *temporally localized*: audios contain multiple events whose relevance varies over time, so whole-clip retrieval can hide highly relevant intervals. Indeed, Fig. 1 shows that progressively partitioning audios from full clips into half- and quarter-length segments exposes substantially stronger prompt–audio alignment. Furthermore, composing new audio from highly relevant segments of existing audios to form new audio naturally has a long history in traditional signal processing (Hunt & Black, 1996; Schwarz, 2007). However, the concatenated sequence is itself the final output, so poorly matched joins remain audible. This makes diffusion a natural complement as SoundWeaver uses audio composition only as a prior from which the diffusion model regenerates the final output.

Theoretical Intuition for Warm-Start Acceleration. T2A diffusion models learn a reverse process $p_\theta(x_{t-1} \mid x_t, c)$ that transforms $x_T \sim \mathcal{N}(0, I)$ into a clean audio latent x_0 over T denoising steps, conditioned on prompt c. The standard forward process (Song et al., 2021) is $q(x_t \mid x_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\, x_0,\ (1-\bar{\alpha}_t)I\big)$, where $\bar{\alpha}_t$ decreases with t. Applying it to data $x_0 \sim p(x_0 \mid c)$ gives the noised marginal $p_t(x_t \mid c)$. Warm-starting encodes a reference $\tilde{x}_0$ into the model's latent space, samples $x_{t^*} \sim q(x_{t^*} \mid \tilde{x}_0)$, and runs only the reverse steps from t^* to 0, skipping $T - t^*$ steps. Let $\Delta(\tilde{x}_0) = \mathbb{E}_{x_0 \sim p(x_0|c)} \|\tilde{x}_0 - x_0\|_2^2$ denote the expected distance between the reference and the target distribution.

Proposition 1. *If the reverse process maps $p_t(x_t \mid c)$ to $p(x_0 \mid c)$, its output distribution $\hat{p}(x_0 \mid c)$ when initialized from $q(x_t \mid \tilde{x}_0)$ satisfies*

$$D_{\mathrm{KL}}\big(\hat{p}(x_0 \mid c) \,\|\, p(x_0 \mid c)\big) \le D_{\mathrm{KL}}\big(q(x_t \mid \tilde{x}_0) \,\|\, p_t(x_t \mid c)\big) \le \frac{\bar{\alpha}_t}{2(1-\bar{\alpha}_t)}\,\Delta(\tilde{x}_0). \tag{1}$$

The first inequality follows from the data processing inequality and the second from the convexity of KL divergence (proof in Appendix B). Proposition 1 yields three implications. First, the bound decays with the signal-to-noise ratio $\mathrm{SNR}_t = \bar{\alpha}_t/(1-\bar{\alpha}_t)$: noising more deeply tolerates a worse reference, but skips fewer steps. Second, for an error tolerance ϵ, the shallowest admissible warm-start satisfies $\mathrm{SNR}_{t^*} \le 2\epsilon/\Delta(\tilde{x}_0)$, so a better reference permits skipping more steps, motivating improving the prior itself (§3.2). Third, $\Delta(\tilde{x}_0)$ is unobservable, since it depends on the unknown $p(x_0 \mid c)$; SoundWeaver therefore treats prompt-alignment scores as proxies for reference quality and learns the mapping from these signals to t^* (§3.3).

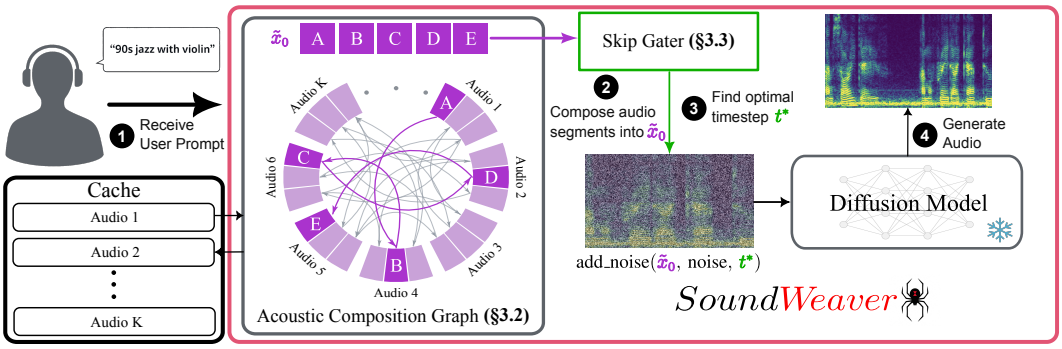


Figure 3: SoundWeaver design overview and request execution flow.

Design Challenges. Guided by this observation, we aim to improve the acoustic prior *before* deciding how aggressively to warm-start from it (Fig. 3). However, this formulation raises two key questions: (i) how do we construct a useful acoustic prior $\hat{x}_0$, especially when no single previous cache matches the request in duration or content, and *(ii)* how much of that prior can be preserved for a given request? SoundWeaver addresses these questions with an *Acoustic Composition Graph* (ACG) that composes reusable acoustic structure across cached audios (§3.2), and an adaptive *Skip Gater* that selects the warm-start timestep t^* (§3.3).

3.2 Acoustic Composition Graph

To warm-start T2A generation, SoundWeaver must construct a semantically relevant acoustic prior $\hat{x}_0$ that reflects the target prompt over the requested duration. Retrieving the single most semantically similar cache assumes the same audio provides useful acoustic structure throughout the generation. In practice, however, the usefulness of a cached audio can vary substantially over time: an audio may contain short intervals that are highly relevant to the target prompt alongside irrelevant content, and its duration may not match the requested duration. Instead, different parts of a prompt may be better represented by different temporal regions of different caches. Thus, useful acoustic structure need not align with the boundaries of any single cached example.

Hence, SoundWeaver constructs $\hat{x}_0$ compositionally from segments across cached audios. Though, this introduces a second requirement: these segments originating from different audios must be joined together without arbitrary acoustic discontinuities. We capture both requirements with an *Acoustic Composition Graph* (ACG), whose nodes correspond to cached audio segments and whose edges represent possible transitions at a desired level of acoustic similarity. Given a prompt and target duration, constructing the prior becomes a structured path-search problem over this graph where the path length is constrained by the requested duration.

Compositional Acoustic Graph Construction. We divide each cached audio into segments of L consecutive audio frames, which form the nodes of the ACG. To determine which segments can be connected, we need a representation that captures acoustic similarity at their boundaries. Direct waveform comparison is poorly suited for identifying alternative continuations since perceptually similar sounds can differ substantially in sample-level phase, noise, and amplitude. Semantic embeddings are also insufficient: they capture how two sounds convey similar content, but not whether they occupy similar local acoustic states. We, therefore, take a cheap tokenizer approach and represent a frame, f, at the start of each segment using EnCodec (Défossez et al., 2022). EnCodec encodes f using residual vector quantization (RVQ) with K codebooks: the first codebook quantizes the frame's latent representation, while each subsequent codebook quantizes the residual left by the preceding ones. A frame is, thus, represented by a token sequence $(c_1, \ldots, c_K)$ that progressively refines its acoustic representation. We denote by $\mathrm{pre}_d(f) = (c_1, \ldots, c_d)$ the prefix formed by the first d codebooks, which provides a natural hierarchy of acoustic similarity from coarse to fine.

From a segment u, a path can continue to the segment that follows u in its source audio, which we denote u^+. It can also switch to another segment v whose first frame has the same d-token prefix as the first frame of u^+. Specifically, writing f_s for the first frame of segment s, we admit an edge

$$u \to v \iff \mathrm{pre}_d(f_v) = \mathrm{pre}_d(f_{u^+}). \tag{2}$$

Matching these boundary-frame prefixes makes v a candidate which is an acoustically compatible continuation of u. Smaller d admits more cross-audio transitions with coarser acoustic matching, while larger d admits fewer transitions with stricter acoustic similarity. To identify these matches efficiently without comparing every pair of segments, SoundWeaver stores the prefix of each segment's first frame in a prefix tree.

Relevance- and Continuity-Aware Path Search. The graph specifies which segments can be connected. We next determine which path best matches the new prompt. We propose an approach inspired by traditional audio processing (Hunt & Black, 1996; Schwarz, 2007) that generates utterances by searching audio segments that minimize a target cost (match to the desired output) plus a concatenation cost (penalizing perceptible boundary jumps) to produce highly natural audio.

We employ FLAM (Wu et al., 2025), a dual text–audio semantic encoder that specializes in scoring short sliding windows against the prompt text. Let $\rho(v)$ denote the aggregate prompt-alignment score of segment v. To prevent audible boundary artifacts, SoundWeaver precomputes an acoustic distance penalty $\omega(u, v)$ for each candidate edge $u \to v$ offline, using the codec's continuous latent embedding z_f:

$$\omega(u, v) = -\frac{1}{\tau}\left\| z_{\text{last}(u)} - z_{\text{last}(\text{pred}(v))} \right\|^2, \tag{3}$$

where $\text{pred}(v)$ denotes the segment immediately preceding v in its source audio, and τ is the median squared distance between adjacent frames. The score does not prohibit cross-audio transitions; rather, it downweights substitutions for which the acoustic context preceding v differs substantially from the end of u. Because audio embeddings are materialized offline and cross-audio penalties can be precomputed offline, online scoring requires only a single prompt text embedding and one matrix-vector dot product across candidate segments.

For a new generation request with target duration T, SoundWeaver searches for a path of $n = \lceil rT/L \rceil$ segments, where r is the codec frame rate, selecting the length-n path that jointly maximizes semantic relevance and acoustic continuity:

$$\max_{v_1 \to \cdots \to v_n} (1-\alpha) \sum_{m=1}^{n} \rho(v_m) + \alpha \sum_{m=2}^{n} \omega(v_{m-1}, v_m), \tag{4}$$

where $\alpha \in [0, 1]$ balances semantic relevance against smoothness. Since equation 4 decomposes into segment-local rewards and first-order transition scores, its global optimum can be found with dynamic programming (DP):

$$R_m(v) = (1-\alpha)\rho(v) + \max_{u \to v}\big[R_{m-1}(u) + \alpha\,\omega(u, v)\big], \qquad R_1(v) = (1-\alpha)\rho(v), \tag{5}$$

where $R_m(v)$ records the maximum cumulative score of an m-segment path ending at v. Our DP solver scales linearly with the requested path length n and the number of admissible graph edges, requiring approximately 2 ms per request in our evaluations (§4.3).

Global Reranking and Prior Assembly. The dynamic program optimizes an additive objective over local segment relevance and pairwise transitions, but this objective does not directly measure the semantic correspondence of the assembled clip as a whole. To address this, we retain the top M paths from the structured search ($M = 16$ by default), assemble each candidate from its source waveform segments, and rerank the resulting clips using whole-clip CLAP similarity. This provides a global semantic signal while evaluating only a small set of high-scoring candidates.

The top-ranked assembled candidate is the acoustic prior $\hat{x}_0$ passed to the Skip Gater (§3.3). Importantly, codec tokens are used for indexing and structured search only. The prior is assembled from the original cached waveforms, avoiding storage overhead and codec reconstruction loss.

3.3 Adaptive Warm-Start Gating

Even for a semantically relevant composed prior, the appropriate amount of reuse varies across requests. Starting from a less-noised prior preserves more acoustic structure and enables skipping more denoising steps, but leaves less opportunity for the model to refine content that deviates from the target prompt. Consequently, the appropriate skip depth depends on the prompt, composed prior, and model, and a fixed similarity threshold cannot capture this variation. Moreover, training

a supervised predictor is often impractical since generation quality lacks reliable golden labels and request distributions evolve over time. We introduce a *Skip Gater* that adaptively selects how many denoising steps to skip for each request.

Contextual skip selection. We formulate warm-start selection as a contextual multi-armed bandit. Each arm $k \in \mathcal{A}$ corresponds to a candidate skip depth, with arms spaced uniformly in log-SNR. This choice follows Proposition 1: warm-start mismatch varies with SNR rather than raw diffusion timestep, so log-SNR spacing provides approximately uniform multiplicative resolution over the mismatch term. Given context x, containing embeddings of the prompt and composed prior together with the denoising budget and serving state, the gater selects the number of denoising steps to skip. We define the reward as:

$$r(x, k) = \Delta E(x, k) + \beta \widetilde{Q}(x, k), \tag{6}$$

where $\Delta E(x, k) \in [0, 1]$ measures the serving benefit of skipping, such as reduction in latency-SLO violation, and $\widetilde{Q}(x, k)$ measures generation quality (e.g., CLAP score). The coefficient β controls the quality–efficiency trade-off (e.g., the developer's preference). Because the serving benefit depends on the current load, the same prior can receive a conservative skip under light load and a more aggressive one when latency reduction is valuable.

Adjacent skip actions are strongly related, so we should avoid learning an independent model for each arm. To this end, we parameterize the expected reward with a Bayesian linear model over the tensor product of a context representation $\phi(x)$ and a continuous arm basis $\psi(k)$:

$$\widehat{R}_w(x, k) = \langle \phi(x) \otimes \psi(k), w \rangle, \qquad w \sim \mathcal{N}(m, \Sigma). \tag{7}$$

The arm basis is defined over log-SNR, allowing observations from one skip level to inform nearby actions. We use Thompson sampling from the Gaussian posterior during exploration and the posterior mean for deterministic user-facing inference.

Learning from noisy and heterogeneous feedback. Absolute quality scores $Q(x, k)$ are difficult to calibrate across prompts: users rarely provide reliable scalar ratings, and automatic proxies such as CLAP provide only an imperfect proxy for perceptual quality. Fortunately, production systems like ChatGPT already collect preference-style feedback, such as thumbs up/down ratings or pairwise comparisons (Yu et al., 2025). Thus we convert the observed quality of action into within-prompt percentile ranks $\widetilde{Q}(x, k) \in [0, 1]$. But still, skip sensitivity varies substantially across prompts. Our analysis of 1k randomly sampled AudioCaps prompts shows that CLAP variance spans nearly two orders of magnitude: simple prompts (e.g., *"Water runs continuously"*) retain quality across wide skip ranges, whereas semantically rich prompts (e.g., *"Dog barking while a cat is meowing..."*) exhibit rapid degradation. To focus learning on decisions that matter, we weight parameter updates by the prompt's quality variance:

$$w_p = \frac{\sigma_p^2}{\bar{\sigma}^2 + \epsilon}, \tag{8}$$

where σ_p^2 is the quality variance across observed skip actions for prompt p, and $\bar{\sigma}^2$ is the mean variance across all prompts. This emphasizes requests for which skip depth materially affects generation quality. The posterior is initialized from offline action sweeps and updated asynchronously from logged outcomes, so learning does not add latency to the generation path. Specifically, when a generation receives negative feedback, we log the prompt and chosen skip depth, then asynchronously regenerate the prompt at every arm off the serving path. This yields a full within-prompt quality profile from which we compute $\widetilde{Q}$ and σ_p^2.

4 EVALUATION

4.1 EXPERIMENTAL SETUP

We evaluate T2A diffusion backbones across architectures and samplers: AudioLDM2 (1.1B) (Liu et al., 2024a) with DDIM (Song et al., 2021), and Stable Audio Open (1.3B) (Evans et al., 2025) with DPM-Solver++ (Lu et al., 2025). We use 50 sampling steps by default, and perform studies on different sampling steps (§4.3). All experiments were done on an NVIDIA A100 GPU. Following prior work (Liu et al., 2023; 2024a), we evaluate on the AudioCaps test set, randomly selecting one caption per clip as the generation prompt.

Table 1: Objective and subjective generation-quality metrics and inference latency across different audio-generation backbones, baselines, and ablations. ^{S} synthetic-audio cache; ^{R} real-audio cache.

Backbone	Method	Objective				Subjective			Latency	
		FD ↓	KL ↓	CLAP ↑	CCDS ↓	PS ↑	OVL ↑	REL ↑	Value (s) ↓	Speedup ↑
AudioLDM 2	Vanilla	30.00	1.56	0.41	0.72	–	–	–	3.95	–
	DeepCache	39.68	1.73	0.38	0.78	0.30	–	–	1.68	2.3×
	MoDM	25.54	1.78	0.39	0.78	0.37	–	–	2.81	1.4×
	SoundWeaver ^{S}	23.81	1.41	0.44	0.75	0.41	0.43	0.61	2.12	1.9×
	SoundWeaver ^{R}	20.53	1.50	0.44	0.72	0.40	0.68	0.70	2.67	1.5×
	w/o ACG	22.63	1.68	0.41	0.82	0.37	–	–	2.62	1.5×
	w/o SG	24.08	1.45	0.42	0.79	0.36	–	–	2.63	1.5×
Stable Audio Open	Vanilla	35.04	1.98	0.29	0.65	–	–	–	14.66	–
	MoDM	36.99	1.88	0.29	0.67	0.37	–	–	8.41	1.7×
	SoundWeaver ^{S}	27.88	1.72	0.36	0.69	0.50	0.30	0.43	6.10	2.4×
	SoundWeaver ^{R}	27.48	1.72	0.34	0.64	0.42	0.70	0.59	3.92	3.7×
	w/o ACG	29.30	1.83	0.33	0.81	0.38	–	–	4.39	3.3×
	w/o SG	27.96	1.65	0.33	0.72	0.43	–	–	4.69	3.1×

We consider two cache sources: (i) *real-audio cache* taken from open-access Clotho v2 audios Drossos et al. (2019), and (ii) the *synthetic cache* containing audio generated by the evaluated model from the corresponding Clotho captions. Thus, the two audio sources differ only in quality. All ablation studies and analysis use the *real-audio cache* variant unless stated otherwise. Additional implementation and setup details are provided in Appendix A.

Baselines. We compare SoundWeaver against the following advances:

- *Vanilla*: The original T2A model generates from Gaussian noise using its standard sampler.
- *DeepCache* (Ma et al., 2024): A training-free method that accelerates U-Net diffusion models by reusing cached high-level feature maps across adjacent denoising steps.
- *MoDM* (Xia et al., 2025): A mixture-of-diffusion method warm-starting a small model from the cached outputs of a large model.

Metrics. We evaluate generation quality using human judgments, LLMs, and objective metrics:

- *Human evaluation*: Winrate of SoundWeaver variants over Vanilla in randomized pairwise comparisons for *Overall Quality (OVL)* and prompt *Relevance (REL)* (e.g. 0.7 means SoundWeaver wins 70% of the time).
- *LLM-based evaluation*: For scalable evaluation, we use Gemini-3.7-Flash as a pairwise audio judge. Following existing advances (Wang et al., 2025; Zheng et al., 2023), we report a *Preference Score (PS)*, the winrate as a fraction when comparing outputs on identical prompts.
- *Objective metrics*: *CLAP* (Elizalde et al., 2023) measures text–audio semantic alignment; *KL divergence (KL)* (Kullback & Leibler, 1951) compares acoustic-event posteriors; *Fréchet Distance (FD)* (Fréchet, 1906), measures distributional similarity to reference audio; and *CLAP Conditional Diversity Score (CCDS)* Novack et al. (2025a) measures pairwise similarity of the outputs generated using two different seeds.

4.2 Main Results

SoundWeaver improves latency and quality. As shown in Table 1, SoundWeaver achieves 1.48–3.74× latency speedups on top of each model's native sampler while consistently improving quality. The real-audio cache yields the largest quality gains, which we attribute to the richer perceptual detail in real audios, and comparable CCDS indicates that output diversity is preserved. These gains hold under online serving (Fig. 6): on a one-hour deployment driven by request traces from DiffusionDB (Wang et al., 2022) with audio prompts and durations used from WavCaps Mei et al. (2024), and arrival rates scaled to our hardware, warm-starting reduces per-request NFEs and yields 1.06–11.82× end-to-end speedups.

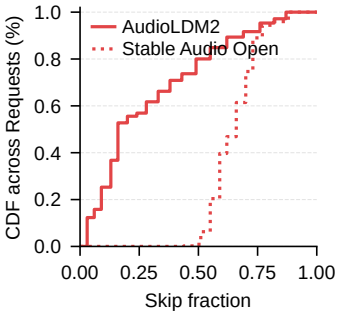


Figure 4: SoundWeaver adaptively skips denoising steps.

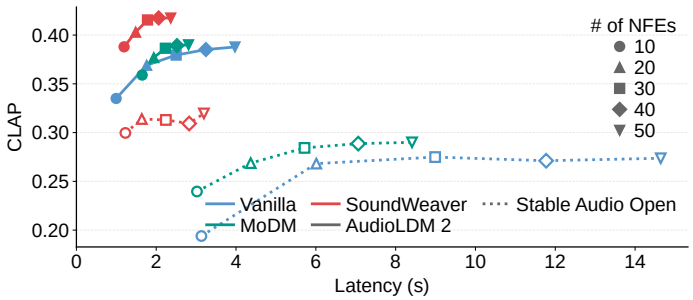
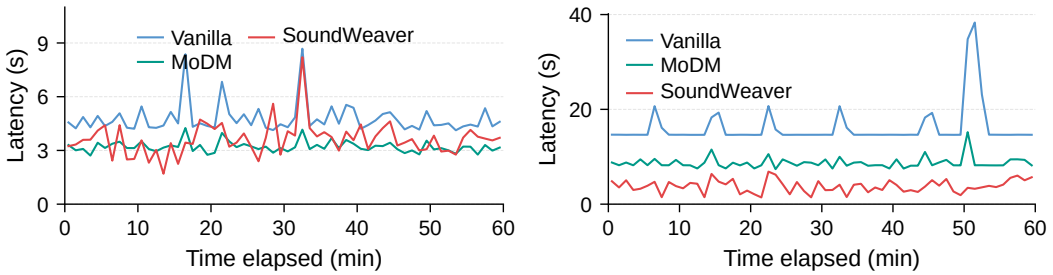


Figure 5: SoundWeaver unlocks new quality-efficiency frontiers across sampling budgets (NFEs) in a training-free manner.



(a) AudioLDM 2 (b) Stable Audio Open

Figure 6: SoundWeaver improves online serving latency in our one-hour deployments.

SoundWeaver adapts skip depth across requests. Figure 4 shows substantial variation in the selected skip ratios across prompts. This wide distribution confirms that the Skip Gater does not collapse to a fixed policy, but adjusts the amount of reuse to each request; the corresponding quality results in Table 1 show that these latency savings are achieved without sacrificing generation quality. Stable Audio Open exhibits higher skip ratios overall, largely because its sampler allocates more steps to high noise regions. So, a warm start at the same SNR removes a larger fraction of the trajectory than under AudioLDM2's DDIM schedule, consistent with Proposition 1.

SoundWeaver introduces negligible overhead. Prior construction and warm-start selection add only $\sim$0.1s per request: ACG search takes a median of 68–79ms, while the Skip Gater requires 27–29ms. Together, this accounts for less than 3% of vanilla generation time on AudioLDM2 and less than 1% on Stable Audio Open, while warm-starting reduces overall latency by 32–73% (Table 1).

4.3 Ablation Studies

Performance Breakdown. We dive into the contributions of the Acoustic Composition Graph (ACG) and adaptive Skip Gater (SG) in Table 1. Removing ACG replaces compositional trajectory search with whole-clip nearest-neighbor retrieval using CLAP score. This consistently degrades generation quality; for example, FD increases from 20.53 to 22.63 on AudioLDM2 and from 27.48 to 29.30 on Stable Audio Open. These results indicate that composing prompt-relevant segments across audios produces stronger warm-start priors than relying on a single retrieved clip. Removing SG instead replaces the contextual bandit with a fixed similarity-threshold policy following prior warm-start approaches (Agarwal et al., 2024; Xia et al., 2025). At comparable serving latency, the adaptive gater yields consistently better quality, showing the benefit of request-dependent skip selection over a static policy.

Impact of Sampling Budget (NFEs). Figure 5 evaluates SoundWeaver across sampling budgets from 10 to 50 steps. SoundWeaver remains effective even at low budgets, where relatively few denoising steps are available to skip. The quality benefit is most pronounced in this regime, suggesting that a structured acoustic prior is particularly valuable when the sampler has limited opportunity to construct the target audio from noise. As the sampling budget increases, the absolute opportunity for skipping also grows, yielding larger latency savings.

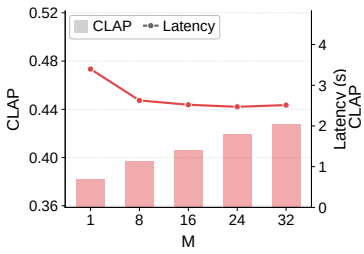
Cache Size (Memory)	CLAP ↑	Latency (s) ↓
100 (0.04 GB)	0.38	3.47
500 (0.20 GB)	0.39	3.05
1K (0.39 GB)	0.42	2.55
5K (1.97 GB)	0.43	2.04

(a) AudioLDM 2

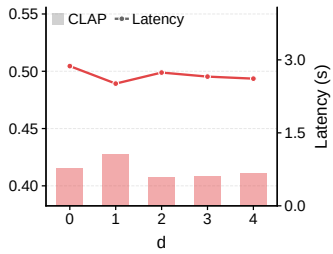
Cache Size (Memory)	CLAP ↑	Latency (s) ↓
100 (0.04 GB)	0.29	4.90
500 (0.20 GB)	0.29	3.96
1K (0.39 GB)	0.30	3.49
5K (1.97 GB)	0.36	3.19

(b) Stable Audio Open

Table 2: Impact of cache pool size on SoundWeaver (WavCaps dataset (Mei et al., 2024)).



(a) Impact of beam size.



(b) Impact of prefix depth.

Figure 7: Sensitivity analysis (AudioLDM 2).

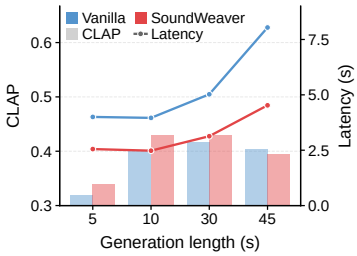


Figure 8: Impact of generation length (AudioLDM 2).

Impact of Cache Pool Size. Table 2 varies the cache from 100 to 5K audio clips. Increasing cache size clearly improves both CLAP and latency. This indicates that larger caches provide more useful acoustic structure for composition, though substantial benefits already emerge at comparatively modest cache sizes. Meanwhile, SoundWeaver does not require an exhaustive audio database to construct rich composite priors, ensuring low overhead.

Impact of Candidate Size M. Figure 7(a) varies $M \in \{1, 8, 16, 24, 32\}$, the number of highest-scoring trajectories retained by graph search for global CLAP reranking. Increasing M improves generation quality by giving the reranker a broader set of compositions from which to select, with gains gradually saturating around $M = 32$. We also observe lower end-to-end latency with larger M, as the resulting higher-quality priors allow the Skip Gater to select more aggressive warm starts.

Impact of Prefix Matching Depth d. Figure 7(b) evaluates the EnCodec prefix depth $d \in \{0, 1, 2, 3\}$, used to establish transition edges between segment boundaries (§3.2). Setting $d = 1$ where at least one token has to be matched achieves the best overall generation quality. Stricter prefix constraints ($d \geq 1$) overly restrict candidate branching, yielding a sparser graph that limits its compositional reuse without improving transition naturalness. Whilst at $d = 0$ creating edges without an acoustic filter is also harmful. End-to-end latency remains essentially flat across different values of d, as the dynamic programming search overhead is only $\sim$2 ms and is completely dominated by downstream diffusion.

Impact of Generation Length. Figure 8 shows the impact of target audio duration across 5, 10, 30, and 45 seconds. On AudioLDM 2, SoundWeaver maintains consistent CLAP improvements of $\sim 0.01 \pm 0.02$ over vanilla generation across varied horizons, while providing substantial latency savings which increase with duration. For Stable Audio Open, latency and quality remains largely constant across durations due to its architecture generating fixed 45-second contexts natively before temporal truncation.

5 CONCLUSION

This paper introduces SoundWeaver, a training-free framework for accelerating T2A diffusion through cross-example acoustic structure reuse. For a new request, SoundWeaver constructs a prompt-conditioned acoustic prior from reusable segments of previously observed audio and warm-starts diffusion from an intermediate noise level. Its acoustic graph enables compositional reuse across audios while accounting for semantic relevance and transition compatibility, and an adaptive Skip Gater determines how aggressively the resulting prior can be reused for each request.

AI USE STATEMENT

In this work, we used Gemini Flash 3.7 to get the preference score (PS) documented in Table 1. Claude and OpenAI models were additionally used to assist in proposing and refining hypotheses; coding, designing, and critiquing methodology and experiments; formulating mathematical claims; and helping write the paper and proofs. The authors verified all AI-assisted work. We take responsibility for the final paper.

The prompt used to get the preference scores is as follows:

Gemini Audio Evaluation Prompt

You are an expert audio quality evaluator. You will be given two audio clips generated from the same text prompt and need to choose the audio file that sounds better. Your evaluation should consider **consistency**, **intelligibility**, **naturalness**, **absence of artifacts**, **timbral richness**, **fidelity**, and overall **quality**. Avoid any position biases and ensure that the order in which the audio were presented does not influence your decision. Do not allow the duration of the audio to influence your evaluation. Be as objective as possible.

You should start with your evaluation by comparing the two audios and provide a short rationale about which one you think is better. After providing your rationale, you should output the final verdict by strictly following this seven-point Likert scale:

Compare the two audio clips and score the difference on a scale from -3 to 3:

Score	Meaning
$+3$	Audio 1 is much better than Audio 2
$+2$	Audio 1 is noticeably better
$+1$	Audio 1 is slightly better
0	Both are equal / indistinguishable
-1	Audio 2 is slightly better
-2	Audio 2 is noticeably better
-3	Audio 2 is much better than Audio 1

Respond strictly in this format:
`[Rationale]:` *short rationale, less than 200 words*
`[Score]:` *integer from* -3 *to* $+3$

ETHICS STATEMENT

SoundWeaver warm-starts generation from previously observed audio, so traces of cached audios may persist in outputs. We mitigate this by fundamentally using fine acoustic units through our ACG, conservatively tuning our Skip Gater to avoid excessive skips, in addition to reporting output diversity (CCDS) throughout. Our real-audio cache uses only Creative Commons–licensed Clotho v2 audios, and SoundWeaver remains highly effective with a fully synthetic cache generated by the model itself (Table 1) as well. We encourage researchers and developers interested in SoundWeaver to only use audio licensed for reuse. Listening-study participants were anonymous volunteers.

REPRODUCIBILITY STATEMENT

SoundWeaver is training-free and built entirely on publicly available pretrained models: AudioLDM 2 and Stable Audio Open as diffusion backbones, the 24 kHz EnCodec model for acoustic tokenization, and FLAM and CLAP for prompt–audio scoring. Section 3 specifies each component of the method, including the edge-admission rule for the Acoustic Composition Graph (Eq. 2), the transition penalty (Eq. 3), the path-search objective and its dynamic-programming solution (Eqs. 4–5), whole-clip reranking, and the Skip Gater's reward, Bayesian parameterization, and variance-weighted updates (Eqs. 6–8). Appendix A lists all implementation details and hyperparameters, including samplers, classifier-free guidance scales, segment length L, prefix depth d, trade-off weight α, candidate count M, and how the MoDM and baselines were adapted to audio. Appendix B states the assumptions behind Proposition 1 and gives its complete proof. All datasets are public: we eval-

uate on the AudioCaps test set, build real and synthetic caches from Clotho v2, use WavCaps for the cache-size study and serving prompts, and derive request arrival traces from DiffusionDB; Section 4 describes prompt sampling, cache construction, and all evaluation metrics. The full prompt used for the LLM-based preference judge is included in the AI Use Statement.

REFERENCES

Shubham Agarwal, Subrata Mitra, Sarthak Chakraborty, Srikrishna Karanam, Koyel Mukherjee, and Shiv Kumar Saini. Approximate caching for efficiently serving Text-to-Image diffusion models. In *21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 24)*, pp. 1173–1189, Santa Clara, CA, April 2024. USENIX Association. ISBN 978-1-939133-39-7. URL https://www.usenix.org/conference/nsdi24/presentation/agarwal-shubham.

Yatong Bai, Trung Dang, Dung Tran, Kazuhito Koishida, and Somayeh Sojoudi. ConsistencyTTA: Accelerating diffusion-based text-to-audio generation with consistency distillation. In *Interspeech 2024*, pp. 3285–3289, 2024. doi: 10.21437/Interspeech.2024-1333.

Yuzhu Chen, Fengxiang He, Shi Fu, Xinmei Tian, and Dacheng Tao. Adaptive time-stepping schedules for diffusion models. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset, 2019. URL https://arxiv.org/abs/1910.09387.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations, 2023. URL https://arxiv.org/abs/2309.05767.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pp. 1–5, 2025.

Jiarui Fang, Jinzhe Pan, Xibo Sun, Aoyu Li, and Jiannan Wang. xdit: an inference engine for diffusion transformers (dits) with massive parallelism, 2024. URL https://arxiv.org/abs/2411.01738.

Maurice Fréchet. Sur quelques points du calcul fonctionnel. *Rendiconti del Circolo Matematico di Palermo*, 22:1–72, 1906.

Andrew J. Hunt and Alan W. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pp. 373–376, 1996. doi: 10.1109/ICASSP.1996.541110.

S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. ISSN 00034851. URL http://www.jstor.org/stable/2236703.

Xiquan Li, Junxi Liu, Yuzhe Liang, Zhikang Niu, Wenxi Chen, and Xie Chen. Meanaudio: Fast and faithful text-to-audio generation with mean flows, 2025. URL https://arxiv.org/abs/2508.06098.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7353–7363. IEEE, 2025a.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. *Proceedings of the International Conference on Machine Learning*, pp. 21450–21474, 2023.

Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2871–2883, 2024a. doi: 10.1109/TASLP.2024.3399607.

Huadai Liu, Rongjie Huang, Yang Liu, Hengyuan Cao, Jialei Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. Audiolcm: Efficient and high-quality text-to-audio generation with minimal inference steps. MM '24, pp. 7008–7017, New York, NY, USA, 2024b. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681072. URL https://doi.org/10.1145/3664647.3681072.

Huadai Liu, Jialei Wang, Rongjie Huang, Yang Liu, Heng Lu, Zhou Zhao, and Wei Xue. FlashAudio: Rectified flow for fast and high-fidelity text-to-audio generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13694–13710. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.acl-long.673.

Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations (ICLR)*, 2022.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan LI, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 5775–5787. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/260a14acce2a89dad36adc8eefe7c59e-Paper-Conference.pdf.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022.

Taehong Moon, Moonseok Choi, Eunggu Yun, Jongmin Yoon, Gayoung Lee, Jaewoong Cho, and Juho Lee. A simple early exiting framework for accelerated sampling in diffusion models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36212–36228. PMLR, 2024. URL https://proceedings.mlr.press/v235/moon24a.html.

Zachary Novack, Zach Evans, Zack Zukowski, Josiah Taylor, CJ Carr, Julian Parker, Adnan Al-Sinan, Gian Marco Iodice, Julian McAuley, Taylor Berg-Kirkpatrick, and Jordi Pons. Fast text-to-audio generation with adversarial post-training. In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 1–5, 2025a. doi: 10.1109/WASPAA66052.2025.11230941.

Zachary Novack, Ge Zhu, Jonah Casebeer, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. Presto! distilling steps and layers for accelerating music generation. In *International Conference on Learning Representations (ICLR)*, 2025b.

Yaqi Qiao, Ping He, Songrun Xie, Ayush Barik, Chensong Zhang, Zhengzhong Tu, and Fan Lai. Flashdiff: Efficient regional execution and scheduling for diffusion model serving, 2026. URL https://arxiv.org/abs/2607.12121.

Koichi Saito, Dongjun Kim, Takashi Shibuya, Chieh-Hsin Lai, Zhi Zhong, Yuhta Takida, and Yuki Mitsufuji. SoundCTM: Unifying score-based and consistency models for full-band text-to-sound generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations (ICLR)*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Diemo Schwarz. Corpus-based concatenative synthesis. *IEEE Signal Processing Magazine*, 24: 92–104, 2007. URL https://api.semanticscholar.org/CorpusID:14214995.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=St1giarCHLP.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32211–32252. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23a.html.

Hui Wang, Jinghua Zhao, Yifan Yang, Shujie Liu, Junyang Chen, Yanzhe Zhang, Shiwan Zhao, Jinyu Li, Jiaming Zhou, Haoqin Sun, Yan Lu, and Yong Qin. Speechllm-as-judges: Towards general and interpretable speech quality evaluation, 2025. URL https://arxiv.org/abs/2510.14664.

Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv:2210.14896 [cs]*, 2022. URL https://arxiv.org/abs/2210.14896.

Yusong Wu, Christos Tsirigotis, Ke Chen, Cheng-Zhi Anna Huang, Aaron Courville, Oriol Nieto, Prem Seetharaman, and Justin Salamon. FLAM: Frame-wise language-audio modeling. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 67719–67740. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/wu25ab.html.

Yuchen Xia, Divyam Sharma, Yichao Yuan, Souvik Kundu, and Nishil Talati. Modm: Efficient serving for image generation via mixture-of-diffusion models. ASPLOS '26, pp. 163–182, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400721656. doi: 10.1145/3760250.3762220. URL https://doi.org/10.1145/3760250.3762220.

Mu Yang, Bowen Shi, Matthew Le, Wei-Ning Hsu, and Andros Tjandra. Audiobox TTA-RAG: Improving Zero-Shot and Few-Shot Text-To-Audio with Retrieval-Augmented Generation. In *Interspeech 2025*, pp. 1243–1247, 2025. doi: 10.21437/Interspeech.2025-1032.

Yifan Yu, Yu Gan, Nikhil Sarda, Lillian Tsai, Jiaming Shen, Yanqi Zhou, Arvind Krishnamurthy, Fan Lai, Hank Levy, and David Culler. Ic-cache: Efficient large language model serving via in-context caching. In *Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles*, SOSP '25, pp. 375–398. ACM, October 2025. doi: 10.1145/3731569.3764829. URL http://dx.doi.org/10.1145/3731569.3764829.

Yi Yuan, Haohe Liu, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. Retrieval-augmented text-to-audio generation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 581–585. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447898. URL https://doi.org/10.1109/ICASSP48485.2024.10447898.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

A Implementation Details and Hyperparameters

Diffusion Model and Sampler Configurations. We configure SoundWeaver across two primary text-to-audio diffusion backbones: AudioLDM 2 (1.1B) (Liu et al., 2024a) and Stable Audio Open (1.3B) (Evans et al., 2025). Both backbones default to 50 sampling steps ($N_{\text{step}} = 50$). AudioLDM 2 uses the DDIM sampler (Song et al., 2021) with a classifier-free guidance (CFG) scale of 3.5, while Stable Audio Open employs DPM-Solver++ (Lu et al., 2025) with a CFG scale of 7.0.

Graph Construction and Path Search. To construct the Acoustic Composition Graph (ACG), each cached audio is encoded using the pre-trained 24 kHz, 12 kbps EnCodec model (Défossez et al., 2022) The tokenized sequence is split into non-overlapping segments of length $L = 150$ frames, each serving as an individual graph node. A directed edge is instantiated between segment u and v whenever the boundary frames share an identical EnCodec residual vector quantization prefix of depth $d = 1$.

During structured dynamic programming search, the trade-off hyperparameter balancing prompt semantic relevance against the continuous latent transition penalty is set to $\alpha = 0.25$. The top $M = 16$ candidate paths extracted by dynamic programming are fully assembled from raw source waveforms and reranked using whole-clip CLAP similarity to determine the final acoustic prior $\tilde{x}_0$.

Baseline Adaptations to Audio.

- **MoDM Adaptation:** To adapt the vision-based MoDM framework (Xia et al., 2025) to audio generation, we replace visual similarity backbones with CLAP text-audio embedding encoders (Elizalde et al., 2023). We calibrate a static, step-wise skip policy by sweeping AudioCaps training prompts across candidate skip depths and identifying the lowest skip fraction that matches vanilla generation fidelity for each similarity bucket.
- **DeepCache Adaptation:** The algorithm was directly implemented over our baselines and the default parameters worked well for AudioLDM2: selected skip branch at layer 0 and cache interval for updating the cache to be every 3 steps.

Subjective Evaluation

- *Preference Score*: we compare our generations with the vanilla generations of the models. To get the final win rate, we convert each positive score to be a win, and each negative score to be a loss, while ties are discarded and tally up the wins for each model.
- *Human Evaluation*: Twenty-two volunteers participated in the evaluation. Each survey included 12 randomly selected prompts from the AudioCaps test set. For each prompt, participants compared two audio samples: one generated by either SoundWeaver ^{S} or SoundWeaver ^{R} and one generated by the vanilla baseline. The two SoundWeaver variants were equally represented, and audio presentation order was randomized. Participants answered two questions per prompt: "Which audio has better perceptual quality (naturalness, clarity, and fewer audible artifacts)?" for overall quality (OVL), and "Which audio best matches the prompt?" for prompt relevance (REL). We report OVL and REL as the fraction of pairwise comparisons won by each model.

B Proof of Proposition 1

Fix a prompt c and a timestep t. Let K_t denote the Markov kernel induced by running the model's reverse sampler from time t to 0, so that initializing the sampler with a distribution μ over x_t yields the output distribution $K_t\mu$ over x_0. This covers both stochastic samplers and deterministic ODE samplers such as DDIM (Song et al., 2021), for which K_t is the push forward under a deterministic map. By assumption, the reverse process is exact on the noised data marginal:

$$K_t\, p_t(\cdot \mid c) = p(\cdot \mid c). \tag{9}$$

Warm-starting from reference $\tilde{x}_0$ initializes the sampler with $q(x_t \mid \tilde{x}_0)$, so the output distribution is $\hat{p}(\cdot \mid c) = K_t\, q(\cdot \mid \tilde{x}_0)$.

First inequality. By the data processing inequality, applying the same Markov kernel to two distributions cannot increase the KL divergence between them:

$$D_{\mathrm{KL}}\big(\hat{p}(x_0 \mid c) \,\|\, p(x_0 \mid c)\big) = D_{\mathrm{KL}}\big(K_t\, q(\cdot \mid \tilde{x}_0) \,\|\, K_t\, p_t(\cdot \mid c)\big) \le D_{\mathrm{KL}}\big(q(x_t \mid \tilde{x}_0) \,\|\, p_t(x_t \mid c)\big). \tag{10}$$

Second inequality. The noised marginal is a mixture of Gaussians,

$$p_t(x_t \mid c) = \int q(x_t \mid x_0)\, p(x_0 \mid c)\, dx_0. \tag{11}$$

Since KL divergence is convex in its second argument,

$$D_{\mathrm{KL}}\big(q(x_t \mid \tilde{x}_0) \,\|\, p_t(x_t \mid c)\big) \le \mathbb{E}_{x_0 \sim p(x_0 \mid c)}\Big[D_{\mathrm{KL}}\big(q(x_t \mid \tilde{x}_0) \,\|\, q(x_t \mid x_0)\big)\Big]. \tag{12}$$

Both distributions inside the expectation are Gaussians with shared covariance $(1-\bar{\alpha}_t)I$ and means $\sqrt{\bar{\alpha}_t}\,\tilde{x}_0$ and $\sqrt{\bar{\alpha}_t}\,x_0$. The KL divergence between Gaussians with equal covariance $\sigma^2 I$ and means m_1, m_2 is $\|m_1 - m_2\|_2^2/(2\sigma^2)$, so

$$D_{\mathrm{KL}}\big(q(x_t \mid \tilde{x}_0) \,\|\, q(x_t \mid x_0)\big) = \frac{\bar{\alpha}_t}{2(1-\bar{\alpha}_t)}\, \|\tilde{x}_0 - x_0\|_2^2. \tag{13}$$

Taking the expectation over $x_0 \sim p(x_0 \mid c)$ gives

$$D_{\mathrm{KL}}\big(q(x_t \mid \tilde{x}_0) \,\|\, p_t(x_t \mid c)\big) \le \frac{\bar{\alpha}_t}{2(1-\bar{\alpha}_t)}\, \Delta(\tilde{x}_0), \tag{14}$$

which completes the proof. □

Shallowest admissible warm-start. Requiring the bound to be at most a tolerance ϵ gives $\mathrm{SNR}_t\, \Delta(\tilde{x}_0)/2 \le \epsilon$, i.e., $\mathrm{SNR}_t \le 2\epsilon/\Delta(\tilde{x}_0)$. Since SNR_t decreases monotonically in t, the set of admissible timesteps is an interval $[t_{\min}, T]$, and the shallowest admissible warm-start t^* is its smallest element. A smaller $\Delta(\tilde{x}_0)$ raises the SNR threshold and therefore lowers t^*, skipping more steps.

Remark: the bound and output diversity. Expanding Δ around the conditional mean $\mu_c = \mathbb{E}[x_0 \mid c]$ gives

$$\Delta(\tilde{x}_0) = \|\tilde{x}_0 - \mu_c\|_2^2 + \mathbb{E}_{x_0 \sim p(x_0 \mid c)} \|x_0 - \mu_c\|_2^2. \tag{15}$$

The second term is the intrinsic variance of the target distribution and does not depend on the reference. Hence the bound remains positive even for an ideal reference: initializing from a single Gaussian $q(x_t \mid \tilde{x}_0)$ rather than the full mixture $p_t(x_t \mid c)$ concentrates outputs around the reference. This quantifies the diversity cost of warm-starting and motivates measuring output diversity (CCDS) alongside quality in §4.