ONE-STEP IS ENOUGH TO JUDGE: HYBRID NOISE SKIMMING FOR FULL-BAND TEXT-TO-AUDIO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-audio (TTA) systems based on latent diffusion and flow matching achieve high fidelity in the full-band scenarios, but their iterative sampling nature hinders practical deployment in real-world applications. Although few-step generation methods such as MeanFlow substantially reduce the number of function evaluations (NFEs), two fundamental challenges still remain in extremely low-NFE models. The first is unreliability: in extremely low NFE environments, quality of generated samples becomes highly sensitive to the initial noise. The second is quality inconsistency: increasing the number of ODE steps does not guarantee monotonic improvements in general sample quality. We propose a hybrid noise skimming framework to overcome both issues through better initial noise selection in few-step generation. The proposed framework combines two complementary components: a learnable Noise Booster amortizes the initial noise selection at the distribution level by refining random noise toward more favorable regions and Single-Step Search which improves the sample quality as instance-wise adaptation using only a single additional search step. The learnable Noise Booster restores the reliability and raises the quality floor by refining initial noise and in contrast to increasing ODE steps, enlarging the search budget yields monotonic improvement in sample quality. Experiments at 44.1 kHz show that our framework achieves SOTA performance with only a one-step generation. The project page is available at: `https://github.com/briskmuntjack/muntjack-20262027`

1 INTRODUCTION

Text-to-audio (TTA) generation, which synthesizes desired audio signal from user text prompt input, has emerged as a promising research field with broad and practical applications such as sound design for visual content creation. Diffusion and flow-matching based models become as a dominant generative paradigm for TTA, owing to their training stability and capability to generate diverse high-fidelity audio. To alleviate the high computation cost as raw waveform generation, recent studies commonly adopt latent generative modeling in a VAE-compressed latent space while preserving perceptual quality (Liu et al., 2023a; Huang et al., 2023; Liu et al., 2024a; Ghosal et al., 2023; Liu et al., 2025; Zhao et al., 2025).

Conventional diffusion and flow-matching based models follow a multi-step trajectory from the noise distribution to data distribution through multiple intermediate states. However, full-band audio generation task as 44.1kHz still requires tens to hundreds of number of function evaluations (NFEs) to achieve high acoustic fidelity despite the latent generative modeling reduces the computational complexity per step. Moreover, the total inference cost scales with both the number of generation steps and number of generated samples, the computational overhead becomes prohibitive under large-scale generation settings, substantially limits practical deployment.

To address this inference bottleneck, many one- or few-step generation approaches such as consistency modeling, shortcut learning, MeanFlow have been proposed (Song et al., 2023; Kim et al., 2024; Luo et al., 2023; Geng et al., 2026a). MeanFlow, a representative method for few-step generation, substantially reduces the generation gap between single-step generation and multi-step generation by learning the average velocity field over the entire velocity trajectory. Previous works have

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.

demonstrated that semantic alignment and perceptual quality can vary depending on the stochastic initial state in conditional generative models (Xu et al., 2025b; Eyring et al., 2024; Tang et al., 2025). Especially extremely low-NFE generation provides highly limited opportunities to recover from the unfavorable initial states (Eyring et al., 2026; Jeon & Lee, 2026). A natural remedy is to increase the generation steps to improve generation quality, allowing progressive opportunities to refine the generation trajectory. These observations show two issues on extremely low-NFE generation: strong sensitivity to the initial noise and low performance consistency to the generation steps.

Motivated by these observations, we propose a hybrid noise skimming framework that reframes the inference-stage noise initialization problem through amortized and adaptive perspective. Instead of depending on expensive instance-wise search as best-of-N method (Ma et al., 2025; Jung et al., 2026), the proposed method shifts the search space by distribution-level refinement and reduces the instance-level uncertainty by performing Single-Step Search. Our learnable Noise Booster transforms a random initialization towards favorable regions of the conditional noise-space by leveraging the knowledge learned from the training data. Even after the initial noise is guided to the favorable regions, there still remains the room to further suitable noise selection at inference time. To complement the amortized refinement at minimal additional cost, we introduce a Single-Step Search that performs lightweight noise selection at inference time. Rather than modifying the generation trajectory itself as common in low-NFE methods, our search method can efficiently explore nearby high-quality solutions and achieve progressively improved performance. We build our framework on an x-prediction MeanFlow backbone (Li & He, 2026; Geng et al., 2026a), which provides an efficient foundation for extremely low-NFE generation under full-band audio scenario.

Extensive experiments on 44.1 kHz full-band TTA demonstrate that proposed framework consistently improves both semantic alignment and perceptual quality under extremely low-NFE generation and state-of-the-art performance in single-step generation. Under single-step generation, the proposed framework improves the CLAP score by 23.3% and contents usefulness (CU) score by 16.2% and improves the audio quality metric PQ score by 5.8%. In addition, we further conduct comprehensive ablation studies to analyze the design choices. Specifically, we investigate which ranking metric is most appropriate to improve the general quality and compare conventional best-of-N search using full generation steps with our Single-Step Search outputs. Our result shows that Single-Step Search preserves search effectiveness while considerably reducing the inference cost.

Our Contributions:

- We introduce a hybrid noise skimming framework that integrates amortized distribution-level refinement and instance-wise Single-Step Search.
- We identify non-monotonic performance across generation steps under extremely low NFE and mitigate it along with initial noise sensitivity.
- Our proposed framework substantially improves the semantic alignment and perceptual quality of generated audio without large-scale brute-force noise search.
- We release our code and the first full-band TTA model built on x-prediction MeanFlow to facilitate future research field.

2 Related Work

2.1 Efficient Diffusion and Few-Step Text-to-Audio Generation

Latent-space generation and fast sampling are widely studied to improve the efficiency of diffusion- (Ho et al., 2020) and flow-matching-based (Lipman et al., 2023) text-to-audio (TTA) generation. AudioLDM (Liu et al., 2023a), Make-An-Audio (Huang et al., 2023), AudioLDM2 (Liu et al., 2024a), and Tango (Ghosal et al., 2023) adopt latent diffusion models (LDMs) (Rombach et al., 2022) for TTA. TangoFlux (Hung et al., 2026) introduces latent flow matching for 44.1 kHz TTA generation. FlashAudio (Liu et al., 2025) and AudioTurbo (Zhao et al., 2025) further accelerate TTA sampling using Rectified Flow (Liu et al., 2023b) and rectified diffusion, respectively.

To enable more efficient one- or few-step generation, several consistency-based approaches are proposed, including the Consistency Model (CM) (Song et al., 2023), Consistency Trajectory Model (CTM) (Kim et al., 2024), and Latent Consistency Model (LCM) (Luo et al., 2023). Consisten-

cyTTA (Bai et al., 2024), AudioLCM (Liu et al., 2024b), and SoundCTM (Saito et al., 2025) apply these approaches to TTA and achieve one- or few-step audio generation. MeanFlow (Geng et al., 2026a) enables single-step generation by learning the average velocity over a finite time interval, and MeanAudio (Li et al., 2026b) applies MeanFlow to TTA. JiT (Li & He, 2026) demonstrates the benefits of x-prediction, which directly predicts clean data.

2.2 Preference and Reward based Post-Training for Text-to-Audio Generation

Preference- and reward-based post-training is widely explored to improve text alignment and perceptual quality in TTA generation. BATON (Liao et al., 2024) uses human-preference-based reward learning to improve audio integrity, temporal consistency, and text alignment. Tango 2 (Majumder et al., 2024) applies Direct Preference Optimization (DPO) (Rafailov et al., 2023) to synthetic preference pairs that capture event presence and temporal ordering. T2A-Feedback (Wang et al., 2025) uses fine-grained AI feedback on event occurrence and sequence as well as acoustic quality, improving both prompt adherence and audio quality. For flow-based TTA, TangoFlux introduces CLAP-Ranked Preference Optimization (CRPO) (Hung et al., 2026). Stable Audio Open Small (Novack et al., 2025) uses ARC post-training to improve prompt adherence while reducing sampling latency.

2.3 Initial-Noise Adaptation and Inference-Time Scaling

The effect of initial noise on generation quality is also actively studied. DNO (Tang et al., 2025) and ReNO (Eyring et al., 2024) improve generation alignment by optimizing the initial noise with reward guidance. Noise Hypernetworks (Eyring et al., 2026) and LENS (Jeon & Lee, 2026) amortize this noise adaptation into learnable transformations, reducing the optimization overhead at inference time. In parallel, Ma et al. (2025) studies reward-verifier-guided inference-time scaling through Random Search, Zero-Order Search, and Search over Paths, and shows that additional search compute can improve diffusion generation quality. SCORE (Jung et al., 2026) extends this idea to TTA and improves audio-text alignment and perceptual quality using reward-based Best-of-N selection and evolutionary search.

3 Proposed Frameworks

We propose the *Hybrid Noise Skimming Framework*, which improves few-step text-to-audio generation by allocating part of the compute budget to the initial noise instead of using it solely for additional sampling steps. As shown in Figure 1, our framework builds on x-prediction MeanFlow (x-MF), which can generate a clean latent even with one-step inference. Conditioned on the text prompt, the *Noise Booster* shifts each noise candidate toward a prompt-dependent region that the frozen generator maps to higher-reward audio. Given N (two or more) Gaussian noise candidates, the *Single-Step Search* (SSS) performs one-step inference for each boosted candidate to preview and score its generated output, and then selects the most promising initial noise. Finally, extremely low-NFE generation is performed only from the selected noise. Accordingly, the search and final generation stages require a total of $N + K$ NFE. We use $K = 2$, which provides a favorable trade-off between generation quality and computational efficiency for x-MF in our setting. The complete budget rule and inference algorithm are given in Appendix A.1.

3.1 Noise Booster

The Noise Booster $B_{\boldsymbol{\phi}}$ performs amortized refinement of the initial-noise distribution. Given a Gaussian noise $\boldsymbol{z} \in \mathbb{R}^{L\times C}$ sampled from the standard prior and a text prompt $\boldsymbol{c}$, it predicts a prompt-conditioned residual correction,

$$\boldsymbol{z}' = \boldsymbol{z} + B_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{c}), \tag{1}$$

using the residual parameterization of noise hypernetworks (Eyring et al., 2026). This transformation induces a prompt-conditioned proposal distribution over initial noises rather than optimizing a single initialization independently for every inference instance.

The Noise Booster is trained across captions such that samples from this refined distribution are more likely to lie in regions that the frozen generator maps to high-reward audio. In this sense, the Noise Booster performs *distribution-level refinement*: it narrows the initial-noise search space

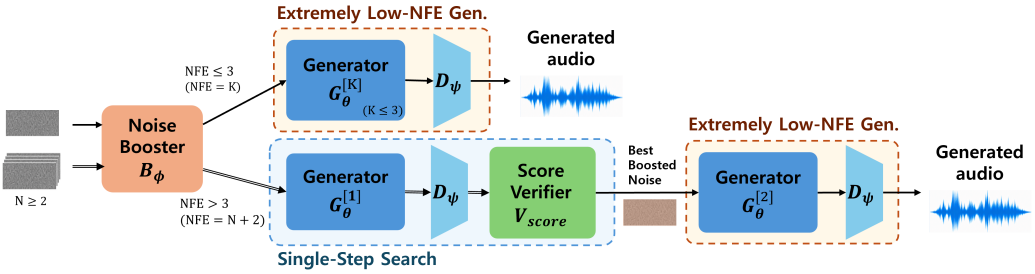


Figure 1: Overview of the proposed Hybrid Noise Skimming framework. When NFE ≤ 3, search is skipped and the boosted noise is directly used for K-step generation $G_\theta^{[K]}$, where $K = \text{NFE}$. When NFE > 3, the Noise Booster refines N Gaussian noise into favorable noise candidates, and Single-Step Search verifies them with one-step generation $G_\theta^{[1]}$ to select the best one. The boosted noise is then used for two-step final generation $G_\theta^{[2]}$.

toward favorable regions and increases the quality of the candidate pool before explicit search is performed. Importantly, it does not eliminate the variation among individual candidates. Single-Step Search therefore complements the Noise Booster by performing instance-wise selection among samples drawn from the refined distribution.

We implement the Noise Booster B_ϕ as a lightweight conditional diffusion transformer (DiT) (Peebles & Xie, 2023). It reuses the text conditioning of the generator and predicts only the residual noise update, while the few-step generator remains frozen. Architectural details, initialization, and computational overhead are provided in Appendix A.2.

3.2 Single-Step Search

Single-Step Search (SSS) performs instance-level refinement at inference time by selecting the most promising initialization among N boosted candidates. Although the Noise Booster shifts the initial-noise distribution toward favorable regions, individual samples drawn from the refined distribution can still exhibit instance-level variation, which motivates this additional selection stage. Rather than fully generating every candidate, the SSS verifies each candidate using only a single-step generation. We therefore refer to it as *single-step search*: each candidate requires 1 NFE during search, after which only the selected candidate is used for final few-step generation.

The key enabler of the SSS is the strong one-step generation capability of x-MF. By adopting an x-prediction parameterization, x-MF substantially outperforms conventional MeanFlow in one-step generation quality, as demonstrated in our ablation studies. This strong one-step performance makes one-step preview $G_\theta^{[1]}(\boldsymbol{z}_i'; \boldsymbol{c})$ a reliable proxy for assessing the quality of each initial noise, enabling the SSS to identify promising candidates without running their full-step trajectories. Moreover, its one-step output directly provides a clean-latent that can be decoded and verified.

Search complexity. A conventional full-step search evaluates all N candidates using the complete K-step generation trajectory, whereas SSS performs one evaluation for each candidate preview and runs the final K-step generation only once from the selected noise. Accordingly, their generator costs are

$$C_{\text{full}} = NK, \qquad C_{\text{HNS}} = N + K, \tag{2}$$

respectively. In SSS, only the N preview evaluations scale with the number of searched candidates, reducing the candidate-dependent cost from NK to N and thereby enabling a substantially larger candidate pool under the same inference budget.

One-step preview and scoring. For each boosted noise $\boldsymbol{z}_i'$, we obtain a one-step preview $\boldsymbol{a}_i^{(1)} = D_\psi\left(G_\theta^{[1]}(\boldsymbol{z}_i'; \boldsymbol{c})\right)$ and verify two complementary aspects of text-to-audio generation. Text–audio relevance is measured by the CLAP cosine similarity (Wu et al., 2023), $s_i^{\text{c}} = R_{\text{clap}}(\boldsymbol{a}_i^{(1)}, \boldsymbol{c})$, while

perceptual quality is measured by the production-quality (PQ) score of Audiobox Aesthetics (Tjandra et al., 2025), $s_i^{\mathrm{q}} = R_{\mathrm{pq}}(\boldsymbol{a}_i^{(1)})$.

Rank-mean selection. Because CLAP similarity score and PQ have different numerical scales and within-prompt distributions, directly combining their raw scores requires metric-specific calibration. Instead, we aggregate their within-pool rankings. Let $\mathrm{rk}(\boldsymbol{s})_i \in [1, N]$ denote the ascending rank of candidate i, with the best candidate assigned rank N. We select

$$i^* = \arg\max_i \ \tfrac{1}{2}\Big(\mathrm{rk}(\boldsymbol{s}^{\mathrm{c}})_i + \mathrm{rk}(\boldsymbol{s}^{\mathrm{q}})_i\Big). \tag{3}$$

This rank-based aggregation balances semantic relevance and perceptual quality without requiring calibration across verifier scales. Alternative verifier and aggregation rules are examined in Appendix A.3.

From one-step preview to few-step generation. The one-step preview is used only to evaluate the initial noise and is not returned as the final output. After selecting i^*, the corresponding boosted noise $\boldsymbol{z}'_{i^*}$ is reused as the starting point for final few-step generation. The effectiveness of this procedure therefore depends on whether candidate rankings obtained from one-step previews remain sufficiently informative for the corresponding few-step outputs. We empirically validate this correspondence by comparing one-step preview search with full-step search in the ablation studies.

3.3 Extremely Low-NFE Generation

Our *Hybrid Noise Skimming Framework* uses x-MF as the generation model for both one-step preview and final extremely low-NFE generation. Unlike conventional MeanFlow, which directly predicts the interval-average velocity, x-MF predicts a clean-latent estimate $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, r, t \mid \boldsymbol{c}, \omega)$. The resulting transition from time t to r is

$$\boldsymbol{z}_r = \frac{r}{t}\boldsymbol{z}_t + \left(1 - \frac{r}{t}\right)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, r, t \mid \boldsymbol{c}, \omega). \tag{4}$$

This x-prediction formulation enables strong generation performance in the extremely low-NFE regime while supporting iterative refinement with the same generator.

After SSS selects the most promising boosted noise $\boldsymbol{z}'_{i^*}$, the final audio is generated as $\boldsymbol{a}^* = D_\psi\Big(G_\theta^{[K]}(\boldsymbol{z}'_{i^*}; \boldsymbol{c})\Big)$. We use $K = 2$ for the final few-step generation following SSS. Our ablation studies show that the 2-step generation provides the largest marginal improvement, whereas further increasing the number of steps yields limited or non-monotonic gains. This motivates allocating additional inference budget to candidate search rather than extending the generation trajectory. The complete K-step sampler and training objectives of x-MF are provided in Appendix A.4.

3.4 Training

The proposed framework consists of the x-MF generator and the Noise Booster model. SSS is not a learning-based model and does not require any trainable parameters. We first train the x-MF generator, freeze its parameters, and then train the Noise Booster model.

Stage 1: Extremely low-NFE generator. The x-MF generator is pretrained with x-prediction flow matching and subsequently fine-tuned with the improved MeanFlow (iMF) objective (Geng et al., 2026b). During iMF fine-tuning, the guidance scale ω is explicitly incorporated as a conditioning variable, allowing the guidance strength to remain adjustable at inference while requiring only one network evaluation per sampling step. The x-prediction parameterization, K-step sampler, and training objectives are detailed in Appendix A.4.

Stage 2: Noise Booster. With the generator, VAE decoder, and reward models frozen, only Noise Booster $B_{\boldsymbol{\phi}}$ is optimized. For an initial noise $\boldsymbol{z}$ and caption $\boldsymbol{c}$, we first define the one-step generated audio from the boosted noise as $\boldsymbol{a}_{\boldsymbol{\phi}}^{(1)} = D_\psi\Big(G_\theta^{[1]}(\boldsymbol{z} + B_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{c}); \boldsymbol{c})\Big)$. The Noise Booster is trained using the same verifiers employed in SSS as reward functions,

$$\mathcal{L}_{\mathrm{NB}}(\boldsymbol{\phi}) = -\,\mathbb{E}_{\boldsymbol{z},\boldsymbol{c}}\Big[w_{\mathrm{c}}\widetilde{R}_{\mathrm{c}}(\boldsymbol{a}_{\boldsymbol{\phi}}^{(1)}, \boldsymbol{c}) + w_{\mathrm{q}}\widetilde{R}_{\mathrm{q}}(\boldsymbol{a}_{\boldsymbol{\phi}}^{(1)})\Big] \quad + \lambda\mathbb{E}_{\boldsymbol{z},\boldsymbol{c}}\left[\frac{1}{LC}\|B_{\boldsymbol{\phi}}(\boldsymbol{z}, \boldsymbol{c})\|_2^2\right], \tag{5}$$

where $\widetilde{R}_{\mathrm{c}}$ and $\widetilde{R}_{\mathrm{q}}$ denote the standardized rewards derived from the CLAP and PQ verifiers, respectively. The ℓ_2 regularization constrains the magnitude of the residual noise update. Because both rewards are reference-free, the Noise Booster is trained using captions only and does not require paired reference audio. Reward standardization, loss weights, and optimization details are provided in Appendix A.5.

4 EXPERIMENTS

4.1 DATASETS

We train on AudioCaps (Kim et al., 2019) and WavCaps (Mei et al., 2024), two publicly available audio–text datasets. AudioCaps provides human-written captions for 10 seconds AudioSet clips while WavCaps provides 403,050 clips collected from multiple audio sources with ChatGPT-assisted captions. After removing blacklist captions in WavCaps that overlap with AudioCaps and cleaning unavailable audio data, we obtain 434,180 text caption–audio pairs (45,395 from AudioCaps and 388,785 from WavCaps); no ground-truth audio beyond this set is used at any stage, and the Noise Booster (Stage 2) uses only the captions. We use the AudioCaps test split for evaluation, where each clip has five human-written captions. We treat each caption as a separate prompt yields 4,428 test pairs from 887 retrievable clips, and report all objective results on these pairs. For human evaluation, we select 100 captions with at most one caption per clip to ensure diversity in sound-event composition and temporal structure. The selected set spans 1 to 4+ sound events and contains diverse temporal patterns and sound categories. We use the original human-written captions without any modification (Appendix C).

4.2 EXPERIMENTAL SETTINGS

Our model adopts the TangoFlux architecture (Hung et al., 2026) for the generator and a DiT (Peebles & Xie, 2023) based network for the Noise Booster. The generator consists of a 518M-parameter Flux (Black Forest Labs, 2024) architecture with 6 MMDiT and 18 DiT blocks, a frozen FLAN-T5-large text encoder (Chung et al., 2024), and the pretrained Stable Audio Open VAE (Evans et al., 2025). The pretrained VAE encodes 10 s, 44.1 kHz stereo audio into a 215×64 latent representation. The Noise Booster contains 6 DiT blocks (Peebles & Xie, 2023) with 44M parameters; further architectural details are provided in Appendix A.2.

We pretrain x-MF with the x-prediction flow-matching objective (Li & He, 2026) for 80 epochs and fine-tune it with the iMF objective (Geng et al., 2026b) for 40 epochs, using a total batch size of 96. The learning rates are 2.5×10^{-4} and 3×10^{-5} for pretraining and fine-tuning, respectively. The Noise Booster is trained for 30 epochs with the generator, VAE, and verifiers frozen, using AdamW (Loshchilov & Hutter, 2019) with a learning rate of 1.25×10^{-4} and batch size 96. Further optimization and sampling details are provided in Appendix A.5. At inference, we use a guidance scale of $\omega = 3.5$ and $K = 2$ steps for final generation. For $\mathrm{NFE} \in 4, 8, 16, 32, 64, 128, 256$, the remaining $N = \mathrm{NFE} - 2$ evaluations are allocated to Single-Step Search. The 1- and 2-NFE settings use only the Noise Booster. Additional details are provided in Appendix A.3.

We evaluate the competitive models introduced in section 1 and 2, including TangoFlux, Resonate, GenAU-L, AudioLCM, ConsistencyTTA, MeanAudio-L, SoundCTM, and SAO small-ARC. We use the official checkpoints and model & inference configurations for all baselines. NFE counts all generator evaluations, including both conditional and unconditional passes when classifier-free guidance is used.

4.3 MAIN RESULTS

4.3.1 OBJECTIVE EVALUATION

We evaluate audio fidelity and diversity using FD and IS with PANNs (Kong et al., 2020), semantic alignment using KL, LAION-CLAP (Wu et al., 2023), and Qwen3-Omni-based Audio Question Answering (AQA) (Xu et al., 2025a), and perceptual quality using the Audiobox Aesthetics (Tjandra et al., 2025) scores: Production Quality (PQ), Content Enjoyment (CE), Content Usefulness (CU),

Table 1: Objective comparison with representative TTA models on the AudioCaps test set.

Model	NFE	SR	FD↓	IS↑	KL↓	CLAP↑	PQ↑	CE↑	CU↑	PC↑	AQA↑
Diffusion, Flow TTA											
GenAU-L	200	16	**10.03**	**15.63**	1.51	0.317	5.69	3.39	5.00	3.17	0.862
TangoFlux	50	44.1	16.32	15.13	1.37	0.320	5.75	3.56	5.02	3.23	0.910
Resonate	50	44.1	11.80	12.83	1.38	0.406	5.75	3.48	4.99	3.32	**0.911**
Accelerated TTA											
AudioLCM	2	16	21.79	8.85	1.61	0.242	5.53	3.26	4.87	3.00	0.819
ConsistencyTTA	1	16	19.56	9.88	1.48	0.262	5.85	3.50	4.94	3.50	0.842
MeanAudio-L	1	16	11.02	12.84	1.34	0.317	5.52	3.36	4.91	3.03	0.882
SoundCTM-UNet	2	16	15.71	9.29	1.50	0.252	5.66	3.31	4.89	3.30	0.836
SoundCTM-DiT	1	44.1	22.13	10.92	1.69	0.346	5.65	3.23	4.87	3.06	0.779
SAO small-ARC	8	44.1	47.92	9.23	2.42	0.181	6.28	3.55	5.40	2.72	0.656
Proposed	1	44.1	14.35	10.86	1.41	0.391	5.84	3.49	4.96	**3.52**	0.865
	2		13.44	11.70	1.41	0.396	5.88	3.54	5.01	3.41	0.872
	4		13.08	12.06	1.38	0.407	5.98	3.59	5.10	3.40	0.878
	8		12.74	12.76	1.35	0.427	6.13	3.66	5.24	3.39	0.885
	16		12.79	12.96	1.34	0.438	6.22	3.72	5.34	3.40	0.886
	32		12.35	13.25	**1.30**	0.445	6.29	3.76	5.41	3.38	0.890
	64		12.29	13.21	1.31	**0.450**	**6.35**	**3.79**	**5.46**	3.38	0.889

Table 2: Human evaluation results. OVL = overall quality, REL = relevance to the caption.

Model	NFE	OVL↑	REL↑
TangoFlux	50	2.87 ± 0.12	$\mathbf{3.73} \pm 0.11$
SAO small-ARC	8	2.99 ± 0.12	2.55 ± 0.13
MeanAudio-L	1	2.66 ± 0.11	3.04 ± 0.13
SoundCTM-DiT	1	2.50 ± 0.11	2.51 ± 0.12
Proposed	1	2.76 ± 0.11	3.21 ± 0.13
	16	$\mathbf{3.00} \pm 0.11$	3.42 ± 0.12

and Production Complexity (PC). KL measures semantic correspondence with the reference audio, whereas CLAP and AQA evaluate alignment with the text prompt. For more details, see Appendix B.

Overall, our model achieves strong full-band 44.1 kHz generation at only 1 NFE, while allocating additional NFE budget to candidate search further improves performance across the evaluation metrics. Table 1 highlights four key observations. (1) Our framework achieves strong extremely low-NFE performance even for full-band 44.1 kHz stereo generation. At only 1 NFE, the Noise Booster alone achieves a CLAP score of 0.391, outperforming the evaluated 16-kHz few-step baselines and substantially surpassing the 44.1 kHz SoundCTM-DiT; notably, it already exceeds the 50-NFE TangoFlux and 200-NFE GenAU-L in text–audio alignment. (2) With modest additional compute, our framework remains competitive with substantially more expensive multi-step systems. At 8 NFE, it achieves a CLAP score of 0.427, exceeding all evaluated baselines, while providing strong FD, KL, PQ, and AQA despite requiring far fewer generator evaluations than TangoFlux, Resonate, and GenAU-L (50–200 NFE). (3) The framework scales effectively when additional inference compute is allocated to searching more initial-noise candidates rather than taking additional sampling steps. From 1 to 64 NFE, FD decreases from 14.35 to 12.29, while CLAP and PQ increase from 0.391 to 0.450 and from 5.84 to 6.35, respectively, with the largest gains obtained in the small-to-moderate budget regime. (4) These improvements extend beyond the verifier objectives: metrics not directly used for candidate selection, including FD, IS, KL, CE, CU, and AQA, also improve overall as the search budget increases. This suggests that Single-Step Search identifies broadly better initializations rather than merely exploiting the CLAP and PQ verifier scores.

4.3.2 HUMAN EVALUATION

We conduct a blind listening test on 100 AudioCaps captions using six systems: our method at 1 and 16 NFE, TangoFlux (Hung et al., 2026) (50 NFE), MeanAudio-L (Li et al., 2026b) (1

NFE), SoundCTM-DiT (Saito et al., 2025) (1 NFE), and Stable Audio Open small with ARC post-training (Novack et al., 2025) (8 NFE). 24 listeners rate overall quality (OVL) and caption relevance (REL) on a 5-point Mean Opinion Score (MOS) scale. Experimental details are described in Appendix C.

Table 2 reports the listening-test results, which are consistent with the objective evaluation. Even at 1 NFE, our proposed framework remains competitive with higher-NFE systems such as TangoFlux and SAO small-ARC and performs favorably among the accelerated baselines, achieving higher OVL and REL scores than MeanAudio-L and SoundCTM-DiT. Allocating the additional inference budget to Single-Step Search improves OVL from 2.76 to 3.00 and REL from 3.21 to 3.42. These results indicate that the improvements observed in the objective metrics are also reflected in human perceptual evaluation, demonstrating promising performance even in the extremely low-NFE regime.

4.4 Ablation Studies

We conduct a series of ablation studies to further understand the contribution of each component and the design choices underlying our framework. We first evaluate the incremental effect of the proposed components on top of the x-MF. We then examine several design choices on the verifier, the number of latent generation steps, and the prediction target.

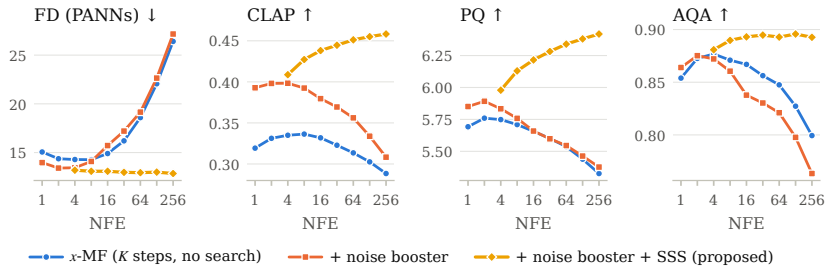


Figure 2: Incremental contribution of the proposed components across NFE budgets.

Component-wise ablation To evaluate the contribution of each proposed component, we progressively add the component to the x-MF until the full model is obtained. As shown in Figure 2, all evaluation metrics show a consistent tendency across the ablations. The x-MF improves up to 4 NFEs, but its performance starts to deteriorate beyond that point. The Noise Booster follows a similar pattern, despite its improved performance in low-NFE region. This demonstrates the non-monotonic behavior of few-step generation models and shows that the Noise Booster alone does not alter this behavior. In contrast, incorporating Single-Step Search reverses this trend. Starting from 4 NFEs where search-enabled configuration is active, performance improves monotonically on all metrics as the NFE budget increases. These results suggest that the two components have complementary benefits rather than redundancy.

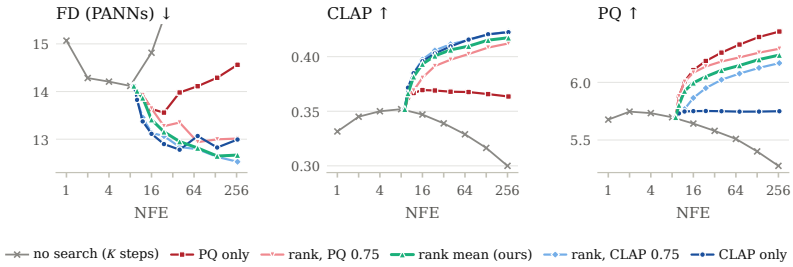


Figure 3: Comparison of verifier criteria for best noise selection.

Verifier design for noise selection Figure 3 compares several verifier criteria for noise selection. We consider PQ and CLAP score as a individual verifier metrics and weighted ranking function and equally weighted rank-aggregation rules. Selecting a single metric as a verifier, every ranking

strategy improves the corresponding metric but degrades the other. Averaging the ranks of the two metrics removes the metric dependence and yields consistent improvements across both perceptual quality and audio-text semantic alignment. Based on these results, we adopt the rank-mean ensemble of PQ and CLAP score as the verifier in our proposed framework.

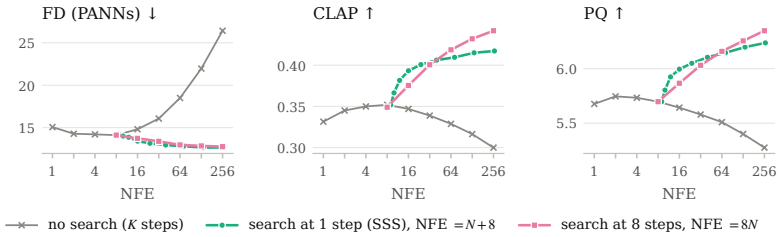


Figure 4: Comparison of search performance where different steps for latent generation are used.

Efficiency of latent generation step We examine whether full-step generation is necessary for verifying noise candidates. Specifically, we compare one-step latent generation and $K = 8$ step generation trajectory. Figure 4 shows the consistent trend across all evaluation metrics, performance improves whether single step generation or full step generation. Notably, in the low-NFE generation steps, single-step verification exhibits even steeper improvement as the search budget increases. We therefore adopt single-step latent generation for verification, which preserves search effectiveness and performance improvement while reducing the computational cost of candidate selection.

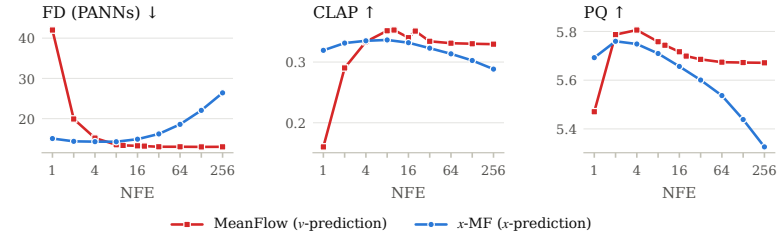


Figure 5: Comparison of target prediction.

Effect of target prediction: Velocity- versus x-prediction. We further investigate the effect of the prediction target by comparing the performance of parameterization method as v-prediction and x-prediction parameterizations before MeanFlow distillation. As shown in Figure 5, in the full-band TTA setting, v-prediction in low-NFE regions suffers from significantly larger performance degradation than x-prediction model, particularly up to 4 NFEs. These results may suggest that velocity-based parameterization is suboptimal for full-band text-to-audio generation under extremely low-NFE budgets. Since our scenario is high-quality generation with very few NFE steps, we adopt x-prediction for all remaining experiments.

5 Conclusions

In this paper, we propose the hybrid noise skimming framework that exploits distribution-level initial noise refinement and lightweight Single-Step Search under extremely low-NFE text-to-audio generation model. By addressing both of the initial noise sensitivity and inefficient inference-time search, the proposed framework shows desirable quality on diverse evaluation metric even under very low-NFE regime. Overall, the proposed framework reach SOTA text-audio alignment and perceptual quality among evaluated systems at 1 NFE by exploiting the amortized Noise Booster and Single-Step search without expensive compute budget on inference stage. Further, we reveal the monotonic performance improvement by additional search steps.

AI USE STATEMENT

In this work, we used generative AI tools for proofreading and correcting English grammar and typographical errors. We have not used generative AI tools for cleaning the datasets and interpret results and all the CRITICAL processes to the research, such as developing conceptual frameworks, formulating mathematical claims for our methodology, designing experiments to validate the proposed method, revising the experimental design based on the observed results, and interpreting results Generating synthetic datasets is not applicable to this work. Additionally, we used generative AI tools for brainstorming and searching for information and proposing a title for a research paper. We have reviewed all AI-assisted work. We reviewed 20 title suggestions generated by AI, selected and revised those that best reflected the content of our study, and used them to formulate the final title. In addition, we manually verified the research papers identified by LLMs and read the original papers to confirm their relevance and content. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our research uses publicly available audio datasets only. Human evaluation is conducted solely for assessing generated quality, and no personally identifiable or sensitive information is collected. Text-to-audio generation generally poses a lower risk of identity impersonation or voice spoofing. However, generated audio may still be misused to misrepresent acoustic events. We therefore acknowledge these potential misuse scenarios and encourage responsible use of the proposed system. Our work focuses on improving generation efficiency and quality rather than enabling such applications. We encourage responsible use of the proposed model and careful consideration of dataset licenses and downstream deployment settings.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we publicly release both the source code and the trained model in this work. Our model is trained exclusively on publicly available open-source datasets, which reduces barriers related to accessibility and licensing issues. We also provide extensive performance analysis in the Appendix and we hope these materials would support future researchers in this community.

REFERENCES

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.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911, 2025.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4171–4186, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3987.

Luca Eyring, Shyamgopal Karthik, Alexey Dosovitskiy, Nataniel Ruiz, and Zeynep Akata. Noise hypernetworks: Amortizing test-time compute in diffusion models. *Advances in Neural Information Processing Systems*, 38:95429–95462, 2026.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *Advances in Neural Information Processing Systems*, 38:75460–75482, 2026a.

Zhengyang Geng, Yiyang Lu, Zongze Wu, Eli Shechtman, J. Zico Kolter, and Kaiming He. Improved mean flows: On the challenges of fastforward generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30467–30476, 2026b.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction-tuned LLM and latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 3590–3598, 2023. doi: 10.1145/3581783.3612348.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. CNN architectures for large-scale audio classification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 13916–13932. PMLR, 2023.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. In *International Conference on Learning Representations*, volume 2026, pp. 150793–150816, 2026.

Haewon Jeon and Si-Hyeon Lee. Lens: Low-frequency eigen noise shaping for efficient diffusion sampling. *arXiv preprint arXiv:2605.07253*, 2026.

Jaemin Jung, Jaehun Kim, Inkyu Shin, and Joon Son Chung. Score: Scaling audio generation using standardized composite rewards. *IEEE Signal Processing Letters*, 33:3391–3395, 2026.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ode trajectory of diffusion. In *International Conference on Learning Representations*, 2024.

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, 2020.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. In *Interspeech*, 2022.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36115–36125, 2026.

Xiquan Li, Junxi Liu, Wenxi Chen, Haina Zhu, Ziyang Ma, and Xie Chen. Resonate: Reinforcing text-to-audio generation via online feedback from large audio language models. *arXiv preprint arXiv:2603.11661*, 2026a.

Xiquan Li, Junxi Liu, Yuzhe Liang, Zhikang Niu, Wenxi Chen, and Xie Chen. Meanaudio: Fast and faithful text-to-audio generation with mean flows. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14378–14393, 2026b.

Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li. BATON: Aligning text-to-audio model with human preference feedback. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 4542–4550, 2024. doi: 10.24963/ijcai.2024/502.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

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. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 21450–21474. PMLR, 2023a.

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: Text-to-audio generation with latent consistency models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7008–7017, 2024b. doi: 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*, pp. 13694–13710, 2025. doi: 10.18653/v1/2025.acl-long.673.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Inference-time scaling for diffusion models beyond scaling denoising steps. *arXiv preprint arXiv:2501.09732*, 2025.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 564–572, 2024. doi: 10.1145/3664647.3681688.

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. doi: 10.1109/TASLP.2024.3419446. URL https://doi.org/10.1109/TASLP.2024.3419446.

Zachary Novack, Zach Evans, Zack Zukowski, Josiah Taylor, CJ Carr, Julian Parker, Adnan Al-Sinan, Gian Marco Iodice, Julian McAuley, Taylor Berg-Kirkpatrick, et al. 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. IEEE, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. *OpenAI Technical Report*, 2018.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-2338.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. doi: 10.1109/CVPR52688.2022.01042.

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.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning (ICML)*, 2023.

Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 58905–58930. PMLR, 2025.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al. Meta audiobox aesthetics: Unified automatic assessment for speech, music and sound. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025.

Zehan Wang, Ke Lei, Chen Zhu, Jiawei Huang, Sashuai Zhou, Luping Liu, Xize Cheng, Shengpeng Ji, Zhenhui Ye, Tao Jin, and Zhou Zhao. T2a-feedback: Improving basic capabilities of text-to-audio generation via fine-grained ai feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 23535–23547, 2025. doi: 10.18653/v1/2025.acl-long.1147.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025a.

Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Good seed makes a good crop: Discovering secret seeds in text-to-image diffusion models. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 3024–3034, 2025b.

Junqi Zhao, Jinzheng Zhao, Haohe Liu, Yun Chen, Lu Han, Xubo Liu, Mark D. Plumbley, and Wenwu Wang. Audioturbo: Fast text-to-audio generation with rectified diffusion. *arXiv preprint arXiv:2505.22106*, 2025.

A Details of the Hybrid Noise Skimming Framework

A.1 Budget allocation and complete inference algorithm

For a total budget of E generator computations (i.e., $\mathrm{NFE} = E$), the proposed framework operates in two regimes. When $E \leq 3$, search is skipped and a single boosted noise is directly used for K-step generation, where $K = E$. When $E > 3$, the remaining $E - 2$ NFE are used for one-step candidate previews, while two evaluations are reserved for final generation:

$$(N, K) = \begin{cases} (1,\ E) & E \leq 3 \quad \text{(Noise Booster + K-step generation)}, \\ (E-2,\ 2) & E > 3 \quad \text{(Noise Booster + SSS + 2-step generation)}. \end{cases} \tag{6}$$

Accordingly, the total NFE is

$$\mathrm{NFE} = \begin{cases} K & E \leq 3, \\ N + 2 & E > 3. \end{cases} \tag{7}$$

Thus, candidate selection becomes active from $E = 4$, where $N = 2$. Algorithm 1 summarizes the complete inference procedure.

Algorithm 1 Hybrid Noise Skimming inference

Require: caption c, budget E, generator G_θ, Noise Booster B_ϕ, decoder D_ψ, verifiers R_{clap}, R_{pq}

$(N, K) \leftarrow$ Equation 6
Draw $z_1, \ldots, z_N \sim \mathcal{N}(0, I)$
for $i = 1, \ldots, N$ **do**
    $z'_i \leftarrow z_i + B_\phi(z_i, c)$
end for
if $E > 3$ **then**
    for $i = 1, \ldots, N$ **do**
        $a_i^{(1)} \leftarrow D_\psi\Big(G_\theta^{[1]}(z'_i; c)\Big)$
        $s_i^c \leftarrow R_{\mathrm{clap}}(a_i^{(1)}, c); \quad s_i^q \leftarrow R_{\mathrm{pq}}(a_i^{(1)})$
    end for
    $i^* \leftarrow \arg\max_i \frac{1}{2}\left(\mathrm{rk}(s^c)_i + \mathrm{rk}(s^q)_i\right)$
else
    $i^* \leftarrow 1$
end if
$a^* \leftarrow D_\psi\Big(G_\theta^{[K]}(z'_{i^*}; c)\Big)$
return a^*

A.2 Noise Booster architecture

Figure 6 illustrates the architecture of the Noise Booster network. The network follows the DiT architecture (Peebles & Xie, 2023), and consists of multi-head self-attention and pointwise feed-forward layers with adaptive layer normalization. To incorporate the prompt condition $\boldsymbol{c}$, each DiT block includes a multi-head cross-attention layer. Unlike the standard DiT formulation, we omit the diffusion or flow-matching timestep embedding, as the Noise Booster directly refines the initial noise without learning a time-dependent generation trajectory. Learned positional embeddings (Devlin et al., 2019; Radford et al., 2018) are applied to the input noise tokens $\boldsymbol{z}_i$ before being processed by the DiT stack. Our implementation uses a lightweight standalone network, in contrast to the generator-coupled LoRA parameterization of Noise Hypernetworks (Eyring et al., 2026), and is architecturally closer to the standalone transformer used in LENS (Jeon & Lee, 2026).

A.3 Single-step search implementation

Preview. The one-step preview $G_1(\boldsymbol{z}'_i; \boldsymbol{c})$ is the same sampler call as the final generation with $K = 1$, evaluated in mini-batches of candidates for the same caption with the duration condition fixed to 10 s. The preview is decoded with the VAE, resampled to 48 kHz mono for CLAP and to 16 kHz mono for Audiobox Aesthetics, and scored without any trimming.

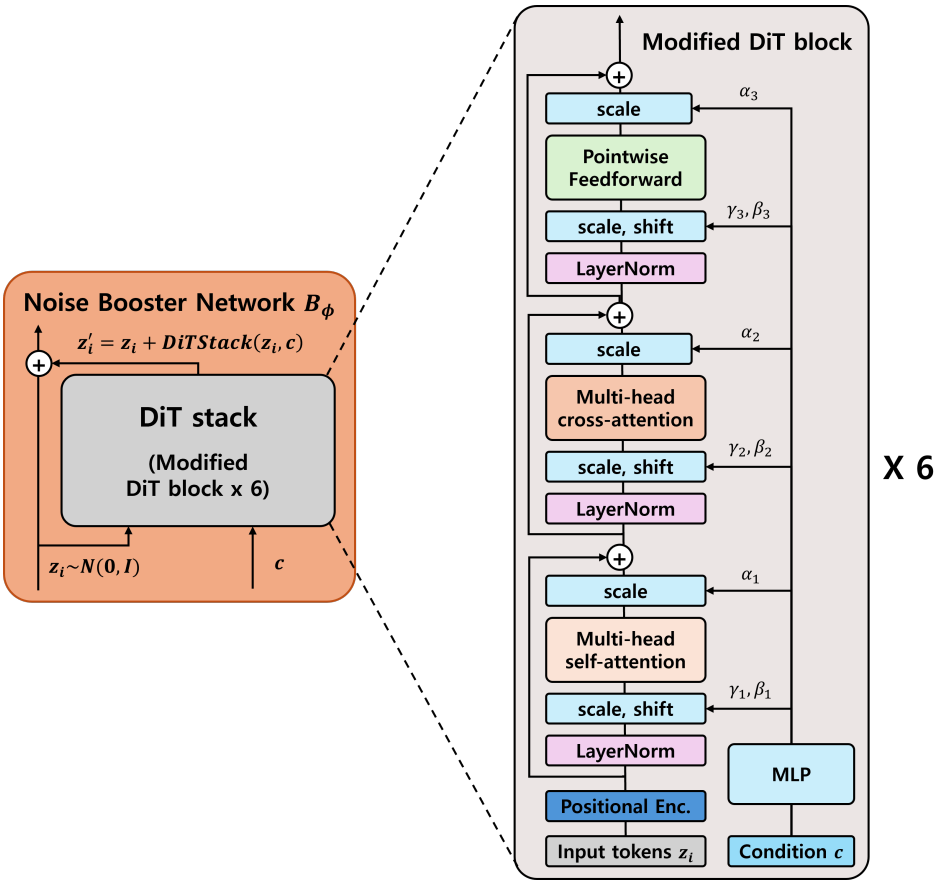


Figure 6: Architecture of the Noise Booster network. A stack of six modified DiT blocks refines the initial Gaussian noise, with each block incorporating prompt-conditioned cross-attention.

Ranking. Ranks are computed by sorting the N scores in ascending order and assigning positions $1, \ldots, N$; exact ties receive the mean of the positions they occupy. The rank-mean of Equation 3 is maximized with the lowest index taken on ties. We also evaluated CLAP-only, PQ-only, weighted rank means, a rank product, and gating rules (top half by one metric, then $\arg\max$ of the other) on a validation split and use the equal-weight rank mean throughout.

Cost accounting. Following prior work on test-time scaling of diffusion models (Ma et al., 2025), we report cost as the number of generator evaluations, so the search contributes N NFE and the final generation K. Each candidate additionally requires one VAE decode and one pass of each reward encoder; these are fixed per-candidate costs that do not depend on K and are shared by any method that scores candidates in the audio domain.

A.4 x-PREDICTION MEANFLOW SAMPLER AND TRAINING

MeanFlow formulation. Let $\boldsymbol{x} \in \mathbb{R}^{L \times C}$ be an audio latent and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Flow matching (Lipman et al., 2023; Liu et al., 2023b) defines

$$\boldsymbol{z}_t = (1-t)\boldsymbol{x} + t\boldsymbol{\epsilon}, \qquad \boldsymbol{v}_t = \frac{\mathrm{d}\boldsymbol{z}_t}{\mathrm{d}t} = \boldsymbol{\epsilon} - \boldsymbol{x}, \qquad t \in [0,1]. \tag{8}$$

MeanFlow (Geng et al., 2026a) models the average velocity over an interval $[r, t]$,

$$\boldsymbol{u}(\boldsymbol{z}_t, r, t) = \frac{1}{t-r}\int_r^t \boldsymbol{v}(\boldsymbol{z}_\tau, \tau)\,\mathrm{d}\tau, \qquad \boldsymbol{z}_r = \boldsymbol{z}_t - (t-r)\boldsymbol{u}(\boldsymbol{z}_t, r, t). \tag{9}$$

We parameterize the network with a clean-latent prediction $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}$ and recover the average velocity as

$$\boldsymbol{u}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, r, t \mid \boldsymbol{c}, \omega) = \frac{\boldsymbol{z}_t - \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, r, t \mid \boldsymbol{c}, \omega)}{t}. \tag{10}$$

Substituting Equation 10 into Equation 9 yields the x-prediction update in Equation 4. The clean-latent parameterization also keeps the network output in the range of data latents rather than the noise-dominated velocity target at large t (Li & He, 2026). More importantly for our framework, setting $r = 0$ makes the network prediction itself a complete latent, enabling one-step candidate scoring.

Sampler and choice of steps. For a uniform grid $t_k = 1 - k/K$, the sampler iterates

$$\boldsymbol{z}_{t_{k+1}} = \frac{t_{k+1}}{t_k}\boldsymbol{z}_{t_k} + \Big(1 - \frac{t_{k+1}}{t_k}\Big)\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}\big(\boldsymbol{z}_{t_k}, t_{k+1}, t_k \mid \boldsymbol{c}, \omega\big), \qquad k = 0, \dots, K-1, \tag{11}$$

starting from $\boldsymbol{z}_{t_0} = \boldsymbol{z}$, and we write $G_K(\boldsymbol{z}; \boldsymbol{c}) = \boldsymbol{z}_{t_K}$. Each step uses one network evaluation. In our model, the second step provides the largest marginal improvement, while additional steps show diminishing and eventually negative returns on the tracked metrics. Moreover, because the Noise Booster is optimized through the one-step generator, its gain transfers most strongly to small K. We therefore use $K = 2$ for the final sample and spend additional inference budget on search candidates.

Flow-matching pretraining. The generator is first pretrained by fixing $r = t$ and minimizing

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t,\boldsymbol{x},\boldsymbol{\epsilon}}\Big\| \frac{\boldsymbol{z}_t - \hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\boldsymbol{z}_t, t, t \mid \boldsymbol{c})}{t} - (\boldsymbol{\epsilon} - \boldsymbol{x})\Big\|_2^2, \tag{12}$$

with a logit-normal distribution for t (Esser et al., 2024). The caption is dropped with probability p_{uncond} so that an unconditional mode $\hat{\boldsymbol{x}}_{\boldsymbol{\theta}}(\cdot \mid \varnothing)$ is available.

iMF fine-tuning with guidance conditioning. We then fine-tune on intervals $r < t$ using the improved MeanFlow (iMF) objective (Geng et al., 2026b), which incorporates classifier-free guidance (Ho & Salimans, 2022) with an explicit conditioning scale ω. Differentiating Equation 9 with respect to t gives

$$\boldsymbol{u}(\boldsymbol{z}_t, r, t) = \boldsymbol{v}(\boldsymbol{z}_t, t) - (t - r)\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{u}(\boldsymbol{z}_t, r, t). \tag{13}$$

For each example, we draw (t, r) and $\omega \sim p(\omega) \propto \omega^{-1}$ on $[\omega_{\min}, \omega_{\max}]$, evaluate the conditional and unconditional boundary velocities $\boldsymbol{v}_{\boldsymbol{c}}$ and $\boldsymbol{v}_{\varnothing}$ at $r = t$ without gradient, and construct

$$\boldsymbol{v}_g = (\boldsymbol{\epsilon} - \boldsymbol{x}) + \Big(1 - \frac{1}{\omega}\Big)(\boldsymbol{v}_{\boldsymbol{c}} - \boldsymbol{v}_{\varnothing}), \tag{14}$$

$$\mathcal{L}_{\mathrm{iMF}} = \mathbb{E}\Big[\,\mathrm{sg}(w)\Big\|\boldsymbol{u}_{\boldsymbol{\theta}} + (t - r)\,\mathrm{sg}\Big(\frac{\mathrm{d}\boldsymbol{u}_{\boldsymbol{\theta}}}{\mathrm{d}t}\Big) - \boldsymbol{v}_g\Big\|_2^2\Big]. \tag{15}$$

The total derivative is computed with a single Jacobian–vector product along the predicted trajectory, $\mathrm{sg}(\cdot)$ denotes stop-gradient, and w is the adaptive MeanFlow weight. At a fixed point, the conditioning variable ω corresponds to the CFG scale, so guidance remains adjustable at test time without requiring a separate unconditional network evaluation at each sampling step.

A.5 Noise Booster training details

Table 3: Noise Booster training configuration.

Item	Value
Frozen generator	x-MF (Stage 1 of section 3.4), $\omega = 3.5$
Rewards	LAION-CLAP cosine (48 kHz mono), Audiobox Aesthetics PQ (16 kHz mono)
Reward weights	$w_{\mathrm{c}} = w_{\mathrm{q}} = 0.5$
Standardization	(μ, σ): CLAP $(0.314,\ 0.079)$, PQ $(5.587,\ 0.528)$
Regularizer	ℓ_2 on the update, $\lambda = 0.5$
Data	training captions only (434k), no reference audio
Optimizer	AdamW, lr 1.25×10^{-4}, constant with 1,000 warm-up steps
Batch / steps	96 batchs, 30 epochs, grad-norm clip 1.0

Table 3 gives the configuration used for Equation 5. The standardization statistics are the mean and the *within-prompt* standard deviation of each reward, measured over an offline pool of 64 random

noises for each of 500 captions with the frozen one-step generator; the within-prompt spread is the relevant scale because the Noise Booster and the search both operate on candidates of the same caption.

B Evaluation Details

Distribution and classifier metrics. FD, KL, and IS are computed with the `av-benchmark` toolkit (Cheng et al., 2025). FD is reported with PANNs embeddings (Kong et al., 2020) in the main text; the toolkit also provides FD with PaSST (Koutini et al., 2022) and VGGish (Hershey et al., 2017) embeddings (the latter often called FAD), which we report in the supplementary tables. KL is the mean KL divergence between the PANNs (or PaSST) class posteriors of each generated clip and its paired ground-truth clip, and IS is the Inception score over the same posteriors. Each embedding network receives audio at its native rate (16 kHz for PANNs and VGGish, 32 kHz for PaSST), resampled and downmixed by the toolkit. CLAP is the cosine similarity between audio and caption embeddings of LAION-CLAP (Wu et al., 2023) (`music_speech_audioset` checkpoint, 48 kHz mono).

Aesthetic metrics. Production quality (PQ), content enjoyment (CE), content usefulness (CU), and production complexity (PC) are the four axes of Audiobox Aesthetics (Tjandra et al., 2025), computed with the official `audio-aes` tool on 16 kHz mono audio and averaged over the test set. PQ is the axis used by the verifier and the booster reward; CE, CU, and PC are never optimized.

AQA. AQA follows the audio question-answering protocol of Resonate (Li et al., 2026a): the judge `Qwen3-Omni-30B-A3B-Instruct` (Xu et al., 2025a) receives the generated audio (16 kHz) together with the prompt "Does this audio contain the sound of *caption*? Please only answer yes or no, without other explanations.", and the score is $\sigma(\log p(\text{yes}) - \log p(\text{no}))$ at the answer position, averaged over the 4,428 test pairs. The judge shares no parameters with CLAP, the aesthetics model, or the generator.

Cost accounting. NFE counts generator evaluations only. For our method it is N one-step previews + final two-steps; for baselines it is the number of sampling steps, doubled when classifier-free guidance requires a separate unconditional pass. VAE decoding and reward scoring during search are not counted (see Appendix A.3).

C Human Evaluation Details

Caption selection. The 100 captions are drawn from the AudioCaps test set, keeping at most one caption per source clip (deduplicated by YouTube id). The selection is balanced across several factors, including the number of audio events (single- and multi-event scenes), temporal structure (e.g., parallel and sequential events), audio category (e.g., animal, mechanical, transportation, environmental sounds).

Models and stimuli. Table 4 summarizes the generation settings of the models compared in the human evaluation. All models generate 10 s clips, one per caption. Audio clips are cropped or zero-padded to exactly 10 s, downmixed to mono (three models are stereo, while the others are mono; leaving the channel count native would give half of the models a spatial cue), normalized to -23 LUFS integrated loudness.

Table 4: Generation settings of the models compared in the human evaluation.

System	Checkpoint	Steps	NFE	Output
TangoFlux (Hung et al., 2026)	CRPO	25	50	44.1 kHz stereo
MeanAudio-L (Li et al., 2026b)	L-Full	1	1	16 kHz mono
SoundCTM-DiT (Saito et al., 2025)	DiT v2	1	1	44.1 kHz mono
SAO small-ARC (Novack et al., 2025)	ARC	8	8	44.1 kHz stereo
Proposed (1 NFE)	x-MF + booster	1	1	44.1 kHz stereo
Proposed (16 NFE)	+ SSS, $N = 14$, $K = 2$	2	16	44.1 kHz stereo

Assignment and interface. Each rater receives 15 captions, chosen so that (i) the number of sound events follows the pool proportions through a cumulative largest-remainder allocation over the four event-count bins, (ii) the number of audio events and temporal type are balanced softly, and (iii) captions with the fewest ratings so far are preferred, which keeps exposure uniform across the pool.