SCORE: Scaling audio generation using Standardized COmposite REwards

Jaemin Jung†, Jaehun Kim†, Inkyu Shin, and Joon Son Chung

***Abstract*—The goal of this paper is to enhance Text-to-Audio (T2A) generation at inference, focusing on generating realistic audio that precisely aligns with text prompts. Despite the rapid advancements, existing approaches often fail to achieve a reliable balance between semantic alignment and perceptual quality. We address this by adopting Inference-Time Scaling for the first time in T2A generation, and propose a multi-reward guidance that places each criterion on equal footing. By standardizing each reward to zero mean and unit variance before a weighted summation, the method reduces the scale mismatch that otherwise demands weight tuning, providing stable guidance at equal weights while leaving the weight as an interpretable control over the desired aspect. Moreover, we introduce a new text-audio alignment metric using an audio language model for more robust evaluation. Empirically, our method improves both semantic alignment and perceptual quality, achieving improved performance over naive generation and existing reward guidance techniques. Synthesized samples are available at https://mm.kaist.ac.kr/projects/score/.**

***Index Terms*—text-to-audio generation, inference-time scaling**

I. Introduction

TEXT-TO-AUDIO (T2A) generation aims to synthesize high-quality audio that accurately matches a text prompt [1]. Diffusion-based approaches [2]–[6] have driven rapid progress [7]–[16]: Stable Audio [9] couples latent diffusion with text and onset conditioning to generate variable-length audio, and EzAudio [10] combines a Diffusion Transformer [17] with classifier-free-guidance rescaling for improved controllability.

However, even advanced T2A models often fail to reliably satisfy the two key aspects: **text-audio alignment** and **perceptual quality**. This issue is particularly pronounced in diffusion-based T2A, where stochastic denoising produces high output variance for a single prompt, hindering reliable generation of the desired sound. This variance, while beneficial for creative diversity, significantly hinders the ability to reliably generate a specific, desired sound.

A promising, training-free solution is Inference-Time Scaling (ITS) [18], which allocates additional computation at inference to guide the generation towards superior outputs, such as higher sample quality and alignment with text. Existing ITS techniques include Best-of-N (BON) selection [19], [20], reward-guided particle sampling [21], and evolutionary search (EvoSearch) [22]. Although ITS has been studied for language, image, and video generation [19], [21], [23]–[29], its application to T2A remains unexplored.

Moreover, existing ITS methods have made limited progress in reward guidance, a crucial component that steers the overall ITS process. Many of them rely on a single reward model as guidance, which can introduce "verifier hacking", a decrease in performance due to over-reliance on a single objective [18]. Others utilize an aggregation of multiple rewards by converting different values into ranks [30], but treating criteria with equal importance under-utilizes the complementary strengths of distinct rewards.

In this paper, we pioneer the application of ITS to T2A generation and conduct a comprehensive empirical study of different methods in this domain. Furthermore, to address the limitations of prior techniques that rely on rank aggregation or weighted sums, we propose Standardized Composite Reward (SCORE), a standardized multi-reward guidance strategy designed to balance two objectives crucial to T2A generation: audio quality and text-audio alignment. Specifically, standardizing each reward to zero mean and unit variance places heterogeneous rewards on comparable numerical scales, so that equal weights already balance the objectives. The combination weight α then serves as an interpretable control, letting users steer generation toward a desired attribute rather than tuning weights to offset mismatched reward scales. Unlike rank aggregation, which discards reward magnitudes, or a naive weighted sum, which is skewed by mismatched scales, SCORE preserves full-precision reward signals while keeping the objectives balanced. Moreover, we introduce an Audio Language Model (ALM) [31]–[34] based evaluation metric named AQAScore, and assess the text-audio alignment beyond the scope of the reward models of interest. Motivated by VQAScore [35], we ask the model with the generated audio and its text prompt and obtain the probability of producing the "yes" token, serving as a quantitative measure of alignment.

Our experiments demonstrate that SCORE significantly enhances the generation quality of T2A model compared to naive generation, improving CLAP and PQ by 10.8% and 10.6%, respectively. We further verify that SCORE effectively reduces the risk of "verifier hacking" and outperforms existing reward aggregation methods in both text-audio alignment and audio quality. Finally, we show that the generation quality faithfully reflects the adjustment of combination weights, offering a principled approach to prioritize specific attributes of interest.

This work was supported by IITP grants funded by the Korean government (MSIT, RS-2025-02263169, Detection and Prediction of Emerging and Undiscovered Voice Phishing).

†These authors contributed equally to this work.

Jaemin Jung, Jaehun Kim and Joon Son Chung are with School of Electrical Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Republic of Korea (e-mail: jjm5811@kaist.ac.kr; kjaehun@kaist.ac.kr; joonson@kaist.ac.kr).

Inkyu Shin is with ByteDance Seed, USA (e-mail: dlsrbgg33@gmail.com)

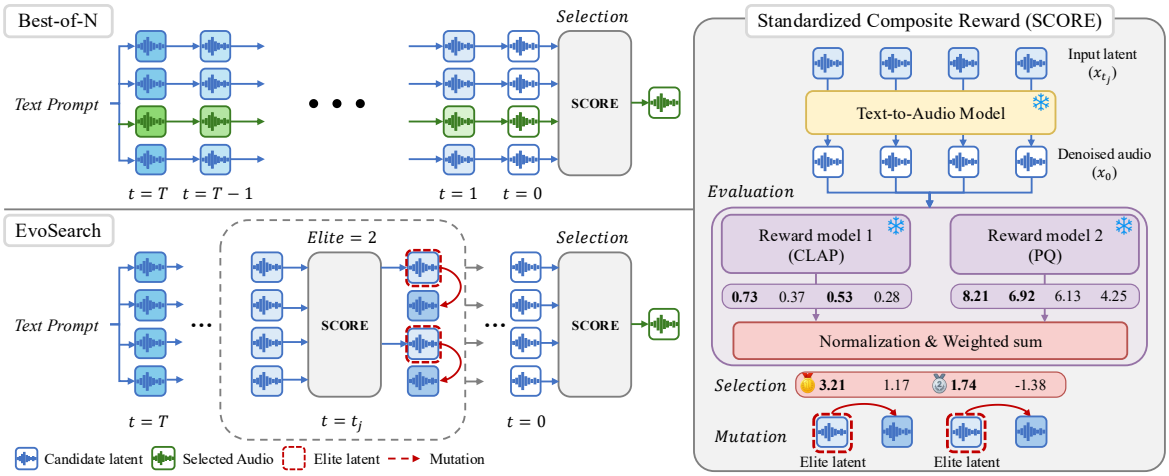


Fig. 1: **Overview of our Inference-Time Scaling framework for Text-to-Audio generation.**

II. Method

A. Inference-Time Scaling for Audio Generation

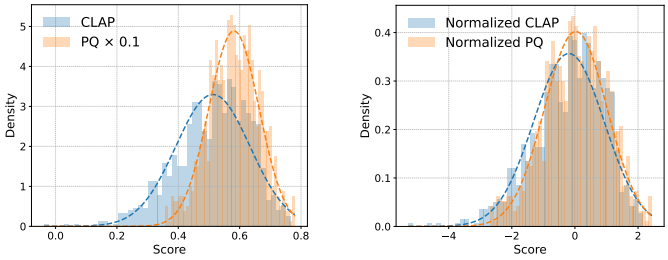
We instantiate ITS for T2A with two established frameworks, BON and EvoSearch (Fig. 1). Both denoise multiple noise latents and select the best final audio using pretrained reward models, differing in where the guidance is applied. BON scores only the final outputs, whereas EvoSearch additionally guides the intermediate latents: at predefined evolution steps $\mathcal{T}_{evo} = \{t_1, \ldots, t_k\}$, each latent is decoded and scored, the top "elite" latents are kept, and the rest are replaced by mutated copies of the elites, where small noise is added for diversity. Repeating this evaluation, selection, mutation loop refines the candidates, and the final audio is selected as in BON. Because EvoSearch applies guidance throughout denoising, it exposes the effect of the reward more strongly than BON.

B. Multi Reward Models

To avoid the aforementioned single-reward "verifier hacking", our method targets two complementary aspects of audio generation: text-audio alignment and perceptual realism. We use the CLAP [36] score as a reward for text-audio alignment, and adopt the Production Quality (PQ) score from Audiobox-Aesthetics [37] as a reliable proxy for human perceptual quality. However, utilizing multiple reward models poses a critical challenge, because each signal originates from a unique distribution. For instance, the CLAP score ranges from 0 to 1, whereas the PQ score ranges from 0 to 10 and exhibits a markedly different distribution, as shown in Fig. 2. This mismatch in distribution can bias the reward guidance. Therefore, we standardize each reward to reduce scale imbalance before aggregation, which is further discussed in the next section.

C. Standardized COmposite REward (SCORE)

Existing methods reduce scale imbalance by converting rewards into ranks [18], [30] or by a heuristic weighted summation [38], [39]. Ranks discard reward magnitudes, while a weighted sum over differently distributed rewards must fold the scale mismatch into its weights: as illustrated in Fig. 2, even rewards linearly rescaled to the same range can differ significantly in distribution, conflating scale correction with genuine preference and requiring a heuristic search to recover good performance.



(a) Raw distributions (b) Normalized distributions

Fig. 2: **Comparison of reward score distributions.** (a) Raw scores of CLAP and PQ show a large distributional gap even after scaling. (b) Standardization with train-set statistics aligns reward means and variances, reducing scale imbalance.

We instead keep full-precision rewards and remove this nuisance with Standardized Composite Reward (SCORE), which standardizes each reward to zero mean and unit variance before a weighted summation. With all rewards on comparable numerical scales, equal weighting assigns equal coefficients to the standardized rewards, and the weight α more directly controls the relative emphasis between objectives. Specifically, we first obtain the raw rewards using each reward model at the evolution timestep t_i. Following EvoSearch, raw reward is defined as:

$$s(x_{t_i}) = E_{x_0 \sim p_0(x_0|x_{t_i})}[r(x_0)|x_{t_i}], \tag{1}$$

where x_0 is the denoised audio from intermediate latent x_{t_i} and $r(\cdot)$ is the reward function. Next, we standardize each raw reward using the mean (μ) and standard deviation (σ) computed for each reward model on the training set:

$$z(x_{t_i}) = \frac{s(x_{t_i}) - \mu}{\sigma}. \tag{2}$$

The standardized scores are then combined via a weighted sum to produce the final reward score for the intermediate state:

$$reward(x_{t_i}) = \alpha \cdot z_{\text{CLAP}}(x_{t_i}) + (1-\alpha) \cdot z_{\text{PQ}}(x_{t_i}), \tag{3}$$

TABLE I: **Performance of EzAudio on the AudioCaps test set.** C and P indicate the CLAP and PQ reward models, respectively. *AQA* denotes AQAScore. The best results are in bold and the second-best results are underlined.

Methods	Scheme	Setup	Text-Consistency		Perceptual Quality			Distribution Metrics	
			CLAP ↑	AQA (%) ↑	PQ ↑	CU ↑	CE ↑	FD ↓	IS ↑
Naive sampling	-	-	0.65±0.04	90.05±0.47	5.73±0.24	5.02±0.21	3.51±0.12	18.87±0.37	11.32±0.24
Best-of-N	Single reward	C	<u>0.72</u>±0.01	91.57±1.36	5.81±0.01	5.07±0.02	3.65±0.03	<u>15.75</u>±0.55	11.63±0.14
	Single reward	P	0.67±0.01	89.49±0.45	6.32±0.02	<u>5.55</u>±0.05	<u>3.84</u>±0.02	17.74±0.78	11.71±0.22
	Rank aggregation	C + P	0.70±0.02	91.11±0.75	6.21±0.01	5.45±0.08	3.81±0.03	16.89±0.75	11.85±0.08
	Weighted sum	C + P	0.71±0.01	91.21±0.18	6.17±0.01	5.41±0.04	3.80±0.03	17.16±0.61	11.79±0.18
	SCORE (α=0.50)	C + P	0.71±0.01	91.43±0.24	6.21±0.02	5.44±0.04	3.81±0.03	17.24±0.63	11.86±0.13
EvoSearch	Rank aggregation	C + P	0.71±0.00	91.23±0.52	6.16±0.06	5.35±0.05	3.73±0.02	15.92±0.30	<u>12.03</u>±0.12
	Weighted sum	C + P	<u>0.72</u>±0.02	91.04±0.27	6.27±0.02	5.45±0.02	3.77±0.01	16.05±0.67	11.86±0.17
	SCORE (α=0.50)	C + P	<u>0.72</u>±0.00	<u>91.63</u>±0.53	<u>6.34</u>±0.02	5.53±0.01	3.82±0.01	<u>15.75</u>±0.40	11.93±0.21
	SCORE (α=0.25)	C + P	0.69±0.00	90.61±0.40	**6.47**±0.02	**5.66**±0.01	**3.86**±0.01	16.54±0.10	**12.04**±0.30
	SCORE (α=0.75)	C + P	**0.74**±0.00	**92.35**±0.63	6.09±0.04	5.29±0.04	3.70±0.02	**15.16**±0.26	11.86±0.26

where the weight $\alpha \in [0, 1]$ sets the relative emphasis of the two objectives. Because both terms are standardized, $\alpha = 0.5$ is balanced by default, and moving α away from 0.5 trades one objective for the other predictably (Section IV-B); α is therefore an interpretable preference control, not a hyperparameter tuned for performance. For instance, raising α favors text-audio alignment through the CLAP reward. We note that this process is not limited to two rewards; any additional reward model of interest can be incorporated.

D. Text-Audio Alignment Evaluation with ALM

Despite the growing interest in matching the generated audio to the text prompt, there is a limited number of relevant evaluation metrics. To gauge the performance of the proposed method from different perspectives, we introduce AQAScore, a novel text-audio alignment metric leveraging Audio Flamingo 3 [31], a pre-trained ALM. Specifically, we adopt the method used in VQAScore [35], where the text-image alignment is measured by asking a Vision Language Model a query *"Does this image contain {prompt}?"* and calculating the probability of generating *"yes"*. For text-audio alignment evaluation, we feed the Audio Flamingo 3 model an audio clip with the prompt *"Does this audio contain {prompt}?"* and obtain the *"yes"* probability.

III. Experiments

A. Experimental Setup

Implementation Details. We conduct experiments using two T2A models: EzAudio [10] and TangoFlux [40]. We use pretrained CLAP [36] and Audiobox-Aesthetics [37] as reward models. We compute reward statistics from 48,596 samples in the AudioCaps [41] training split and use the same statistics for both backbones. For the weighted-sum baseline, we fix the PQ weight at 1 and increase the CLAP weight from 2 to 12 in increments of 2, selecting 10:1 as the best-performing ratio. We evaluate on the AudioCaps test split, consisting of 944 audio-caption pairs, using two sampling frameworks, BON and EvoSearch. We set the population size to N=10 with 100 denoising steps for EzAudio and N=20 with 50 denoising steps for TangoFlux. For EvoSearch, we apply evolution at the initial noise step and after 20 denoising steps, using an elite size of 2 and a fixed population size. We report the mean and standard deviation over three runs with distinct random seeds. Experiments are conducted on NVIDIA RTX 4090 GPUs.

Evaluation Metrics. Audio samples are evaluated with various metrics to gauge the performance in different aspects. For generation quality, we use Frechet Distance (FD) and Inception Score (IS) [42]. In addition, three metrics proposed in Audiobox-Aesthetics are measured to further analyze the audio quality: Production Quality (PQ), Content Usage (CU), and Content Enjoyment (CE). We use CLAP similarity as well as the proposed AQAScore for text consistency.

IV. Results

A. Main Results

Limitations on Single-Objective Guidance. We first compare the performance of different rewarding schemes with BON method. As shown in Table I, ITS utilizing a single reward model shows a clear trade-off across different quality attributes. Compared to naive sampling, using the CLAP model as a reward improves the CLAP score by 10.8% but increases the PQ score by only 1.4%. On the other hand, using the PQ model as a single reward improves all Audiobox-Aesthetics metrics by approximately 10%, but marginally improves CLAP metric by 3%. This clearly shows the over-reliance on a single reward model, and the issue is effectively addressed by utilizing multiple reward models. Using rank aggregation and the proposed SCORE, the generated audio improves in both CLAP and PQ, by 9.2% and 8.4%, respectively, while also improving the other metrics. This demonstrates that using multiple rewards in our method effectively mitigates over-specialization to a single reward.

Effectiveness of Standardized Composite Reward. The effect of the proposed SCORE is more pronounced in the EvoSearch framework. As detailed in Table I, SCORE with equal weights on the two reward models ($\alpha = 0.5$) improves CLAP by 10.8% and PQ by 10.6% over naive sampling. Moreover, the improvement is demonstrated beyond the scope of the rewards, resulting in a 10.2% relative increase in CU and 8.8% in CE. We demonstrate the advantage of our method over existing reward aggregation approaches through consistent improvements across most evaluation metrics. We attribute this to

TABLE II: **CMOS evaluations on EzAudio.** Negative values indicate that listeners prefer SCORE over the baseline. All values are presented with 95% confidence intervals.

Comparing Method	Text-Consistency	Audio-Quality
vs. Naive sampling	-0.60 ± 0.63	-0.47 ± 0.36
vs. Rank aggregation	-0.10 ± 0.65	-0.22 ± 0.29

full-precision rewards giving finer guidance than ordinal ranks, and to standardization removing the distribution mismatch that still hampers the weighted summation. The gain is smaller under BON, where SCORE acts only at the final selection; EvoSearch's iterative refinement of intermediate latents lets the composite reward steer generation more strongly.

Evaluation with AQAScore. To evaluate the text-audio alignment without potential bias introduced by CLAP reward, we gauge the performance with the proposed AQAScore. While AQAScore broadly follows the trend of CLAP, it reveals discrepancies. For example, BON with PQ reward attains a lower AQAScore than naive sampling despite a higher CLAP score, serving as evidence of "verifier hacking". Such observations provide an additional perspective on alignment evaluation, and we validate that AQAScore faithfully reflects the effectiveness of SCORE.

Subjective Evaluation. We further evaluate audio quality and text consistency via a comparative MOS (CMOS) study with 20 subjects, each rating SCORE against a baseline on 10 samples.[1] We adopt CMOS to assess the *relative* performance against each baseline, as widely done for generative audio systems [43], [44], where a negative score indicates listeners prefer SCORE. As shown in Table II, listener ratings favor SCORE over naive generation (-0.60 alignment, -0.47 quality) and also over the rank-based aggregation method (-0.10 alignment, -0.22 quality). This corroborates the trend in Table I and shows SCORE is perceptually preferred even over a competitive ITS scaling baseline.

B. Ablation Study

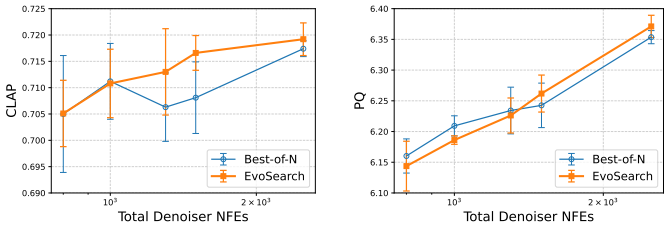
Effectiveness of Combination Weight. We sweep the combination weight α to confirm that, once rewards are standardized, it behaves as a clean preference control rather than a value to be optimized: it trades the two objectives predictably without collapsing either. As shown in Table I, setting $\alpha = 0.25$ to prioritize audio quality significantly boosts all quality-related metrics (PQ, CU, CE) at only a slight cost to textual alignment. Conversely, prioritizing text alignment by setting $\alpha = 0.75$ achieves the highest CLAP score among all methods, while the PQ score remains highly competitive, successfully avoiding the quality collapse observed in the single-objective case. Finally, a balanced weight of $\alpha = 0.5$ produces a faithful equilibrium between the two objectives, as demonstrated by its strong performance across all metrics.

Effect of Number of Function Evaluations. To verify the scalability of the proposed method, we apply SCORE to two ITS methods while progressively increasing the number of function evaluations (NFE). As shown in Fig. 3, both methods achieve higher CLAP and PQ scores as NFE increases. Moreover, the error bars (standard deviation over three seeds) shrink with larger NFE, indicating reduced variability across random seeds and more robust generation under sampling randomness. Overall, these results suggest that SCORE scales reliably with increased inference compute, yielding both improved performance and more stable generations.

[1]All participants provided informed consent prior to participating in the subjective evaluation.

TABLE III: **Performance of TangoFlux on the AudioCaps test set.** Text and perceptual gains denote average relative improvements over naive sampling.

Methods	Setup	Text-Consistency		Perceptual Quality			Improvement (%)	
		CLAP ↑	AQA ↑	PQ ↑	CU ↑	CE ↑	Text	Perceptual
Naive	-	0.62	91.66	5.87	5.18	3.68	–	–
Single reward	C	0.74	93.25	5.87	5.15	3.65	+10.54	-0.46
Single reward	P	0.65	91.96	6.69	5.90	4.04	+2.58	+12.55
SCORE (Ours)	C + P	0.72	93.30	6.49	5.71	3.95	+8.96	+9.38



(a) CLAP score (b) Audiobox-Aesthetics - PQ

Fig. 3: **Scaling performance of SCORE-guided Best-of-N and EvoSearch on EzAudio.** A larger population N yields higher CLAP and PQ scores. Error bars represent the standard deviation across three random seeds.

Cross-Architecture Generalization. To further assess the generalization ability of SCORE, we conduct additional experiments on a different T2A backbone, TangoFlux [40], under the EvoSearch setup. As shown in Table III, single-reward guidance exhibits a clear trade-off between text consistency and perceptual quality. Optimizing CLAP improves text-related metrics but slightly degrades perceptual quality, while optimizing PQ substantially improves perceptual metrics but provides only limited gains in text consistency. In contrast, SCORE achieves balanced improvements in both aspects, yielding gains of $+8.96\%$ in text consistency and $+9.38\%$ in perceptual quality over the naive baseline. These results suggest that SCORE is not specific to EzAudio, but can generalize across T2A architectures.

V. Conclusion

This paper pioneers the extension of Inference-Time Scaling to Text-to-Audio generation and presents SCORE, a standardized composite reward framework. By standardizing and combining multiple reward signals, SCORE balances text-audio alignment and perceptual quality while reducing the risk of "verifier hacking", and the combination weight offers controllable trade-offs between the two. Extensive experiments show consistent improvements across all evaluation metrics, establishing SCORE as a principled and flexible approach for multi-objective guidance in diffusion-based audio generation.

REFERENCES

[1] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi, "Audiogen: Textually guided audio generation," in *Proc. ICLR*, 2023. [Online]. Available: https://openreview.net/forum?id=CYK7RfcOzQ4

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. NeurIPS*, 2020.

[3] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. ICLR*, 2021. [Online]. Available: https://arxiv.org/abs/2010.02502

[4] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *Proc. ICLR*, 2021. [Online]. Available: https://arxiv.org/abs/2011.13456

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR*, 2022.

[6] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022. [Online]. Available: https://arxiv.org/abs/2207.12598

[7] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: text-to-audio generation with latent diffusion models," in *Proc. ICML*, 2023. [Online]. Available: https://proceedings.mlr.press/v202/liu23f.html

[8] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, 2024.

[9] Z. Evans, C. Carr, J. Taylor, S. H. Hawley, and J. Pons, "Fast timing-conditioned latent audio diffusion," in *Proc. ICML*, 2024. [Online]. Available: https://proceedings.mlr.press/v235/evans24a.html

[10] J. Hai, Y. Xu, H. Zhang, C. Li, H. Wang, M. Elhilali, and D. Yu, "Ezaudio: Enhancing text-to-audio generation with efficient diffusion transformer," in *Proc. Interspeech*, 2025.

[11] J. Jung, J. Ahn, C. Jung, T. D. Nguyen, Y. Jang, and J. S. Chung, "Voicedit: Dual-condition diffusion transformer for environment-aware speech synthesis," in *Proc. ICASSP*, 2025.

[12] R. Huang, J. Huang, D. Yang, Y. Ren, L. Liu, M. Li, Z. Ye, J. Liu, X. Yin, and Z. Zhao, "Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models," in *Proc. ICML*, 2023.

[13] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction guided latent diffusion model," in *Proc. ACM MM*, 2023.

[14] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan *et al.*, "Audiobox: Unified audio generation with natural language prompts," *arXiv preprint arXiv:2312.15821*, 2023. [Online]. Available: https://arxiv.org/abs/2312.15821

[15] R. Valle, R. Badlani, Z. Kong, S.-g. Lee, A. Goel, S. Kim, J. F. Santos, S. Dai, S. Gururani, A. Aljafari *et al.*, "Fugatto 1: Foundational generative audio transformer opus 1," in *Proc. ICLR*, 2025, pp. 4325–4353.

[16] S.-g. Lee, Z. Kong, A. Goel, S. Kim, R. Valle, and B. Catanzaro, "Etta: Elucidating the design space of text-to-audio models," in *Proc. ICML*, 2025.

[17] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proc. ICCV*, 2023.

[18] N. Ma, S. Tong, H. Jia, H. Hu, Y.-C. Su, M. Zhang, X. Yang, Y. Li, T. Jaakkola, X. Jia *et al.*, "Scaling inference time compute for diffusion models," in *Proc. CVPR*, 2025.

[19] E. Xie, J. Chen, Y. Zhao, J. Yu, L. Zhu, C. Wu, Y. Lin, Z. Zhang, M. Li, J. Chen *et al.*, "Sana 1.5: Efficient scaling of training-time and inference-time compute in linear diffusion transformer," in *Proc. ICML*, 2025.

[20] S. Lifshitz, S. A. McIlraith, and Y. Du, "Multi-agent verification: Scaling test-time compute with multiple verifiers," in *Proc. COLM*, 2025. [Online]. Available: https://openreview.net/forum?id=LriQ3NY9uL

[21] H. Yang, F. Tang, M. Hu, Q. Yin, Y. Li, Y. Liu, Z. Peng, P. Gao, J. He, Z. Ge *et al.*, "Scalingnoise: Scaling inference-time search for generating infinite videos," *arXiv preprint arXiv:2503.16400*, 2025. [Online]. Available: https://arxiv.org/abs/2503.16400

[22] H. He, J. Liang, X. Wang, P. Wan, D. Zhang, K. Gai, and L. Pan, "Scaling image and video generation via test-time evolutionary search," in *Proc. NeurIPS*, 2025. [Online]. Available: https://openreview.net/forum?id=CFlOUNWsaP

[23] I. Puri, S. Sudalairaj, G. Xu, K. Xu, and A. Srivastava, "Rollout roulette: A probabilistic inference approach to inference-time scaling of llms using particle-based monte carlo methods," in *Proc. NeurIPS*, 2025.

[24] B. Brown, J. Juravsky, R. Ehrlich, R. Clark, Q. V. Le, C. Ré, and A. Mirhoseini, "Large language monkeys: Scaling inference compute with repeated sampling," in *Proc. COLM*, 2025. [Online]. Available: https://openreview.net/forum?id=0xUEBQV54B

[25] C. Snell, J. Lee, K. Xu, and A. Kumar, "Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning," in *Proc. ICLR*, 2025, pp. 10 131–10 165.

[26] Y. Wu, Z. Sun, S. Li, S. Welleck, and Y. Yang, "Inference scaling laws: An empirical analysis of compute-optimal inference for LLM problem-solving," in *Proc. ICLR*, 2025, pp. 55 824–55 845.

[27] S. Li, K. Kallidromitis, A. Gokul, A. Koneru, Y. Kato, K. Kozuka, and A. Grover, "Reflect-DiT: Inference-Time Scaling for Text-to-Image Diffusion Transformers via In-Context Reflection," in *Proc. ICCV*, 2025.

[28] R. Singhal, Z. Horvitz, R. Teehan, M. Ren, Z. Yu, K. McKeown, and R. Ranganath, "A general framework for inference-time scaling and steering of diffusion models," in *Proc. ICML*, 2025. [Online]. Available: https://proceedings.mlr.press/v267/singhal25b.html

[29] F. Liu, H. Wang, Y. Cai, K. Zhang, X. Zhan, and Y. Duan, "Video-t1: Test-time scaling for video generation," in *Proc. ICCV*, 2025.

[30] Y. Jin, Z. Ye, Z. Tian, H. Liu, Q. Kong, Y. Guo, and W. Xue, "Inference-time scaling for diffusion-based audio super-resolution," in *Proc. AAAI*, 2026.

[31] S. Ghosh, A. Goel, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle, and B. Catanzaro, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," in *Proc. NeurIPS*, 2025.

[32] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," in *Proc. ICLR*, 2024, pp. 16 607–16 629.

[33] J. Tian, S.-g. Lee, Z. Kong, S. Ghosh, A. Goel, C.-H. H. Yang, W. Dai, Z. Liu, H. Ye, S. Watanabe *et al.*, "Ualm: Unified audio language model for understanding, generation and reasoning," in *Proc. ICLR*, 2026, pp. 24 563–24 585.

[34] J. Xu, Z. Guo, H. Hu, Y. Chu, X. Wang, J. He, Y. Wang, X. Shi, T. He, X. Zhu *et al.*, "Qwen3-omni technical report," *arXiv preprint arXiv:2509.17765*, 2025. [Online]. Available: https://arxiv.org/abs/2509.17765

[35] Z. Lin, D. Pathak, B. Li, J. Li, X. Xia, G. Neubig, P. Zhang, and D. Ramanan, "Evaluating text-to-visual generation with image-to-text generation," in *Proc. ECCV*, 2024.

[36] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in *Proc. ICASSP*, 2023.

[37] A. Tjandra, Y.-C. Wu, B. Guo, J. Hoffman, B. Ellis, A. Vyas, B. Shi, S. Chen, M. Le, N. Zacharov *et al.*, "Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound," *arXiv preprint arXiv:2502.05139*, 2025. [Online]. Available: https://arxiv.org/abs/2502.05139

[38] L. Eyring, S. Karthik, K. Roth, A. Dosovitskiy, and Z. Akata, "Reno: Enhancing one-step text-to-image models through reward-based noise optimization," in *Proc. NeurIPS*, 2024.

[39] S. Kim, M. Kim, and D. Park, "Test-time alignment of diffusion models without reward over-optimization," in *Proc. ICLR*, 2025, pp. 87 238–87 272.

[40] C.-Y. Hung, N. Majumder, Z. Kong, A. Mehrish, A. A. Bagherzadeh, C. Li, R. Valle, B. Catanzaro, and S. Poria, "Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization," in *Proc. ICLR*, 2026, pp. 150 793–150 816.

[41] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. NAACL*, 2019.

[42] S. Barratt and R. Sharma, "A note on the inception score," *arXiv preprint arXiv:1801.01973*, 2018. [Online]. Available: https://arxiv.org/abs/1801.01973

[43] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023. [Online]. Available: https://arxiv.org/abs/2301.02111

[44] K. Shen, Z. Ju, X. Tan, Y. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers," in *Proc. ICLR*, 2024, pp. 698–722.