ACCELERATING DIFFUSION-BASED TEXT-TO-AUDIO GENERATION WITH CONSISTENCY DISTILLATION

Yatong Bai[*1,2] *Trung Dang*[2] *Dung Tran*[2] *Kazuhito Koishida*[2] *Somayeh Sojoudi*[1]

[1]University of California, Berkeley [2]Applied Sciences Group, Microsoft Corporation

ABSTRACT

Diffusion models power a vast majority of text-to-audio (TTA) generation methods. Unfortunately, these models suffer from slow inference speed due to iterative queries to the underlying denoising network, thus unsuitable for scenarios with inference time or computational constraints. This work modifies the recently proposed consistency distillation framework to train TTA models that require only a single neural network query. In addition to incorporating classifier-free guidance into the distillation process, we leverage the availability of generated audio during distillation training to fine-tune the consistency TTA model with novel loss functions in the audio space, such as the CLAP score. Our objective and subjective evaluation results on the AudioCaps dataset show that consistency models retain diffusion models' high generation quality and diversity while reducing the number of queries by a factor of 400.

***Index Terms*—** Diffusion models, Consistency models, Audio generation, Generative AI, Neural networks

1. INTRODUCTION

Text-to-audio (TTA) generation has recently gained significant popularity [1, 2, 3, 4, 5, 6, 7, 8, 9]. This task involves generating audio based on a user-provided textual prompt. TTA models have rapidly improved and demonstrated the ability to produce diverse, precise, and high-quality audio. Many existing TTA models are based on latent diffusion models (LDM) [10], which have gained popularity in various applications due to their superior generation quality. However, they suffer from slow inference speed as they require iterative queries to the underlying neural network. Such limitations pose challenges in scenarios with time or computation constraints.

This work proposes a novel approach to accelerate diffusion-based TTA models. The proposed method is based on consistency distillation (CD) [11], which distills a pre-trained diffusion model into a consistency model that only requires a single neural network query per generation. Our approach leverages classifier-free guidance (CFG) [12], which has been shown to significantly enhance text-conditioned generative model performance, by incorporating it into the CD process. We explore three different approaches for using CFG: direct guidance, fixed guidance, and variable guidance. To our knowledge, this is the first work to extend CD to CFG models.

Moreover, leveraging the generated audio that is only available during consistency distillation training, we propose fine-tuning the consistency TTA model with audio space loss functions to further improve the audio quality and the audio-text correspondence and use the CLAP score as an example loss function. In contrast, back-propagation from the audio is prohibitively expensive for diffusion models due to the recurrent diffusion process.

Our experiments on the AudioCaps dataset show that the single-step consistency model is comparable with the 400-step diffusion model across five objective metrics, subjective audio quality, and audio-text correspondence. We encourage the reader to listen to our generated examples at `consistency-tta.github.io/demo`.

The paper is structured as follows. Section 2 reviews the related literature, including diffusion models and acceleration techniques. Section 3 outlines our proposed methods to accelerate diffusion-based TTA models. Section 4 discusses our experimental results. Additional discussions and details are presented in the appendix. Throughout this paper, vectors and matrices are denoted as bold symbols whereas scalars use regular symbols.

2. BACKGROUND AND RELATED WORK

2.1. Diffusion models

Diffusion models [13, 14], known for their diverse, high-quality generations, have rapidly gained popularity among various conditional and unconditional generation tasks in vision and audio fields [15, 16, 3, 17]. In the vision domain, while pixel-level diffusion (e.g., EDM [16]) performs well on small image sizes, generating larger images usually requires latent diffusion models (LDMs) [10], where the diffusion process takes place in a latent space. In the audio domain, generative model applications can be further categorized into speech, music, and in-the-wild audio generation. This paper considers the in-the-wild audio setting, where the goal is to generate diverse samples covering a variety of real-world sounds. While some works considered autoregressive models [8] or Mel-space diffusion [9], LDMs have emerged as the dominant TTA approach [1, 2, 3, 4, 5, 6, 7].

The intuition of diffusion models is to gradually recover a clean sample from a noisy sample. During training, Gaussian noise is progressively added to a ground-truth sample $\boldsymbol{z}_0$, forming a continuous diffusion trajectory. At the end of the trajectory, the noisy sample becomes indistinguishable from pure Gaussian noise. This trajectory is then discretized into N time steps, where the noisy sample at each step is denoted as $\boldsymbol{z}_n$ for $n = 1, \ldots, N$. In each training iteration, a random step n is selected, and a Gaussian noise with variance depending on n is injected into the clean sample to produce $\boldsymbol{z}_n$. A denoising neural network, often a U-Net [18], is optimized to recover the noise distribution from the noisy sample. During inference, Gaussian noise is used to initialize the last noisy sample $\hat{\boldsymbol{z}}_N$, where $\hat{\boldsymbol{z}}_n$ denotes the predicted sample at step n. The diffusion model generates a clean sample $\hat{\boldsymbol{z}}_0$ by iteratively querying the denoising network, producing the sequence $\hat{\boldsymbol{z}}_{N-1}, \ldots, \hat{\boldsymbol{z}}_0$.

2.2. Accelerating diffusion model inference

Diffusion models suffer from high generation latency and expensive inference computation due to iterative queries to the denoising network. To this end, several methods have been proposed to reduce

*Work done during internship at Microsoft. Correspondence email `yatong_bai@berkeley.edu`.

the number of model queries. Such methods are mostly presented for image generation tasks and can be grouped into two main categories: improved differential equation solvers and distillation methods.

Improved differential equation solvers can reduce the number of inference steps N of existing diffusion models without additional training. Examples include DDIM [19], Euler [20], Heun, DPM [21, 22], and PNDM [23]. The best solvers can reduce N to 10-50 from the hundreds required by vanilla inference using DDPM [14].

On the other hand, distillation methods, where a pre-trained diffusion model serves as the teacher and a student model is trained to simulate multiple teacher steps in a single step, have been shown to reduce the number of denoising steps to below 10. One representative method is progressive distillation (PD) [24], which iteratively halves the number of diffusion steps. While PD can reduce the number of steps to only a few, the single-step capability is unideal, and the repetitive distillation procedure can be time-consuming. To this end, consistency distillation [11] has been proposed. The training goal of CD is to reconstruct the noiseless image within a single step from an arbitrary step on the teacher model's diffusion trajectory. Note that both PD and CD were proposed for *unconditional* image generation. Recently, consistency models have also been applied to speech and singing voice synthesis [25]. Unlike our work, where the consistency model operates in a latent space, the models in [25] operate on Mel spectrograms and do not consider CFG. For the text-conditioned in-the-wild audio generation task in this work, there are additional considerations, which we discuss in Section 3.

2.3. Classifier-free guidance

CFG [12] is a simple yet effective method for adjusting the text conditioning strength for guided generation problems, significantly improving the performance of existing diffusion-based TTA models. CFG obtains two noise estimations from the denoising network in the diffusion model – one with text conditioning (denoted as $\boldsymbol{v}_{\text{cond}}$) and one without (by masking the text embeddings, denoted as $\boldsymbol{v}_{\text{uncond}}$). The guided estimation, denoted by $\boldsymbol{v}_{\text{cfg}}$, is obtained via

$$\boldsymbol{v}_{\text{cfg}} = w \cdot \boldsymbol{v}_{\text{cond}} + (1 - w) \cdot \boldsymbol{v}_{\text{uncond}}, \tag{1}$$

where the scalar $w \geq 0$ is the guidance strength. When w is between 0 and 1, CFG interpolates the conditioned and unconditioned estimations. When w is greater than 1, CFG becomes an extrapolation. For example, for TANGO, $w = 3$ produces the best overall result [1].

Since CFG is external to the denoising network in diffusion models, it makes distillating guided models more complex than their unguided counterparts. The authors of [26] outlined a two-stage pipeline for performing PD on a CFG classifier. The first stage absorbs CFG into the denoising network by letting the student network take w as an additional input. The second stage performs conventional PD on top of the stage-1 student. During both training stages, the CFG strength w is randomized, and the resulting distilled network allows for selecting w during inference.

3. CONSISTENCY DISTILLATION FOR TTA

We select TANGO [1] as the distillation teacher model due to its high performance. However, we highlight that most of the innovations in this paper can also be applied to other diffusion-based TTA models.

3.1. Overall setup

Similar to TANGO, our model has four components: a conditional U-Net, a text encoder that processes the textual prompt, a VAE encoder-decoder pair that converts the Mel spectrogram to and from the U-Net latent space, and a HiFi-GAN vocoder [27] that produces time-domain audio waveform from the Mel spectrogram. We only train the U-Net and freeze other components.

During training, the Mel spectrogram of the audio is processed by the VAE encoder to produce a latent representation, and the prompt is transformed by the text encoder into a text embedding. They are given to the conditional U-Net as the input and the condition. The VAE decoder and the HiFi-GAN are not used.

During inference, the text embedding is used to guide the U-Net to reconstruct a latent audio representation. The Mel spectrogram and waveform are recovered by the VAE decoder and the HiFi-GAN vocoder, respectively. The VAE encoder is not used.

3.2. Consistency distillation

The goal of CD is to learn a student U-Net $f_{\text{S}}(\cdot, \cdot, \cdot)$ from the diffusion U-Net module in the teacher TTA model $f_{\text{T}}(\cdot, \cdot, \cdot)$. The architecture of f_{S} is the same as the f_{T}, taking three inputs: the noisy latent representation $\boldsymbol{z}_n$, the corresponding time step n, and the text embedding $\boldsymbol{e}_{\text{te}}$. Furthermore, the parameters in f_{S} are initialized using f_{T} information.

The goal for the student U-Net is to generate a realistic latent audio representation within a single forward pass, directly producing an estimated clean example $\hat{\boldsymbol{z}}_0$ based on $\boldsymbol{z}_n$, where $n \in \{0, \ldots, N\}$ is an arbitrary step along the diffusion trajectory [11, Algorithm 2]. The risk function to be minimized for achieving this goal is

$$\mathbb{E}_{\substack{(\boldsymbol{z}_0, \boldsymbol{e}_{\text{te}}) \sim \mathcal{D} \\ n \sim \text{U}_{\text{int}}(1, N)}} \Big[d\Big(f_{\text{S}}(\boldsymbol{z}_n, n, \boldsymbol{e}_{\text{te}}), f_{\text{S}}(\hat{\boldsymbol{z}}_{n-1}, n-1, \boldsymbol{e}_{\text{te}}) \Big) \Big], \tag{2}$$

where $d(\cdot, \cdot)$ is a distance measurement, $\mathcal{D}$ is the training dataset, $\text{U}_{\text{int}}(1, N)$ denotes the discrete uniform distribution supported over the set $\{1, \ldots, N\}$, and $\hat{\boldsymbol{z}}_{n-1} = \text{solve} \circ f_{\text{T}}(\boldsymbol{z}_n, n, \boldsymbol{e}_{\text{te}})$ is the teacher diffusion model's estimation for $\boldsymbol{z}_{n-1}$. Here, $\text{solve} \circ f_{\text{T}}$ denotes the composite function of the teacher denoising U-Net and the solver that converts the U-Net raw output to the estimation of the previous time step. We use the ℓ_2 distance in this latent space as $d(\cdot, \cdot)$, with additional discussions in Appendix A.3. Intuitively, this risk measures the expected distance between the student's reconstructions from two adjacent time steps on the diffusion trajectory.

The authors of [11] used the Heun solver for querying the teacher diffusion model during distillation and adopted "Karras noise schedule", a discretization scheme that unevenly selects the time steps on the diffusion trajectory. In Section 4, we empirically investigate multiple solvers and noise schedules.

3.3. Consistency distillation with classifier-free guidance

Since CFG is crucial to the conditional generation quality, we consider three methods for incorporating it into the distilled model.

Direct Guidance directly performs CFG on the consistency model output by applying (1). Since this method naïvely extrapolates or interpolates on the consistency model $\boldsymbol{z}_0$ prediction, the CFG operation will likely move the prediction outside the manifold of realistic latent representations.

Fixed Guidance Distillation aims to distill from the diffusion model coupled with CFG using a fixed guidance strength w. Specifically, the training risk function is still (2), but $\hat{\boldsymbol{z}}_{n-1}$ is replaced with the estimation after CFG. Now, $\hat{\boldsymbol{z}}_{n-1}$ becomes $\text{solve} \circ f_{\text{T}}^{\text{cfg}}(\boldsymbol{z}_n, n, \boldsymbol{e}_{\text{te}}, w)$, where the guided teacher output $f_{\text{T}}^{\text{cfg}}$ is

$$f_{\text{T}}^{\text{cfg}}(\boldsymbol{z}_n, n, \boldsymbol{e}_{\text{te}}, w) = \\ w \cdot f_{\text{T}}(\boldsymbol{z}_n, n, \varnothing) + (1 - w) \cdot f_{\text{T}}(\boldsymbol{z}_n, n, \boldsymbol{e}_{\text{te}}),$$

Table 1. Ablate various guidance weights, distillation techniques, solvers, noise schedules, training lengths, loss weights, and initialization. "CFG w" represents the guidance weight; "# queries" indicates the number of neural network queries during inference. U-Net modules have 557M parameters, except in variable guidance models (559M). Distillation runs are 40 epochs; inference uses FP32 precision.

# queries (↓)	Solver	Noise schedule	CFG w	Guidance method	Min-SNR	Initialization	FAD (↓)	FD (↓)	KLD (↓)
1	DDIM	Uniform	1	-	✗	Unguided	13.48	45.75	2.409
	Heun	Karras					10.97	50.19	2.425
2	DDIM	Uniform	3	Direct guidance	✗	Unguided	8.565	38.67	2.015
	Heun	Karras					7.421	39.36	1.976
1	Heun	Karras	3	Fixed guidance distillation	✗	Unguided	5.702	33.18	1.494
		Uniform			✗	Unguided	4.168	28.54	1.384
		Uniform			✓	Unguided	3.766	27.74	1.443
		Uniform			✗	Guided	3.859	27.79	1.421
1	Heun	Uniform	3	Variable guidance distillation	✓	Guided	3.956	28.27	1.442
			4				3.180	27.92	1.394
			6				2.975	28.63	1.378

with $\varnothing$ denoting the masked language token. Here, w should be fixed to the value corresponding to the best teacher generation quality.

Variable Guidance Distillation is similar to fixed guidance distillation, but with randomized guidance strength w during distillation, so that w can be adjusted during inference. To make the student network compatible with adjustable w, we add a w-encoding condition branch to f_{S} (which now has four inputs). We use Fourier encoding for w following [26] and merge the embedding into f_{S} similarly to the time step embedding. Each training iteration samples a random guidance strength w via the uniform distribution supported on $[0, 6)$.

The latter two methods are related to the two-stage distillation procedure outlined in [26], with details described in Appendix A.2.

3.4. Min-SNR training loss weighing strategy

The literature has proposed to improve diffusion models by using the truncated signal-noise ratio (SNR) to weigh the training loss at each time step n for diffusion models, and the Min-SNR strategy [28] is one of the latest examples. The specific calculation of Min-SNR depends on the parameterization of the diffusion model. Specifically, diffusion models can be trained to predict the clean example $\boldsymbol{z}_0$, the additive noise $\boldsymbol{\epsilon}$, or the noise velocity $\boldsymbol{v}$. The Min-SNR weighting formulation is different for the three parameterizations.

This work investigates whether the Min-SNR strategy also improves CD. Since consistency models predict the clean sample $\boldsymbol{z}_0$, we use the Min-SNR formulation for $\boldsymbol{z}_0$-predicting diffusion models, which is $\omega(n) = \min\{\mathrm{SNR}(t_n), \gamma\}$, where $\omega(n)$ is the loss weight for the n^{th} time step, $\mathrm{SNR}(t)$ is the SNR at time t, t_n is the time corresponding to the n^{th} time step, and γ is a constant defaulted to 5. For the Heun solver used in most of our experiments, $\mathrm{SNR}(t)$ is the inverse of the additive Gaussian noise variance at time t.

3.5. End-to-end fine-tuning with CLAP

Since our consistency TTA model produces audio in a single neural network query, we can optimize auxiliary losses operating in the audio space along with the latent-space CD loss to improve the audio quality and semantics. On the contrary, since a diffusion model has an iterative inference process, optimizing such a model by back-propagating from the audio resembles the training of a recurrent neural network, which is known to be expensive and challenging. This work uses the CLAP score [29] as an example of fine-tuning loss function. The CLAP score, denoted by CS, is defined as:

$$\mathrm{CS}(\hat{\boldsymbol{x}}, \boldsymbol{x}) = \max\left\{100 \times \frac{\mathbf{e}_{\hat{\boldsymbol{x}}} \cdot \mathbf{e}_{\boldsymbol{x}}}{\|\mathbf{e}_{\hat{\boldsymbol{x}}}\| \cdot \|\mathbf{e}_{\boldsymbol{x}}\|}, 0\right\}, \tag{3}$$

where $\hat{\boldsymbol{x}}$ is the generated audio waveform, $\boldsymbol{x}$ is the reference (ground-truth waveform or textual prompt), and $\mathbf{e}_{\hat{\boldsymbol{x}}}$ and $\mathbf{e}_{\boldsymbol{x}}$ are the corresponding embeddings extracted by the CLAP model.

We select the CLAP score due to its superior embedding quality arising from the diverse training tasks and datasets, as well as its consideration of audio-text correspondence. Since the CD training loss (2) does not use ground truth information, optimizing this score provides valuable feedback to the consistency model.

4. EXPERIMENTS

4.1. Dataset and experiment settings

The experiments in this work use AudioCaps [30], a popular dataset for in-the-wild audio generation. AudioCaps is a collection of human-captioned YouTube audio, each instance having a length of at most ten seconds. Our AudioCaps copy contains 882 test instances and 45,260 training instances. Like several existing works [1, 3], the core generative U-Net of our models is trained only on the AudioCaps training set, leaving larger datasets for future work.

While we explicitly use TANGO [1] as the baseline, our methods apply to diffusion-based TTA models in general. We select FLAN-T5-Large [31] as the text encoder and use the same checkpoint as [1]. For the VAE and the HiFi-GAN, we use the checkpoint pretrained on AudioSet released by the authors of [3] as in [1]. For faster training and inference, we shrink the U-Net from 866M parameters used in [1] to 557M. All consistency models are distilled from this smaller TANGO model, which performs similarly to the checkpoint from [1] (Table 2). Additional details about our model, training setup, and evaluation are shown in Appendix A.3 and A.4.

4.2. Objective evaluation results

Our objective evaluation considers five metrics: FAD, FD, KLD, $\mathrm{CLAP_A}$, and $\mathrm{CLAP_T}$. Specifically, FAD is the Fréchet distance between generated and ground-truth audio embeddings extracted by the VGGish model [32], FD is the Fréchet distance between the embeddings extracted by PANN [33], and KLD is Kullback-Leibler divergence between the PANN embeddings. $\mathrm{CLAP_A}$ and $\mathrm{CLAP_T}$ are the CLAP scores with respect to the ground-truth audio waveform and the textual prompt. We use the CLAP checkpoint from [34] trained on LAION-Audio-630k [34], AudioSet [35], and music data.

We first ablate the performance of the consistency TTA generation model under various training settings, with the results presented in Table 1. Note that "guided initialization" refers to initializing the consistency model with a guidance-aware diffusion model, whereas "unguided initialization" refers to initializing with the unmodified

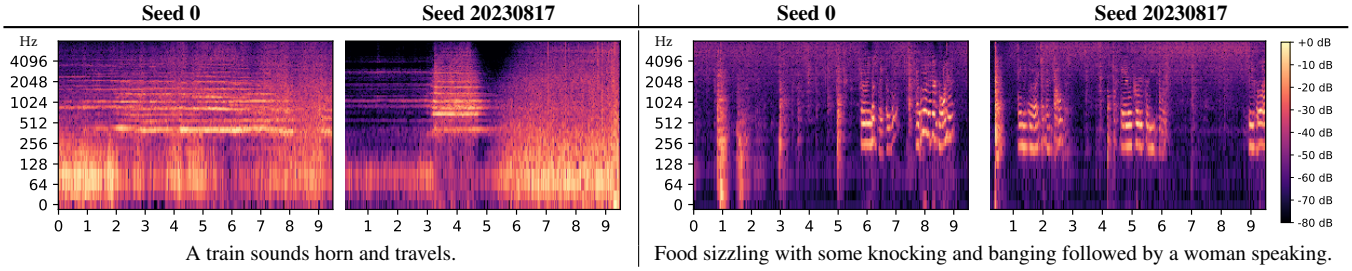
Table 2. Compare consistency models to the diffusion baselines. Distillation runs are extended to 60 epochs for better performance; CLAP-fine-tuning uses 10 additional epochs. All CD runs use the Heun teacher solver, uniform noise schedule, variable guidance distillation, guided initialization, Min-SNR loss weights, and BF16 inference precision. Bold numbers indicate the best results among consistency models.

	U-Net # params	CFG w	# queries (↓)	CLAP_T (↑)	CLAP_A (↑)	FAD (↓)	FD (↓)	KLD (↓)
AudioLDM-L reported in [3]	739M	2	400	-	-	2.08	27.12	1.86
TANGO reported in [1]	866M	3		-	-	1.59	24.53	1.37
TANGO [1] tested by us	866M	3		24.10	72.85	1.631	20.11	1.362
Our TANGO model	557M	3		24.57	72.79	1.908	19.57	1.350
Consistency model without CLAP fine-tuning	559M	3	1	21.00	71.39	3.202	22.04	1.411
		4		22.05	72.08	2.610	21.71	1.373
		5		22.50	72.30	2.575	22.08	**1.354**
Consistency model with CLAP fine-tuning	559M	3	1	24.44	72.39	**2.182**	**20.44**	1.368
		4		24.69	**72.54**	2.406	20.97	1.358
		5		**24.70**	72.53	2.626	21.33	1.356

Table 3. Compare the human evaluation results of consistency and diffusion models. Bold numbers are defined same as Table 2.

	U-Net # params	# queries (↓)	CLAP fine-tuning	CFG w	Human Quality (↑)	Human Corresp (↑)	CLAP_T (↑)	CLAP_A (↑)	FAD (↓)	FD (↓)	KLD (↓)
Diffusion	557M	400	✗	3	4.136	4.064	24.57	72.79	1.908	19.57	1.350
Consistency	559M	1	✗	5	**3.902**	4.010	22.50	72.30	2.575	22.08	**1.354**
			✓	4	3.830	**4.064**	**24.69**	**72.54**	**2.406**	**20.97**	1.358
Ground-truth audio	-	-	-	-	4.424	4.352	26.71	100.0	0.000	0.000	0.000

Table 4. The generated audio noticeably varies with different random seeds. The horizontal axis is time in seconds.



TANGO teacher weights. Table 1 demonstrates that distilling with fixed or variable guidance significantly improves the performance over direct or no guidance. In terms of the teacher solver used during distillation, with $N = 18$ discretization steps as in [11], the more accurate Heun solver is advantageous over the simpler DDIM solver. Moreover, the uniform noise schedule is preferred over the Karras schedule (see Appendix A.1 for a detailed discussion). We also observe that the Min-SNR weights and the guided initialization improve the FD and FAD but slightly sacrifice the KLD.

Table 2 compares the consistency TTA models with the diffusion baseline models. On top of the best consistency model, we perform end-to-end CLAP fine-tuning, co-optimizing three loss components: the consistency loss (2), CLAP_A, and CLAP_T. Table 2 demonstrates that fine-tuning further improves all objective metrics except KLD. Furthermore, the gap between the best consistency and diffusion models is small for all quality metrics, with the FD and KLD even surpassing the reported numbers from [1] and [3].

Note that the diffusion baseline models use 200 steps following [3, 1], each step requiring two noise estimations due to CFG, amounting to 400 total network queries per generation. Thus, with minimal performance drop, the proposed consistency model reduces the number of U-Net queries by a factor of 400.

4.3. Subjective evaluation results

Finally, we conduct subjective evaluations in two aspects: overall audio quality and audio-text correspondence. For each subject, we use 25 generated audio clips from the same set of prompts together with those from ground-truth samples. We instructed 20 evaluators to rate the audio clips on a scale of 1 to 5 for each aspect. Other details can be found in Appendix A.4. We further confirm that the consistency model produces audios close to those of the diffusion model in terms of subjective evaluation scores. Moreover, optimizing the CLAP scores improves the text-audio correspondence score, which supports our assumption that CLAP_T provides closed-loop feedback to help align the generated audio with the prompt.

4.4. Diversity of generated audio

We also observe that different random seeds, i.e., different initial Gaussian latent for the consistency TTA model, generate noticeably different audio, confirming that consistency models produce diverse generations like diffusion models. We present two example prompts from the CLAP-finetuned model in Table 4 to illustrate this diversity.

5. CONCLUSION

This work proposes an approach to accelerate the core module of diffusion-based TTA models hundreds of times based on consistency distillation. The delicate distillation procedure that emphasizes CFG achieves this vast acceleration with minimal generation quality reduction, enabling diverse and realistic in-the-wild audio generation within one neural network query. The differentiability of the resulting model allows for end-to-end fine-tuning, unlocking possibilities for further improving the training method of such models.

6. REFERENCES

[1] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," *arXiv preprint arXiv:2304.13731*, 2023.

[2] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu, "Diffsound: Discrete diffusion model for text-to-sound generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[3] 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," *arXiv preprint arXiv:2301.12503*, 2023.

[4] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley, "AudioLDM 2: Learning holistic audio generation with self-supervised pretraining," *arXiv preprint arXiv:2308.05734*, 2023.

[5] 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," *arXiv preprint arXiv:2301.12661*, 2023.

[6] Jiawei Huang, Yi Ren, Rongjie Huang, Dongchao Yang, Zhenhui Ye, Chen Zhang, Jinglin Liu, Xiang Yin, Zejun Ma, and Zhou Zhao, "Make-an-Audio 2: Temporal-enhanced text-to-audio generation," *arXiv preprint arXiv:2305.18474*, 2023.

[7] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal, "Any-to-any generation via composable diffusion," *arXiv preprint arXiv:2305.11846*, 2023.

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

[9] Seth Forsgren and Hayk Martiros, "Riffusion - stable diffusion for real-time music generation," *URL https://riffusion.com*, 2022.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[11] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever, "Consistency models," in *International Conference on Machine Learning*, 2023.

[12] Jonathan Ho and Tim Salimans, "Classifier-free diffusion guidance," in *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021.

[13] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, 2015.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, 2020.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[16] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine, "Elucidating the design space of diffusion-based generative models," in *Advances in Neural Information Processing Systems*, 2022.

[17] Qingqing Huang, Daniel S Park, Tao Wang, Timo I Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, et al., "Noise2Music: Text-conditioned music generation with diffusion models," *arXiv preprint arXiv:2302.03917*, 2023.

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[19] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[20] Leonhard Euler, *Institutionum calculi integralis*, vol. 1, impensis Academiae imperialis scientiarum, 1824.

[21] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu, "DPM-solver: A fast ODE solver for diffusion probabilistic model sampling in around 10 steps," in *Advances in Neural Information Processing Systems*, 2022.

[22] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu, "DPM-solver++: Fast solver for guided sampling of diffusion probabilistic models," *arXiv preprint arXiv:2211.01095*, 2022.

[23] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao, "Pseudo numerical methods for diffusion models on manifolds," in *International Conference on Learning Representations*, 2022.

[24] Tim Salimans and Jonathan Ho, "Progressive distillation for fast sampling of diffusion models," in *International Conference on Learning Representations*, 2021.

[25] Zhen Ye, Wei Xue, Xu Tan, Jie Chen, Qifeng Liu, and Yike Guo, "Comospeech: One-step speech and singing voice synthesis via consistency model," *arXiv preprint arXiv:2305.06908*, 2023.

[26] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans, "On distillation of guided diffusion models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

[27] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae, "HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis," in *Advances in Neural Information Processing Systems*, 2020.

[28] Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo, "Efficient diffusion training via min-snr weighting strategy," *arXiv preprint arXiv:2303.09556*, 2023.

[29] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "CLAP: learning audio concepts from natural language supervision," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[30] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "AudioCaps: Generating captions for audios in the wild," in *Conference of the North American Chapter of the Association for Computational Linguistics*, 2019.

[31] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al., "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[32] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "CNN architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[33] 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*, vol. 28, pp. 2880–2894, 2020.

[34] 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*, 2023.

[35] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017.

[36] Brian McFee, "ResamPy: efficient sample rate conversion in python," *Journal of Open Source Software*, vol. 1, no. 8, pp. 125, 2016.

[37] Yao-Yuan Yang, Moto Hira, Zhaoheng Ni, Anjali Chourdia, Artyom Astafurov, Caroline Chen, Ching-Feng Yeh, Christian Puhrsch, David Pollack, Dmitriy Genzel, Donny Greenberg, Edward Z. Yang, Jason Lian, Jay Mahadeokar, Jeff Hwang, Ji Chen, Peter Goldsborough, Prabhat Roy, Sean Narenthiran, Shinji Watanabe, Soumith Chintala, Vincent Quenneville-Bélair, and Yangyang Shi, "TorchAudio: Building blocks for audio and speech processing," *arXiv preprint arXiv:2110.15018*, 2021.

A. ADDITIONAL DISCUSSIONS AND DETAILS

A.1. Additional discussions regarding teacher solver

Table 1 presents the generation quality of the consistency model f_{S} distilled with various solver settings, confirming our selection of the Heun solver and the uniform schedule. While [11] agreed that the Heun solver achieved better results, the authors suggested using the Karras schedule. Our explanation of this discrepancy is that TANGO was trained using the uniform schedule, whereas the teacher models in [11] were trained with the Karras schedule. It is likely beneficial to use the same scheduler to train the diffusion model.

We also compared the TANGO inference quality when coupled with various solvers (fixing $N = 18$ inference steps), including DDPM, DDIM, Euler, Heun, and DPM++(2S), and confirmed that the Heun solver with a uniform schedule yielded the best FAD and FD metrics. DPM++(2S) achieved a better KLD, but the difference is negligible.

A.2. Relationship to two-stage progressive distillation

Unlike PD in [26], which requires iteratively halving the number of diffusion steps, CD in our method reduces the required inference step to one within a single training process. As a result, the two distillation stages proposed in [26] can be merged. Specifically, stage-2 distillation can be performed without stage 1, provided that the step of querying the stage-1 model is replaced by querying the original teacher model with CFG. Merging stage 1 and stage 2 then results in our "variable guidance distillation" method discussed in Section 3.3. Subsequently, stage 1 becomes optional since it only serves to provide a guidance-aware initialization to stage 2.

A.3. Model and training details

We noticed that the audio resampling implementation has a major influence on some metrics, with FAD being especially sensitive. To ensure high training quality and fair evaluation, we use ResamPy [36] for all resampling procedures unless the resampling step needs to be differentiable. Specifically, CLAP fine-tuning requires differentiable resampling, and we use TorchAudio [37] instead.

The structure of our 557M-parameter U-Net is similar to the 866M U-Net used in [1], with the only modification being reducing the "block out channels" from $(320, 640, 1280, 1280)$ to $(256, 512, 1024, 1024)$. All CD runs use two 48GB-VRAM GPUs, with a total batch size of 12 and five gradient accumulation steps. The optimizer is AdamW with a 10^{-4} weight decay, and the learning rate is 10^{-5} for CD and 10^{-6} for CLAP fine-tuning. The "CD target network" (see [11] for details) is an exponential model average (EMA) copy with a 0.95 decay rate. We also maintain an EMA copy with a 0.999 decay rate for the reported experiment results. All training uses BF16 numerical precision.

Regarding the distance measure $d(\cdot, \cdot)$ introduced in (2), the authors of [11] considered several options for $d(\cdot, \cdot)$ for image generation tasks and concluded that using LPIPS (an evaluation metric that embeds the generated image and calculates the weighted distance in several feature spaces) as the optimization objective produced higher generation quality than using the pixel-level ℓ_2 or ℓ_1 distance. However, since our latent diffusion model already operates in a latent feature space, we simply use the ℓ_2 distance in this latent space.

A.4. Evaluation details

While the maximal audio length of the AudioCaps dataset is 10.00 seconds and the U-Net module of the TTA models is trained to generate 10.00-second latent audio representations, the HiFi-GAN vocoder produces 10.24-second audio. We observe that this mismatch negatively impacts the generation quality. Specifically, the final 0.24 seconds of the generated audio is empty, and there are slight distorting artifacts near the end of the 10-second useful portion. To this end, for the objective evaluation results in Tables 2 and 3, we truncate the generated audio to 9.70 seconds. Table 1 uses the full 10.24 seconds. The ground-truth reference waveforms are not truncated.

The human evaluation results in Table 3 are based on 20 evaluators each rating 25 audio clips per model, forming 500 samples per model. For each evaluator, the three models and the ground truth use the same set of prompts (the prompts vary across evaluators). Each evaluator rates each audio on a scale of 1-5, with rating criteria defined in the evaluation form. To ensure evaluation fairness, the model type generating each waveform is not disclosed to the evaluator, and the generations of the models are shuffled. We find it extremely challenging for a human to distinguish the outputs from the three generative models, with many ground truth waveforms also indistinguishable. An example evaluation form is available at `consistency-tta.github.io/evaluation`.

B. ACKNOWLEDGMENTS

We sincerely appreciate the contributions to human evaluation from Chih-Yu Lai, Mo Zhou, Afrina Tabassum, You Zhang, Sara Abdali, Uros Batricevic, Yinheng Li, Asing Chang, Rogerio Bonatti, Sam Pfrommer, Ziye Ma, Tanvir Mahmud, Eli Brock, Tanmay Gautam, Jingqi Li, Brendon Anderson, Hyunin Lee, and Saeed Amizadeh.