Text-to-Audio Generation using Instruction-Tuned LLM and Latent Diffusion Model

Deepanway Ghosal[‡], **Navonil Majumder**[‡], **Ambuj Mehrish**[‡], **Soujanya Poria**[‡]
[‡] DeCLaRe Lab, Singapore University of Technology and Design, Singapore
deepanway_ghosal@mymail.sutd.edu.sg
{navonil_majumder,ambuj_mehrish,sporia}@sutd.edu.sg

: https://github.com/declare-lab/tango
: https://tango-web.github.io/

Abstract

The immense scale of the recent large language models (LLM) allows many interesting properties, such as, instruction- and chain-of-thought-based fine-tuning, that has significantly improved zero- and few-shot performance in many natural language processing (NLP) tasks. Inspired by such successes, we adopt such an instruction-tuned LLM FLAN-T5 as the text encoder for text-to-audio (TTA) generation—a task where the goal is to generate an audio from its textual description. The prior works on TTA either pre-trained a joint text-audio encoder or used a non-instruction-tuned model, such as, T5. Consequently, our latent diffusion model (LDM)-based approach (TANGO) outperforms the state-of-the-art AudioLDM on most metrics and stays comparable on the rest on AudioCaps test set, despite training the LDM on a 63 times smaller dataset and keeping the text encoder frozen. This improvement might also be attributed to the adoption of audio pressure level-based sound mixing for training set augmentation, whereas the prior methods take a random mix.

Preprint. Under review.

1 Introduction

Following the success of automatic text-to-image (TTI) generation [28–30], many researchers have also succeeded in text-to-audio (TTA) generation [16, 17, 38] by employing similar techniques as the former. Such models may have strong potential use cases in the media production where the creators are always looking for novel sounds that fit their creations. This could be especially useful in prototyping or small-scale projects where producing the exact sound could be infeasible. Beyond this, these techniques also pave the path toward general-purpose multimodal AI that can simultaneously recognize and generate multiple modalities.

To this end, the existing works use a large text encoder, such as, `RoBERTa` [18] and `T5` [27], to encode the textual description of the audio to be generated. Subsequently, a large transformer decoder or a diffusion model generates the audio prior, which is subsequently decoded by a pre-trained VAE, followed by a vocoder. We instead assume that replacing the text encoder with an instruction-tuned large language model (LLM) would improve text understanding and overall audio generation without any fine-tuning, due to its recently discovered gradient-descent mimicking property [3]. To augment training samples, the existing methods take a randomly generated combination of audio pairs, along with the concatenation of their descriptions. Such a mixture does not account for the overall pressure level of the source audios, potentially leading to a louder audio overwhelming the quieter one. Thus, we employ a pressure level-based mixing method, as suggested by Tokozume et al. [35].

Our model (TANGO) [1] is inspired by latent diffusion model (LDM) [30] and AudioLDM [17] models. However, instead of using CLAP-based embeddings, we used a large language model (LLM) due to its powerful representational ability and fine-tuning mechanism, which can help learn complex concepts in the textual description. Our experimental results show that using an LLM greatly improves text-to-audio generation and outperforms state-of-the-art models, even when using a significantly smaller dataset. In the image generation literature, the effects of LLM has been studied before by Saharia et al. [32]. However, they considered T5 as the text encoder which is not pre-trained on instruction-based datasets. FLAN-T5 [2] is initialized with a T5 checkpoint and fine-tuned on a dataset of 1.8K NLP tasks in terms of instructions and chain-of-thought reasoning. By leveraging instruction-based tuning, FLAN-T5 has achieved state-of-the-art performance on several NLP tasks, matching the performance of LLMs with billions of parameters.

In Section 3, we empirically show that TANGO outperforms AudioLDM and other baseline approaches on most of the metrics on AudioCaps test set under both objective and subjective evaluations, despite training the LDM on a 63 times smaller dataset. We believe that if TANGO is trained on a larger dataset such as AudioSet (as Liu et al. [17] did), it would be able to provide even better results and improve its ability to recognize a wider range of sounds.

The overall contribution of this paper is threefold:

1. We do not use any joint text-audio encoder—such as CLAP—for guidance. Liu et al. [17] claim that CLAP-based audio guidance is necessary during training for better performance. We instead use a frozen instruction-tuned pre-trained LLM FLAN-T5 with strong text representation capacity for text guidance in both training and inference.

2. AudioLDM needed to fine-tune RoBERTa [18] text encoder to pre-train CLAP. We, however, keep FLAN-T5 text encoder frozen during LDM training. Thus, we find that LDM itself is capable of learning text-to-audio concept mapping and composition from a 63 times smaller training set, as compared to AudioLDM, given an instruction-tuned LLM.

3. To mix audio pairs for data augmentation, inspired by Tokozume et al. [35], we consider the pressure levels of the audio pairs, instead of taking a random combination as the prior works like AudioLDM. This ensures good representations of both source audios in the fused audio.

[1] The acronym TANGO stands for `Text-to-Audio using iNstruction Guided diffusiOn` and was suggested by `ChatGPT`. The word TANGO is often associated with music [37] and dance [36]. According to Wikipedia [36], "Tango is a partner dance and social dance that originated in the 1880s along the Río de la Plata, the natural border between Argentina and Uruguay." The image above resembles the TANGO dance form and was generated by prompting `Dalle-V2` with `"A couple dancing tango with musical notes in the background"`

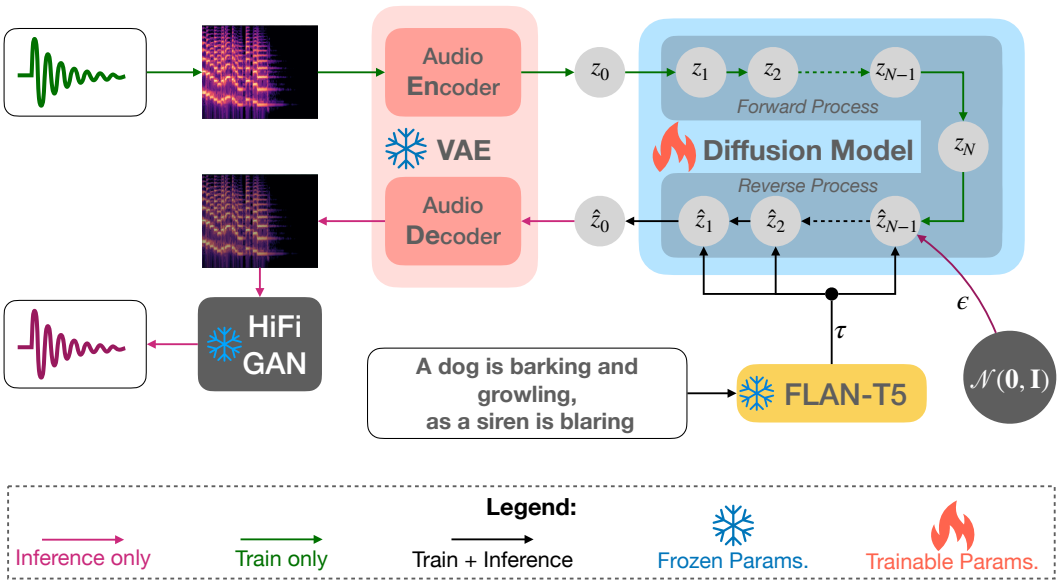

Figure 1: Overall architecture of TANGO.

2 Method

TANGO, as depicted in Fig. 1, has three major components: i) textual-prompt encoder, ii) latent diffusion model (LDM), and iii) mel-spectogram/audio VAE. The textual-prompt encoder encodes the input description of the audio. Subsequently, the textual representation is used to construct a latent representation of the audio or audio prior from standard Gaussian noise, using reverse diffusion. Thereafter the decoder of the mel-spectogram VAE constructs a mel-spectogram from the latent audio representation. This mel-spectogram is fed to a vocoder to generate the final audio.

2.1 Textual-Prompt Encoder

We use the pre-trained LLM FLAN-T5-LARGE (780M) [2] as the text encoder (E_{text}) to obtain text encoding $\tau \in \mathbb{R}^{L \times d_{text}}$, where L and d_{text} are the token count and token-embedding size, respectively. Due to the pre-training of FLAN-T5 models on a large-scale chain-of-thought- (CoT) and instruction-based dataset, Dai et al. [3] posit that they are able to learn a new task very well from the in-context information by mimicking gradient descent through attention weights. This property is missing in the older large models, such as `RoBERTa` [18] (used by Liu et al. [17]) and `T5` [27] (used by Kreuk et al. [16]). Considering each input sample a distinct task, it might be reasonable to assume that the gradient-descent mimicking property could be pivotal in learning the mapping between textual and acoustic concepts without fine-tuning the text encoder. The richer pre-training may also allow the encoder to better emphasize the key details with less noise and enriched context. This again may lead to the better transformation of the relevant textual concepts into their acoustics counterparts. Consequently, we keep the text encoder frozen, assuming the subsequent reverse diffusion process (see Section 2.2) would be able to learn the inter-modality mapping well for audio prior to construction. We also suspect that fine-tuning E_{text} may degrade its in-context learning ability due to gradients from the audio modality that is out of distribution to the pre-training dataset. This is in contrast with Liu et al. [17] that fine-tunes the pre-trained text encoder as a part of the text-audio joint-representation learning (CLAP) to allow audio prior reconstruction from text. In Section 3, we empirically show that such joint-representation learning may not be necessary for text-to-audio transformation.

2.2 Latent Diffusion Model for Text-Guided Generation

The latent diffusion model (LDM) [30] is adapted from Liu et al. [17], with the goal to construct the audio prior z_0 (see Section 2.5) with the guidance of text encoding τ. This essentially reduces to approximating the true prior $q(z_0|\tau)$ with parameterized $p_\theta(z_0|\tau)$.

LDM can achieve the above through forward and reverse diffusion processes. The forward diffusion is a Markov chain of Gaussian distributions with scheduled noise parameters $0 < \beta_1 < \beta_2 < \cdots < \beta_N < 1$ to sample noisier versions of z_0:

$$q(z_n|z_{n-1}) = \mathcal{N}(\sqrt{1-\beta_n}z_{n-1}, \beta_n\mathbf{I}), \tag{1}$$

$$q(z_n|z_0) = \mathcal{N}(\sqrt{\overline{\alpha}_n}z_0, (1-\overline{\alpha}_n)\mathbf{I}), \tag{2}$$

where N is the number of forward diffusion steps, $\alpha_n = 1 - \beta_n$, and $\overline{\alpha}_n = \prod_{i=1}^{n} \alpha_n$. Song et al. [34] show that Eq. (2) conveniently follows from Eq. (1) through reparametrization trick that allows direct sampling of any z_n from z_0 via a non-Markovian process:

$$z_n = \sqrt{\overline{\alpha}_n}z_0 + (1-\overline{\alpha}_n)\epsilon, \tag{3}$$

where the noise term $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The final step of the forward process yields $z_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

The reverse process denoises and reconstructs z_0 through text-guided noise estimation ($\hat{\epsilon}_\theta$) using loss

$$\mathcal{L}_{DM} = \sum_{n=1}^{N} \gamma_n \mathbb{E}_{\epsilon_n \sim \mathcal{N}(\mathbf{0},\mathbf{I}), z_0} ||\epsilon_n - \hat{\epsilon}_\theta^{(n)}(z_n, \tau)||_2^2, \tag{4}$$

where z_n is sampled from Eq. (3) using standard normal noise ϵ_n, τ is the text encoding (see Section 2.1) for guidance, and γ_n is the weight of reverse step n [5], taken to be a measure of signal-to-noise ratio (SNR) in terms of $\alpha_{1:N}$. The estimated noise is used to reconstruct z_0:

$$p_\theta(z_{0:N}|\tau) = p(z_N) \prod_{n=1}^{N} p_\theta(z_{n-1}|z_n, \tau), \tag{5}$$

$$p_\theta(z_{n-1}|z_n, \tau) = \mathcal{N}(\mu_\theta^{(n)}(z_n, \tau), \tilde{\beta}^{(n)}), \tag{6}$$

$$\mu_\theta^{(n)}(z_n, \tau) = \frac{1}{\sqrt{\alpha_n}}[z_n - \frac{1-\alpha_n}{\sqrt{1-\overline{\alpha}_n}}\hat{\epsilon}_\theta^{(n)}(z_n, \tau)], \tag{7}$$

$$\tilde{\beta}^{(n)} = \frac{1-\bar{\alpha}_{n-1}}{1-\bar{\alpha}_n}\beta_n. \tag{8}$$

The noise estimation $\hat{\epsilon}_\theta$ is parameterized with U-Net [31] with a cross-attention component to include the text guidance τ. In contrast, AudioLDM [17] uses audio as the guidance during training. During inference, they switch back to text guidance, as this is facilitated by pre-trained joint text-audio embedding (CLAP). We did not find audio-guided training and pre-training CLAP to be necessary, as argued in Section 2.1.

2.3 Augmentation

Many text-to-image [25] and text-to-audio [16] works have shown the efficacy of training with fusion-based augmented samples to improve cross-modal concept-composition abilities of the diffusion network. Therefore, we synthesize additional text-audio pairs by superimposing existing audio pairs on each other and concatenating their captions.

Unlike Liu et al. [17] and Kreuk et al. [16], to mix audio pairs, we do not take a random combination of them. Following Tokozume et al. [35], we instead consider the human auditory perception for fusion. Specifically, the audio pressure level G is taken into account to ensure that a sample with high pressure level do not overwhelm the sample with low pressure level. The weight of an audio sample (x_1) is calculated as a relative pressure level (see Fig. 2 in the appendix for its distribution)

$$p = (1 + 10^{\frac{G_1 - G_2}{20}})^{-1}, \tag{9}$$

where G_1 and G_2 are pressure levels of two audio samples x_1 and x_2, respectively. This ensures good representation of both audio samples, post mixing.

Furthermore, as pointed out by Tokozume et al. [35], the energy of a sound wave is proportional to the square of its amplitude. Thus, we mix x_1 and x_2 as

$$\text{mix}(x_1, x_2) = \frac{px_1 + (1-p)x_2}{\sqrt{p^2 + (1-p)^2}}. \tag{10}$$

2.4 Classifier-Free Guidance

To guide the reverse diffusion process to reconstruct the audio prior z_0, we employ a classifier-free guidance [6] of text input τ. During inference, a guidance scale w controls the contribution of text guidance to the noise estimation $\hat{\epsilon}_\theta$, with respect to unguided estimation, where empty text is passed:

$$\hat{\epsilon}_\theta^{(n)}(z_n, \tau) = w\epsilon_\theta^{(n)}(z_n, \tau) + (1-w)\epsilon_\theta^{(n)}(z_n). \tag{11}$$

We also trained a model for which the text guidance was randomly dropped for 10% of the samples during training. We found this model to perform equivalently to a model for which text guidance was always used for all samples.

2.5 Audio VAE and Vocoder

Audio variational auto-encoder (VAE) [12] compresses the mel-spectogram of an audio sample, $m \in \mathbb{R}^{T \times F}$, into an audio prior $z_0 \in \mathbb{R}^{C \times T/r \times F/r}$, where C, T, F, r are the number of channels, number of time-slots, number of frequency-slots, and compression level, respectively. The LDM (see Section 2.2) reconstructs the audio prior $\hat{z}_0$ using input-text guidance τ. The encoder and decoder are composed of ResUNet blocks [14] and are trained by maximizing evidence lower-bound (ELBO) [12] and minimizing adversarial loss [8]. We adopt the checkpoint of audio VAE provided by Liu et al. [17]. Thus, we use their best reported setting, where C and r are set to 8 and 4, respectively.

As a vocoder to turn the audio-VAE decoder-generated mel-spectogram into an audio, we also use HiFi-GAN [13] as Liu et al. [17].

3 Experiments

3.1 Datasets and Training

Text-to-Audio Generation. We perform our main text-to-audio generation experiments on the AudioCaps dataset [11]. The dataset contains 45,438 audio clips paired with human-written captions for training. The validation set contains 2,240 instances. The audio clips are ten seconds long and were collected from YouTube videos. The clips were originally crowd-sourced as part of the significantly larger AudioSet dataset [4] for the audio classification task.

We train our LDM using only the paired (text, audio) instances from the AudioCaps dataset. We use the AudioCaps test set as our evaluation data. The test set contains five human-written captions for each audio clip. We use one caption for each clip chosen at random following Liu et al. [17] for consistent evaluation with their work. The randomly chosen caption is used as the text prompt, using which we generate the audio signal from our model.

Audio VAE and Vocoder. We use the audio VAE model from Liu et al. [17]. This VAE network was trained on the AudioSet, AudioCaps, Freesound[2], and BBC Sound Effect Library[3] (SFX) datasets. Longer audio clips in Freesound and BBC SFX were truncated to the first thirty seconds and then segmented into three parts of ten seconds each. All audio clips were resampled in 16KHz frequency for training the VAE network. We used a compression level of 4 with 8 latent channels for the VAE network.

We also use the vocoder from Liu et al. [17] for audio waveform generation from the mel spectrogram generated by the VAE decoder. The vocoder is a HiFi-GAN [13] network trained on the AudioSet dataset. All audio clips were resampled at 16KHz for training the vocoder network.

[2]https://freesound.org/

[3]https://sound-effects.bbcrewind.co.uk

Model, Hyperparameters, and Training Details We freeze the FLAN-T5-LARGE text encoder in TANGO and only train the parameters of the latent diffusion model. The diffusion model is based on the Stable Diffusion U-Net architecture [30, 31] and has a total of 866M parameters. We use 8 channels and a cross-attention dimension of 1024 in the U-Net model.

We use the AdamW optimizer [19] with a learning rate of 3e-5 and a linear learning rate scheduler for training. We train the model for 40 epochs on the AudioCaps dataset and report results for the checkpoint with the best validation loss, which we obtained at epoch 39. We use four A6000 GPUs for training TANGO, where it takes a total of 52 hours to train 40 epochs, with validation at the end of every epoch. We use a per GPU batch size of 3 (2 original + 1 augmented instance) with 4 gradient accumulation steps. The effective batch size for training is 3 (instance) $*4$ (accumulation) $*4$ (GPU) $= 48$.

3.2 Baseline Models

In our study, we examine three existing models: DiffSound by Yang et al. [38], AudioGen by Kreuk et al. [16], and AudioLDM by Liu et al. [17]. AudioGen and DiffSound use text embeddings for conditional generative training, while AudioLDM employs audio embeddings to avoid potential noise from weak textual descriptions in the paired text-audio data. AudioLDM uses audio embeddings from CLAP and asserts that they are effective in capturing cross-modal information. The models were pre-trained on large datasets, including AudioSet, and fine-tuned on the AudioCaps dataset, before evaluation, for enhanced performance. Thus, comparing them to our model TANGO would not be entirely fair.

Despite being trained on a much smaller dataset, our model TANGO outperformed the baselines that were trained on significantly larger datasets. We may largely attribute this to the use of LLM FLAN-T5. Therefore, our model TANGO sets itself apart from the three existing models, making it an exciting addition to the current research in this area.

It is important to note that the `AudioLDM-L-Full-FT` checkpoint from Liu et al. [17] was not available for our study. Therefore, we used the `AudioLDM-M-Full-FT` checkpoint, which was released by the authors and has 416M parameters. This checkpoint was fine-tuned on both the AudioCaps and MusicCaps datasets. We performed a subjective evaluation using this checkpoint in our study. We attempted to fine-tune the `AudioLDM-L-Full` checkpoint on the AudioCaps dataset. However, we were unable to reproduce the results reported in Liu et al. [17] due to a lack of information on the hyperparameters used.

Our model can be compared directly to `AudioLDM-L` since it has almost the same number of parameters and was trained solely on the AudioCaps dataset. However, it is worth noting that Liu et al. [17] did not release this checkpoint, which made it impossible for us to conduct a subjective evaluation of its generated samples.

3.3 Evaluation Metrics

Objective Evaluation. In this work, we used two commonly used objective metrics: Frechet Audio Distance (FAD) and KL divergence. FAD [10] is a perceptual metric that is adapted from Fechet Inception Distance (FID) for the audio domain. Unlike reference-based metrics, it measures the distance between the generated audio distribution and the real audio distribution without using any reference audio samples. On the other hand, KL divergence [38, 16] is a reference-dependent metric that computes the divergence between the distributions of the original and generated audio samples based on the labels generated by a pre-trained classifier. While FAD is more related to human perception, KL divergence captures the similarities between the original and generated audio signals based on broad concepts present in them. In addition to FAD, we also used Frechet Distance (FD) [17] as an objective metric. FD is similar to FAD, but it replaces the VGGish classifier with PANN. The use of different classifiers in FAD and FD allows us to evaluate the performance of the generated audio using different feature representations.

Subjective Evaluation. Following Liu et al. [17] and Kreuk et al. [16], we ask six human evaluators to assess two aspects — overall audio quality (OVL) and relevance to the input text (REL) – of 30 randomly-selected baseline- and TANGO-generated audio samples on a scale from 1 to 100. The evaluators were proficient in the English language and instructed well to make a fair assessment.

Table 1: The comparison between TANGO and baseline TTA models. *FT* indicates the model is fine-tuned on the Audiocaps (AC) dataset. The AS and AC stand for AudioSet and AudiocCaps datasets respectively. We borrowed all the results from [17] except for AudioLDM-L-Full which was evaluated using the model released by the authors on Huggingface. Despite the LDM being trained on a much smaller dataset, TANGO outperforms AudioLDM and other baseline TTA models as per both objective and subjective metrics. ‡ indicates the results are obtained using the checkpoints released by Liu et al. [17].

Model	Datasets	Text	#Params	Objective Metrics			Subjective Metrics	
				FD ↓	KL ↓	FAD ↓	OVL ↑	REL ↑
Ground truth	–	–	–	–	–	–	91.61	86.78
DiffSound [38]	AS+AC	✓	400M	47.68	2.52	7.75	–	–
AudioGen [16]	AS+AC+8 others	✓	285M	–	2.09	3.13	–	–
AudioLDM-S	AC	✗	181M	29.48	1.97	2.43	–	–
AudioLDM-L	AC	✗	739M	27.12	1.86	2.08	–	–
AudioLDM-M-Full-FT‡	AS+AC+2 others	✗	416M	26.12	**1.26**	2.57	79.85	76.84
AudioLDM-L-Full‡	AS+AC+2 others	✗	739M	32.46	1.76	4.18	78.63	62.69
AudioLDM-L-Full-FT	AS+AC+2 others	✗	739M	**23.31**	1.59	1.96	–	–
TANGO	AC	✓	866M	24.52	1.37	**1.59**	**85.94**	**80.36**

3.4 Results and Analysis

Main Results. We report our main comparative study in Table 1. We comapre our proposed method TANGO with DiffSound [38], AudioGen [16] and various configurations of AudioLDM [17]. AudioLDM obtained best results with 200 sampling steps from the LDM during inference. For a fair comparison, we also use 200 inference steps in TANGO and in our additional AudioLDM experiments. We used a classifier-free guidance scale of 3 for TANGO. AudioLDM used a guidance scale among {2, 2.5, 3} in their various experiments.

TANGO achieves new state-of-the-art results for objective metrics when trained only on the AudioCaps dataset, with scores of 24.52 FD, 1.37 KL, and 1.59 FAD. This is significantly better than the most direct baseline `AudioLDM-L`, which also used only the AudioCaps dataset for LDM training. We attribute this to the use of FLAN-T5 as text encoder in TANGO. We also note that TANGO matches or beats the performance of `AudioLDM-*-FT` models, which used significantly (~ 65 times) larger datasets for LDM training. The `AudioLDM-*-FT` models used two phases of LDM training – first on the collection of the four datasets, and then only on AudioCaps. TANGO is thus far more sample efficient as compared to the `AudioLDM-*-FT` model family.

TANGO also shows very promising results for subjective evaluation, with an overall audio quality score of 85.94 and a relevance score of 80.36, indicating its significantly better audio generation ability compared to AudioLDM and other baseline text-to-audio generation approaches.

Effect of Inference Steps and Classifier-Free Guidance. The number of inference steps and the classifier-free guidance scale are of crucial importance for sampling from latent diffusion models [34, 6]. We report the effect of varying number of steps and varying guidance scale for audio generation in AudioCaps in Table 2. We found that a guidance scale of 3 provides the best results for TANGO. In the left part of Table 2, we fix the guidance scale of 3 and vary the number of steps from 10 to 200. The generated audio quality and resultant objective metrics consistently become better with more steps. Liu et al. [17] reported that the performance for AudioLDM plateaus at around 100 steps, with 200 steps providing only marginally better performance. However, we notice a substantial improvement in performance when going from 100 to 200 inference steps for TANGO, suggesting that there could be further gain in performance with more inference steps.

We report the effect of varying guidance scale with a fixed 100 steps in the right half of Table 2. The first row uses a guidance scale of 1, thus effectively not applying classifier-free guidance at all during inference. Not surprisingly, the performance of this configuration is poor, lagging far behind the classifier-free guided models across all the objective measures. We obtain almost similar results with a guidance scale of 2.5 and better FD and KL with a guidance scale of 5. We obtain the best FAD metric at a guidance scale of 3 and the metric becomes poorer with larger guidance.

Table 2: Effect on the objective evaluation metrics with a varying number of inference steps and classifier-free guidance.

Model	Varying Steps					Varying Guidance				
	Guidance	Steps	FD ↓	KL ↓	FAD ↓	Steps	Guidance	FD ↓	KL ↓	FAD ↓
TANGO	3	10	45.12	1.66	11.38	100	-	35.76	2.02	6.22
		20	31.38	1.39	4.52		2.5	26.32	1.39	1.97
		50	25.33	**1.27**	2.13		3	26.13	1.37	**1.87**
		100	26.13	1.37	1.87		5	**24.28**	**1.28**	2.32
		200	**24.52**	1.37	**1.59**		10	26.10	1.31	3.30

Temporal Sequence Modelling. We analyze how TANGO and AudioLDM models perform audio generation when the text prompt contains multiple sequential events. Consider the following examples: *A toy train running as a young boy talks followed by plastic clanking then a child laughing* contains three separate sequential events, whereas *Rolling thunder with lightning strikes* contains only one. We segregate the AudioCaps test set using the presence of temporal identifiers – *while, before, after, then, followed* – into two subsets, one with multiple events and the other with single event. We show the objective evaluation results for audio generation on these subsets in Table 3. TANGO achieves the best FD and FAD scores for both multiple events and single event instances. The best KL divergence score is achieved by the `AudioLDM-M-Full-FT` model. We conjecture that the larger corpus from the four training datasets in AudioLDM could be more helpful in improving the reference-based KL metric, unlike the reference-free FD and FAD metrics.

Table 3: Objective evaluation results for audio generation in the presence of multiple events or a single event in the text prompt in the AudioCaps test set. The multiple events and single event subsets collectively constitute the entire AudioCaps test set. It should be noted that FD and FAD are corpus-level non-linear metrics, and hence the FD and FAD scores reported in Table 1 are not average of the subset scores reported in this table.

Model	Datasets	Multiple Events			Single Event		
		FD ↓	KL ↓	FAD ↓	FD ↓	KL ↓	FAD ↓
AudioLDM-L-Full	AS+AC+2 others	43.65	1.90	3.77	35.39	1.66	5.24
AudioLDM-M-Full-FT		34.57	**1.32**	2.45	29.40	**1.21**	3.27
TANGO	AC	**33.36**	1.45	**1.75**	**28.59**	1.30	**2.04**

Performance against Number of Labels. Recall that the AudioCaps dataset was curated from the annotations of the audio classification task in the AudioSet dataset. The text prompts in AudioCaps can thus be paired with the discrete class labels of AudioSet. The AudioSet dataset contains a total of 632 audio event classes. For instance, *A woman and a baby are having a conversation* and its corresponding audio clip has the following three labels: *Speech, Child speech kid speaking, Inside small room*. We group instances having one label, two labels, and multiple (two or more) labels in AudioCaps and evaluate the generated audios across the objective metrics. We report the result of the experiment in Table 4. TANGO outperforms AudioLDM models across all the objective metrics for audio generation from texts with one label or two labels. For texts with multiple labels, AudioLDM achieves a better KL divergence score and TANGO achieves better FD and FAD scores. Interestingly, all the models achieve consistently better FD and KL scores with progressively more labels, suggesting that such textual prompts are more effectively processed by the diffusion models.

Categorical Modelling. The class labels in AudioSet can be arranged hierarchically to obtain the following top-level categories: i) Human sounds, ii) Animal sounds, iii) Natural sounds, iv) Sounds of Things, v) Channel, environment, background sounds, vi) Source-ambiguous sounds, and vii) Music. We map the class labels in AudioCaps to the seven main categories listed above. The Music category is very rare in AudioCaps and the rest either appear on their own or in various combinations with others. We select the most frequently occurring category combinations and analyze the performance of various models within the constituting AudioCaps instances in Table 5. The performance of the two models is pretty balanced across the FD and KL metrics, with TANGO being better in

Table 4: Performance of audio generation in AudioCaps for texts containing one, two, or multiple (two or more) labels. Each text in AudioCaps has its corresponding multi-category labels from AudioSet. We use these labels to segregate the AudioCaps dataset into three subsets.

Model	Datasets	One Label			Two Labels			Multiple Labels		
		FD ↓	KL ↓	FAD ↓	FD ↓	KL ↓	FAD ↓	FD ↓	KL ↓	FAD ↓
AudioLDM-L-Full	AS+AC+2 others	48.11	2.07	4.71	44.93	1.90	4.09	34.94	1.68	4.59
AudioLDM-M-Full-FT		46.44	1.85	3.77	39.01	1.29	3.52	26.74	**1.10**	2.62
TANGO	AC	**40.81**	**1.84**	**1.79**	**35.09**	**1.56**	**2.53**	**26.05**	1.24	**1.96**

Table 5: Performance of `AudioLDM-M-Full FT` and TANGO for the most frequently occurring categories in AudioCaps dataset. CEB indicates the Channel, environment, and background sounds category.

Human	Animal	Natural	Things	CEB	FD ↓		KL ↓		FAD ↓	
					AudioLDM	TANGO	AudioLDM	TANGO	AudioLDM	TANGO
✓	✗	✗	✗	✗	38.15	**34.06**	1.01	**0.99**	2.81	**2.13**
✗	✓	✗	✗	✗	78.62	**77.78**	**1.82**	1.92	**4.28**	4.62
✓	✓	✗	✗	✗	**61.91**	70.32	**0.89**	1.29	6.32	**5.19**
✗	✗	✓	✗	✗	**51.61**	57.75	**1.89**	1.96	6.75	**5.15**
✗	✗	✗	✓	✗	35.60	**33.13**	**1.35**	1.43	5.42	**3.40**
✗	✗	✓	✓	✗	55.06	**42.00**	1.46	**1.12**	6.57	**3.89**
✓	✗	✗	✓	✗	**37.57**	39.22	**1.11**	1.34	3.26	**3.18**
✗	✗	✗	✓	✓	54.25	**52.77**	1.43	**1.33**	11.49	**9.26**

some, and AudioLDM in others. However, TANGO achieves better FAD scores in all but one group, with large improvements in (human, animal), (natural), (things), and (natural, things) categories.

4 Related Works

Diffusion Models. Recent years have seen a surge in diffusion models as a leading approach for generating high-quality speech [1, 15, 24, 25, 9, 7]. These models utilize a fixed number of Markov chain steps to transform white noise signals into structured waveforms. Among them, FastDiff has achieved remarkable results in high-quality speech synthesis [7]. By leveraging a stack of time-aware diffusion processes, FastDiff can generate speech samples of exceptional quality at an impressive speed, 58 times faster than real-time on a V100 GPU, making it practical for speech synthesis deployment. It surpasses other existing methods in end-to-end text-to-speech synthesis. Another noteworthy probabilistic model for audio synthesis is DiffWave [15], which is non-autoregressive and generates high-fidelity audio for various waveform generation tasks, including neural vocoding conditioned on mel spectrogram, class-conditional generation, and unconditional generation. DiffWave delivers speech quality that is on par with the powerful WaveNet vocoder [23] while synthesizing audio much faster. Diffusion models have emerged as a promising approach for speech processing, particularly in speech enhancement [20, 33, 26, 21]. Recent advancements in diffusion probabilistic models have led to the development of a new speech enhancement algorithm that incorporates the characteristics of noisy speech signals into the forward and reverse diffusion processes [22]. This new algorithm is a generalized form of the probabilistic diffusion model, known as the conditional diffusion probabilistic model. During its reverse process, it can adapt to non-Gaussian real noises in the estimated speech signal, making it highly effective in improving speech quality. In addition, Qiu et al. [26] propose SRTNet, a novel method for speech enhancement that incorporates the diffusion model as a module for stochastic refinement. The proposed method comprises a joint network of deterministic and stochastic modules, forming the "enhance-and-refine" paradigm. The paper also includes a theoretical demonstration of the proposed method's feasibility and presents experimental results to support its effectiveness, highlighting its potential in improving speech quality.

Text-to-Audio Generation. The field of text-to-audio generation has received limited attention until recently [16, 38]. In Yang et al. [38], a text encoder is used to obtain text features, which are then processed by a non-autoregressive decoder to generate spectrogram tokens. These tokens are fed to a vector quantized VAE (VQ-VAE) to generate mel spectrograms that are used by a vocoder to generate audio. The non-autoregressive decoder is a probabilistic diffusion model. In addition, Yang et al. [38] introduced a novel data augmentation technique called the mask-based text generation

strategy (MBTG), which masks out portions of input text that do not represent any event, such as those indicating temporality. The aim of MBTG is to learn augmented text descriptions from audio during training. Although this approach seems promising, its fundamental limitation is the lack of diversity in the generated data, as it fails to mix different audio samples. Later, Kreuk et al. [16] proposed a correction to this method, mixing audio signals according to random signal-to-noise ratios and concatenating the corresponding textual descriptions. This approach allows for the generation of new (text, audio) pairs and mitigates the limitations of Yang et al. [38]. Unlike Yang et al. [38], the architecture proposed in Kreuk et al. [16] uses a transformer encoder and decoder network to autoregressively generate audio tokens from text input.

Recently, Liu et al. [17] proposed AudioLDM, which translates the Latent Diffusion Model of text-to-visual to text-to-audio generation. They pre-trained VAE-based encoder-decoder networks to learn a compressed latent representation of audio, which was then used to guide a diffusion model to generate audio tokens from text input. They found that using audio embeddings instead of text embeddings during the backward diffusion process improved conditional audio generation. During inference time, they used text embeddings for text-to-audio generation. Audio and text embeddings were obtained using pre-trained CLAP, which is the audio counterpart of CLIP embeddings used in the original LDM model.

5 Limitations

TANGO is not always able to finely control its generations over textual control prompts as it is trained only on the small AudioCaps dataset. For example, the generations from TANGO for prompts *Chopping tomatoes on a wooden table* and *Chopping potatoes on a metal table* are very similar. *Chopping vegetables on a table* also produces similar audio samples. Training text-to-audio generation models on larger datasets is thus required for the model to learn the composition of textual concepts and varied text-audio mappings. In the future, we plan to improve TANGO by training it on larger datasets and enhancing its compositional and controllable generation ability.

6 Conclusion

In this work, we investigate the effectiveness of the instruction-tuned model, FLAN-T5, for text-to-audio generation. Specifically, we use the textual embeddings produced by FLAN-T5 in the latent diffusion model to generate mel-spectrogram tokens. These tokens are then fed to a pre-trained variational auto-encoder (VAE) to generate mel-spectrograms, which are later used by a pre-trained vocoder to generate audio. Our model achieved superior performance under both objective and subjective evaluations compared to the state-of-the-art text-to-audio model, AudioLDM, despite using only 63 times less training data. We primarily attribute this performance improvement to the representational power of FLAN-T5, which is due to its instruction-based tuning in the pre-training stage. In the future, we plan to investigate the effectiveness of FLAN-T5 in other audio tasks, such as, audio super-resolution and inpainting.

References

[1] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020.

[2] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL `https://arxiv.org/abs/2210.11416`.

[3] Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can gpt learn in-context? language models secretly perform gradient descent as meta-optimizers. *ArXiv*, abs/2212.10559, 2022.

[4] 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 *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[5] 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, 2023.

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

[7] Rongjie Huang, Max WY Lam, Jun Wang, Dan Su, Dong Yu, Yi Ren, and Zhou Zhao. Fastdiff: A fast conditional diffusion model for high-quality speech synthesis. *arXiv preprint arXiv:2204.09934*, 2022.

[8] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, 2016.

[9] Myeonghun Jeong, Hyeongju Kim, Sung Jun Cheon, Byoung Jin Choi, and Nam Soo Kim. Diff-tts: A denoising diffusion model for text-to-speech. *arXiv preprint arXiv:2104.01409*, 2021.

[10] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *INTERSPEECH*, pages 2350–2354, 2019.

[11] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *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)*, pages 119–132, 2019.

[12] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[13] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33:17022–17033, 2020.

[14] Qiuqiang Kong, Yin Cao, Haohe Liu, Keunwoo Choi, and Yuxuan Wang. Decoupling magnitude and phase estimation with deep resunet for music source separation. In *International Society for Music Information Retrieval Conference*, 2021.

[15] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

[16] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre D'efossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *ArXiv*, abs/2209.15352, 2022.

[17] Haohe Liu, Zehua Chen, Yiitan Yuan, Xinhao Mei, Xubo Liu, Danilo P. Mandic, Wenwu Wang, and MarkD . Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *ArXiv*, abs/2301.12503, 2023.

[18] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[19] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[20] Yen-Ju Lu, Yu Tsao, and Shinji Watanabe. A study on speech enhancement based on diffusion probabilistic model. In *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 659–666, 2021.

[21] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7402–7406, 2022. doi: 10.1109/ICASSP43922.2022.9746901.

[22] Yen-Ju Lu, Zhong-Qiu Wang, Shinji Watanabe, Alexander Richard, Cheng Yu, and Yu Tsao. Conditional diffusion probabilistic model for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7402–7406. IEEE, 2022.

[23] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

[24] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[25] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, Mikhail Kudinov, and Jiansheng Wei. Diffusion-based voice conversion with fast maximum likelihood sampling scheme. *arXiv preprint arXiv:2109.13821*, 2021.

[26] Zhibin Qiu, Mengfan Fu, Yinfeng Yu, LiLi Yin, Fuchun Sun, and Hao Huang. Srtnet: Time domain speech enhancement via stochastic refinement. *arXiv preprint arXiv:2210.16805*, 2022.

[27] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *ArXiv*, abs/2102.12092, 2021.

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022.

[30] 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*, pages 10684–10695, 2022.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

[32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[33] Joan Serrà, Santiago Pascual, Jordi Pons, R Oguz Araz, and Davide Scaini. Universal speech enhancement with score-based diffusion. *arXiv preprint arXiv:2206.03065*, 2022.

[34] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020.

[35] Yuji Tokozume, Yoshitaka Ushiku, and Tatsuya Harada. Learning from between-class examples for deep sound recognition. *CoRR*, abs/1711.10282, 2017. URL `http://arxiv.org/abs/1711.10282`.

[36] Wikipedia. Tango. `https://en.wikipedia.org/wiki/Tango`, 2021. [Online; accessed 21-April-2023].

[37] Wikipedia. Tango music. https://en.wikipedia.org/wiki/Tango_music, 2021. [Online; accessed 21-April-2023].

[38] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*, 2022.

Appendix

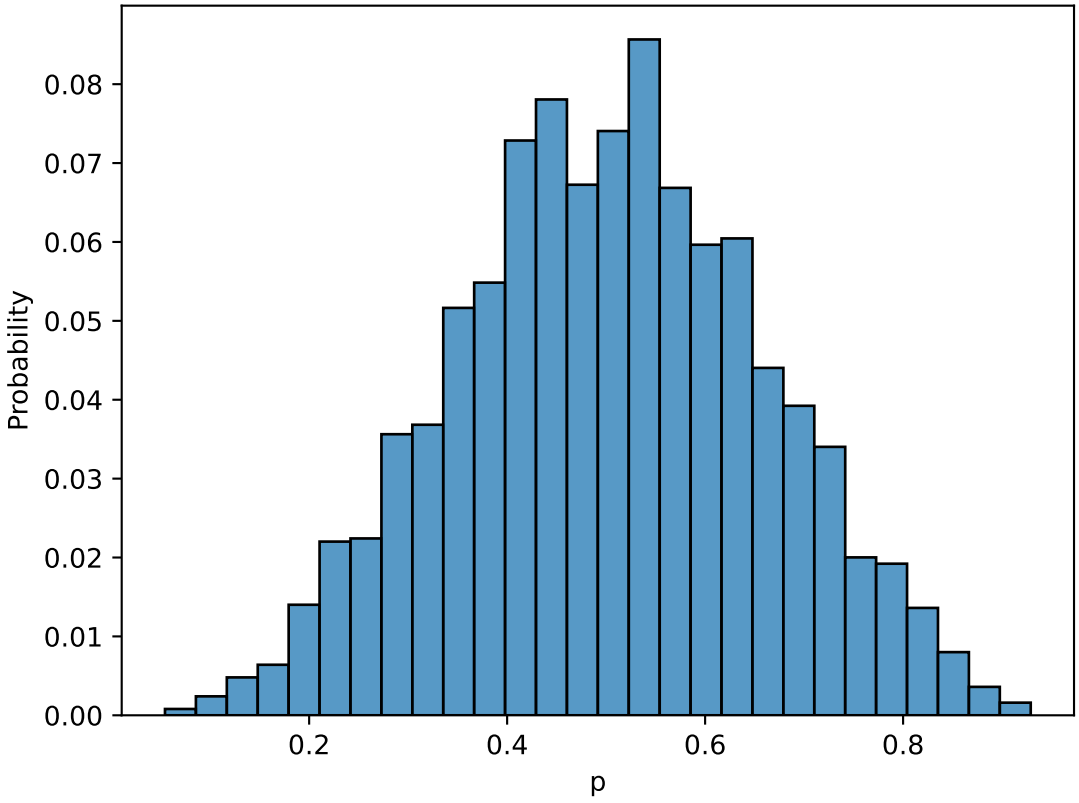

Figure 2: Distribution of relative pressure level (see Eq. (9)) across the augmented samples.