

Improving Text-To-Audio Models with Synthetic Captions

Zhifeng Kong^{*1}, Sang-gil Lee^{*1}, Deepanway Ghosal², Navonil Majumder², Ambuj Mehrish², Rafael Valle¹, Soujanya Poria², Bryan Catanzaro¹

¹NVIDIA ²Singapore University of Technology and Design
{zkong, sanggill, rafaelvalle}@nvidia.com, sporia@sutd.edu.sg

Abstract

It is an open challenge to obtain high quality training data, especially captions, for text-to-audio models. Although prior methods have leveraged *text-only language models* to augment and improve captions, such methods have limitations related to scale and coherence between audio and captions. In this work, we propose an audio captioning pipeline that uses an *audio language model* to synthesize accurate and diverse captions for audio at scale. We leverage this pipeline to produce a dataset of synthetic captions for AudioSet, named AF-AudioSet, and then evaluate the benefit of pre-training text-to-audio models on these synthetic captions. Through systematic evaluations on AudioCaps and MusicCaps, we find leveraging our pipeline and synthetic captions leads to significant improvements on audio generation quality, achieving a new *state-of-the-art*.

Index Terms: Text-to-Audio, Text-to-Music, Audio Captioning.

1. Introduction

There has been great progress in generative models that generate audio given text descriptions. These models are called text-to-audio (TTA) models [1, 2, 3, 4, 5], and have great potential in a wide range of tasks such as music composition, interactive art, media creation, and education. They also play a critical role in building general purpose multimodal models and agents that can understand and simulate the world in multiple modalities.

Training large-scale text-to-audio models, however, is very challenging. In contrast to the text-to-image domain where there are millions of high-quality samples available [6], there are much fewer high-quality training samples (i.e. audio and caption pairs) in the text-to-audio domain.¹ Meanwhile, the benefits of scaling both compute and data, especially during the pre-training phase, have been emphasized in recent research [7, 8, 9]. In consonance with these findings, this paper shows that pre-training on *high quality* datasets, even if they are synthetic, can drastically improve the quality of text-to-audio models.

In order to create a large dataset for pre-training, prior methods either transform tags and labels into natural language [10, 11], or augment audio and captions through mixing and concatenation [2, 3]. These approaches are limited because they require transforming pre-existing metadata, which can be of low quality and result in inconsistencies between the transformed metadata and the audio.

In this paper, we propose an alternative approach to obtain high quality audio captions that can be produced at scale and

that are based on the audio content. Our approach uses a pre-trained audio language model to automatically caption audio in the wild. Our approach does not require annotation nor metadata associated with audio and, as such, it can be easily scaled-up.

Automatically captioning audios in the wild, however, has several major challenges. First, the audio language model needs to generalize well to a wide range of audio contents. Second, the generated captions need to be diverse. Finally, given the variability in the quality of generated captions, a mechanism is needed to rank and filter generated captions. We address the first challenge by adopting the recently proposed Audio Flamingo chat model [12] trained on diverse dialogues. We ensure that the captions are diverse by generating captions on the diverse AudioSet dataset [13]. Last, to promote the accuracy of the generated captions, we filter them based on their CLAP similarities with the corresponding audios [14]. With these strategies in place, we are able to generate a large, diverse and high quality dataset of synthetic captions called AF-AudioSet.

We use text-to-audio (AudioCaps [15]) and text-to-music (MusicCaps [5]) benchmarks to evaluate the benefits of using our method and synthetic captions dataset during pre-training. We systematically study different data filtering and combination strategies, model sizes, as well as commonly used architectural designs based on Tango [3]. We find the optimal pre-training recipes to be consistent across many settings, and with these recipes, we are able to achieve the *state-of-the-art* audio generation quality on both benchmarks. To the best of our knowledge, this is the first systematic study to create large-scale high-quality synthetic captions using audio language models and verify their effectiveness in improving text-to-audio models.²

In summary, our contributions are as follows:

- We propose a data labeling pipeline to generate large-scale high-quality synthetic captions for audio.
- We introduce AF-AudioSet: a large, diverse, and high-quality synthetic caption dataset produced with our pipeline.
- We obtain *state-of-the-art* models on text-to-audio and text-to-music through pre-training on AF-AudioSet, and conduct systematic study across a variety of settings.

2. Related work

2.1. Diffusion-based Text-to-Audio Generation

The research community has made significant progress in diffusion-based [16, 17] text-to-audio generation models, with

^{*}Equal contribution.

¹Public datasets contain about 0.5K hours of audio and about 100K captions. It is challenging to train large-scale models on these data.

²AF-AudioSet: https://github.com/NVIDIA/audio-flamingo/blob/main/labeling_machine. Demos: <https://huggingface.co/spaces/declare-lab/Tango-AF>; <https://huggingface.co/spaces/declare-lab/Tango-Music-AF>. Checkpoints: <https://huggingface.co/declare-lab/tango-af-ac-ft-ac>; <https://huggingface.co/declare-lab/tango-music-af-ft-mc>.

recent examples including AudioLDM [1, 18], Make-An-Audio [2], and Tango [3, 19]. These models use a pre-trained text encoder (e.g., CLAP [20, 14], T5 [21], or FLAN-T5 [22]) to obtain text embeddings, and a pre-trained variational autoencoder (VAE) [23] to obtain latent features of audio. Similar to latent diffusion models (LDM) [17], the diffusion decoder is trained to generate the audio latent features conditioned on the text embeddings. The generated latent is decoded to a mel spectrogram representation using the VAE, followed by a neural vocoder [24, 25] that converts the mel spectrogram into waveform.

2.2. Training Data Augmentation

Obtaining diverse, large-scale, and high-quality training data, specially captions, is one of the major challenges in training high-quality text-to-audio models. Especially, a very limited amount of accurate audio-caption pairs are available. Conversely, it is possible to leverage human annotators to produce high quality captions such as AudioCaps [15] and MusicCaps [5]. However, such datasets are very small, e.g. less than 10,000 samples, making it challenging to train large text-to-audio models.

The current main approach to augment audio captions is to use a large language model to rephrase tags and labels into short captions [10]. While this approach can scale-up captions to some extent, it is limited by the existence and quality of the metadata. Other approaches focus on augmenting the audio data by concatenating or mixing two samples to form new samples [1, 2, 3]. Though these methods can improve concept-composition capabilities, combining captions is a non-trivial task given that the combination of sounds can result in different captions.³

In this paper, we introduce an alternative approach where we label audio based on an audio language model. Our approach can generate high-quality captions as it listens to the audio contents, and can be scaled-up as it does not require any metadata to be provided. We generate over 600K diverse captions on AudioSet, and find that it can effectively enhance the generation quality of text-to-audio models. To the best of our knowledge, this is the first study that uses an audio language model to create synthetic captions and use them to train and improve text-to-audio models.

2.3. Audio Captioning Models

There are several audio-language models that can generate audio captions: Pengi [26], LTU [27], Qwen-Audio [28], Salmonn [29], and Audio Flamingo [12]. They use different methods to extract audio features and integrate these features into a large language model. Qwen-Audio and Salmonn are more focused on speech related tasks, while Pengi, LTU, and Audio Flamingo are more focused on non-speech audio understanding. Audio Flamingo in addition provides a chat model trained on diverse dialogues, which can generate more natural and diverse descriptions. Therefore, we use this model in our data synthesis pipeline.

3. Method and Experimental Setup

In this section, we introduce our captioning method and the large-scale synthetic dataset which we call *AF-AudioSet*. We also introduce our text-to-audio pretraining and finetuning setup.

3.1. Generating *AF-AudioSet*

We use Audio Flamingo [12] to generate captions for audio in the unbalanced training set of AudioSet [13]. Audio Flamingo

³Imagine captions for the sounds of racing cars and people screaming with and without the sound of gun shots.

Table 1: Number of captions and audio in *AF-AudioSet* available at different τ , the CLAP similarity threshold.

| τ | 35% | 40% | 45% | 50% |
|------------|---------|---------|---------|--------|
| # Captions | 696,079 | 366,018 | 164,756 | 61,225 |
| # Audios | 331,421 | 188,537 | 91,923 | 37,220 |

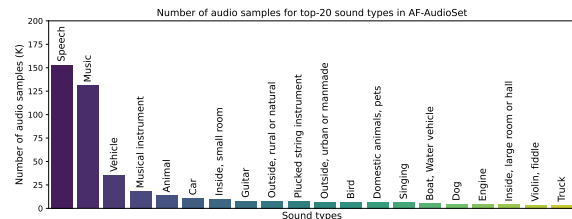


Figure 1: Distribution of sound types in *AF-AudioSet*.

has two series of models. The foundation model is trained on a number of benchmarking datasets including captioning, question-answering, and classification. The chat model is further finetuned on dialogues with more diverse questions and instructions. We investigated both models and found that the synthetic captions from the chat model are more natural and diverse, and therefore decided to use the chat model. Specifically, we prompt the model with the following instruction: “Can you briefly describe what you hear in this audio?”. During inference, we generate 20 captions per audio by sampling Audio Flamingo with top- $k = 50$ and top- $p = 95\%$.

Given that there is variation in the quality of the generated captions and that we want to promote captions with higher quality, we use the CLAP similarity [14] between the caption and the audio to rank and filter the synthetic captions. The similarity is computed as $\cos(\mathbf{v}_{\text{text}}, \mathbf{v}_{\text{audio}})$, where \mathbf{v}_{text} is the CLAP text embedding and $\mathbf{v}_{\text{audio}}$ is the CLAP audio embedding. We then store the Top-3 most correlated captions for each audio, and remove captions whose cosine similarities are $< 35\%$. We call the filtered synthetic caption dataset *AF-AudioSet*. In Table 1 we demonstrate the number of captions and audio available at different CLAP similarity thresholds. We demonstrate the distribution of sound types in Figure 1.

3.2. Text-to-Audio Setup

To systematically study the effect of pretraining text-to-audio models on *AF-AudioSet*, we use the Tango model [3] with a variety of experimental choices. Tango has three major components: a frozen audio VAE from AudioLDM [1], the FLAN-T5 text encoder [22], and a latent diffusion model that models the latent space of the audio VAE and is conditioned on the text embeddings. The HiFi-GAN vocoder [24] then turns generated mel spectrogram into waveform.

Model size. We consider three model sizes: small, medium, and large, each with different number of channels. The medium one is the same as [3] with 866M parameters. The large one has 1.93B parameters, and the small one has 217M parameters. Following [1], we also investigate replacing the FLAN-T5 text encoder [22] with the CLAP text encoder [14], followed by FiLM [30] conditioning layers – which we call Tango-CLAP.

Pretraining dataset. We study subsets of *AF-AudioSet* with the four CLAP thresholds τ shown in Table 1. By changing τ values we can investigate the trade-off between size and quality in our synthetic data. We also use the same audio captioning pipeline to generate additional captions for AudioCaps – which can be seen as data augmentation – and investigate the effect of

pretraining on this augmented dataset. Finally, we investigate mixing synthetic and real data during pretraining.

Tasks. We run experiments on both text-to-audio on AudioCaps [15] and text-to-music on MusicCaps [5]. Specifically, we finetune (pretrained or non-pretrained) Tango models on the train split of either dataset, and run evaluation on their test split.

Metrics. We report the Fréchet Distance (FD) [31], Fréchet Audio Distance (FAD) [32], Inception Score (IS) [33] with PANNs audio classifier backbone [34], and CLAP similarity [14] with the 630k-best checkpoint.⁴

Training Setup. In all experiments, we use 8 A100 GPUs to train the models. We pretrain with a batchsize of 128 for 100K iterations, and finetune with a batchsize of 48 for 40 epochs. The optimization method follows Tango [3].

4. Experiments

In this section, we aim to answer the following questions: 1) Does pretraining on AF-AudioSet improve generation quality? 2) What is the optimal quality-size trade-off (i.e., τ)? 3) What are the best recipes for different, text encoders, model sizes, and downstream tasks? 4) Does mixing synthetic and real captions during pretraining improve generation quality?

4.1. Pretraining Leads to SOTA Text-to-Audio Quality

Our text-to-audio results on AudioCaps are shown in Table 2. Tango-FT-AC refers to the *baseline* Tango without pretraining, and Tango-Full-FT-AC refers to the one pretrained on Tango-PromptBank [3]. Tango-AF-FT-AC refers to Tango pre-trained on AF-AudioSet ($\tau = 0.45$), and Tango-AF&AC-FT-AC refers to Tango pre-trained on AF-AudioSet ($\tau = 0.45$) + AudioCaps. After pretraining, these models are finetuned on AudioCaps. We find that pretraining on AF-AudioSet leads to a systematic improvement over the Tango baseline, especially in IS. We also find that pretraining on AF-AudioSet + AudioCaps results in further improvements, especially in FD. Positively, our best results also outperform recent state-of-the-art results.

The text-to-music results on MusicCaps are shown in Table 3. Tango-FT-MC refers to the *baseline* Tango without pretraining, and Tango-Full-FT-MC refers to the one pretrained on Tango-PromptBank [3]. TangoMusic-AF-FT-MC refers to Tango pre-trained on AF-AudioSet ($\tau = 0.35$). All models are then finetuned on MusicCaps. After pretraining on AF-AudioSet, the model significantly improves on all metrics and outperforms recent state-of-the-art baselines.

We summarize the results as the major finding of this paper:

Pretraining Tango on AF-AudioSet can lead to state-of-the-art text-to-audio and text-to-music generation quality.

4.2. Trade-off between Synthetic Captions Quality and Size

The CLAP threshold τ controls the trade-off between caption quality and data size in AF-AudioSet. A larger τ leads to a smaller subset (see Table 1) but the remaining captions are more correlated to the audio given CLAP as a similarity score. In Figures 2 and 3, we plot the evaluation metrics on AudioCaps and MusicCaps with different τ and using the medium-sized Tango. On AudioCaps, $\tau = 0.45$ leads to the best results and significantly outperforms the baseline Tango without pretraining. The results monotonically improve as τ increases from 0.35 to

⁴AudioLDM [1] suggested that FD is preferred over FAD as FD uses a higher quality audio classifier (PANNs) [34].

Table 2: Evaluation results on AudioCaps. Pretraining on AF-AudioSet (Tango-AF-FT-AC) leads to consistent improvement over the non-pretrained one (Tango-FT-AC). Pretraining on a mix of AF-AudioSet and AudioCaps (Tango-AF&AC-FT-AC) further improves and leads to SOTA text-to-audio generation quality. [†] indicates the numbers are taken from their original papers. [‡] indicates the numbers are taken from [19].

| Model | FD ↓ | CLAP ↑ | IS ↑ |
|----------------------|--------------------|--------------------------|--------------------|
| AudioLDM-L-Full [1] | 23.31 [†] | - | - |
| Make-an-Audio [2] | 18.32 [†] | 0.454 | 7.29 [†] |
| CoDi [35] | 22.90 [†] | - | 8.77 [†] |
| ConsistencyTTA [36] | 20.97 [†] | 0.496 | 8.88 [†] |
| Auffusion [37] | 21.99 [†] | 0.539 | 10.57 [†] |
| Tango-FT-AC [3] | 19.84 | 0.500 | 9.06 |
| Tango-Full-FT-AC [3] | 18.93 [†] | 0.539[‡] | 7.86 [‡] |
| Tango-AF-FT-AC | 19.06 | 0.503 | 10.87 |
| Tango-AF&AC-FT-AC | 17.19 | <u>0.527</u> | 11.04 |

Table 3: Evaluation results on MusicCaps. Pretraining on AF-AudioSet (TangoMusic-AF-FT-MC) significantly outperforms the non-pretrained one (Tango-FT-MC) and leads to SOTA text-to-music generation quality.

| Model | FD ↓ | FAD ↓ | IS ↑ |
|------------------------|--------------|-------------|-------------|
| MusicGen (medium) [38] | 35.52 | 5.02 | 1.94 |
| AudioLDM-2 [18] | 22.08 | 3.83 | 2.17 |
| Tango-FT-MC | 47.47 | 7.88 | 1.85 |
| Tango-Full-FT-MC | 38.19 | 6.83 | 2.71 |
| TangoMusic-AF-FT-MC | 21.84 | 1.99 | 2.21 |

0.45. However, there is result degradation when τ changes from 0.45 to 0.5, indicating the subset with $\tau = 0.5$ may be too small for pretraining. On MusicCaps, as τ increases, FD and FAD become slightly worse while IS becomes slightly better. However, the differences are small, and all results are significantly better than the baseline Tango model without pretraining. We think that $\tau = 0.35$ leads to the best generation quality as it has the best FD and FAD. We summarize the recipes below:

$\tau = 0.45$ leads to the best results on AudioCaps.
 $\tau = 0.35$ leads to the best results on MusicCaps.

4.3. AF-AudioSet versus TangoPromptBank

We also investigate pretraining on TangoPromptBank [3] – a collected pretraining set with more than 1M samples – and finetune on both benchmarks. The resulting models are called Tango-Full-FT-AC(or MC), and results are in Tables 2 and 3. We find pretraining on AF-AudioSet (Tango-AF-*) can match the generation quality as pretraining on TangoPromptBank on AudioCaps and significantly outperforms it on MusicCaps. The results indicate that AF-AudioSet has high quality captions, and pretraining on this smaller yet higher quality set can lead to similar or better results. We summarize our findings below:

Pretraining on high-quality datasets, even if they are synthetic and smaller, can lead to similar or better generation quality.

4.4. The Effect of Text Encoder and Tango Size

We investigate the optimal filtering threshold τ for Tango-CLAP, where we replace the FLAN-T5 text encoder with CLAP. The results in Figure 4 are very similar to Tango and the best results occur at $\tau = 0.45$.

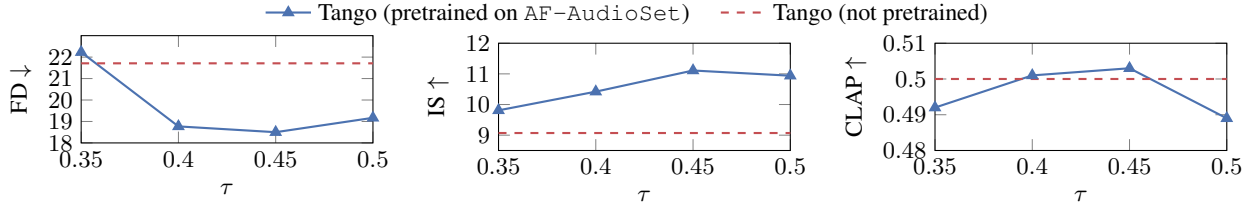


Figure 2: Evaluation results on **AudioCaps** with different CLAP thresholds of **AF-AudioSet**. The model is **Tango (medium)** finetuned on **AudioCaps**. $\tau = 0.45$ leads to the best results overall and significant improvements over the non-pretrained one.

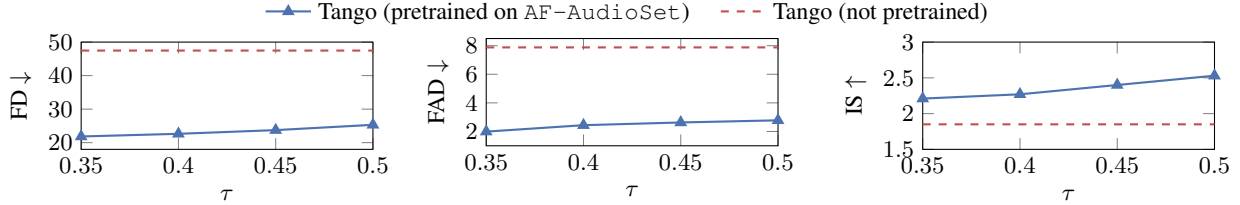


Figure 3: Evaluation results on **MusicCaps** with different CLAP thresholds of **AF-AudioSet**. The model is **Tango (medium)** finetuned on **MusicCaps**. $\tau = 0.35$ leads to the best FD and FAD and significant improvements over the non-pretrained one.

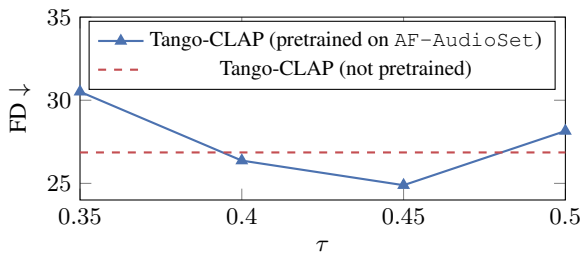


Figure 4: Evaluation results on **AudioCaps** with different CLAP thresholds of **AF-AudioSet**. The model is **Tango-CLAP (medium)** finetuned on **AudioCaps**. The results are similar to **Tango** in Figure 2.

We then study the effect of our synthetic data on different Tango model sizes: small, medium, and large, and compare the results with and without pretraining. Figures 5 and 6 show results on **AudioCaps** and **MusicCaps**. For all sizes, pretraining leads to significant improvements, indicating that pretraining on **AF-AudioSet** is an efficient and versatile strategy to improve audio generation quality. We summarize our findings below:

The recipes and conclusions in Section 4.2 also apply to other model conditioning architectures and model sizes.

4.5. The Effect of Mixed pretraining Sets

Finally, we study the effect of several mixed pretraining sets, where we combine synthetic and real captions during pretraining. First, we look at combining **AF-AudioSet** and **AudioCaps**. The results are in Table 2, with model name **Tango-AF&AC-FT-AC**. The results show that combining both datasets during pretraining improves generation quality. Then, we look at augmenting **AudioCaps** with synthetic captions generated with our pipeline described in Section 3. The results for this setting are: FD= 19.59, CLAP= 0.507, and IS= 9.85. The results show that simply augmenting **AudioCaps** leads to consistently better generation quality. We summarize our findings below:

Combining synthetic and real data during pretraining can lead to further improvements on generation quality.

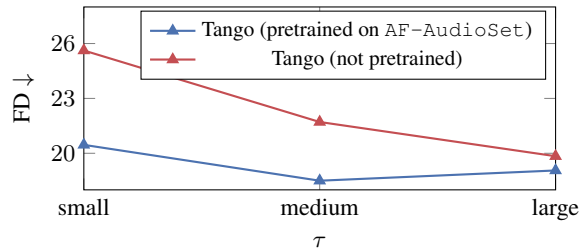


Figure 5: Evaluation results on **AudioCaps** with different model sizes. The model is **Tango** pre-trained on **AF-AudioSet** with $\tau = 0.45$ and finetuned on **AudioCaps**. The improvement by pretraining is clear across all model sizes.

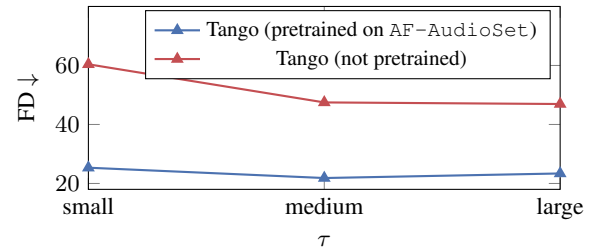


Figure 6: Evaluation results on **MusicCaps** with different model sizes. The model is **Tango** pre-trained on **AF-AudioSet** with $\tau = 0.35$ and finetuned on **MusicCaps**. The improvement by pretraining is clear across all model sizes.

5. Discussion

We expect that the quality of synthetic captions will improve, as audio language models become larger and more powerful – including audio captioning models and contrastive audio-text embeddings. An important future direction is to investigate a better synthesis pipeline to further improve diversity and accuracy of synthetic captions. Another important future direction is to investigate better pretraining strategies.

6. References

- [1] 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 *International Conference on Machine*

- Learning. PMLR, 2023, pp. 21 450–21 474.
- [2] 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 *International Conference on Machine Learning*. PMLR, 2023, pp. 13 916–13 932.
 - [3] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, “Text-to-audio generation using instruction-tuned llm and latent diffusion model,” *arXiv preprint arXiv:2304.13731*, 2023.
 - [4] 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 *The Eleventh International Conference on Learning Representations*, 2022.
 - [5] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi *et al.*, “Musiclm: Generating music from text,” *arXiv preprint arXiv:2301.11325*, 2023.
 - [6] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, “Laion-400m: Open dataset of clip-filtered 400 million image-text pairs,” *arXiv preprint arXiv:2111.02114*, 2021.
 - [7] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, “Scaling laws for neural language models,” *arXiv preprint arXiv:2001.08361*, 2020.
 - [8] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark *et al.*, “Training compute-optimal large language models,” *arXiv preprint arXiv:2203.15556*, 2022.
 - [9] A. Hägele, E. Bakouch, A. Kosson, L. B. Allal, L. Von Werra, and M. Jaggi, “Scaling laws and compute-optimal training beyond fixed training durations,” *arXiv preprint arXiv:2405.18392*, 2024.
 - [10] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, “Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research,” *arXiv preprint arXiv:2303.17395*, 2023.
 - [11] J. Melechovsky, Z. Guo, D. Ghosal, N. Majumder, D. Herremans, and S. Poria, “Mustango: Toward controllable text-to-music generation,” *arXiv preprint arXiv:2311.08355*, 2023.
 - [12] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, “Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities,” *arXiv preprint arXiv:2402.01831*, 2024.
 - [13] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 776–780.
 - [14] 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 *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
 - [15] C. D. Kim, B. Kim, H. Lee, and G. 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)*, 2019, pp. 119–132.
 - [16] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.
 - [17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-resolution image synthesis with latent diffusion models,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.
 - [18] 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 Transactions on Audio, Speech, and Language Processing*, 2024.
 - [19] N. Majumder, C.-Y. Hung, D. Ghosal, W.-N. Hsu, R. Mihalcea, and S. Poria, “Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization,” 2024.
 - [20] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, “Clap learning audio concepts from natural language supervision,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.
 - [21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, “Exploring the limits of transfer learning with a unified text-to-text transformer,” *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.
 - [22] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, “Scaling instruction-finetuned language models,” *Journal of Machine Learning Research*, vol. 25, no. 70, pp. 1–53, 2024.
 - [23] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
 - [24] J. Kong, J. Kim, and J. Bae, “Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis,” *Advances in neural information processing systems*, vol. 33, pp. 17 022–17 033, 2020.
 - [25] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “Bigvgan: A universal neural vocoder with large-scale training,” in *The Eleventh International Conference on Learning Representations*, 2022.
 - [26] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, “Pengi: An audio language model for audio tasks,” *Advances in Neural Information Processing Systems*, vol. 36, pp. 18 090–18 108, 2023.
 - [27] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. R. Glass, “Listen, think, and understand,” in *The Twelfth International Conference on Learning Representations*, 2023.
 - [28] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” *arXiv preprint arXiv:2311.07919*, 2023.
 - [29] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, M. Zejun, and C. Zhang, “Salmonn: Towards generic hearing abilities for large language models,” in *The Twelfth International Conference on Learning Representations*, 2023.
 - [30] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, “Film: Visual reasoning with a general conditioning layer,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.
 - [31] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.
 - [32] D. Roblek, K. Kilgour, M. Sharifi, and M. Zuluaga, “Fr’echet audio distance: A reference-free metric for evaluating music enhancement algorithms,” in *Proc. Interspeech*, 2019, pp. 2350–2354.
 - [33] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved techniques for training gans,” *Advances in neural information processing systems*, vol. 29, 2016.
 - [34] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. 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.
 - [35] Z. Tang, Z. Yang, C. Zhu, M. Zeng, and M. Bansal, “Any-to-any generation via composable diffusion,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
 - [36] Y. Bai, T. Dang, D. Tran, K. Koishida, and S. Sojoudi, “Accelerating diffusion-based text-to-audio generation with consistency distillation,” *arXiv preprint arXiv:2309.10740*, 2023.
 - [37] J. Xue, Y. Deng, Y. Gao, and Y. Li, “Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation,” *arXiv preprint arXiv:2401.01044*, 2024.
 - [38] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, “Simple and controllable music generation,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.