# Taming Data and Transformers for Audio Generation

**Moayed Haji-Ali**[1,2,*] **Willi Menapace**[2] **Aliaksandr Siarohin**[2] **Guha Balakrishnan**[1]

**Sergey Tulyakov**[2] **Vicente Ordonez**[1]

[1]Rice University [2]Snap Inc.

Project Webpage: https://taming-snap.github.io/

## Abstract

Generating ambient sounds and effects is a challenging problem due to data scarcity and often insufficient caption quality, making it difficult to employ large-scale generative models for the task. In this work, we tackle the problem by introducing two new models. First, we propose AutoCap, a *high-quality* and *efficient* automatic audio captioning model. We show that by leveraging metadata available with the audio modality, we can substantially improve the quality of captions. AutoCap reaches CIDEr score of $83.2$, marking a $3.2\%$ improvement from the best available captioning model at *four times* faster inference speed. We then use AutoCap to caption clips from existing datasets, obtaining $761,000$ audio clips with high-quality captions, forming the *largest* available audio-text dataset. Second, we propose GenAu, a scalable transformer-based audio generation architecture that we scale up to 1.25B parameters and train with our new dataset. When compared to state-of-the-art audio generators, GenAu obtains significant improvements of $15.7\%$ in FAD score, $22.7\%$ in IS, and $13.5\%$ in CLAP score, indicating significantly improved quality of generated audio compared to previous works. This shows that the *quality* of data is often as important as its quantity. Besides, since AutoCap is fully automatic, new audio samples can be added to the training dataset, unlocking the training of even larger generative models for audio synthesis.

## 1 Introduction

Generative models have revolutionized the field of content creation, enabling the generation of high-quality natural images [91, 94, 92, 87, 37], vivid videos [41, 110, 114, 89, 80], and intricate 3D shapes [11, 98, 60]. The domain of audio synthesis has undergone comparable advancement [43, 42, 65, 124, 33, 95, 83, 125, 26, 67, 115, 35], with three broad areas of study. Speech synthesis is the first, and probably the oldest problem in the domain, dating back to the eighteenth century [13, 117]. Modern text-to-speech synthesis adapts diffusion models used originally to synthesize images [50, 105, 34, 7]. The second domain is text-to-music synthesis, which has seen similar advances [2, 55, 68, 57, 88, 28, 84, 27, 133, 52, 132, 82]. The success in these two subdomains rests on two key pillars: (i) the availability of high-quality large-scale datasets containing text-to-speech, text-to-music data, and (ii) the development of scalable generative modeling methods [40, 100]. Indeed, there is an abundance of recordings of public speeches, performances, and musical pieces.

The world of audio synthesis is incomplete without the third category—ambient audio generation [65, 66, 43, 42, 30, 53], which is the main focus of this work. Unlike, the first two groups, ambient sound generation does not have a large corpus of accurately annotated samples readily available [46, 19]. It is fairly straightforward to collect text-to-speech data, by transcribing existing speeches. For ambient

[*]Work partially done during an internship at Snap Inc.

Preprint. Under review.

sounds, the task is substantially more challenging. Consider imagining a sound of "a peaceful early morning by the sea", or the sound of "rocks falling on a wooden floor". Going from an ambient sound to a caption is even more challenging as different sounds can suit the same visual scene. This is exploited in animation, where there are no real sounds to work with, and the entire sound scene is created artificially. The second challenge of ambient sound synthesis is the lack of available data. For visual modalities, such as images and videos [123, 81], huge datasets are collected from the Internet, filtered, processed, and annotated. Often such media is accompanied by raw description and metadata, which is used to train reliable automatic captioning models [9]. This is in contrast to the audio modality, where data sourced from video platforms like YouTube is often heavily edited or predominantly containing speech and music. Previous efforts to extract ambient sound clips from YouTube have shown a near $99\%$ rejection rate, making it impractical to compile a large-scale dataset.

In this work, we address the challenges of ambient sound synthesis from two sides. First, we introduce AutoCap, an efficient and high-quality audio captioning model that leverages visual information to reliably and automatically caption a large-scale dataset consisting of 761k audio clips. Second, we propose GenAu, a latent audio generator, based on a scalable FIT-architecture [80, 8].

Current human-captioned audio datasets are limited, containing fewer than 51k text-audio pairs in total. This significantly impacts the training of current captioning models, making them more susceptible to overfitting and reducing their ability to generalize effectively. With this in mind, we build AutoCap, a state-of-the-art Automatic Audio Captioner (AAC). First, we refine the commonly used encoder-decoder transformer design based on a pretrained BART [58] model by introducing a Q-Former [59] that learns to summarize the encoded audio tokens into four times fewer tokens. By reducing the number of input tokens to the BART model, we speed up inference—an important step towards large-scale audio captioning—and provide better alignment with the original BART token representation due to the Q-Former additional capacity compared to simple projection layers used in previous work [49]. Second, we propose to use metadata and captions derived from video content to aid the captioning process and in this way, remedy the data scarcity problem affecting audio captioning methods. Critically, we augment the encoder inputs to assume both audio features and a set of descriptive textual metadata including audio title and a caption derived from the visual modality. This dual-input approach not only allows our model to achieve state-of-the-art performance on AudioCaps [46], marking a $3.2\%$ improvement in CIDEr score, but it also helps reduce the domain gap with in-the-wild audios, improving model robustness. Equipped with this method, we obtain 761,000 audio clips from the major existing audio datasets and pair each audio sample with a descriptive synthetic caption. More importantly, since AutoCap is fully automatic, new audio samples can be easily added to the training set, paving the way to scaling the data beyond current constraints.

To adapt audio generative models for larger scale training, we introduce GenAu, a scalable transformer-based architecture that achieves *significant* improvements over state-of-the-art audio generation models. Our approach introduces key architectural modifications over existing audio latent diffusion models [65, 43, 30, 42]. First, we train an efficient 1D-VAE [42] to transform a Mel-Spectrogram representation to a sequence of tokens and search for the optimal latent space for audio generation. Second, we recognize that audio grows fast temporally, and therefore, an efficient architecture that can handle larger sequences of tokens is needed. In particular, we employ a transformer architecture in the denoising backbone where we modify the FIT transformer [8] to generate audio in the latent space. Lastly, we extend the proposed FIT architecture to incorporate text conditioning through a dual encoder strategy. This involves an instruction-finetuned language model, FLAN-T5 [14], and an audio-centric CLAP encoding [49]. This adaptation significantly improves the model's overall performance, achieving $15.7\%$ better FAD, $22.7\%$ higher Inception Score, and $13.5\%$ improvement in CLAP score, demonstrating superior audio-text alignment and audio generation quality.

In summary, this work introduces: (i) AutoCap, a state-of-the-art audio captioner tailored towards the annotation of data at a large scale, which leverages audio metadata to improve accuracy and robustness, and a Q-Former to improve inference time and reduce overfitting; (ii) a large-scale dataset comprising 761k audio clips paired with synthetic captions derived from the proposed audio captioner and constituting the largest audio-text dataset currently available; (iii) GenAu, a novel audio generator based on a scalable transformer architecture specifically adapted to the audio domain. Our model achieves significantly improved quality when compared to the previous state-of-the-art.

## 2 Related Work

**Automatic Audio Captioning (AAC).** The goal of AAC is to produce natural language descriptions for given audio content. Most recent AAC methods [16, 118, 96, 101, 44, 15, 54, 122, 130] employ encoder-decoder transformer architectures, where an encoder receiving the audio signal produces a representation that is used by the decoder to produce the output caption. WavCaps [77] employs the CNN14 [51] and HTSAT [6] audio encoders and uses a pretrained BART [58] language decoder. CoNeTTE [135] proposes an audio encoder based on the ConvNeXt architecture and uses a vanilla transformer decoder [108] trained from scratch. Recently, EnCLAP [49] proposes the joint use of two audio representations in the form of CLAP [21] sequence embeddings and a discrete EnCodec [20] audio representation, and uses a pretrained BART model as the language backbone. Other work explores augmentation strategies to counter data scarcity [47, 135, 127]. *Liu et al*.[70] recently proposed to leverage the visual information using a pre-trained visual encoder to address sound ambiguities, reporting improvements. BART-Tags [32] generates captions conditioned on a sequence of predicted AudioSet tags. Our method uses audio metadata and visual information as additional conditioning signals and leverages a lightweight Q-Former [59] model that summarizes the audio feature to improve captioning speed and reduce model overfitting.

**Text-Audio Datasets.** The performance of text-audio models [134, 61, 17, 73, 18, 99, 22, 72, 102, 31, 12, 129], including AAC, is currently hindered by the lack of high-quality large-scale paired audio text data. The two main existing datasets are AudioCaps [46] and Clotho [19], comprising only 46k and 6k respectively of human-captioned audio clips. LAION-Audio [119] consists of 630k audio samples with raw descriptions, but annotation is highly noisy. WavCaps [77] proposes a filtering procedure based on ChatGPT [1] to collect 400k audio clips and weakly caption them based on the noisy descriptions alone. While weak-captioning does improve downstream metrics, it is suboptimal because it does not consider the audio signal. A recent work [43] explored a knowledge distillation approach that leverages data labels and a pre-trained audio captioner and retriever. In our work we collect 761k audio clips and produce captions using our state-of-the-art captioning method, resulting in a larger dataset with captions that are more aligned to the audio content.

**Text-conditioned audio generation.** The current state-of-the-art text-to-audio generation methods widely adopt diffusion models [126, 53, 65, 66, 42, 30, 25, 111, 53]. AudioLDM [65] makes use of a latent diffusion model conditioned on CLAP embeddings, avoiding the need for the textual modality at training time. AudioLDM 2 [66] introduces a general representation of audio unifying the tasks of music, speech, and sound effects generation. Similarly, Audiobox [111] generates audio across different modalities such as speech and sound effects. StableAudio [25] introduces timing embeddings to allow the generation of long audios up to 95s. Recent work also explored controllable audio generation [97, 121, 79, 85, 131, 62, 64], visual-conditioned audio generation [116, 78, 112], and more recently join audio-video generation [103, 104, 120, 38, 106, 107, 4, 48, 113, 75, 10]. In this work, we show that improvements to data captioning quality and size, and the adoption of scalable architecture designs lead to state-of-the-art text-to-audio generation performance.

## 3 Method

In this section, we describe our approach to high-quality text-to-audio generation. Sec. 3.1 describes our method for producing accurate audio captions using our AutoCap model, Sec. 3.2 describes our strategy for building our large-scale captioned audio dataset, including data selection and re-captioning, lastly Sec. 3.3 describes our scalable audio generation architecture for our GenAu model.

### 3.1 Automatic Audio Captioning

The availability of textual descriptions that are highly aligned with the modality to be generated, coupled with large-scale datasets is of critical importance in achieving high-quality generation results [3]. In this context, automatic image and video captioning have been successfully employed to improve text-to-image and text-to-video models by augmenting and improving the quality of the training data for these models [3, 9]. Despite the large quantity of available non-captioned audio data, current state-of-the-art Automatic Audio Captioning (AAC) methods have not been used to improve the quality of the training data of text-to-audio models [77, 135, 49]. We posit that this is in part due to the lower quality of the produced captions for existing models compared to similar models available

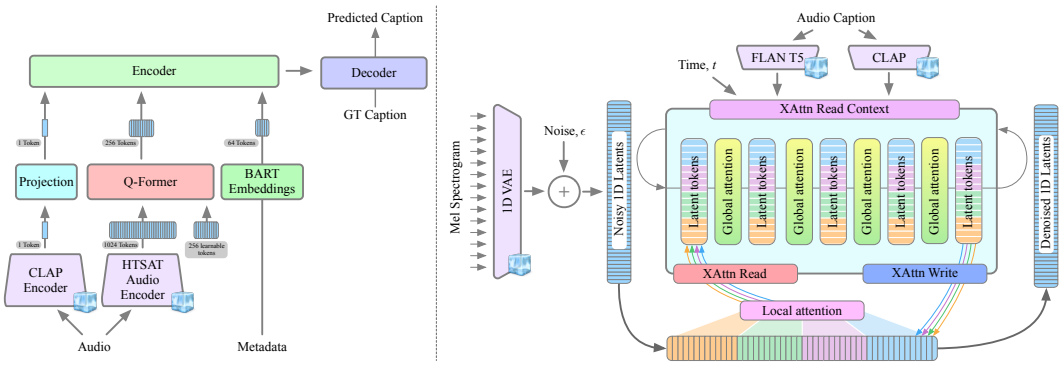


Figure 1: **(Left) Overview of AutoCap.** We employ frozen CLAP [49] and HTSAT [6] encoders to produce the audio representation. We then compact this representation into 4x fewer tokens using a Q-Former [59] module. This representation, along with tokens derived from pertinent metadata, is processed by a pretrained BART encoder-decoder model to generate the final caption.
**(Right) Overview of GenAu.** Following latent diffusion models, we use a frozen 1D-VAE to convert a Mel-Spectrogram into latent sequences, which are then divided into groups and processed using 'local' attention layers based on the FIT architecture [8]. 'Read' and 'write' layers, implemented as cross-attention, facilitate information transfer between input latents and *learnable* latent tokens. Finally, 'global' attention layers on *latent tokens* allow for global communication across all groups.

for other modalities. AAC methods have generally adopted an encoder-decoder transformer design, where an audio encoder (*e.g.* [20, 51, 6]) is responsible for producing an output representation that can be taken as input by a decoder model to produce an output audio caption. While audio-text pairs are difficult to obtain from web data, audio files from many sources are still commonly associated with metadata that might be relevant for captioning such as raw user descriptions, or captions derived from a related modality (*i.e.* accompanied visual information). Motivated by this observation and the benefits of better-aligned captions, we propose AutoCap, an audio captioning model where the encoded audio representation is augmented with metadata. Fig. 1 presents an overview of AutoCap.

We consider a dataset of audio signals paired with a corresponding caption $\langle \mathbf{a}, \mathbf{y} \rangle$ and metadata represented as a set of token sequences $\{\mathbf{m}_j\}_{j=1}^{j=M}$. Inspired by state-of-the-art AAC methods [77, 135, 49], we employ an encoder-decoder sequence-to-sequence model. We start by computing a global feature representation of the audio:

$$\mathbf{x}_{\text{clap}} = \mathcal{P}_{\text{clap}}(\mathcal{E}_{\text{clap}}(\mathbf{a})), \tag{1}$$

where $\mathcal{P}_{\text{clap}}$ is a learnable projection layer and $\mathcal{E}_{\text{clap}}$ is the audio encoder of a pretrained CLAP model [21]. Then we compute a separate local feature representation of the input audio:

$$\mathbf{x}_{\text{audio}} = \mathcal{Q}(\mathcal{E}_{\text{a}}(\mathbf{a})), \tag{2}$$

where $\mathcal{Q}$ is a Q-Former [59] that outputs a compact sequence audio representation and $\mathcal{E}_{\text{a}}$ is the HTSAT [6] audio encoder that produces a time-aligned representation. The Q-Former efficiently learns 256 latent tokens, which serve as keys in cross-attention layers with the input features, thereby condensing the audio input features into 256 tokens. Metadata sequences $\mathbf{m}_i$ are then embedded using the embedding layer of the pretrained encoder-decoder model to obtain corresponding embedding sequences $\mathbf{x}_{\text{meta}_i}$. We represent the input audio and metadata as the following input sequence:

$$\mathbf{x} = \mathbf{x}_{\text{clap}}\ \texttt{[boa]}\ \mathbf{x}_{\text{audio}}\ \texttt{[eoa]}\ \texttt{[bom]}_1\ \mathbf{x}_{\text{meta}_1}\ \texttt{[bom]}_1\ \ldots\ \texttt{[bom]}_M\ \mathbf{x}_{\text{meta}_M}\ \texttt{[bom]}_M, \tag{3}$$

where $\texttt{[boa]}\texttt{[eoa]}$ represent beginning and end of audio sequence embeddings $\mathbf{x}_{\text{audio}}$, and $\texttt{[bom]}_i$, $\texttt{[bom]}_i$ represent beginning and end of metadata embeddings $\mathbf{x}_{\text{meta}_i}$. Then this input sequence is used to obtain an output predicted caption $\hat{\mathbf{y}}$ as follows:

$$\hat{\mathbf{y}} = \mathcal{D}_{\text{t}}(\mathcal{E}_{\text{t}}(\mathbf{x})), \tag{4}$$

where $\mathcal{E}_{\text{t}}$ is an audio encoder that produces a representation of the input signal $\mathbf{x}$ which is fed to the decoder $\mathcal{D}_{\text{t}}$. We adopt a pretrained BART transformer model [58] as our encoder-decoder. Finally, we train our model using a standard cross-entropy loss over next token predictions:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{T}\sum_{t=1}^{t=T} \log p(\mathbf{y}_t | \mathbf{y}_{1:t-1}, \mathbf{x}). \tag{5}$$

Table 1: Overview of the employed dataset sources and audio clips counts for each of them.

| Data Source | Num. of pairs |
|---|---|
| AudioSet | 339,387 |
| VGGSounds | 126,905 |
| Freesounds | 262,300 |
| BBC Sound Effects | 31,201 |
| SoundBible | 1,320 |
| Our dataset | 761,113 |

To avoid degrading the quality of the pretrained BART and audio encoder models, we adopt a two-stage training procedure. In Stage 1, both the audio encoders and BART model are kept frozen, thus allowing the Q-Former, projection layers, and newly introduced delimiter tokens to align to the existing BART input representation. In this stage, we pretrain the model using a larger dataset of weakly-labeled audio clips. In Stage 2, we unfreeze all BART model parameters apart from the embedding layer and finetune the model on Audiocaps dataset at a lower learning rate to make the captioning style align more closely to the target dataset. This training strategy effectively leverages the larger, weakly-labeled dataset while minimizing the knowledge drift in the pretrained BART.

### 3.2 Data Selection and Re-captioning

Generative models in the image and video domains have shown benefits from increased quantities of data and improved quality of captions. In the audio domain, however, the major human-annotated audio-text datasets, namely AudioCaps [46] and Clotho [19], comprise a total of only 51k audio clips combined. Leveraging our trained AutoCap model, we aim to significantly expand the number of such videos by pairing them with high-quality synthetic captions.

We start dataset construction by selecting the AudioSet [29] and VGG-Sound [5] datasets as ideal candidates for re-captioning as each audio clip is derived from a video with an associated title and description. Additionally, since these datasets were constructed to be audio-oriented, they tend to have good alignment between the visual and audio components. To enrich the metadata information, we consider captions derived from the video modality using the video captioning method of *Chen et al*. [9]). In addition, we filter AudioSet to exclude content labeled as 'music' and 'speech' as we are more interested in the generation of audio effects. We sample 10-second audio clips randomly from each selected video. Furthermore, we select Freesound, BBC Sound Effects, and SoundBible as additional data sources. We input their associated title to the audio captioning model and leave the additional video caption metadata empty as no video modality is present. Tab. 1 summarizes the data sources employed to build our dataset. In total we collect and re-caption 761,000 audio clips, forming the largest available dataset of audio with paired captions.

We report in Tab. 2 a comparison of our collected dataset with the most popular audio text datasets. Our dataset possesses the largest number of audio clips and the longest total duration. Unlike WavCaps, whose captions are produced using the raw data and without considering the audio modality, our dataset is captioned by leveraging additional metadata derived from the video modality while maintaining an audio-centric approach. Compared to LAION-Audio-630k, the largest available dataset in terms of text-audio pairs, we offer an increased number of audio clips and more than double the total duration. Moreover, our dataset presents higher quality captions compared to LAION-Audio captions which are directly acquired from the raw data and thus present a high level of noise.

### 3.3 Scalable Text-2-Audio Generation

We design our audio generation pipeline, GenAu, as a latent diffusion model. Fig. 1 (right) shows an overview of our proposed model. In the following section, we describe in detail the structure of our latent variational autoencoder (VAE) and the latent diffusion model.

**Latent VAE.** Directly modeling waveform audio data is complex due to the high data dimensionality of audio signals. Instead, we replace the waveform with a Mel-spectrogram representation and use a VAE to further reduce its dimensionality, following prior work [79, 43]. We note that the once generated, a Mel-spectrogram representation can be decoded back to a waveform through the use

Table 2: Comparative overview of the main audio-language datasets.

| Dataset | # Text-Audio Pairs | Duration (h) | Text source |
|---|---|---|---|
| AudioCaps | 52,904 | 144.94 | Human |
| Clotho | 5,929 | 37.00 | Human |
| MACS | 3,537 | 9.83 | Human |
| WavText5K | 4,072 | 23.20 | Online raw-data |
| SoundDescs | 32,979 | 1,060.4 | Online raw-data |
| LAION-Audio-630K | 633,526 | 4,325.39 | Online raw-data |
| WavCaps | 403,050 | 7,567.92 | Processed raw-data |
| Ours | 761,113 | 8,763.12 | Automatic re-captioning |

of an audio vocoder [50]. However, commonly-used 2D autoencoder designs [65, 66, 79], are not well suited to the Mel-spectrogram representation, as the separation between the Mel channels is non-linear and thus not well suited for 2D convolutions. We instead opt for a 1D-VAE design based on 1D convolutions similar to a recent work [42], with the additional benefit of a more compact audio representation. We train our 1D-VAE model using a combination of reconstruction, adversarial, and KL regularization losses following [23].

**Latent diffusion model.** Following the latent diffusion paradigm, we generate audio by training a diffusion model in the latent space of the 1D-VAE. Transformer models currently attain state-of-the-art performance in audio generation [42]. To improve model scalability, we propose to use an efficient transformer architecture due to its success in handling long-range interactions as in video generation [8, 80]. In particular, we adopt the FIT architecture of *Menapace et al*. [80] which was originally proposed to work in the pixel space, and revise it for the latent space of the audio modality.

Given a 1D input $\mathbf{x}$, we first apply a projection operation to produce a sequence of input patch tokens. We then apply a sequence of FIT blocks to the input patches where each block divides patch tokens into contiguous groups of a predefined size. A set of *local* self-attention layers are then applied separately to each group to avoid the quadratic computational complexity of attention computation. Differently from the video domain [80] where the high input dimensionality makes the *local* layers excessively expensive, we found them to be beneficial for audio generation. To further reduce the amount of computation while maintain long-range interaction, each block considers a small set of latent tokens. First, a *read* operation implemented as a cross-attention layer transfers information from the patches to the latent tokens. Later, a series of *global* self-attention operations are applied to the latent tokens, allowing information-sharing between different groups. Finally, a *write* operation implemented as a cross-attention layer transfers information from the latent tokens back to the patches. Due to the reduced number of latent tokens when performing the global self-attention, computational requirements of the model are reduced with respect to a vanilla transformer design [108].

To condition the generation on an input prompt, we use a pretrained FLAN-T5 model [14] and a CLAP [21] text encoder to produce the their respective embeddings $e_{\mathrm{FLAN}}$ and $e_{\mathrm{CLAP}}$, which we concatenate with the diffusion timestep $t$ to form the input conditioning signal $c$. We insert an additional cross attention operation inside each FIT block immediately before the 'read' operation that makes latent tokens attend to the conditioning. Moreover, we use a conditioning on dataset ID to adapt the generation style to different types of datasets.

We follow a standard linear noise scheduler and train the model using the epsilon diffusion objective:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x},\boldsymbol{\epsilon}} \left\| \mathcal{G}(\mathbf{x}_t, c) - \boldsymbol{\epsilon} \right\|_2^2, \tag{6}$$

where $\mathcal{G}$ is the FIT generator backbone, $\mathbf{x}_t$ is the input with applied noise at diffusion timestep $t$, and $\boldsymbol{\epsilon}$ is noise sampled in $N(0, 1)$ with the same shape as the input.

## 4 Experiments

We structure the experiments section as follows: Sec. 4.1 evaluates AutoCap by quantitatively comparing it to previous work. Sec. 4.2 demonstrates the capabilities of GenAu through quantitative comparisons. For both, we discuss training details, baselines, metrics, results, and ablations.

Table 3: Captioning evaluation results on AudioCaps test split for various models. AS: AudioSet, AC: AudioCaps, WC: WavCaps, CL: Clotho, MA: Multi-Annotator Captioned Soundscapes.

| Model | Pretraining Data | BLEU1 | BLEU4 | ROUGE$_L$ | METEOR | CIDEr | SPICE | SPIDEr |
|---|---|---|---|---|---|---|---|---|
| ACT [76] | AS | 64.7 | 25.2 | 46.8 | 22.2 | 67.9 | 16.0 | 42.0 |
| V-ACT [71] | - | 69.8 | 28.1 | 49.4 | 23.7 | 71.1 | 17.2 | 44.2 |
| BART-tags [32] | AS | 69.9 | 26.6 | 49.3 | 24.1 | 75.3 | 17.6 | 46.5 |
| AL-MixGEN [47] | - | 70.0 | 28.9 | 50.2 | 24.2 | 76.9 | 18.1 | 47.5 |
| ENCLAP-Large [49] | - | - | - | - | 25.5 | 80.2 | 18.8 | 49.5 |
| HTSAT-BART [122] | - | 67.5 | 27.2 | 48.3 | 23.7 | 72.1 | 16.9 | 44.5 |
| HTSAT-BART [122] | AC+CL+WC | 70.7 | 28.3 | 50.7 | 25.0 | 78.7 | 18.2 | 48.5 |
| CNext-trans [135] | - | - | - | - | - | - | - | 46.6 |
| CNext-trans [135] | AC+CL+MA+WC | - | - | - | 25.2 | 80.6 | 18.4 | 49.5 |
| AutoCap (audio) | AC | 70.0 | 28.0 | 51.7 | 24.6 | 77.3 | 18.2 | 47.8 |
| AutoCap (audio+text) | AC | 72.1 | 28.6 | 51.5 | **25.6** | 80.0 | 18.8 | 49.4 |
| AutoCap (audio) | AC+CL+WC | **73.1** | 28.1 | **52.0** | **25.6** | 80.4 | **19.0** | 49.7 |
| AutoCap (audio+text) | AC+CL+WC | 72.3 | **29.7** | 51.8 | 25.3 | **83.2** | 18.2 | **50.7** |

### 4.1 Automatic Audio Captioning

**Training dataset and details.** We train our captioning model in two stages. During stage 1, we pretrain on a large weakly labeled dataset of 634,208 audio clips, constructed from AudioSet [29], Freesound, BBC Sound Effects, SoundBible, AudioCaps [46], and Clotho [19] datasets. We use the ground truth captions from AudioCaps and Clotho dataset, WavCaps captions for Freesound, SoundBible, and BBC Sound Effects, and handcrafted captions through a template leveraging the provided ground truth class labels for AudioSet. We use the title provided with each clip, and pre-compute video captions using a pretrained Panda70M model [9] for the clips that are associated with video modality and pass an empty string otherwise. We pretrain the model for 20 epochs with a learning rate of 1e-4. We keep the audio encoder and pretrained BART frozen during this stage. In Stage 2, we fine-tune the model for 20 epochs only on AudioCaps using a learning rate of 1e-5. We randomly sample 10-second clips at 32KHz for all of our captioning experiments.

**Baselines.** We compare with ACT [76], V-ACT [71], BART-tags [32], AL-MixGEN [47], ENCLAP [49], HTSAT-BART [122] and CNext-trans [135]. Among these baselines, ENCLAP and CNext-trans achieve the best performance. ENCLAP benefits from a stronger audio encoder and the use of a CLAP representation for additional guidance. CNext-trans trains a lightweight transformer instead of fine-tuning a pretrained language model to reduce overfitting.

**Metrics and evaluation.** We report results using the the established BLEU1 [86], BLEU2 [86], ROUGE [63], Meteor [56], CIDEr [109], and SPIDEr [69] metrics. We evaluate our method on the AudioCaps test split using the last checkpoint. We use only 876 clips as some videos were deleted since the original data release. We follow the same evaluation pipeline as baselines and include their reported results. Results that were not provided in these publications are excluded from our analysis.

Table 4: Captioning ablation study on AudioCaps evaluating key captioning model components.

| Model | METEOR ↑ | CIDEr ↑ | SPICE ↑ | SPIDEr ↑ |
|---|---|---|---|---|
| Ours | **25.3** | **83.2** | 18.2 | **50.7** |
| - w/o CLAP | **25.3** | 80.7 | **18.4** | 49.6 |
| - w/o Stage 2 | 24.2 | 75.6 | 17.3 | 46.5 |
| - w/o Stage 1 | 22.6 | 59.6 | 15.4 | 37.5 |
| - Unfreeze Word Embedding | 22.5 | 82.6 | 18.1 | 50.4 |

**Results.** In Tab. 3 we report the quantitative comparison. Our method outperforms previous methods on all metrics, achieving notable improvements in the CIDEr and BLUE1 scores, with values of 83.2 and 73.1, respectively. We found that incorporating metadata significantly enhances the CIDEr scores but slightly reduces the SPICE scores. This trade-off likely results from the enhanced descriptive detail brought by the metadata, which, while enriching the content, introduces noise that may compromise

Table 5: GenAu results on AudioCaps test split.

| Model | Prams | FD ↓ | IS ↑ | FAD ↓ | $\text{CLAP}_{LAION}$ ↑ | $\text{CLAP}_{MS}$ ↑ |
|---|---|---|---|---|---|---|
| GroundTruth | - | - | - | - | 0.251 | 0.671 |
| AudioLDM-L [65] | 739M | 37.89 | 7.14 | 5.86 | - | 0.429 |
| AudioLDM 2-L [66] | 712M | 32.50 | 8.54 | 5.11 | 0.212 | 0.621 |
| TANGO [30] | 866M | 26.13 | 8.23 | 1.87 | 0.185 | 0.597 |
| TANGO 2 [74] | 866M | 19.77 | 8.45 | 2.74 | 0.264 | 0.590 |
| Make-An-Audio [43] | 453M | 27.93 | 7.44 | 2.59 | 0.207 | 0.621 |
| Make-An-Audio 2 [42] | 937M | **15.34** | 9.58 | 1.27 | 0.251 | 0.645 |
| GenAu-Small | 493M | <u>15.54</u> | <u>11.10</u> | **1.07** | <u>0.271</u> | **0.674** |
| GenAu-Large | 1.25B | 16.51 | **11.75** | <u>1.21</u> | **0.285** | <u>0.668</u> |

Table 6: GenAu ablation results on AudioCaps. Training is performed on AudioCaps only, or on our full dataset with or without AutoCap re-captioning.

| Model | Training Data | Re-Captioning | FD ↓ | FAD ↓ | IS ↑ |
|---|---|---|---|---|---|
| GenAu-Small | AC | N/A | 15.96 | 1.21 | 9.55 |
| GenAu-Large | AC | N/A | 16.89 | 1.51 | 10.25 |
| GenAu-Small | Full | | 17.02 | 1.85 | 10.52 |
| GenAu-U-Net | Full | ✓ | 25.57 | 1.98 | 9.54 |
| GenAu-Small | Full | ✓ | **15.54** | **1.07** | 11.10 |
| GenAu-Large | Full | ✓ | 16.51 | 1.21 | **11.75** |

the model's semantic precision. In addition, AudioCaps is labeled based on audio information alone, thus evaluation penalizes the description of information that can not be deduced with certainty from the audio modality only, such as the specific type of object producing a rustling sound. Compared to ENCLAP-Large [49], and CoNeTTE (CNext-trans) [135], we find the captions produced by our method to be more descriptive and precise with a better temporal understanding. ENCLAP-Large often misses important details and exhibits lower temporal accuracy. CNext-trans, while accurate, often produces short captions that lack details. We include qualitative comparisons in the project *Website*. Moreover, AutoCap is *four times* faster than ENCALP, producing a caption for a 10-second clip in 0.28 seconds, compared to ENCALP which takes 1.12 seconds. Furthermore, we observe consistent improvements when pretraining on a large scale of weakly-labeled data during first stage, validating the effectiveness of our training strategy in benefiting from a larger, weakly-labeled dataset.

**Ablations.** In Tab. 4, we ablate model design choices. We observe the use of the CLAP embedding to bring a 2.5 points increase in the CIDEr score. We also validate that when not performing Stage 2 training, which involves finetuning of the BART [58] model, performance degrades on all metrics, a finding we attribute to the necessity of adapting BART's decoder to the sentence structure typical of AudioCaps. A more severe degradation in performance is observed if Stage 1 is not performed, with the misaligned representation between the encoder and the decoder causing catastrophic forgetting in the language model. Finally, if BART word embeddings are finetuned in Stage 2 instead of being kept frozen, we observe a slight performance degradation.

### 4.2 Text-2-Audio Generation

**Training dataset and details.** We use our best-performing captioning model to re-caption the WavCaps dataset. In addition, we obtain 339,387 videos from AudioSet and 126,905 videos from VGGSounds, totaling 761,113 clips. For those obtained from sound-only platforms, we input an empty string as the video caption. For full details of the data sources of our training dataset, please refer to Tab. 1. We additionally use Clotho and AudioCaps training datasets with their ground truth caption. To stay consistent with baselines, we train at 16kHz resolution. We use a patch size of 1 and a group size of 32. We use LAMB optimizer [128] with a LR of 5e-3. We train for 220k steps and choose the checkpoint with the highest IS, at step 210k and 207k for the large and small model.

**Baselines.** We compare with TANGO 1 & 2, [30], AudioLDM 1 & 2 [65, 66], and Make-An-Audio 1 & 2 [43, 42]. Both AudioLDM and Make-an-Audio train a UNet-based latent diffusion models [92] on Mel-Spectrogram representation of the audio, by regarding the Mel-Spectrogram as a single channel image, and use a pretrained CLAP encoder to condition the generation on an input prompt. TANGO proposed to use FLAN-T5 [14] as the text encoder and reported significant improvements. AudioLDM-2 and Make-an-Audio-2 proposed to use a dual encoder strategy of a T5 [90] and CLAP encoder. AudioLDM-2 focused on extending the generation and conditioning to various domains. Specifically, they use language of audio (LOA) to condition the generation on images, audio, or transcripts and train their model for music and speech generation. Make-an-Audio-2 proposes to use a 1D VAE representation and employ a feed-forward Transformer-based model to replace the UNet. Recently, Tango-2 proposed to use instruction fine-tuning on a synthetic dataset to enhance the temporal understanding. In our experiments, we focus on text-conditioned natural audio generation and generate 10s clips at a resolution of 16Khz. We only provide a qualitative comparison with StableAudio [25] in the project webpage since it reports different metrics without any checkpoints.

**Metrics.** We compare the performance of our method with baselines using the standard Frechet Distance (FD), Inception score (IS), and CLAP score on Audioset test dataset, containing 964 samples. These metrics use pretrain PANNs [51] model features. Some prior work also reported the Frechet Distance results using the VGGish network [39], denoted as (FAD) [45]. Additionally, to compute the CLAP score, some prior work [66] used CLAP from LAION [119], while others [74, 43, 42] used CLAP from Microsoft [21]. To avoid confusion, we report both metrics and denote them as $\text{CLAP}_{LAION}$ and $\text{CLAP}_{MS}$, respectively. We follow the same evaluation protocols of AudioLDM [65] and use the AudioLDM evaluation package to compute the metrics. Due to inconsistencies in evaluation pipelines and varying results for the same baselines reported in different studies, we have chosen to recompute all metrics using the official checkpoints to ensure consistent comparisons.

**Results.** In Tab. 5, we report evaluation results. Our method achieves superior performance compared to the state-of-the-art methods in terms of IS, FAD, and $\text{CLAP}_{LAION}$ scores, marking an improvement of $22.7\%$, $15.7\%$, and $13.5\%$, respectively. This shows that GenAu can produce high audio quality and achieve better semantic alignment with the conditioning text. We also find that, while the smaller model can learn the target distribution better, as noted by a lower FD and FAD score, the larger model achieves better IS and $\text{CLAP}_{LAION}$, indicating that more scaling to the audio generator might result in even more enhancement of the audio quality and its alignment with the input prompt.

**Ablations.** We evaluate our main method variations in Tab. 6. We first test the effect of scaling up the dataset size. When compared with training the small model on AC only, we observe that training the small model on the full dataset without re-captioning produces improvements in IS but degrades both FD and FAD. Instead, when training on the full dataset after re-captioning, we notice consistent and significant improvements on all metrics for both the small and large models. This demonstrates that scaling data introduces significant improvements, but only if the annotation quality is adequately curated. Second, scaling up the model size gives noticeable improvements, especially in audio quality whether the model was trained on the full dataset or only AC. Finally, we compare our small model against a U-Net [93] baseline at a similar computational complexity. We notice significant improvements across all metrics enabled by our scalable transformer architecture.

## 5 Conclusion

We take a holistic approach to improving the quality of existing audio generators. starting by addressing the scarcity of large-scale captioned audio datasets, we build a state-of-the-art audio captioning method, AutoCap, which leverages audio metadata to collect a dataset of $761,000$ annotated audio clips. We then built a latent diffusion model based on a scalable transformer architecture which we train on our re-captioned dataset to obtain AutoCap, a state-of-the-art model for audio generation. Our approach not only enhances audio generation but also broadens potential applications. Since AutoCap is fully automated, it can be used to obtain an audio dataset that order of magnitudes larger than the current available ones. Additionally, since AutoCap could be used for captioning current video datasets, it enables novel applications text-to-audio-video joint generation, a more natural and desired choice of video generation. Moreover, our scalable audio generator, GenAu, could be extended to other domains such as speech and music.

## References


[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text. *arXiv*, 2023.

[3] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions.

[4] Gehui Chen, Guan'an Wang, Xiaowen Huang, and Jitao Sang. Semantically consistent video-to-audio generation using multimodal language large model, 2024.

[5] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[6] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[7] Li-Wei Chen, Shinji Watanabe, and Alexander I. Rudnicky. A vector quantized approach for text to speech synthesis on real-world spontaneous speech. *ArXiv*, 2023.

[8] Ting Chen and Lala Li. Fit: Far-reaching interleaved transformers. *arXiv*, 2023.

[9] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[10] Ziyang Chen, Daniel Geng, and Andrew Owens. Images that sound: Composing images and sounds on a single canvas, 2024.

[11] Yen-Chi Cheng, Hsin-Ying Lee, Sergey Tulyakov, Alexander G Schwing, and Liang-Yan Gui. SDFusion: Multimodal 3d shape completion, reconstruction, and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4456–4465, 2023.

[12] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms, 2024.

[13] Kratzenstein Christian, G. *Tentamen resolvendi problema*. Academiae scientiarvm, 1781.

[14] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, 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. *arXiv*, 2022.

[15] Matéo Cousin, Étienne Labbé, and Thomas Pellegrini. Multilingual Audio Captioning using machine translated data. *arXiv e-prints*, art. arXiv:2309.07615, September 2023. doi: 10.48550/arXiv.2309.07615.

[16] Soham Deshmukh, Benjamin Elizalde, Dimitra Emmanouilidou, Bhiksha Raj, Rita Singh, and Huaming Wang. Training audio captioning models without audio, 2023.

[17] Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. Pengi: An audio language model for audio tasks, 2024.

[18] Soham Deshmukh, Rita Singh, and Bhiksha Raj. Domain adaptation for contrastive audio-language models, 2024.

[19] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020.

[20] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv*, 2022.

[21] 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 (ICASSP)*. IEEE, 2023.

[22] Benjamin Elizalde, Soham Deshmukh, and Huaming Wang. Natural language supervision for general-purpose audio representations, 2024.

[23] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis, 2021.

[24] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv*, 2024.

[25] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion. In *International Conference on Machine Learning (ICML)*, 2024.

[26] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons. Fast timing-conditioned latent audio diffusion, 2024.

[27] Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Long-form music generation with latent diffusion, 2024.

[28] Zhengcong Fei, Mingyuan Fan, and Junshi Huang. Music consistency models, 2024.

[29] Jort F. Gemmeke, Daniel P. W. 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 (ICASSP)*, 2017.

[30] Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3590–3598, 2023.

[31] Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand, 2024.

[32] Félix Gontier, Romain Serizel, and Christophe Cerisara. Automated audio captioning by fine-tuning bart with audioset tags. In *DCASE 2021 - 6th Workshop on Detection and Classification of Acoustic Scenes and Events*, 2021.

[33] Wenhao Guan, Kaidi Wang, Wangjin Zhou, Yang Wang, Feng Deng, Hui Wang, Lin Li, Qingyang Hong, and Yong Qin. Lafma: A latent flow matching model for text-to-audio generation, 2024.

[34] Haohan Guo, Hui Lu, Xixin Wu, and Helen M. Meng. A multi-scale time-frequency spectrogram discriminator for gan-based non-autoregressive tts. In *Proc. INTERSPEECH 2022*, 2022.

[35] Zhifang Guo, Jianguo Mao, Rui Tao, Long Yan, Kazushige Ouchi, Hong Liu, and Xiangdong Wang. Audio generation with multiple conditional diffusion model, 2023.

[36] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv*, 2023.

[37] Moayed Haji-Ali, Guha Balakrishnan, and Vicente Ordonez. ElasticDiffusion: Training-free Arbitrary Size Image Generation through Global-Local Content Separation. *arXiv*, abs/2311.18822, 2023.

[38] Akio Hayakawa, Masato Ishii, Takashi Shibuya, and Yuki Mitsufuji. Discriminator-guided cooperative diffusion for joint audio and video generation, 2024.

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

[40] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[41] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[42] 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*, 2023.

[43] 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. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[44] Marek Kadlčík, Adam Hájek, Jürgen Kieslich, and Radosław Winiecki. A whisper transformer for audio captioning trained with synthetic captions and transfer learning, 2023.

[45] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms, 2019.

[46] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

[47] Eungbeom Kim, Jinhee Kim, Yoori Oh, Kyungsu Kim, Minju Park, Jaeheon Sim, Jinwoo Lee, and Kyogu Lee. Exploring train and test-time augmentations for audio-language learning. *arXiv preprint arXiv:2210.17143*, 2022.

[48] Gwanghyun Kim, Alonso Martinez, Yu-Chuan Su, Brendan Jou, José Lezama, Agrim Gupta, Lijun Yu, Lu Jiang, Aren Jansen, Jacob Walker, and Krishna Somandepalli. A versatile diffusion transformer with mixture of noise levels for audiovisual generation, 2024.

[49] Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo. Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[50] 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 (NeurIPS)*, 2020.

[51] 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*, 2019.

[52] Junghyun Koo, Gordon Wichern, Francois G. Germain, Sameer Khurana, and Jonathan Le Roux. Smitin: Self-monitored inference-time intervention for generative music transformers, 2024.

[53] 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 *The Eleventh International Conference on Learning Representations*, 2023.

[54] Etienne Labbé, Thomas Pellegrini, and Julien Pinquier. Killing two birds with one stone: Can an audio captioning system also be used for audio-text retrieval?, 2023.

[55] Max W. Y. Lam, Qiao Tian, Tang-Chun Li, Zongyu Yin, Siyuan Feng, Ming Tu, Yuliang Ji, Rui Xia, Mingbo Ma, Xuchen Song, Jitong Chen, Yuping Wang, and Yuxuan Wang. Efficient neural music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[56] Alon Lavie and Abhaya Agarwal. Meteor: an automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, 2007.

[57] Jean-Marie Lemercier, Simon Rouard, Jade Copet, Yossi Adi, and Alexandre Défossez. An independence-promoting loss for music generation with language models, 2024.

[58] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[59] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning (ICML)*, 2023.

[60] Muheng Li, Yueqi Duan, Jie Zhou, and Jiwen Lu. Diffusion-sdf: Text-to-shape via voxelized diffusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[61] Yiming Li, Xiangdong Wang, and Hong Liu. Audio-free prompt tuning for language-audio models, 2023.

[62] Jinhua Liang, Huan Zhang, Haohe Liu, Yin Cao, Qiuqiang Kong, Xubo Liu, Wenwu Wang, Mark D. Plumbley, Huy Phan, and Emmanouil Benetos. Wavcraft: Audio editing and generation with large language models, 2024.

[63] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004.

[64] Haohe Liu, Ke Chen, Qiao Tian, Wenwu Wang, and Mark D. Plumbley. Audiosr: Versatile audio super-resolution at scale, 2023.

[65] 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. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[66] 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.

[67] Huadai Liu, Rongjie Huang, Yang Liu, Hengyuan Cao, Jialei Wang, Xize Cheng, Siqi Zheng, and Zhou Zhao. Audiolcm: Text-to-audio generation with latent consistency models, 2024.

[68] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *AAAI Conference on Artificial Intelligence*, 2021.

[69] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[70] Xubo Liu, Qiushi Huang, Xinhao Mei, Haohe Liu, Qiuqiang Kong, Jianyuan Sun, Shengchen Li, Tom Ko, Yu Zhang, Lilian H. Tang, Mark D. Plumbley, Volkan Kılıç, and Wenwu Wang. Visually-Aware Audio Captioning With Adaptive Audio-Visual Attention. In *Proc. INTERSPEECH 2023*, 2023.

[71] Xubo Liu, Qiushi Huang, Xinhao Mei, Haohe Liu, Qiuqiang Kong, Jianyuan Sun, Shengchen Li, Tom Ko, Yu Zhang, Lilian H. Tang, Mark D. Plumbley, Volkan Kılıç, and Wenwu Wang. Visually-aware audio captioning with adaptive audio-visual attention, 2023.

[72] Xubo Liu, Qiuqiang Kong, Yan Zhao, Haohe Liu, Yi Yuan, Yuzhuo Liu, Rui Xia, Yuxuan Wang, Mark D. Plumbley, and Wenwu Wang. Separate anything you describe, 2023.

[73] Rehana Mahfuz, Yinyi Guo, and Erik Visser. Improving audio captioning using semantic similarity metrics, 2023.

[74] Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization, 2024.

[75] Yuxin Mao, Xuyang Shen, Jing Zhang, Zhen Qin, Jinxing Zhou, Mochu Xiang, Yiran Zhong, and Yuchao Dai. Tavgbench: Benchmarking text to audible-video generation, 2024.

[76] Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D. Plumbley, and Wenwu Wang. Audio captioning transformer, 2021.

[77] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv*, 2023.

[78] Xinhao Mei, Varun Nagaraja, Gael Le Lan, Zhaoheng Ni, Ernie Chang, Yangyang Shi, and Vikas Chandra. Foleygen: Visually-guided audio generation, 2023.

[79] Jan Melechovsky, Zixun Guo, Deepanway Ghosal, Navonil Majumder, Dorien Herremans, and Soujanya Poria. Mustango: Toward controllable text-to-music generation, 2024.

[80] Willi Menapace, Aliaksandr Siarohin, Ivan Skorokhodov, Ekaterina Deyneka, Tsai-Shien Chen, Anil Kag, Yuwei Fang, Aleksei Stoliar, Elisa Ricci, Jian Ren, and Sergey Tulyakov. Snap video: Scaled spatiotemporal transformers for text-to-video synthesis, 2024.

[81] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[82] Javier Nistal, Marco Pasini, Cyran Aouameur, Maarten Grachten, and Stefan Lattner. Diff-a-riff: Musical accompaniment co-creation via latent diffusion models, 2024.

[83] Xinlei Niu, Jing Zhang, Christian Walder, and Charles Patrick Martin. Soundlocd: An efficient conditional discrete contrastive latent diffusion model for text-to-sound generation, 2024.

[84] Zachary Novack, Julian McAuley, Taylor Berg-Kirkpatrick, and Nicholas J. Bryan. Ditto: Diffusion inference-time t-optimization for music generation, 2024.

[85] Francesco Paissan, Luca Della Libera, Zhepei Wang, Mirco Ravanelli, Paris Smaragdis, and Cem Subakan. Audio editing with non-rigid text prompts, 2024.

[86] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 2002.

[87] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[88] Emilian Postolache, Giorgio Mariani, Luca Cosmo, Emmanouil Benetos, and Emanuele Rodolà. Generalized multi-source inference for text conditioned music diffusion models, 2024.

[89] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling, 2023.

[90] 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 (JMLR)*, 2022.

[91] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[92] 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 (CVPR)*, 2022.

[93] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[94] 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.

[95] Koichi Saito, Dongjun Kim, Takashi Shibuya, Chieh-Hsin Lai, Zhi Zhong, Yuhta Takida, and Yuki Mitsufuji. Soundctm: Uniting score-based and consistency models for text-to-sound generation, 2024.

[96] Leonard Salewski, Stefan Fauth, A. Sophia Koepke, and Zeynep Akata. Zero-shot audio captioning with audio-language model guidance and audio context keywords, 2023.

[97] Yangyang Shi, Gael Le Lan, Varun Nagaraja, Zhaoheng Ni, Xinhao Mei, Ernie Chang, Forrest Iandola, Yang Liu, and Vikas Chandra. Enhance audio generation controllability through representation similarity regularization, 2023.

[98] Jaehyeok Shim, Changwoo Kang, and Kyungdon Joo. Diffusion-based signed distance fields for 3d shape generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20887–20897, June 2023.

[99] Fangxun Shu, Lei Zhang, Hao Jiang, and Cihang Xie. Audio-visual llm for video understanding, 2023.

[100] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv*, 2020.

[101] Arvind Krishna Sridhar, Yinyi Guo, Erik Visser, and Rehana Mahfuz. Parameter efficient audio captioning with faithful guidance using audio-text shared latent representation, 2023.

[102] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models, 2024.

[103] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context, interleaved, and interactive any-to-any generation, 2023.

[104] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion, 2023.

[105] Ye Tao, Chaofeng Lu, Meng Liu, Kai Xu, Tianyu Liu, Yunlong Tian, and Yongjie Du. A fast and high-quality text-to-speech method with compressed auxiliary corpus and limited target speaker corpus. In *International Conference on Language Resources and Evaluation*, 2024.

[106] Zeyue Tian, Zhaoyang Liu, Ruibin Yuan, Jiahao Pan, Xiaoqiang Huang, Qifeng Liu, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Vidmuse: A simple video-to-music generation framework with long-short-term modeling, 2024.

[107] Danial Samadi Vahdati, Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm. Beyond deepfake images: Detecting ai-generated videos, 2024.

[108] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[109] Ramakrishna Vedantam, C. Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[110] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual descriptions. In *International Conference on Learning Representations*, 2022.

[111] Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. Audiobox: Unified audio generation with natural language prompts, 2023.

[112] Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: A lightweight solution for vision-to-audio generation by connecting foundation models, 2023.

[113] Kai Wang, Shijian Deng, Jing Shi, Dimitrios Hatzinakos, and Yapeng Tian. Av-dit: Efficient audio-visual diffusion transformer for joint audio and video generation, 2024.

[114] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. Lavie: High-quality video generation with cascaded latent diffusion models, 2023.

[115] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation with rectified flow matching, 2024.

[116] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation with rectified flow matching, 2024.

[117] Kempelen Wolfgang. *Mechanismus der menschlichen Sprache nebst der Beschreibung seiner sprechenden Maschine*. 1791.

[118] Shih-Lun Wu, Xuankai Chang, Gordon Wichern, Jee weon Jung, François Germain, Jonathan Le Roux, and Shinji Watanabe. Improving audio captioning models with fine-grained audio features, text embedding supervision, and llm mix-up augmentation, 2024.

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

[120] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners, 2024.

[121] Manjie Xu, Chenxing Li, Duzhen zhang, Dan Su, Wei Liang, and Dong Yu. Prompt-guided precise audio editing with diffusion models, 2024.

[122] Yaoxun Xu, Hangting Chen, Jianwei Yu, Qiaochu Huang, Zhiyong Wu, Shixiong Zhang, Guangzhi Li, Yi Luo, and Rongzhi Gu. Secap: Speech emotion captioning with large language model, 2023.

[123] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[124] Jinlong Xue, Yayue Deng, Yingming Gao, and Ya Li. Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation, 2024.

[125] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, Zhou Zhao, Shinji Watanabe, and Helen Meng. Uniaudio: An audio foundation model toward universal audio generation, 2023.

[126] 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.

[127] Zhongjie Ye, Yuqing Wang, Helin Wang, Dongchao Yang, and Yuexian Zou. Featurecut: An adaptive data augmentation for automated audio captioning. In *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 313–318. IEEE, 2022.

[128] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes, 2020.

[129] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding, 2023.

[130] Yiming Zhang, Xuenan Xu, Ruoyi Du, Haohe Liu, Yuan Dong, Zheng-Hua Tan, Wenwu Wang, and Zhanyu Ma. Zero-shot audio captioning using soft and hard prompts, 2024.

[131] Yixiao Zhang, Akira Maezawa, Gus Xia, Kazuhiko Yamamoto, and Simon Dixon. Loop copilot: Conducting ai ensembles for music generation and iterative editing, 2023.

[132] Yixiao Zhang, Yukara Ikemiya, Woosung Choi, Naoki Murata, Marco A. Martínez-Ramírez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Instruct-musicgen: Unlocking text-to-music editing for music language models via instruction tuning, 2024.

[133] Yixiao Zhang, Yukara Ikemiya, Gus Xia, Naoki Murata, Marco A. Martínez-Ramírez, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Musicmagus: Zero-shot text-to-music editing via diffusion models, 2024.

[134] Ge Zhu, Jordan Darefsky, and Zhiyao Duan. Cacophony: An improved contrastive audio-text model, 2024.

[135] Étienne Labbé, Thomas Pellegrini, and Julien Pinquier. Conette: An efficient audio captioning system leveraging multiple datasets with task embedding. *arXiv*, 2023.

# A Limitations.

**Audio captioning.** Sounds emitted by different objects may be remarkably similar (*i*.e. a waterfall and heavy rain, or a can and a motorcycle engine). In these situations, if metadata does not have sufficient information, the audio captioning model may not be able to correctly disambiguate sounds. In addition, we find the model may not capture accurately time dependencies between different sounds in the audio and correctly disentangle foreground from background sounds. Additionally, since our captioning model is only finetuned on AudioCap, and Audicaps has a limited vocabulary (less than 1000 words), our model also tends to produce repetitive words and captions.

**Audio generation.** As we train the model to generate natural sound effects, our audio generation model underperforms in more specific domains where more targeted models might excel such as in music generation or text-to-speech synthesis. Additionally, even though we train on a large dataset, the vocabulary of the paired texts is limited making the model struggle to generate accurate audio for long and descriptive prompts.

# B Training Details

We train our audio generation model using the LAMB optimizer with a learning rate of $0.005$ using a cosine schedule, a weight decay of $0.1$ and a dropout factor of $0.1$. For the small model we train for 100k steps using a batch size of 512. For the large model variant, we train for 50k steps using a batch size of 960. For the last steps of model training, rather than training on the full dataset, we fintune it on AudioCaps only as it constitutes the dataset with captions of highest quality.

# C Additional Evaluation

In this section, we present additional evaluation details and results which are complemented by our *Website*.

## C.1 Evaluation details

While the established practice in the evaluation of audio captioning methods is to report the results on the test set using the checkpoint that performs best on the validation subset, prior work [135, 49] reported high instability of the metrics on the validation subset and weak correlation between the validation and test performance, making the model's results vary significantly for different seeds. To alleviate this, ENCLAP [49] selects around five best performing validation checkpoints and report their best results on the test set. CNext-trans [135] uses FENSE score to pick the best validation checkpoint. Our model, thanks to the two-stage training paradigm, significantly reduces this instability and we observe steady performance gains as training progresses. Therefore, we report the results at convergence, specifically after 20 epochs.

## C.2 Additional Captioning Evaluation

In Tab. 7 we show qualitative results of the captions produced by our method and compare it with state-of-the-art AAC methods. See the *Website* for qualitative results accompanied by the original audio. While ENCLAP [49] and CoNeTTE [135] tend to produce shord captions, our method produces the most descriptive captions, capturing the most amount of elements from the ground truth, an important capability to allow high-quality audio generation [3].

## C.3 Additional Audio Generation Evaluation

In this section, we report additional evaluation results and ablations on the task of audio generation.

In Tab. 8, we evaluate fundamental architectural choices in the design of our scalable FIT model. When removing either the Flan-T5 or CLAP encodings, we notice a steady reduction in all metrics. When increasing the number of latent tokens we also notice a steady improvement in performance as more compute is allocated to the model. Similarly, increasing the patch size to 2 results in a performance decrease under all metrics due to the reduced amount of allocated computation.

Table 7: Qualitative comparison of captioning results on the AudioCaps dataset. See the *Website* for qualitative results accompanied by the respective audio.

| Method | Caption |
|---|---|
| Ground Truth | *A man talking as ocean waves trickle and splash while wind blows into a microphone* |
| Ours | *A man speaks as wind blows and water splashes* |
| CoNeTTE [135] | *A man is speaking and wind is blowing* |
| ENCLAP [49] | *A man is speaking and wind is blowing* |
| Ground Truth | *An adult male speaks, birds chirp in the background, and many insects are buzzing* |
| Ours | *Birds chirp in the distance, followed by a man speaking nearby, after which insects buzz nearby* |
| CoNeTTE [135] | *A man speaking with birds chirping in the background.* |
| ENCLAP [49] | *Birds are chirping and a man speaks* |
| Ground Truth | *A telephone dialing tone followed by a plastic switch flipping on and off* |
| Ours | *A telephone dialing followed by a series of plastic clicking then plastic clanking before plastic thumps on a surface* |
| CoNeTTE [135] | *A telephone ringing followed by a beep.* |
| ENCLAP [49] | *A telephone dialing followed by a series of electronic beeps* |
| Ground Truth | *A running train and then a train whistle* |
| Ours | *A train moves getting closer and a horn is triggered* |
| CoNeTTE [135] | *A train horn blows and a steam whistle is blowing* |
| ENCLAP [49] | *A train running on railroad tracks followed by a train horn blowing as wind blows into a microphone* |
| Ground Truth | *A child is speaking followed by a door moving* |
| Ours | *A child speaks followed by a loud crash and a scream* |
| CoNeTTE [135] | *A woman speaking followed by a door opening and closing.* |
| ENCLAP [49] | *A young girl speaks followed by a loud bang* |
| Ground Truth | *Water splashing as a baby is laughing and birds chirp in the background* |
| Ours | *A baby laughs and splashes, and an adult female speaks* |
| CoNeTTE [135] | *A baby is laughing and people are talking.* |
| ENCLAP [49] | *A baby laughs and splashes in water* |
| Ground Truth | *Leaves rustling in the wind with dogs barking and birds chirping* |
| Ours | *Birds chirp in the distance, and then a dog barks nearby* |
| CoNeTTE [135] | *A dog is barking and a person is walking.* |
| ENCLAP [49] | *Birds chirp and a dog barks* |
| Ground Truth | *Tapping followed by water spraying and more tapping* |
| Ours | *Some light rustling followed by a clank then water pouring* |
| CoNeTTE [135] | *A toilet is flushed and water is running.* |
| ENCLAP [49] | *A faucet is turned on and runs* |

In Tab. 9, we ablate the 1D-VAE bottleneck size in terms of reconstruction loss and performance of a subsequently trained latent audio diffusion model, in terms of FAD, FD and IS. Similarly to the phenomenon observed in the image and video generation domain [36, 24], we observe that larger number of channels allocated to the latent space result in lower reconstruction losses, but make the latent space more complex, hindering generation quality. We adopt 64 1D-VAE channels for all our experiments.

## D Computational resources

We train our models on A100 80GB GPUs. Our capioning model is trained 9 hours on 8 A100 80GB GPUs. Our largest audio generation model is trained for 48h on 48 A100 80GB GPUs.

Table 8: Ablation of different FIT architectural variations in terms of patch size number of latent tokens and adopted text encoders on the AudioCaps dataset.

| Tokens | Patch size | FLAN-T5 | CLAP | FD ↓ | FAD ↓ | IS ↑ |
|---|---|---|---|---|---|---|
| 256 | 1 | ✓ | ✓ | **16.45** | **1.29** | **10.26** |
| 256 | 1 | | ✓ | 17.41 | 1.39 | 10.0 |
| 256 | 1 | ✓ | | 20.47 | 1.86 | 8.89 |
| 384 | 1 | | ✓ | 17.41 | 1.39 | 10.0 |
| 192 | 1 | | ✓ | 18.0.1 | 2.01 | 8.91 |
| 128 | 1 | | ✓ | 25.56 | 1.77 | 7.49 |
| 256 | 2 | ✓ | ✓ | 18.53 | 1.70 | 9.0 |

Table 9: Ablation of different 1D-VAE designs on audio generation on the AudioCaps dataset.

| Channels | Recon. loss | FAD ↓ | FD ↓ | IS ↑ |
|---|---|---|---|---|
| 64 | 0.159 | **1.29** | **16.45** | **10.26** |
| 128 | 0.107 | 1.43 | 16.78 | 10.11 |
| 256 | **0.064** | 1.80 | 18.63 | 9.43 |