CLIPSynth: Learning Text-to-audio Synthesis from Videos using CLIP and Diffusion Models

Hao-Wen Dong[1,2]* Gunnar A. Sigurdsson[1] Chenyang Tao[1] Jiun-Yu Kao[1] Yu-Hsiang Lin[1]
Anjali Narayan-Chen[1] Arpit Gupta[1] Tagyoung Chung[1] Jing Huang[1] Nanyun Peng[1,3] Wenbo Zhao[1]

[1]*Amazon Alexa AI* [2]*University of California San Diego* [3]*University of California, Los Angeles*

Abstract

We propose CLIPSynth, a self-supervised text-queried sound synthesis model that can be trained with unlabeled videos in the wild. During training, the CLIPSynth model first projects an image (a video frame) to a text-image embedding space using the contrastive language-image pretraining (CLIP) model, and then synthesizes a mel spectrogram using a diffusion model conditioned on the image embedding. At inference time, we perform a zero-shot modality transfer by projecting a text query to the same text-image embedding space, and synthesize the text embedding into a mel spectrogram. We evaluate CLIPSynth on the MUSIC and VGG-Sound datasets with both objective evaluation metrics and a subjective listening test. Our experimental results show that CLIPSynth can generate realistic instrumental and generic sounds relevant to the input text queries. Moreover, CLIPSynth outperforms a retrieval-based baseline on MUSIC in terms of the Fréchet audio distance.

1. Introduction

Prior work has shown that modern machine learning models can learn text-to-sound synthesis using a large amount of audio-text pairs as training data [9, 15]. However, we argue that humans do not learn the sounds of an object this way. Unlike machines, humans incorporate multi-sensory inputs and learn the sounds of an object by associating the visual and auditory inputs [1, 2]. For example, by watching a cat meowing, humans can associate the meowing sound to the "sounding object," i.e., the cat here. Meanwhile, humans learn that this object is called a "cat" elsewhere. Motivated by this observation, we explore a more human-like approach for text-queried sound synthesis. In this paper, we propose a new self-supervised model for text-queried sound synthesis that uses naturally occurring learning signals such as videos more effectively without additional human annotations.

Inspired by [4], we propose to learn the desired text-audio correspondence by leveraging the image modality as a bridge. We adopt the contrastive language-image pretraining (CLIP) model [13] to handle the text-image correspondence and learn the image-audio correspondence from unlabeled videos, which are easier to acquire than paired text-audio data. Built upon this idea, we propose the CLIPSynth model for self-supervised text-queried sound synthesis. The proposed model consists of a query encoder followed by a synthesizer: the query encoder is a frozen CLIP model that encodes an image frame from a video to the joint text-image embedding space at training time; the synthesizer is a conditional denoising diffusion model [5, 12] that learns to generate the mel spectrogram of the audio of the same video conditioned on the query embedding. To the best of our knowledge, the proposed CLIPSynth model represents the *first self-supervised text-queried sound synthesis model.*

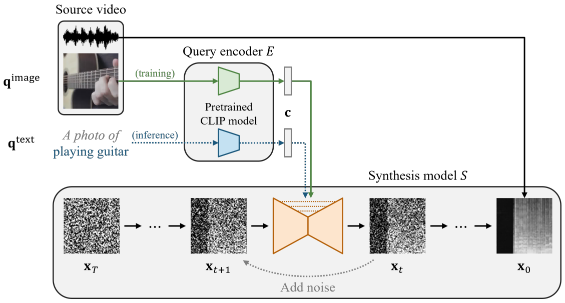

Figure 1. An illustration of the proposed CLIPSynth model.

To evaluate the proposed model, we compute objective evaluation metrics and conduct a subjective listening test following [9, 15]. We compare our proposed models against several variants and baseline models on the MUSIC [16] and VGG-Sound [3] datasets. We adopt the Hifi-GAN model [8] to synthesize the generated spectrograms into waveforms. Our experimental results show that the proposed model can generate realistic instrumental and generic sounds relevant to the input text queries. Moreover, CLIPSynth outperforms a retrieval-based baseline on MUSIC in terms of Fréchet audio distance.

Our contributions can be summarized as follows:

- We propose the first text-queried sound synthesis model that can be trained using only unlabeled videos.
- We show we can learn audio-textual correspondence using relatively easy-to-acquire audio-visual data by leveraging the zero-shot modality transfer capability of CLIP.

Audio samples can be found on our demo website.[1]

[*] Work done during an internship at Amazon.

[1] https://salu133445.github.io/clipsynth/

2. Method

2.1. CLIPSynth

Our architecture is inspired by CLIPSep [4], a framework designed for sound separation that learns the text-audio correspondence from unlabeled videos with the help of CLIP. As illustrated in Figure 1, the proposed framework consists of a query encoder and a conditional synthesizer. The query encoder E encodes the input query $\mathbf{q}$ into a condition vector $\mathbf{c} = E(\mathbf{q})$. The synthesizer S synthesizes the mel spectrogram $\tilde{\mathbf{x}}_0$ conditioned on the condition vector $\mathbf{c}$. We adopt the CLIP model [13] as the query encoder, and use the improved denoising diffusion model [12] as the synthesizer. During training, we feed an image $\mathbf{q}^{\text{image}}$ (e.g., a photo of a guitar) extracted from a video to the encoder, and the synthesizer is trained to predict the mel spectrogram of the associated audio. t inference time, instead of an image, we feed a text query $\mathbf{q}^{\text{text}}$ (e.g., "`a photo of playing guitar`") that describes the sound we want to synthesize to the query encoder. Leveraging the joint language-image embedding of the pretrained CLIP model, the query embedding from text input should be close to the embeddings of the images corresponding to the text; i.e., $E(\mathbf{q}^{\text{image}}) \approx E(\mathbf{q}^{\text{text}})$.

Instead of modeling raw waveforms directly, we propose to first synthesize mel spectrograms with the diffusion model introduced above, and then generate the synthesized mel spectrograms into waveforms using the Hifi-GAN model [8]. This way, we can base the diffusion model on a conditional U-Net [14] following [5], where the condition vector $\mathbf{c}$ is concatenated to the output features of each U-Net layer and then passed to the next layer. Moreover, we feed the texts to the CLIP text encoder in the form of "`a photo of {query}`" to reduce the modality gap as suggested by [13].

Implementation We used the pretrained CLIP model provided by [13] as it is without any further finetuning. Following [12], we used the hybrid learning objective and cosine noise schedule to train all the diffusion models. We used 4K diffusion steps during training and 1K steps during inference. We trained each model using a single NVIDIA A100 GPU. For the MUSIC dataset, we trained all the models for 100K steps, which took half a day, as the model tend to converge in 100K steps. For the VGG-Sound dataset, we trained all the models for 500K steps, which took two days. We trained the Hifi-GAN models for 500K steps on both datasets, which took three days. We used the AdamW optimizer for training the CLIPSynth model using the hyperparameters suggested by [12]. We trained the Hifi-GAN models from scratch using the implementation and hyperparameters provided by [8].

3. Experimental Setup

Data. Our training and test data consists of the MUSIC [16] and VGG-Sound [3] datasets. MUSIC contains more than 1,164 full-length instrument-playing videos downloaded from YouTube, covering 21 instrument classes. Since the MUSIC dataset contains full-length videos, we sliced the videos into 10-second chunks for training, and we ended up with 19,809 10-sec videos (55 hours in total) after the preprocessing. VGG-Sound contains 199,467 10-sec YouTube videos across 310 classes, and 166,702 videos (463 hours in total) remained usable after the pre-processing. Overall, VGG-Sound is more diverse yet noisy than MUSIC, and videos in VGG-Sound often contain much off-screen noise, including narration and background noise. For the preprocessing, we used a sampling rate of 16,000 Hz. For the spectrogram computation, we used a filter length of 2,048, a hop length of 1,024 and a window size of 2,048 in the short-time Fourier transform (STFT). We used 64 Mel bands so that we have 64-by-64 mel spectrograms. Each mel spectrogram encodes 4.16 seconds of audio.

Evaluation metrics. For the objective evaluation metrics, we follow Diffsound [15] and AudioGen [9] and use Fréchet Audio Distance (FAD), which has been shown to correlate well with human auditory perception [7]. FAD computes the Fréchet distance between distributions of deep features extracted by a pretrained audio classifier on the machine-generated samples against real audio samples. A lower FAD suggests that the machine-generated samples are more similar to the ground truth samples. In addition, following [11], we compute the Fréchet Inception Distance (FID) on the generated mel spectrograms.

Baseline models. We wanted to compare our proposed model with the Diffsound [15] and AudioGen [9] models. However, the code released in [15] is incomplete and we cannot reproduce the results. The authors in [9] has not released the code at the time of submitting this work. Thus, we compare our proposed model against several baselines we implemented. First, we note that the proposed CLIPSynth model can be trained in multiple ways. In addition to the fully self-supervised proposed CLIPSynth model which we train on video frames in the wild, we consider two variants:

- **CLIPSynth-Text** shares the same network architecture as CLIPSynth but is trained on text queries.
- **CLIPSynth-Hybrid** has the same network architecture as CLIPSynth but is trained on both text and image queries.

Note that both CLIPSynth-Text and CLIPSynth-Hybrid require audio-text pairs for training. Hence, they are *not* self-supervised models. Moreover, we consider the following baseline models:

- **CLIPRetriever** finds the image that is closest to the input text query in the CLIP embedding space and returns the associated audio of that image. Note that this is a *retrieval-based* model, not a generative model.

In addition, we also include the FAD values for Hifi-GAN reconstructed audios, which are obtained by extracting mel

Table 1. Results of the objective evaluation. The colors indicate a lower or higher FID/FAD than that of CLIPSynth.

Model	Generative	Unlabeled data only	Query Type		MUSIC		VGG-Sound	
			Training	Test	FAD↓	FID↓	FAD↓	FID↓
CLIPSynth (proposed)	✓	✓	Image	Text	6.30	40.12	8.68	34.63
CLIPSynth-Text	✓	×	Text	Text	10.32	22.00	6.78	27.50
CLIPSynth-Hybrid	✓	×	Image+Text	Text	6.21	22.62	5.83	25.88
CLIPSynth	✓	✓	Image	Image	2.41	19.30	5.49	24.56
SpecVQGAN [6]	✓	✓	Image	Image	33.45*	-	7.70*	-
CLIPSynth-Text	✓	×	Text	Image	25.96	47.92	8.92	38.44
CLIPSynth-Hybrid	✓	×	Image+Text	Image	4.92	20.52	5.89	25.88
CLIPRetriever (retrieval-based)	×	×	-	Text	10.36	-	2.43	-
Hifi-GAN reconstructions	×	-	-	-	2.64	-	4.09	-

*We used a pretrained model trained on VGG-Sound released by the authors since we could not reproduce their results when training the model from scratch.

spectrograms of the ground truth audios and converting them back to waveforms using Hifi-GAN.

4. Results

4.1. Quantitative results

To quantitatively evaluate the proposed system against the baseline models, we converted the class names into text queries using a query template for the text-queried models. Specifically, we use the query template "`a photo of playing {query}`" on MUSIC and "`a photo of {query}`" on VGG-Sound . For the image-queried models, we randomly extracted frames from the test videos as the image queries. We then used these randomly sampled queries to generate 512 samples for each model as this size has been shown sufficient to produce a stable FAD score [7]. Finally, we computed the FAD score between the set of generated audio samples and the set of all audio tracks in the entire test set (10% of the dataset). Similarly, we computed the FID score between the set of generated mel spectrograms (treated as images) and the set of mel spetrograms (treated as images) of all test audio samples.

We show in Table 1 the objective evaluation results. In general, we see that a lower FAD score, which is computed on the generated waveforms, usually comes with a lower FID score, which is computed on the generated mel spectrograms. We see that the proposed CLIPSynth model achieves an FAD of 6.30 and 8.68 when tested with text-queries on the MUSIC and VGG-Sound datasets, respectively. Moreover, CLIPSynth outperforms CLIPRetriever, a retrieval-based baseline, on MUSIC. Further, the Hifi-GAN reconstructed audio achieves an FAD score of 2.64 and 4.06 on the MUSIC and VGG-Sound datasets, suggesting that the main performance bottleneck lies in the synthesis model rather than the Hifi-GAN model. CLIPSynth outperforms the SpecVQGAN model [6] on VGG-Sound when tested with image queries.

Further, we observe a significant zero-shot modality transfer gap when the training and test modalities differ. This is consistent with a recent study that shows a significant modality gap inside the learnt language-image embedding space [10]. We see that the CLIPSynth model trained on image queries achieves an FAD of 2.41 and 5.49 when tested with the same modality on the MUSIC and VGG-Sound datasets. However, it only achieves an FAD of 6.30 and 8.68 when tested with text queries, representing an FAD difference of 3.89 and 3.19. Similarly, while the CLIPSynth-Text model, which is trained using text queries, achieves an FAD of 10.32 and 6.78 with text queries on the MUSIC and VGG-Sound datasets, it only achieves an FAD of 25.96 and 8.92 when tested with image queries, representing an FAD difference of 15.64 and 2.14. By training the model with both image and text queries, the CLIPSynth-Hybrid achieves a smaller modality gap of 1.29 and 0.06 in FAD on MUSIC and VGG-Sound as compared to CLIPSynth and CLIPSynth-Text.

4.2. Subjective listening test

In addition to the objective evaluation metrics, we conducted a subjective listening test to assess the performance of our proposed model against several baselines. We recruited 30 evaluators via Amazon Mechanical Turk. Each survey participant was instructed to listen to 10 pairs of randomly selected audio samples generated by two different models using the same text query. In this pairwise A/B test, the survey participant was asked to select the preferred audio samples in terms of *audio quality* (regardless of relevance to the queries), *relevance* (to the queries), and *noise levels*. We considered five text-queried models: CLIPSynth, CLIPSynth-Text, CLIPSynth-Hybrid, and CLIPRetriever. To aggregate the results, we gave a score of 1 whenever a model won in an A/B test and 0 when it lost; both models got a score of 0.5 for a draw. We computed the average score each model received. Moreover, we fed 64-by-128 noise arrays to the model to generate longer music samples of 8.32 seconds

Table 2. Results of the subjective listening test.

Model	Unlabeled data only	Query Type		MUSIC			VGG-Sound		
		Training	Test	Quality↑	Relevance↑	Noise↓	Quality↑	Relevance↑	Noise↓
CLIPSynth (proposed)	✓	Image	Text	0.511	0.473	0.481	0.500	0.388	0.619
CLIPSynth-Text	×	Text	Text	0.405	0.505	0.510	0.405	0.505	0.500
CLIPSynth-Hybrid	×	Image+Text	Text	0.434	0.447	0.531	0.431	0.448	0.547
CLIPRetriever	✓	-	Text	0.724	0.653	0.398	0.750	0.712n	0.297

duration for evaluation purposes.

We report in Table 2 the scores of the subjective listening test. First, we can see that the retrieval-based CLIPRetriever model significantly outperforms the other models on both datasets. Second, the proposed CLIPSynth model outperforms CLIPSynth-Text in terms of audio quality on both datasets. However, on the VGG-Sound dataset, CLIPSynth obtains a low score on the relevance criterion. We hypothesize that this is partly due to the noisiness of VGG-Sound, which poses a challenge in learning the desired text-audio correspondence. However, on the MUSIC dataset, the proposed CLIPSynth model outperforms CLIPSynth-Hybrid in the relevance criterion. We note that CLIPSynth is the only self-supervised model in this comparison, and it was not always beaten by other systems in the A/B tests.

5. Conclusion

We presented CLIPSynth, a new self-supervised model for text-queried sound synthesis. We base the CLIPSynth model on CLIP and a conditional diffusion model to synthesize mel spectrograms from an input text query. We examined the proposed model on the clean MUSIC and noisy VGG-Sound datasets. The subjective and objective evaluations have demonstrated the effectiveness of this approach.

Limitations. We have observed several limitations of the proposed model. First, off-screen sounds pose a challenge in the proposed setting as the model will try to imagine something invisible from the image inputs, which increases the undesired zero-shot modality gap when transitioning from image queries at training time to text queries at inference time. Moreover, the proposed system cannot handle purely audio-relevant queries (e.g., “loud,” “quiet,” “high-pitched” and “low-pitched”) as they have little meaning in the visual domain. Still, we argue that in-the-wild videos are good candidates as training data for learning text-audio or image-audio correspondence as they provide rich information.

Future Work There are several future directions we would like to explore. First, we want to equip our model with the ability to generate different styles of outputs. Second, we want to enable combinatory prompts (e.g., “{query 1} and {query 2}”) and blending tones (e.g., “piano + guitar”). Moreover, prior work has also observed a significant modality gap in multi-modal contrastive representation learning [10]. By incorporating their proposed techniques, we can reduce the modality gap and consequently improve the performance of our proposed model. Finally, the self-supervised learning framework proposed in this work can be scaled to a larger collection of videos in the wild, and we leave this computationally-intense extension to future work.

References

[1] R. Arandjelović and A. Zisserman. Look, listen and learn. In *Proc. ICCV*, 2017.

[2] R. Arandjelović and A. Zisserman. Objects that sound. In *Proc. ECCV*, 2018.

[3] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. VGGSound: A large-scale audio-visual dataset. In *Proc. ICASSP*, 2020.

[4] H.-W. Dong, N. Takahashi, Y. Mitsufuji, J. McAuley, and T. Berg-Kirkpatrick. CLIPSep: Learning text-queried sound separation with noisy unlabeled videos. In *Proc. ICLR*, 2023.

[5] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proc. NeurIPS*, 2020.

[6] V. Iashin and E. Rahtu. Taming visually guided sound generation. In *Proc. BMVC*, 2021.

[7] K. Kilgour, M. Zuluaga, D. Roblek, and M. Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. In *Proc. INTERSPEECH*, 2019.

[8] J. Kong, J. Kim, and J. Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Proc. NeurIPS*, 2020.

[9] F. Kreuk, G. Synnaeve, A. Polyak, U. Singer, A. Défossez, J. Copet, D. Parikh, Y. Taigman, and Y. Adi. AudioGen: Textually guided audio generation. In *Proc. ICLR*, 2023.

[10] W. Liang, Y. Zhang, Y. Kwon, S. Yeung, and J. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Proc. NeurIPS*, 2022.

[11] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley. AudioLDM: Text-to-Audio Generation with Latent Diffusion Models. In *Proc. ICML*, 2023.

[12] A. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *Proc. ICML*, 2019.

[13] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proc. ICML*, 2021.

[14] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. MICCAI*, 2015.

[15] D. Yang, J. Yu, H. Wang, W. Wang, C. Weng, Y. Zou, and D. Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *arXiv preprint arXiv:2207.09983*, 2022.

[16] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. In *Proc. ECCV*, 2018.