

CLIPSynth: Self-supervised Text-queried Sound Synthesis



Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), *Gunnar A Sigurdsson* (/profile?id=~Gunnar_A_Sigurdsson1), *Chenyang Tao* (/profile?id=~Chenyang_Tao1), *Jiun-Yu Kao* (/profile?id=~Jiun-Yu_Kao1), *Yu-Hsiang Lin* (/profile?id=~Yu-Hsiang_Lin2), *Anjali Narayan-Chen* (/profile?id=~Anjali_Narayan-Chen1), *Arpit Gupta* (/profile?id=~Arpit_Gupta1), *Tagyoung Chung* (/profile?id=~Tagyoung_Chung2), *Jing Huang* (/profile?id=~Jing_Huang3), *Nanyun Peng* (/profile?id=~Nanyun_Peng1), *Wenbo Zhao* (/profile?id=~Wenbo_Zhao1)

18 Jan 2023 (modified: 15 Jun 2023) Submitted to ICML 2023 Conference, Area Chairs, Authors, Reviewers, Senior Area Chairs Revisions (/revisions?id=yI5AoKGqcV)

Keywords: sound synthesis, audio synthesis, multimodal learning, self-supervised learning, contrastive learning

TL;DR: We propose a new self-supervised model for text-queried sound synthesis that can be trained using only unlabeled videos

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.

Supplementary Material: zip (/attachment?id=yI5AoKGqcV&name=supplementary_material)

Paper Checklist Guidelines: I certify that all co-authors of this work have read and commit to adhering to the Paper Checklist Guidelines, Call for Papers and Publication Ethics.

Submission Number: 781

◀ ▶ ▾ ▹ ▸ ▹ ▸ ▹ ▸ Fully expand content

Sort: Newest First ☰ ☲ ☱ - = ▾ 🔗

Everyone Program Chairs Authors Area Chairs Senior Area Chairs 15 / 15 replies shown

Reviewers Submitted Reviewer 2Sj4 Reviewer ufcf Reviewer TfCZ Reviewer zQM5

✕

Paper Decision

Decision

✍ Program Chairs (👤 kyunghyun.cho@nyu.edu (/profile?id=kyunghyun.cho@nyu.edu), emma.p.brunskill@gmail.com (/profile?id=emma.p.brunskill@gmail.com), krausea@ethz.ch (/profile?id=krausea@ethz.ch), ebrun@cs.stanford.edu (/profile?id=ebrun@cs.stanford.edu), +8 more (/group/info?id=ICML.cc/2023/Conference/Program_Chairs))

📅 24 Apr 2023, 13:12 (modified: 24 Apr 2023, 14:36) 👁 Program Chairs, Authors

Decision: Reject

Comment:

This paper presents CLIPSynth, a novel text-to-audio generation model trained with unlabeled video datasets. While many of the reviewers agreed that the idea of leveraging an unlabeled video dataset is intriguing, some of the reviewers were unconvinced of the effectiveness or superiority of this approach compared to using text-audio pairs. Specifically, many results on VGG-Sound showed low relevancy, which was also shown in the subjective evaluation. Given the availability of large text-audio pairs datasets and recent research that utilizes them, it is necessary to demonstrate the competitiveness of the proposed method in terms of generation quality.

Despite the author's rebuttal, the reviewers' maintained their initial ratings. Since the major concerns that were pointed out cannot be overlooked, I recommend rejecting this paper.

Message to Area Chairs and Program Chairs

Official Comment

✍ Authors (👤 Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Gunnar A Sigurdsson (/profile?id=~Gunnar_A_Sigurdsson1), Chenyang Tao (/profile?id=~Chenyang_Tao1), Jiun-Yu Kao (/profile?id=~Jiun-Yu_Kao1), +7 more (/group/info?id=ICML.cc/2023/Conference/Submission781/Authors))

📅 19 Mar 2023, 23:43 👁 Program Chairs, Area Chairs, Authors, Senior Area Chairs

Comment:

Dear Area Chairs and Program Chairs,

We believe Reviewer ufcf did not read our paper carefully before writing the review for the following reasons.

- The reviewer commented

The constructed prompts are also strange, it's more natural to use "a sound of playing {query}" rather than "a photo of playing {query}".

However, we explained in Section 4.1 that

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 Radford et al. (2021).

Moreover, we examined the effects of query templates in Section 6.3. Apparently, Reviewer ufcf did not read our paper carefully as no other reviewers raised a similar question.

- The reviewer commented

As a text-guided sound synthesis, authors may compare their work to DiffSound, which needs to be added to the paper.

However, we clearly explained in Section 5.3 that

We wanted to compare our proposed model with the DiffSound (Yang et al., 2022) and AudioGen (Kreuk et al., 2022) models. However, the code released by Yang et al. is incomplete and we cannot reproduce the results.

- The reviewer asked

What's the criterion of these bolds in Table 2?

However, we explained clearly in the caption of Table 2 that

The bold values indicate the best performing model per category.

The other reviewers did not misunderstand any of the above points that we clearly explained in our manuscript. We hope the area chairs and program chairs can take this into consideration.

Thank you!



Checked

Official Comment Area Chair osQs 25 Mar 2023, 09:40

Program Chairs, Area Chairs, Authors, Senior Area Chairs

Comment:

Dear authors,

I've read the review and your comment. I agree that some of the questions were already explained in the paper. I have also recently found out that there could be a potential COI between the reviewer and this submission.

I'll take it into account.



Replying to Checked

Thank you

Official Comment

Authors (Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Gunnar A Sigurdsson (/profile?id=~Gunnar_A_Sigurdsson1), Chenyang Tao (/profile?id=~Chenyang_Tao1), Jiun-Yu Kao (/profile?id=~Jiun-Yu_Kao1), +7 more (/group/info?id=ICML.cc/2023/Conference/Submission781/Authors))

26 Mar 2023, 22:43 Program Chairs, Area Chairs, Authors, Senior Area Chairs

Comment:

We appreciate it!



Official Review of Submission781 by Reviewer ufcf

Official Review Reviewer ufcf 08 Mar 2023, 04:04 (modified: 14 Mar 2023, 07:23)

Program Chairs, Area Chairs, Authors, Reviewer ufcf, Reviewers Submitted, Senior Area Chairs

Revisions (/revisions?id=2YGI1KvT3X)

Summary:

The authors propose a text-queried sound synthesis model that can be trained with unlabeled videos in the wild. 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. Experimental results show that CLIPSynth can generate realistic instrumental and generic sounds relevant to the input text queries.

Strengths And Weaknesses:

Strengths

1. Experiments. The authors present the results with both qualitative and quantitative analysis and demonstrate the out-of-distribution generalization.
2. Clear presentation. CLIPSynth is clearly presented and demonstrated, and the idea is straightforward and easy to understand. Key components are illustrated in a proper way.

Weaknesses

1. Datasets. It's weird to study text-queried sound synthesis on video datasets. Since text-sound pairs data exist, why not use them for experimental studies? The constructed prompts are also strange, it's more natural to use "a sound of playing {query}" rather than "a photo of playing {query}".

2. Baselines. As a text-guided sound synthesis, authors may compare their work to Diffsound, which needs to be added to the paper. Besides, SpecVQGAN is a video-guided sound synthesis paper, which is inappropriate for comparison.
3. Experimental results. What's the criterion of these bolds in Table 2? For VGGSound, the FID (8.68) in CLIPSynth seems worse than that (6.78) in CLIPSynth-Text. It has conclusions different from MUSIC dataset, and can you explain why?
4. Subjective evaluation should have error bars around results, please refer to Imagen or Audiogen and update your results.
5. Novelty. CLIPSynth is a combination of a diffusion generator and text encoder, and I think this adaptation lacks novelty.

Questions:

1. I think it would be better to name it "language-free" instead of "self-supervised."
2. Could the model understand prompts with natural language descriptions? It is unclear since the presented demos are generated by simple words.

Limitations:

/

Ethics Flag: No

Soundness: 2 fair

Presentation: 3 good

Contribution: 2 fair

Rating: 3: Reject: For instance, a paper with technical flaws, weak evaluation, inadequate reproducibility and incompletely addressed ethical considerations.



Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes



Response to Reviewer ufcf

Official Comment

 Authors ( Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Gunnar A Sigurdsson (/profile?id=~Gunnar_A_Sigurdsson1), Chenyang Tao (/profile?id=~Chenyang_Tao1), Jiun-Yu Kao (/profile?id=~Jiun-Yu_Kao1), +7 more (/group/info?id=ICML.cc/2023/Conference/Submission781/Authors))

 20 Mar 2023, 10:35

 Program Chairs, Area Chairs, Authors, Reviewer ufcf, Reviewers Submitted, Senior Area Chairs

Comment:

We thank the reviewer for the insightful feedback. We are glad that the reviewer found our experiments thorough and our paper clearly presented.

Regarding the weaknesses pointed out by the reviewer, we would like to clarify some misunderstandings:

1. *Datasets*. Our main motivation for using unlabeled video datasets is in its scalability. We noticed that the paired text-audio data used in (Kruek et al., 2023) and (Wu et al., 2023) are composed of label-like pseudo sentences. For example, Kruek et al. (2023) used pseudo sentences in the form of “dog bark park”, whereas Wu et al. (2023) used the template “The sound of label-1, label-2, ..., and label-n”. **The only dataset that comes with natural language descriptions is the AudioCaps dataset (Kim et al., 2019), which represents only 3.3% of the whole dataset used by Kruek et al. (2023) and Wu et al. (2023).** We believe **our proposed framework offers a viable option by leveraging unlabeled videos as a proxy to learn the desired text-audio correspondence**, as pointed out by Reviewer 2Sj4. Regarding the chosen query template, as explained in Section 4.1, we used “a photo of plating {query}” as it has been shown by Radford et al. that using the prefix of “a photo of” reduces the modality gap between texts and images. The template “a sound of playing {label}” does not generally make sense for the CLIP model as the word “sound” has little meaning in the visual domain.
2. *Baselines*. As explained in Section 5.3, we intended to compare against Diffsound (Yang et al., 2023) and AudioGen (Kruek et al., 2023). For Diffsound, **the code provided in their GitHub repository was incomplete and we could not reproduce their results by the time when we submitted the paper.** Moreover, the code for AudioGen has not been released till now.

3. *Experimental results.* As explained in the caption, the bold values indicate the best performing model per category. We will remove the bolding in Tables 2 and 4 in the revised manuscript to avoid confusions, as also suggested by Reviewer TfCZ.
4. *Subjective evaluation.* From the feedback of a pilot study, the evaluators reported the difficulty in rating the audio samples with a consistent standard. Hence, we decided to adopt A/B tests to acquire more robust test results. We aggregated the evaluation results by assigning scores to the winning model, and unfortunately error bars can not be computed this way. The results reported in Table 4 are aggregated over 300 pairs of A/B tests, where the detailed pairwise A/B test results are available in Figure 14 in the Appendix.
5. *Novelty.* **We argue that the main novelty of this work is the capability of learning text-queried sound synthesis from unlabeled videos in the wild.** While Dong et al. (2023) has explored learning text-queried sound separation from unlabeled videos, we were the first to extend this concept to sound synthesis, and we have shown the possibilities of learning sound synthesis from unlabeled videos. The proposed framework can be scaled up and trained on the tremendous amount of videos available on the internet. Further, it demonstrates a more human-like machine learning framework for audio synthesis. **We believe that ICML is not only about finding new network architectures, but also novel frameworks and paradigms that might lead to new insights into future research.**

Answers to the questions:

1. We thank the reviewer for the suggestion. We will remove the word “self-supervised” and change our title into “CLIPSynth: Learning Text-queried Sound Synthesis from Unlabeled Videos” in the revised manuscript.
2. No. As explained in Section 4.1, the text query is constructed from the label and a query template (e.g., “a photo of {query}”) throughout our experiments. We would like to point out that enabling natural language input for text-queried sound synthesis is an ongoing challenge that some concurrent work is still exploring, e.g., AudioLDM (Liu et al., 2023) and Make-An-Audio (Huang et al., 2023). Moreover, we argue that our proposed method using unlabeled videos offers a possible workaround to approach this problem without requiring a large amount of paired audio-natural-language data. The current implementation does not exhibit such capability as the CLIP text encoder uses a bag-of-word model which does not handle natural language well. However, the proposed framework is orthogonal to the image-language model, and we would like to explore using a more powerful image-language model that understands natural language in the future.



➔ *Replying to Response to Reviewer ufcf*

References

Official Comment

✍ Authors (👤 Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Gunnar A Sigurdsson (/profile?id=~Gunnar_A_Sigurdsson1), Chenyang Tao (/profile?id=~Chenyang_Tao1), Jiun-Yu Kao (/profile?id=~Jiun-Yu_Kao1), +7 more (/group/info?id=ICML.cc/2023/Conference/Submission781/Authors))

📅 20 Mar 2023, 10:38 👤 Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Comment:

- Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually Guided Audio Generation," *Proc. ICLR*, 2023.
- 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," *Proc. ICASSP*, 2023.
- Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "AudioCaps: Generating Captions for Audios in The Wild," *Proc. NAACL*, 2019.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," *Proc. ICML*, 2022.
- Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu, "DiffSound: Discrete Diffusion Model for Text-to-sound Generation," *arXiv preprint arXiv:2207.09983*, 2023.
- Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick, "CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos," *Proc. ICLR*, 2023.

- Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley, "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," *arXiv preprint arXiv:2301.12503*, 2023.
- Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, Zhou Zhao, "Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models," *arXiv preprint arXiv:2301.12661*, 2023.

Official Review of Submission781 by Reviewer 2Sj4

Official Review ✎ Reviewer 2Sj4 📅 05 Mar 2023, 01:11 (modified: 14 Mar 2023, 06:45)

👁️ Program Chairs, Area Chairs, Authors, Reviewer 2Sj4, Reviewers Submitted, Senior Area Chairs

Summary:

This work presents CLIPSynth, a text-conditional audio synthesis (text-to-sound) which is trained by self-supervised learning (SSL) on unlabeled video. A key contribution of this work is enabling a working text-to-sound system without the paired dataset of text and audio which is scarce (thereby blocking the progress of text-conditional audio generation compared to other domains). This is done by side-stepping the problem through CLIP [Radford+2021] that aligns the image and text representation as shared conditional information and mel spectrograms as the target of a conditional diffusion model, enabling cross-modal querying for text-to-sound.

Strengths And Weaknesses:

Strengths

Although the use of a contrastive pretraining model (i.e., CLIP) for cross-modal querying is now well recognized from the literature and may not seem surprising, CLIPSynth delivers a timely incorporation of the considered approach in the audio generation domain. In my opinion, the fact that the CLIP-based methods are shown to be effective in other domains (image, for example) does not necessarily undermine the findings in this work. Especially given the challenge of the paired data scarcity problem in audio domain, presenting an effective way to leverage unlabeled video as a proxy constitutes a valuable contribution in this field, and future work is highly likely to consider this direction as a viable option to overcome the issue.

Text-to-sound has recently seen a surge of interest along with numerous pioneering works. Although such works are not included in this work due to incompleteness or inaccessibility, a comparative evaluation using the own variants under the unified experimental setup is justified, which is timely and welcomed addition.

Weaknesses

The diffusion model used by CLIPSynth applies 1,000 reverse steps, which can be considerably slower than the other methods. Because the inference latency is of importance in a practical environment, reporting runtime for inference (real-time-factor, for example), model size, and other attributes of CLIPSynth seems appropriate to provide the readers a complete picture on the comparative analysis. Furthermore, the authors can apply advanced samplers that reduce the reverse steps for diffusion models and measure a level of degradation in quality. If the fast samplers for diffusion model can provide satisfactory quality, it can also alleviate the concern on the inference overhead.

In my opinion, the sound quality and text adherence of CLIPSynth manifests a promising proof-of-concept of the purely self-supervised method, but on a flip side, the text adherence seems generally better for the models that leverages text, as evidenced by the "Relevance" score on VGG-Sound and recognized by the authors. I can also sense that CLIPSynth occasionally generates audio that does not adhere to the text description compared to the baselines. Perhaps the authors can consider fine-tuning the model on a few paired data samples, as diffusion models have shown to excel at few-shot adaptation through fine-tuning in various domains.

Questions:

The authors used the pretrained CLIP without any fine-tuning, which bears no issue for the presented approach. I am wondering if the authors have considered fine-tuning the model, perhaps by further integrating the spectrogram as an auxiliary modality, for example. To extend the query-based sound synthesis beyond an image or text prompt, one will require an appropriate encoder that projects the input to the shared latent space. Do the authors have opinion and/or implications on improving the cross-modal projection model?

Recently, there have been concurrent works on text-conditional audio and music synthesis, including AudioLDM[Liu+2023], Make-An-Audio[Huang+2023], and MusicLM[Agostinelli+2023], for example, where several works contain overlapping contributions. While these works are outside of the review window and the comparative experiments are not necessary, in future revisions, it will be helpful to the target audience to present additional discussion and comparison between the concurrent works, followed by emphasizing the unique aspects of this work.

Perhaps "humans do not learn the sounds of an object this way.", "a more human-like approach for text-queried sound synthesis." might seem a stretch; can the authors devise more technical and structured ways to emphasize the necessity of introducing the contrastive cross-modal model?

In my opinion, Figure 6 seems to carry little information, as the reverse diffusion process is now well recognized in the research community.

Limitations:

The manuscript handled limitations of this work adequately: a reduced text adherence for certain scenarios compared to the baselines with direct exposure to text during training, inability to handle off-screen sounds and audio-specific queries.

Ethics Flag: No

Soundness: 3 good

Presentation: 2 fair

Contribution: 3 good

Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes



Response to Reviewer 2Sj4

Official Comment

Authors (Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Gunnar A Sigurdsson (/profile?id=~Gunnar_A_Sigurdsson1), Chenyang Tao (/profile?id=~Chenyang_Tao1), Jiun-Yu Kao (/profile?id=~Jiun-Yu_Kao1), +7 more (/group/info?id=ICML.cc/2023/Conference/Submission781/Authors))

20 Mar 2023, 11:34

Program Chairs, Area Chairs, Authors, Reviewer 2Sj4, Reviewers Submitted, Senior Area Chairs

Comment:

We thank the reviewer for the insightful feedback. We are glad that the reviewer found our proposed framework a valuable contribution in the field.

Regarding the weaknesses pointed out by the reviewer, we would like to respond the insightful comments and suggestions as follows:

1. Regarding the inference speed, it takes 30 seconds for the diffusion model to synthesize a sample for 1000 reverse steps. However, if we do it in parallel with a batch size of 64, it takes only 2 minutes to synthesize all 64 samples. For the Hifi-GAN vocoder, the inference speed is less than 1 second, which is fairly fast compared to the diffusion model. We will include this information in the revised manuscript. Regarding the model size, the diffusion model has 30.4M parameters and it requires 681 MB of GPU memory. We thank the reviewer for the helpful suggestion, and we will include this information in the revised manuscript.
2. Regarding the relevance between the input text query and the generated sound, we would like to emphasize the challenges of the proposed self-supervised learning framework, e.g., off-screen sounds and purely audio-relevant queries as discussed in Section 7. We believe our work would make a positive contribution to the field in **offering an alternative perspective of approaching text-to-audio synthesis using unlabeled videos.**

Answers to the questions:

1. In this paper, we froze the pretrained CLIP model and performed zero-shot modality transfer from image inputs (training) to text inputs (inference). The trained CLIPSynth can take either text or image inputs at test time, but we did not experiment with other modalities on the input side. However, we noticed that in the

concurrent Make-An-Audio paper (Huang et al., 2023), the authors explored X-to-audio generation and demonstrated promising results in personalized text-to-audio generation, audio inpainting and visual-to-audio generation. We note that their proposed approach is orthogonal to our proposed framework, and one can combine the two approaches to enable inputs in other modalities.

- Regarding the concurrent work, please see below for a brief comparison. We would like to emphasize that **our proposed method is the only model that requires no paired text-audio data, which opens a unique direction that can scale up to the tremendous amount of videos available on the internet.** Moreover, unlike MusicLM and Noise2Music, our experiments were performed on two reproducible, open datasets. We will also release the source code to facilitate future research along this direction. We thank the reviewer for the helpful suggestion, and we will add a paragraph in the related work section to discuss the difference between our work and concurrent work.

Model	No paired text-audio data needed	Reproducible, open dataset	Condition encoder	Synthesis model	Spectrogram-based
CLIPSynth (ours)	✓	✓	CLIP	Diffusion	✓
AudioLDM (Liu et al., 2023)	(used for training CLAP)	✓	CLAP	Latent diffusion (on a VAE latent space)	✓
Make-An-Audio (Huang et al., 2023)	(used for training CLAP)	✓	CLAP	Latent diffusion (on a VAE-GAN latent space)	✓
MusicLM (Agostinelli et al., 2023)	(used for training MuLan)		MuLan	w2v-BERT + SoundStream	
Noise2Music (Huang et al., 2023)	(used for training both MuLan and Noise2Music)		T5	Cascaded diffusion	

- We thank the reviewer for pointing this out. By a human-like machine approach, we were referring to the fact that humans do not rely on some “dictionary” for sounds of different objects. Instead, humans learn many sounds from observing the world unsupervisedly (which corresponds to watching YouTube videos in this paper). For example, we may learn that “*there’s an instrument resembling a big violin that can play much lower pitches than a violin*” in one scenario, and later learn that “*such a big-violin-like instrument is called a cello*” elsewhere. Hence, we are interested in building a machine learning framework that can mimic such behavior and learn sound synthesis from the naturally-occurring videos in the wild. We will add more context to the introduction section.

- We will move Table 6 to the Appendix in the revised manuscript, and we will use the saved space to provide more information on the model (speed, model size, etc.) and the comparison against concurrent work.



➔ Replying to Response to Reviewer 2Sj4

References

Official Comment

✍ Authors (👤 Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Gunnar A Sigurdsson (/profile?id=~Gunnar_A_Sigurdsson1), Chenyang Tao (/profile?id=~Chenyang_Tao1), Jiun-Yu Kao (/profile?id=~Jiun-Yu_Kao1), +7 more (/group/info?id=ICML.cc/2023/Conference/Submission781/Authors))

📅 20 Mar 2023, 11:35 🗳 Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Comment:

- Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley, "AudioLDM: Text-to-Audio Generation with Latent Diffusion Models," *arXiv preprint arXiv:2301.12503*, 2023.
- Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, Zhou Zhao, "Make-An-Audio: Text-To-Audio Generation with Prompt-Enhanced Diffusion Models," *arXiv preprint arXiv:2301.12661*, 2023.

- 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 preprint arXiv:2301.11325*, 2023.
- Qingqing Huang, Daniel S. Park, Tao Wang, Timo I. Denk, Andy Ly, Nanxin Chen, Zhengdong Zhang, Zhishuai Zhang, Jiahui Yu, Christian Frank, Jesse Engel, Quoc V. Le, William Chan, Zhifeng Chen, and Wei Han, "Noise2Music: Text-conditioned Music Generation with Diffusion Models," *arXiv preprint arXiv:2302.03917*, 2023.

Official Review of Submission781 by Reviewer TfCZ

Official Review  Reviewer TfCZ  03 Mar 2023, 03:32 (modified: 14 Mar 2023, 06:45)

 Program Chairs, Area Chairs, Authors, Reviewer TfCZ, Reviewers Submitted, Senior Area Chairs

Summary:

In this paper the authors proposed a weakly supervised method for text to audio generation. They use the image modality as a pivot between text and audio, given that (a) CLIP models are available to match text and image (b) video naturally provide pairs of images and audio. Specifically, the author train a diffusion model trained to model the spectrogram of the audio of a video segment, conditioned on the CLIP embedding computed from a frame from the same segment. At test time, they replace the conditioning on the frame by a conditioning of a text using the text encoder of the CLIP model. Finally, they generate a waveform using HiFi GAN.

The authors provide objective metrics (FAD and FID) and subjective metrics (non standard protocol and metric) on two datasets: MUSIC and VGG-Sound. The results show that the proposed method is competitive on the MUSIC dataset. Results on VGG-Sounds are much less promising

Strengths And Weaknesses:

tl;dr: Idea is simple and elegant, but the results on VGG-Sound are far from convincing, with most of the samples provided by the authors having little relevance to the provided text prompt. Subjective evaluations are non standard, and besides without error bars so that it is not easy to derive conclusions from it.

Strength

- idea is simple yet efficient for leveraging weakly supervised video data for text to audio generation.
- reuses a lot of existing components and models (e.g. HiFi GAN, pre-trained CLIP etc.)
- authors provide a number of experimental validations, including subjective studies, comparison with fully supervised models.
- In particular, insight on the modality gap between the image embedding used at train time and the text encoder used at test time is interesting.
- Despite the absence of external baselines, the authors did a good job at providing ablation and comparing with fully supervised variants.

Weaknesses

- potentially limited impact: most videos do have description with them (e.g. on Youtube), and as noted by Kreuk et al. (2022), gathering 4000 hours of audio paired with text is not really a problem.
- Some well studied datasets like Audioset were left out of the study.
- The method proposed really shines on the MUSIC dataset, which seems limited in terms of label complexity (e.g. more categorical than true complex texty labels). On VGG-Sound, it seems it underperforms compared with just using the text conditioning provided with the dataset. Samples on VGG-Sound are completely unconvincing, often bearing no relation to the text prompt provided.
- Some objective metrics like the KL divergence of a pretrained classifier between audio regenerated from a prompt and the ground truth audio is not provided (see [Kreuk et al. 2022]). This would capture how well the semantic content of the audio is reconstructed, which as mentioned just above would likely be quite bad on VGG-Sound.
- The subjective tests is non standard. The authors use A/B testing but eventually just derive a per-model score rather than doing model comparisons. Why not then use more standard MUSHRA or MOS protocols as done in a number of other audio synthesis works ?
- Error bars on the subjective evaluation are missing.

A last remark: the bold per category in the table is quite confusing, given that a number of categories have a single model, e.g. Table 4. This makes the table harder to read. For instance, from far way it makes it look like the proposed method as a good relevance on VGG-Sound, when in fact it is quite bad, which is confirmed when listening to the samples.

Questions:

In Section 5.1, you speak a preprocessing, but do not mention what is consist in.

Provide the error bars for Table 4. Why did the authors not use a standard protocol ?

Limitations:

nothing to remark.

Ethics Flag: No

Soundness: 2 fair

Presentation: 3 good

Contribution: 2 fair

Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.



Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes



Response to Reviewer TfCZ

Official Comment

 Authors ( Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Gunnar A Sigurdsson (/profile?id=~Gunnar_A_Sigurdsson1), Chenyang Tao (/profile?id=~Chenyang_Tao1), Jiun-Yu Kao (/profile?id=~Jiun-Yu_Kao1), +7 more (/group/info?id=ICML.cc/2023/Conference/Submission781/Authors))

 20 Mar 2023, 10:46  Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Comment:

We thank the reviewer for the insightful feedback. We are glad that the reviewer find the idea simple and elegant.

Regarding the weaknesses pointed out by the reviewer, we would like to respond the insightful comments and suggestions as follows:

1. We argue that the clip-level descriptions provided on YouTube are usable in the context of text-queried audio generation. We doubt the possibility of using such clip-level descriptions in learning text-to-sound synthesis due to the little correspondence. For example, most videos in the VGG-Sound datasets are longer than 1 minute, **having a clip-level description of a 1-min long video is not necessarily helpful in learning to synthesize a specific sound that occurs only for 1-5 seconds**. We are aware of the extensive prior work in weakly supervised learning for sound event detection, but there is no existing weakly supervised model for text-to-audio generation in our knowledge. In contrast, our proposed method leverages the naturally-occurring tightly-aligned correspondence between frames and sounds in videos. Moreover, we would like to emphasize that our main motivation for using unlabeled video datasets is in its scalability. We noticed that the paired text-audio data used in (Kruek et al., 2023) and (Wu et al., 2023) are composed of label-like pseudo sentences. For example, Kruek et al. (2023) used pseudo sentences in the form of “dog bark park”, whereas Wu et al. (2023) used the template “The sound of label-1, label-2, ..., and label-n”. **The only dataset that comes with natural language descriptions is the AudioCaps dataset (Kim et al., 2019), which represents only 3.3% of the whole dataset used by Kruek et al. (2023) and Wu et al. (2023)**. We believe **our proposed framework offers a viable option by leveraging unlabeled videos as a proxy to learn the desired text-audio correspondence**, as pointed out by Reviewer 2Sj4.
2. We chose VGG-Sound rather than AudioSet because VGG-Sound is further filtered by audio-visual correspondence, whereas AudioSet contains much more off-screen sounds and background noise. Since our proposed framework requires good enough audio-visual correspondence in the training data, as discussed in Section 7. We will include the reasoning of this decision in our revised manuscript.

3. Regarding the relevance between the input text query and the generated sound, we would like to emphasize the challenges of the proposed self-supervised learning framework, e.g., off-screen sounds and purely audio-relevant queries as discussed in Section 7. We believe our work would make a positive contribution to the field in **offering an alternative perspective of approaching text-to-audio synthesis using unlabeled videos**.
4. We thank the reviewer for this valuable suggestion. Due to the short period of author rebuttal period, we have not had time to compute these evaluation metrics. We will include the KL (Yang et al., 2023) and CLAP score (Hessel et al., 2021; Wu et al., 2023) results in the revised manuscript. Instead of the KL divergence measure used in DiffSound (Yang et al., 2023) and AudioGen (Kruek et al., 2023), we decided to .
5. We thank the reviewer for this insightful feedback. We were not aware of the MUSHRA method commonly used in audio codec evaluation. As for the MOS-based evaluation, from the feedback of a pilot study, the evaluators reported the difficulty in rating the audio samples with a consistent standard. Hence, we decided to adopt A/B tests to acquire more robust test results. We aggregated the evaluation results by assigning scores to the winning model, and unfortunately error bars can not be computed this way. The results reported in Table 4 are aggregated over 300 pairs of A/B tests, where the detailed pairwise A/B test results are available in Figure 14 in the Appendix.
6. Thank you for the valuable suggestion. We will remove the bolding in Tables 2 and 4 in the revised manuscript to avoid confusions.

Answers to the questions:

1. We thank the reviewer for pointing this out. The preprocessing refers to the data downloading and audio/frames extraction from the downloaded videos. We note some videos are no longer available on YouTube and some downloaded videos are corrupted (e.g., missing an audio track). We will provide this information in the revised manuscript.
2. From the feedback of a pilot study, the evaluators reported the difficulty in rating the audio samples with a consistent standard. Hence, we decided to adopt A/B tests to acquire more robust test results. We aggregated the evaluation results by assigning scores to the winning model, and unfortunately error bars can not be computed this way. We note that the results reported in Table 4 are aggregated over 300 pairs of A/B tests, where the detailed pairwise A/B test results are available in Figure 14 in the Appendix.



➔ *Replying to Response to Reviewer TfCZ*

References

Official Comment

✍ Authors (👁 Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Gunnar A Sigurdsson (/profile?id=~Gunnar_A_Sigurdsson1), Chenyang Tao (/profile?id=~Chenyang_Tao1), Jiun-Yu Kao (/profile?id=~Jiun-Yu_Kao1), +7 more (/group/info?id=ICML.cc/2023/Conference/Submission781/Authors))


📅 20 Mar 2023, 10:47 👁 Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Comment:

- Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, "AudioGen: Textually Guided Audio Generation," *Proc. ICLR*, 2023.
- 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," *Proc. ICASSP*, 2023.
- Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "AudioCaps: Generating Captions for Audios in The Wild," *Proc. NAACL*, 2019.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, Yejin Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," *Proc. EMNLP*, 2021.



Official Review of Submission781 by Reviewer zQM5

Official Review  Reviewer zQM5  02 Mar 2023, 05:03 (modified: 14 Mar 2023, 06:45)

 Program Chairs, Area Chairs, Authors, Reviewer zQM5, Reviewers Submitted, Senior Area Chairs

Summary:

This paper proposes a novel self-supervised text-queried sound synthesis model trained with unlabeled videos in the wild. The proposed model utilizes the largely trained language-image pretrained model (CLIP) without any finetuning in order to extract the similar visual embeddings from the textual embedding during inference time. The authors also adopt the diffusion model in order to generate the mel-spectrogram.

Strengths And Weaknesses:

Strength

1. The paper has the adequate contributions and is easy to follow.
2. The authors propose the self-supervised text-queried sound synthesis idea. It is meaningful in regarding the zero-shot modality transfer using the largely trained language-image pretrained model.
3. The authors show the extensive experiment results with manifold audio samples.

Weakness

1. The overall proposed model seems to be the combination of existing methodologies (CLIP and diffusion model), so the proposed model does not seem to be novel in a technical viewpoint.
2. The performances of the proposed model do not seem to be decent compared to different comparing methodologies. I understand that it is due to the modality gap that image and text inherently contain. However, it would be great to explain more discussion regarding the numerical performance in the result section.

Questions:

1. The quantitative results are mostly ablation studies. It is recommended to additionally conduct the experiment comparing with different previous studies. I believe that the code of Diffsound is now available.
2. Is there any reason why the authors have chosen the diffusion model as a generative model? Is the diffusion model superior compared to other generative models such as GAN? Please explain the benefits of diffusion model in the text-to-sound generation technique. Also I am wondering how the performance would change when the model does not adopt the diffusion model.
3. Is it possible to report different metric besides FAD and FID, such as KL and other audio quality assessment metrics?

Limitations:

The authors have adequately addressed the limitations, but the potential negative societal impact is not addressed.

Ethics Flag: No

Soundness: 2 fair

Presentation: 3 good

Contribution: 3 good

Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.



Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes



Response to Reviewer zQM5

Official Comment

 Authors ( Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Gunnar A Sigurdsson (/profile?id=~Gunnar_A_Sigurdsson1), Chenyang Tao (/profile?id=~Chenyang_Tao1), Jiun-Yu Kao (/profile?id=~Jiun-Yu_Kao1), +7 more (/group/info?id=ICML.cc/2023/Conference/Submission781/Authors))

 20 Mar 2023, 10:52  Program Chairs, Area Chairs, Authors, Reviewers Submitted, Senior Area Chairs

Comment:

We thank the reviewer for the insightful feedback. We are glad that the reviewer found our proposed framework a valuable contribution in the field.

Regarding the weaknesses pointed out by the reviewer, we would like to respond the insightful comments and suggestions as follows:

1. **We argue that the main novelty of this work is the capability of learning text-queried sound synthesis from unlabeled videos in the wild.** While Dong et al. (2023) has explored learning text-queried sound separation from unlabeled videos, we were the first to extend this concept to sound synthesis, and we have shown the possibilities of learning sound synthesis from unlabeled videos. The proposed framework can be scaled up and trained on the tremendous amount of videos available on the internet. Further, it demonstrates a more human-like machine learning framework for audio synthesis. **We believe that ICML is not only about finding new network architectures, but also novel frameworks and paradigms that might lead to new insights into future research.**
2. We would like to emphasize that **our proposed method is the first text-to-sound synthesis model that requires no paired text-audio data.** Comparing our proposed model with other models in the same ground is unfair, but we have no choice but to set up these "upper bound" methods in our experiments. Still, we would like to highlight that CLIPSynth, despite using no labeled data, is not overwhelmingly underperforming against CLIPSynth-Text and CLIPSynth-Hybrid. **We believe our work offers an alternative perspective of approaching text-to-audio synthesis using unlabeled videos.**

Answers to the questions:

1. As explained in Section 5.3, we intended to compare against Diffsound (Yang et al., 2023) and AudioGen (Kruek et al., 2023). For Diffsound, **the code provided in their GitHub repository was incomplete and we could not reproduce their results by the time when we submitted the paper.** Moreover, **the code for AudioGen has not been released till now.**
2. We chose the diffusion model as the backbone model because Diffsound (Yang et al., 2023) has demonstrated promising results in using diffusion models for text-to-audio generation. While diffusion models suffer from slow inference speed due to its iterative nature, it has been shown to be easier and more stable to train. Also, note that we can control the quality-speed trade-off by adjusting the number of reverse sampling steps of a diffusion model. Moreover, we note that the proposed framework can be applied to various backbone generative models including GANs. While we did not examine using GANs in this work, we would love to see follow-up work exploring the use of different backbone models.
3. We thank the reviewer for this valuable suggestion. Due to the short period of author rebuttal period, we have not had time to compute these evaluation metrics. We will include the KL (Yang et al., 2023) and CLAP score (Hessel et al., 2021; Wu et al., 2023) results in the revised manuscript.

References:

- Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu, "Diffsound: Discrete Diffusion Model for Text-to-sound Generation," *arXiv preprint arXiv:2207.09983*, 2023.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, Yejin Choi, "CLIPScore: A Reference-free Evaluation Metric for Image Captioning," *Proc. EMNLP*, 2021.
- 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," *Proc. ICASSP*, 2023.
- Hao-Wen Dong, Naoya Takahashi, Yuki Mitsufuji, Julian McAuley, and Taylor Berg-Kirkpatrick, "CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos," *Proc. ICLR*, 2023.

About OpenReview (/about)
Hosting a Venue (/group?
id=OpenReview.net/Support)
All Venues (/venues)
Sponsors (/sponsors)

Frequently Asked Questions
(<https://docs.openreview.net/getting-started/frequently-asked-questions>)
Contact (/contact)
Feedback

[Terms of Service \(/legal/terms\)](#)

[Privacy Policy \(/legal/privacy\)](#)

[OpenReview \(/about\)](#) is a long-term project to advance science through improved peer review, with legal nonprofit status through [Code for Science & Society \(https://codeforscience.org/\)](https://codeforscience.org/). We gratefully acknowledge the support of the [OpenReview Sponsors \(/sponsors\)](#).