CLIPSep: Learning Text-queried Sound Separation with Noisy Unlabeled Videos

(/pdf?id=H-T3F0dMbyj)

Hao-Wen Dong (/profile?id=~Hao-Wen_Dong1), Naoya Takahashi (/profile?id=~Naoya_Takahashi1), Yuki Mitsufuji (/profile?id=~Yuki_Mitsufuji1), Julian McAuley (/profile?id=~Julian_McAuley1), Taylor Berg-Kirkpatrick (/profile?id=~Taylor_Berg-Kirkpatrick1)

Published: 01 Feb 2023, Last Modified: 16 Feb 2023 ICLR 2023 poster Readers:
Everyone Show Bibtex Show Revisions (/revisions?id=H-T3F0dMbyj)

Keywords: universal sound separation, source separation, contrastive language-image pre-training, multi-modal learning, self-supervised learning

TL;DR: A new method the leverages the pretrained CLIP model and noise invariant training for learning text-queried sound separation with only noisy unlabeled videos

Abstract: Recent years have seen progress beyond domain-specific sound separation for speech or music towards universal sound separation for arbitrary sounds. Prior work on universal sound separation has investigated separating a target sound out of an audio mixture given a text query. Such text-queried sound separation systems provide a natural and scalable interface for specifying arbitrary target sounds. However, supervised text-queried sound separation systems require costly labeled audio-text pairs for training. Moreover, the audio provided in existing datasets is often recorded in a controlled environment, causing a considerable generalization gap to noisy audio in the wild. In this work, we aim to approach text-queried universal sound separation by using only unlabeled data. We propose to leverage the visual modality as a bridge to learn the desired audio-textual correspondence. The proposed CLIPSep model first encodes the input query into a query vector using the contrastive language-image pretraining (CLIP) model, and the query vector is then used to condition an audio separation model to separate out the target sound. While the model is trained on image-audio pairs extracted from unlabeled videos, at test time we can instead query the model with text inputs in a zero-shot setting, thanks to the joint language-image embedding learned by the CLIP model. Further, videos in the wild often contain off-screen sounds and background noise that may hinder the model from learning the desired audio-textual correspondence. To address this problem, we further propose an approach called noise invariant training for training a query-based sound separation model on noisy data. Experimental results show that the proposed models successfully learn text-queried universal sound separation using only noisy unlabeled videos, even achieving competitive performance against a supervised model in some settings.

Anonymous Url: I certify that there is no URL (e.g., github page) that could be used to find authors' identity.

No Acknowledgement Section: I certify that there is no acknowledgement section in this submission for double blind review.

Code Of Ethics: I acknowledge that I and all co-authors of this work have read and commit to adhering to the ICLR Code of Ethics

Submission Guidelines: Yes

Please Choose The Closest Area That Your Submission Falls Into: Applications (eg, speech processing, computer vision, NLP)

Reply Type: Paper3561 Decision and Paper3561 Official Review **21 Replies**

Author: everybody Visible To: all readers Hidden From: nobody

[−] Paper Decision

ICLR 2023 Conference Program Chairs

20 Jan 2023 ICLR 2023 Conference Paper3561 Decision Readers: Everyone

Decision: Accept: poster

Metareview: Summary, Strengths And Weaknesses:
The authors present an algorithm for universal sound separation, to extract target audio from a mixture given text or image queries as input. Two systems are described, CLIPSep, which can be used to extract target sound using text queries, and CLIPSep-NIT, which additionally enables training on noisy (offscreen noise) audio. The reviewers agree that the work is novel, and enables training a text-driven separation model on noisy videos. Presented results are also thorough and convincing, along with links to demos.

Justification For Why Not Higher Score:
The paper builds on prior works for image-language and audio separation models, enabling a novel use case (text queried separation) and the ability to train on large scale data without the need for careful supervision. Although multiple reviewers highlighted novelty, there were also some concerns about certain parts of the model being simple extensions (like noise-invariant training). There are also prior works on text-queried audio separation.

Justification For Why Not Lower Score:
Novelty and evaluations were highlighted by all reviewers.

Note From PC:
if the above contains the word "oral" or "spotlight" please see: "oral" presentation means -> notable-top-5% and "spotlight" means -> notable-top-25%. As stated in our emails, we are disassociating presentation type from AC recommendations

Add **Public Comment**

[−] Official Review of Paper3561 by Reviewer LZWB

ICLR 2023 Conference Paper3561 Reviewer LZWB

26 Oct 2022 (modified: 26 Oct 2022) ICLR 2023 Conference Paper3561 Official Review Readers: Everyone

Summary Of The Paper:

Summary

This paper proposes a text-queried universal sound separation model that can be trained on noisy in-the-wild videos (i.e. videos that contain both on-screen and off-screen sounds). Two versions are proposed: CLIPSep and CLIPSep-NIT (CLIPSep with noise invariant training).

CLIPSep: during training, mix audio from two videos. Extract the CLIP embedding of an image frame; from the spectrogram of the audio mixture, predict k masks; predict a k-dim query vector q_i from the CLIP embedding; predict overall mask for source i using query vector q_i to combine across the k masks, with an additional k-dimensional scaling weight w_i and scalar bias b_i; audio is reconstructed using inverse STFT on masked STFT. Training loss is weighted binary cross-entropy between estimated mask and ground-truth mask (so training requires isolated source audio from on-screen-only video). During inference, CLIP embedding is computed from text (assuming this will be close to CLIP embedding of image), and just one mask is predicted for the source described by the text.

CLIPSep-NIT: same as CLIPSep, except that for each of the n sources during training, an additional "noise" mask is predicted, which is an additional query vector that combines the k predicted masks with a noise query vector. Then during training, all permutations of the noise masks added to the source masks are considered, and the permutation

with the minimum error is used. It seems the purpose of the noise masks is to "soak up" sounds not related to the CLIP embedding. At test time, the noise masks are discarded.

Contributions

1. First text-driven separation model (to my knowledge) that can be trained on noisy videos, enabled by the NIT trick.
2. NIT is a contribution, though I feel its novelty is relatively minor, since it's just a constrained version of permutation invariant training (PIT).

Strength And Weaknesses:

Strengths

1. To my knowledge, this is the first method to train text-queried separation on noisy mixtures.
2. The evaluation is done on both MUSIC+ and VGGSound-Clean+, measuring performance on both music separation and universal separation, and these results are convincing.
3. Paper includes link to anonymized demo page, which is convincing.

Weaknesses

1. I think the paper makes the post-selection step required for a MixIT model to be harder than it actually is. For a MixIT-trained model with N outputs, it's pretty easy to pick a source, e.g. with a sound classification network. This setup was actually proposed with a classification-regularized loss in: Wisdom, Scott, Aren Jansen, Ron J. Weiss, Hakan Erdogan, and John R. Hershey. "Sparse, efficient, and semantic mixture invariant training: Taming in-the-wild unsupervised sound separation." In 2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), pp. 51-55. IEEE, 2021. (https://arxiv.org/pdf/2106.00847.pdf (https://arxiv.org/pdf/2106.00847.pdf)) Another advantage of MixIT is that the outputs are more interpretable, compared to models that rely on conditioning, such as the one described in this paper. Thus, I think it may be good to discuss the pros and cons of separate-then-select versus conditional separation in the paper.
2. This statement is a bit incorrect:

   "However, AudioScope still requires a post-selection process if there is more than one predicted on-screen channel."

   The goal of AudioScope is to recover all on-screen sounds in a single channel, which is what the model does: it uses on-screen probabilities as mixing weights across the sources.
3. "where s1, . . . , sn are the n audio sources,": in practice, these are mixtures, right? The model is just assuming that they are single sources. it might be good to refine the terminology here a bit.
4. Some explanation of why k masks are predicted, then combined, would be good. I think this is kind of analogous to the multiple output sources in MixIT, which can be combined for a particular user interface or output goal, e.g. AudioScope combines with on-screen probabilities to get an estimate of on-screen sound.
5. The equation for computing the overall source mask from the k masks is confusing. What does the \odot versus the \cdot mean? If w_i is k-dimensional, I don't see a sum over k, since it's \odot'ed with scalar q_{ij} times \tilde{M}*j. Should this actually be w*{i,j}? Please specify how this is done.
6. The model uses mask-based losses, which, in my own experience, are often suboptimal compared to signal based losses (i.e. computed loss in time domain, backpropping through iSTFT applied to masked mixture STFT). Also, in the NIT loss, adding masks together and applying a ceil of 1 does not exactly correspond to adding signals in the time domain, because of STFT consistency. it would be interesting to try time-domain based losses for this network, and see if that provides any improvement. Also, the architecture in the MixIT paper used mixture consistency, so that output sources sum up to the original input mixture. This might also be a useful constraint on the architecture here.
7. I think best practice for reporting units in decibels is to use only one decimal place. Humans can often not even hear 0.1 dB of difference. Thanks, by the way, for reporting std dev from the mean and median.
8. More explanation of the motivation of NIT would be very welcome. My intuition is that it helps "soak up" extra noise by providing additional output sources, but this might not be right. Please add some explicit discussion of the motivation.

Typos and minor comments

a. "For eaxmple," -> "For example,"

Clarity, Quality, Novelty And Reproducibility:
Clarity: the paper is very clear. I only have minor suggestions for improvement (see weaknesses)

Quality: high quality. Evaluation is solid and compares to relevant baselines. Some nice additional information is provided in the appendices.

Novelty: paper is novel, in that it proposes a text-driven separation method that can be trained on noisy data, and minor novelty in the noise invariant training.

Reproducibility: the code and models are made available.

Summary Of The Review:
Overall, a nice paper that accomplishes training text-driven separation on noisy in-the-wild data. Achieves good performance compared to prior approaches, and qualitative demos are convincing.

Correctness: 4: All of the claims and statements are well-supported and correct.
Technical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.
Empirical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.
Flag For Ethics Review: NO.
Recommendation: 6: marginally above the acceptance threshold
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.

Add **Public Comment**

[+] **Official Comment by Paper3561 Authors • Response to reviwer LZWB part 1/2**

[+] **Official Comment by Paper3561 Authors • Response to reviwer LZWB part ...**

[−] # Official Review of Paper3561 by Reviewer vU71

ICLR 2023 Conference Paper3561 Reviewer vU71

24 Oct 2022 (modified: 03 Dec 2022) ICLR 2023 Conference Paper3561 Official Review Readers: Everyone

Summary Of The Paper:
The paper under review proposes a method of selecting a single sound source from a mixture of sounds in a video via a text description of the visual component of the video. The system can be trained on unlabeled data, aka unsupervised training. This is a novel configuration of using a pre-trained audio-visual correspondence model to allow text queries to select the single audio source to separate from a mixture in the video. Unlike what is claimed in the paper though in section 4.1, work was published this year on querying by text to separate a source from an audio mixture (this is understandable given timing). There is also a contribution of a form of noise invariant training that allows for the model to account for sounds in the mixture that have no correspondence in the video. The results are conducted on test sets, MUSIC and VGGSound-Clean, that have audio collected from the wild (YouTube), however they have been artificially mixed to yield multiple sound sources. The results are competitive with PIT, although PIT has a "post-processing" requirement.

Strength And Weaknesses:
Strengths:

- A new configuration of querying by text to separate out an audio source in a video with sources that have corresponding audio and visual signals.
- Shows performance competitive with state-of-the-art in sound separation

Weaknesses

- Tests are made by artificially combining samples of YouTube videos. Can you conduct test results on naturally occurring mixtures?
- Results report an automatically computed quantitative metric, ie SDR. It is unclear whether how this corresponds to actual user preferences. Since the results are close, could a qualitative survey be conducting comparing the results of PIT with CLIPSep, similar to how they were done in Sound of Pixels using Mechanical Turk?

Clarity, Quality, Novelty And Reproducibility:
The clarity and quality is good and generally well written. It lacks a certain level of final polish to make 1. how it differs previous, comparable work and 2. the findings absolutely clear. Most of the details can be found in the text, but summaries and figures could make it more obvious. For example Figure 4, showing mean SDR for image and text inputs in test, for models training with different modalities. This would be clearer in a table, ie | Test Modality | Train Modality | Image | Text | Both | ClipSep (Image) | 7.5 | 5.5 | ? | ClipSep (Text) | 6.2 | 8.1 | ? | ClipSep (Both) | 8.1 | 8.2 | ? |

- #s are approximately estimated from figure 4. Here one can see how good the model is if the train/test modalities are matched. There's more lost when trained on image and tested on text (unfortunately the main goal of the paper). Using both in train help significantly. Could you test with both? Would be an interesting result.

The paper is novel in a narrow sense, since the field has a lot of work in audio separation via query and addressing unsupervised separation of audio sources. The unsupervised separation of audio by query is similar to the work in:

- Liu et al., Separate What You Describe: Language-Queried Audio Source Separation, Proc Interspeech 2022
  - text queries are used to select a source to separate in audio-only samples
  - the paper under review has the addition of a visual modality to improve the correspondence between text and the input modes.
- Zhao et al. The Sounds of Pixels. ECCV 2108 (cited by paper and base implementation)
  - unsupervised audio-visual source separation in videos with musicians playing music, selection/query by image~
  - the paper under review adds a text query component to select the source to separate out, and a Noise Invariant Training scheme to cope with (audio) noise sources that have no correspondence in the video. it also focuses on unconstrained sound vs only music in Zhao.
- Wisdom et al. Unsupervised Sound Separation Using Mixture Invariant Training
  - unsupervised audio separation, mixture of mixtures invariant training
  - doesn't provide a means to select a single source to extract (separates all sources)

The paper uses publicly presented data sources and published github repositories. The paper should be relatively easy to reproduce.

Minor comments

- are the masks used in the paper binary or ratio? Zhao mentions that both are possible.
- 4th line in Conclusion has a typo "language pre*training*".

Summary Of The Review:
Overall the paper is novel in a narrow sense. It builds on the Sound Of Pixels work but adding a method of textual query. The results are good, demonstrating the approach is viable, however in the opinion of this reviewer, not overwhelming excellent (other reviewers may disagree). It feels more incremental than ground breaking, hence the recommendation to marginally accept.

Correctness: 3: Some of the paper’s claims have minor issues. A few statements are not well-supported, or require small changes to be made correct.

Technical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Empirical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Flag For Ethics Review: Yes, Legal compliance (e.g., GDPR, copyright, terms of use)

Details Of Ethics Concerns:
I am unsure of this, but I believe that this work used YouTube videos as training data and thus requires downloading them which is against YouTube's term of service. There has been a lot of published work though that has used YouTube as a data source such as AudioSet [1] and VoxCeleb [2]. [1] is even from Google, YouTube's parent.

Recommendation: 8: accept, good paper

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.

Add **Public Comment**

[+] **Official Comment by Paper3561 Authors • Response to reviwer vU71 part 2/2**

[+] **Official Comment by Paper3561 Reviewer vU71 • Response to authors**

[+] **Official Comment by Paper3561 Authors • Response to reviwer vU71 part 1/2**

[−] # Official Review of Paper3561 by Reviewer dcmL

ICLR 2023 Conference Paper3561 Reviewer dcmL

24 Oct 2022 (modified: 01 Dec 2022) ICLR 2023 Conference Paper3561 Official Review Readers: Everyone

Summary Of The Paper:

The paper describes a self-supervised way to do sound separation using a frozen pre-trained CLIP, along with video data (assumed to also have audio).

The core method of CLIPSep is shown in Fig 2. During training, they run the frames of two different videos through CLIP in order to independently get embeddings, which are then projected into another space by a learnable projection mapping. In parallel, they add together the audio streams of both videos, encode this as a spectrogram, and then run that through an Audio UNet. They then independently combine the output of the UNet with each video's projections in order to predict an audio mask. That audio mask is compared against the true audio mask for the video in order to get a loss.

Figure 3 expands on CLIPSep and introduces CLIPSep-NIT in order to better account for noisy streams of audio. It's more complicated, but the gist is to create audio masks that account for the noise found in in-the-wild videos. This is patterned after the MixIT approach from Wisdom et al.

They then show that this self-supervised approach can be comparable to supervised datasets on two different tasks involving mixing test VGGSound and eval MUSIC+ with VGGSound.

Strength And Weaknesses:

Strengths:

1. The main strength is that the method is novel. I like this idea a lot and think there's something materially interesting if you ramp up the dataset size.
2. The comparisons are also clear. The tables show the delineations between the models that you compare and I don't have trouble understanding what's going on wrt numbers.

Weaknesses:

1. The explanation of the model feels like some info is left out, notably from where the images are extracted with respect to the audio. As I understand, there is a singular image per video (2 total to be exact), but it's unclear how the audio is determined around that. It can't be instantaneous. Is it 10 seconds around it? Maybe I'm missing it, but this seems important for reproduction.
2. There should be audio samples here. It's hard to truly evaluate what's going on without audio samples. I don't see any such links in the paper.
3. I don't understand at all what is section 4.1. What is the task? I read through it a few times and it's unclear to me what you're actually doing there.

Clarity, Quality, Novelty And Reproducibility:

1. Clarity

- What's up with the Figure 3 graphic? The clarity of this paper would be helped a lot if you made the 2nd half of this better because it's hard to grok what's going on in the text itself. As an example, why is part of it greyed out? If that's supposed to be inference, then it doesn't match w the blue text that describes inference before. Another example is if the greyed out dotted lines from projection --> predicted noise mask are using the black line, very unclear. Then in the dark blue directional arrows from predicted noise mask to the noise invariant training we have a similar issue. Add something text to make this clear, it's unfortunately harming what is an interesting section.
- Please clarify what's going on in 4.1.

2. Quality

- I get that the authors tested all the models on their hybrid approach in 4.2 and it came back w at least the *order* I'd expect. That was cool. However, it does seem strange that they did this mixing of datasets. Is that what other papers are doing? I'm not as familiar w this field as I'd like to be to question that, but it is does seem kind of strange.
- Otherwise, the Quality was good imo.

3. Novelty

- This is where the paper shines. I like the idea a lot and think there is merit in pushing this further. It's an interesting way to create an original interface at test time.

4. Reproducibility:

- There should be more details about the image + audio pairings. I see in the Appendix that they use 4 second audio clips, but where is the image drawn from?
- Also see comment above in Clarity about CLIPSep-NIT.

Summary Of The Review:
Before I increase my score, I would want to see the paper improve significantly wrt clarity, notably section 4.1 and Figure 3. I would also like to see a better explanation for the evaluation approach in 4.2 and perhaps something to support it elsewhere in the literature. If those are satisfied I will increase my score because I do like this paper and think the underlying method deserves to be recognized.

Correctness: 3: Some of the paper's claims have minor issues. A few statements are not well-supported, or require small changes to be made correct.

Technical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Empirical Novelty And Significance: Not applicable

Flag For Ethics Review: NO.

Recommendation: 8: accept, good paper

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

Add **Public Comment**

[+] **Official Comment by Paper3561 Authors • Response to reviewer dcmL**

[+] **Official Comment by Paper3561 Reviewer dcmL • Response**

[−] # Official Review of Paper3561 by Reviewer 8zRs

ICLR 2023 Conference Paper3561 Reviewer 8zRs

23 Oct 2022 (modified: 03 Dec 2022) ICLR 2023 Conference Paper3561 Official Review Readers: Everyone

Summary Of The Paper:
This paper proposes a source separation system for text-queried source separation. The authors propose to train the system with a picture query during training time, however in inference time they use text for the query. In addition to the basic system, they also propose to add a mixit layer at the end of the pipeline to increase the noise robustness of the system.

Strength And Weaknesses:
Strengths:

- The proposed problem is definitely interesting, and I can see the practical applications of this system.
- The results (shared in the link https://dezimynona.github.io/separation/ (https://dezimynona.github.io/separation/)) seems to suggest that the system is doing what is intended.

Weaknesses:

- I think it would have been nice to also compare with a baseline system which uses sentence embeddings as a guide. This paper could be a nice point of comparison https://arxiv.org/pdf/2203.15147.pdf (https://arxiv.org/pdf/2203.15147.pdf). You could have done this comparison in two ways. 1) On your experiments you can directly train this model and compare 2) You could have taken a pretrained systems for both your approach, and the baseline and compare in a zero-shot manner. The VGGSound+None experiment that you have on your demo page is a nice option for this.
- There is little difference between the separation quality of Clipsep and Clipsep+NIT. In some of the examples on your demo page the two methods sound very similar.

Clarity, Quality, Novelty And Reproducibility:
The paper reads well in general. In terms of novelty, due to the fact that this paper proposes a new training methodology which enables training with audio-video pairs, it seems to differentiate itself from the existing papers.

Summary Of The Review:
I think this paper proposes an interesting training methodology. I think it's above the acceptance threshold. My only problem with it is the lack of comparison with text-query-only models. (See my comment above)

Update after rebuttal: The authors provided a BERT based baseline, and I increased my score.

Correctness: 3: Some of the paper's claims have minor issues. A few statements are not well-supported, or require small changes to be made correct.

Technical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Empirical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Flag For Ethics Review: NO.

Recommendation: 8: accept, good paper

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.

Add **Public Comment**

[+] **Official Comment by Paper3561 Authors • Response to reviewer 8zRs**

[+] **Official Comment by Paper3561 Reviewer 8zRs • Thanks for your response**

[+] **Official Comment by Paper3561 Authors • Additional baseline results ...**

[+] **Official Comment by Paper3561 Reviewer 8zRs • Text encoder**

[−] Official Review of Paper3561 by Reviewer CpEt

ICLR 2023 Conference Paper3561 Reviewer CpEt

21 Oct 2022 (modified: 17 Nov 2022) ICLR 2023 Conference Paper3561 Official Review Readers: Everyone

Summary Of The Paper:
CLIPSep demonstrates how a pretrained CLIP model can be used to train a source separation model using unlabeled videos and achieve competitive results in some settings.

Strength And Weaknesses:
Strengths:

- This model shows a path toward training a sound separation model that is text queryable for arbitrary sources and can be trained on unlabeled video data.
- Results are competitive with labeled approaches in some settings.

Weaknesses:

- The goal of this approach is to be able to scale up training on an arbitrary number of in-the-wild videos. However, the model is trained and evaluated only on relatively small and clean datasets. Even when the data is somewhat noisy (e.g., the offscreen noises in VGGSound), the model starts to exhibit difficulties using only text queries. The authors acknowledge these issues in the Discussion section and provide some ideas for improvement, but I'm concerned that we don't know how well the model will actually scale up to in-the-wild video datasets. It's possible that entirely different techniques will end up being needed to get to that level of unlabeled training.
  - *Update*: After discussion with the authors, I realized I misunderstood the scale of VGGSound and how representative it is of "in the wild" audio, so I am much less concerned with how well this technique will scale up.
- Motivation for some of the architecture design choices is not fully explained and alternatives are not fully explored (details below).
  - *Update*: After discussion with the authors, they have updated the paper to explain some of these choices. I found the discussion around "early fusion" vs. "late fusion" particularly interesting.

Clarity, Quality, Novelty And Reproducibility:
Novelty: This is the first work to show training a text-queryable sound separation model trained on unlabeled video data.

Reproducibility: All code and pretrained models will be made available.

Overall clarity is good, but I have a few suggestions:

- Section 2.3: My understanding is that the CLIP model is used as is without any training or finetuning. I think the final sentence of this paragraph could be reworded to make it clear that the part of the model you're optimizing doesn't include CLIP.
- The paper mentions a few times that the model and code is based on Sound-of-Pixels. I realize that the techniques in this paper are different than the SOP approach, but I think it would be helpful to have those differences called out explicitly because important parts are reused.
- For the architecture, I'd like to hear more about the intuition behind having the U-Net output k masks without any conditioning on the separation query. Rather than having the query vectors mix the intermediate masks, why not just condition mask generation on the query?
- Why are the noise heads discarded at test time? Is the intuition that you're training the U-Net to use some of its k masks to specialize in noise and then not be utilized by the query vectors?

Summary Of The Review:
CLIPSep shows a novel approach to training source separation on unlabeled videos. However, I am concerned that the main value to this approach will come from scaling up the training dataset to many in-the-wild videos, but that setting was not attempted in this paper. As shown by the VGGSound results, it's possible that there will be problems with the noisiness of in-the-wild videos that prevent this technique from working without additional insights and modifications.

Update: As mentioned above, I am now much less concerned about the ability of this technique to scale up.

Correctness: 4: All of the claims and statements are well-supported and correct.

Technical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Empirical Novelty And Significance: 3: The contributions are significant and somewhat new. Aspects of the contributions exist in prior work.

Flag For Ethics Review: NO.

Recommendation: 8: accept, good paper

Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

Add **Public Comment**

[+] **Official Comment by Paper3561 Authors • Response to reviewer CpEt part 2/2**

[+] **Official Comment by Paper3561 Authors • Response to Reviewer CpEt part 1/2**

[+] **Official Comment by Paper3561 Reviewer CpEt • Response**

[+] **Official Comment by Paper3561 Authors • Response**