Language-Guided Audio-Visual Source Separation via Trimodal Consistency

Reuben Tan[1] Andrea Burns[1] Arijit Ray[1] Bryan A. Plummer[1] Oriol Nieto [2]
Justin Salamon[2] Bryan Russell[2] Kate Saenko[1]
[1]Boston University, [2]Adobe Research
{rxtan, aburns4, array, bplum, saenko}@bu.edu, {onieto, salamon, brussell}@adobe.com

Abstract

We propose a self-supervised approach for learning to perform audio source separation in videos based on natural language queries, using only unlabeled video and audio pairs as training data. A key challenge in this task is learning to associate the linguistic description of a sound-emitting object to its visual features and the corresponding components of the audio waveform, all without access to annotations during training. To overcome this challenge, we adapt off-the-shelf vision-language foundation models to provide pseudo-target supervision via two novel loss functions and encourage a stronger alignment between the audio, visual and natural language modalities. During inference, our approach can separate sounds given text, video and audio input, or given text and audio input alone. We demonstrate the effectiveness of our self-supervised approach on three audio-visual separation datasets, including MUSIC, SOLOS and AudioSet, where we outperform state-of-the-art strongly supervised approaches despite not using object detectors or text labels during training. We include samples of separated audio in the supplemental material.

1. Introduction

Our everyday audiovisual world is composed of many visible sound sources, often with multiple sources layering on top of one another. For example, consider the video of the guitar and cello musicians playing together in Fig. 1. The two instruments have distinct timbres, and the musicians play non-unison, but complementary melodies. Despite hearing both instruments simultaneously, humans have an innate ability to identify and isolate the melody of a single source object. In this paper, we define the corresponding machine task as follows: given a natural language query that selects a sounding object, such as "person playing a guitar", separate its sound source from the input audio waveform and localize it in the input video, without any supervision.

This task is challenging. First, there is no approach for associating the linguistic description of a sound-emitting

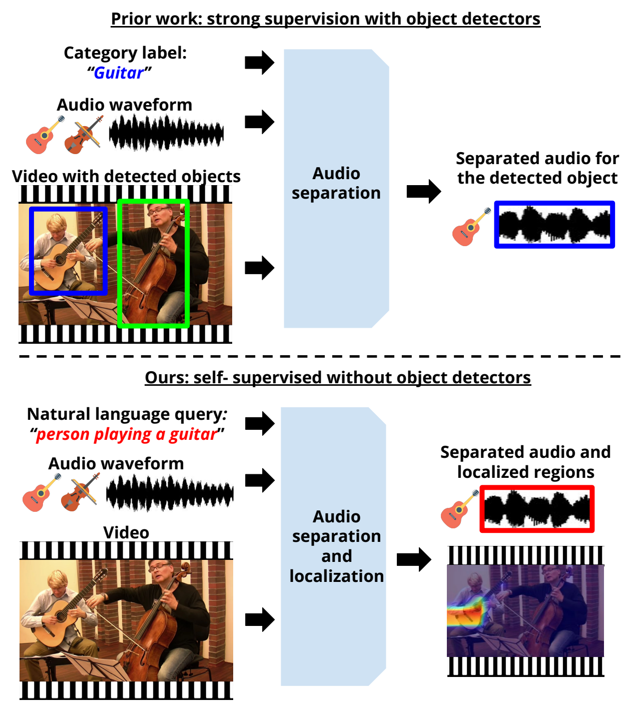


Figure 1. We propose to separate and localize audio sources based on a natural language query, by learning to align the modalities on completely unlabeled videos. In comparison, prior audio-visual sound separation approaches require object label supervision

object to its visual features and the corresponding components of the audio waveform without access to annotations during training. Existing audio-visual methods [5, 12, 39] do not generalize to natural language queries due to their dependence on discrete object class labels. Second, an ideal solution would jointly identify and localize sound-emitting objects in videos as well as separate the corresponding components in the audio waveform without strong supervision. Although prior audio-visual work has demonstrated the benefits of aligning relevant object regions in the video with their corresponding sounds [5, 12], these approaches require strong supervision including object label and bounding box annotations (see Fig. 1 top). Overcoming these challenges would enable important downstream applications including holistic video understanding [31], embodied AI [6], and bidirectional audio-to-video retrieval [36].

To address these challenges, we make the following contributions. First, we propose Audio Visual Separation through Text (AVSeT), a self-supervised approach that leverages large vision-language "foundation" models [17, 24] to provide pseudo-supervision for learning the alignment between the three modalities: audio, video and natural language. Our key insight is to learn a strong *transitive* relation from audio to natural language using vision as an intermediary modality, while preserving the alignment between the visual and natural language modalities embodied by the foundation models. However, just using the visual representations of these foundation models in existing AV separation approaches does not preserve the transitive relationships between the three modalities (Sec. 4.1).

As our second contribution, we introduce two novel multimodal alignment objectives that encourage the learnt audio representations to encode the semantics of captions and infer the latent transitive relation between the three modalities. While natural language provides the generality to express a large and varied range of visual concepts for audio separation in videos, the absence of captions in unlabeled videos during training poses a significant challenge in our self-supervised formulation. To learn the transitive alignment, we adapt a foundation model to extract *latent captions* from unlabeled videos. Intuitively, the latent captions are representations that express the visual concepts present in the videos. Third, we introduce a Multiple Instance Learning formulation to learn to perform audio separation at the video *region* level since we do not have prior information on relevant objects or their locations in the videos during training.

Finally, we demonstrate the effectiveness of our proposed AV-SeT approach through extensive evaluations on the audio source separation task on the SOLOS [21], MUSIC [39], and AudioSet [13] datasets. We show that our self-supervised approach outperforms strongly-supervised state-of-the-art approaches without using labels during training by leveraging the capability of vision-language foundation models. More importantly, we demonstrate that AV-SeT learns to use language queries for audio separation despite not training with ground-truth language supervision.

2. Related Work

Audio-only source separation. The goal of audio-only source separation is to use the aural cues present in the input audio waveform to separate the individual components. Before the advent of deep learning, audio signal processing techniques are conventionally used to address the problem of audio-only source separation, where such approaches have relied on strong assumptions such as the number of sources in the audio waveforms to compute non-negative matrix factorization [3, 8, 35] of audio spectrograms. Deep learning based approaches [5, 12, 39] commonly adopt the self-supervised 'mix-and-separate' strategy where multiple audio sources are combined into a synthetic mixture and the goal is to predict a spectrogram mask to retrieve the queried audio components.

Multimodal source separation. Recent work in audio-text [18, 19] and audio-visual [5, 12, 38, 39] separation also use the 'mix-and-separate' strategy to train a decoder to predict a spectrogram mask based on natural language and video queries, respectively. In the case of the former, state-of-the-art approaches can either accept a single object label [5,12,39] or free-form natural language queries [18,19]. Existing audio-visual source separation approaches often rely on training object detectors with strong supervision from bounding box annotations to localize objects of interest before learning to perform source separation at the object-level. In contrast, our proposed approach does not rely on pretrained object detectors or object labels in the training videos. The Sound of Pixels (SOP) model is the most similar in spirit to our proposed approach since it does not rely on object detectors. Instead, it learns to perform sound separation based on a video-level representation during training. Our AV-SeT approach is also similar to the Voiceformer [26] approach which aims to separate speech from multiple speakers using a combination of audio, visual and language inputs. In contrast to our approach, their approach requires ground-truth text transcripts for training.

Multimodal representation learning. One key challenge in the multimodal setting is grounding information from the text and/or audio modalities in the video frames [2,5,12,33]. Since annotating video datasets is a costly process, many approaches have aimed at understanding the effectiveness of self-supervised pretraining [15, 34] on unlabeled videos. A significant number of existing work aims to utilize complementary information from the audio, text and visual modalities [37] to learn robust representations that can generalize to downstream tasks such as action recognition [1, 23, 30] and text-to-video retrieval [2, 20]. These approaches generally pre-train their encoders on large-scale datasets such as Kinetics-400 / 600 [4] and HowTo100M [20] before finetuning linear probes on top of the pretrained encoders for target tasks and datasets. With the increasing prevalence of multimodal foundation models such as CLIP [24] and ALIGN [17], recent work has also focused on prompting these models to adapt them to new tasks ranging from open-vocabulary object detection [14] to domain adaptation [11] without modifying their parameters.

3. Approach

Given a set of unlabeled videos and their audio tracks, we strive to learn a model with self-supervision that can

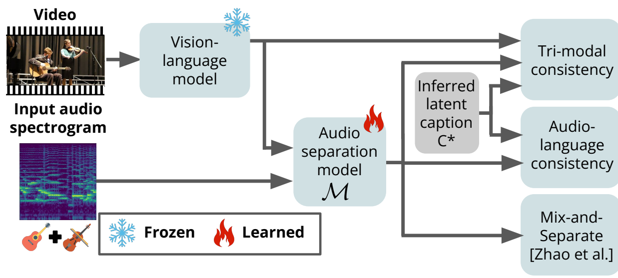


Figure 2. **Training overview.** We train our proposed AV-SeT model $\mathcal{M}$ using three objective functions. We introduce tri-modal and audio-language consistency alignment objectives to learn the latent alignment between the audio, video and natural language modalities. Furthermore, we adopt a prior mask prediction loss [39] to guide the training of our mask prediction decoder.

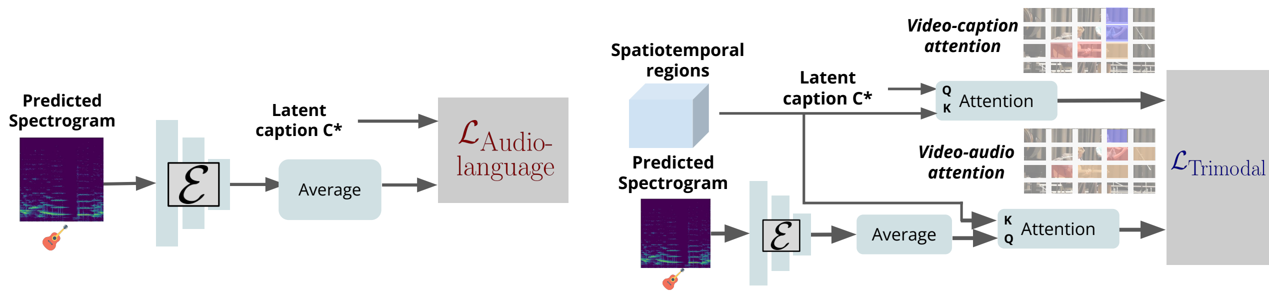
separate an audio source given an input natural language query and possibly an accompanying video. This goal requires a model to learn a strong alignment between the audio, vision, and language modalities. To this end, we propose a novel Audio-Visual Separation through Text (AV-SeT) approach (Figure 2) that learns an effective representation for audio-visual source separation, as well as a strong alignment between the audio and language modalities. To learn this representation, we exploit the strong joint visual-semantic alignment contained in pretrained vision-language foundation models by encouraging our model to learn a direct projection function for audio inputs into their joint embedding space (Section 3.1). Our key insight is to use videos as an intermediary modality to learn a *transitive* alignment between the audio and language modalities. Since there is already an alignment between the visual and language modalities in foundation models, learning a strong correspondence between audio and vision should translate to one between audio and language through the property of transitivity. This concept is integral to learning to reason across all three modalities from unlabeled videos.

Despite the intuitiveness of this idea, we observe that using the visual representations of foundation models in existing audio-visual separation approaches does not help them to learn the above-mentioned transitive alignment (see Table 4). We introduce the novel tri-modal and audio-language consistency alignment objectives (Figure 2). Intuitively, the tri-modal objective encourages the model to maximize the semantic consistency between all three modalities by using videos as the intermediary modality and the audio-language objective helps to further improve the transitive alignment between audio and natural language. Without text annotations, we leverage vision-and-language foundation models to infer latent captions that correspond to the audio components, thus providing pseudo-language supervision (Section 3.1). In addition to the above losses, we adopt the "mix-and-separate" strategy to train our audio separation model, where the audio from multiple videos are mixed and the goal is to separate them using the visual information in the videos. For an input audio waveform A, we apply a Short-Time Fourier Transform (STFT) to obtain its magnitude spectrogram A^S and phase A^{phase}. Our learning algorithm only makes use of A^S, which encodes how the magnitude of the audio frequencies changes over time. We later reconstruct the predicted waveform by applying the inverse STFT on the predicted spectrogram using the original phase A^{phase}. We describe our AV-SeT audio separation model $\mathcal{M}$ in Section 3.2.

3.1. Tri-modal alignment

A conventional method to learn a strong alignment between the audio, language and vision modalities is to maximize the pairwise similarities of their input representations using a contrastive learning formulation [1, 20]. However, this necessitates the presence of labels that indicate ground-truth correspondences between these modalities. In our self-supervised setting, the unlabeled videos only provide noisy labels of audio-visual pairings but do not contain text labels. To circumvent this limitation, we propose to extract *latent captions*, which are *pseudo-words* that encode the visual concepts in videos with language semantics. Prior work [7, 9] has demonstrated that the rich visual-semantic alignment in vision-language foundation models can be adapted to extract latent word representations that convey the semantics of their conditioning images. Inspired by this insight, we introduce the novel idea of using latent captions to identify sound-emitting object candidates in unlabeled videos for training, thereby allowing us to train without prior knowledge of existing objects in the videos.

Specifically, we perform "textual inversion" [7, 9, 10] where we adapt the CLIP language encoder to express the visual concepts of a video in the form of an encoded learnable parameter. Instead of introducing new concepts into the vocabulary of these models which require their parameters to be updated, we search for their latent language representations through a visual reconstruction objective. This objective is based on the key insight that the final learnt latent captions should be expressive enough to convey the visual semantics of the video. While it is also possible to use the visual frame representations as latent captions, we reason that learning them as outputs of the language encoder will help a model generalize better to language queries during inference. For a given video V, we encode its center frame V_{center} as: $f^V_{center} = g^V(V_{center})$, where g^V is the embedding function of the CLIP visual encoder. Then, we replace the token embeddings used for language queries with a learnable parameter p and encode it as: $g^L(p)$, where g^L is the embedding function of the CLIP language encoder. We optimize the weights of p by maximizing the cosine sim-



((a)) Audio-language consistency loss.

((b)) Tri-modal consistency loss.

Figure 3. We introduce the audio-language and tri-modal consistency alignment objectives to facilitate our model in learning a *transitive alignment* between the audio and natural language modalities by using videos as an intermediary modality.

ilarity between f^V_{center} and $g^L(p)$:

$$p^* \in \arg\max_p sim\left(f^V_{center}, g^L(p)\right) \tag{1}$$

where $sim(x, y) = x^T y/(\|x\|\|y\|)$ and $||.||$ denotes the L_2 norm operator. Let $C^* = g^L(p^*)$ be the latent caption. We illustrate this operation in the supplemental. In theory, C^* is analogous to a description of the frame and thus serves as an effective replacement for ground-truth annotations.

Our audio-language and tri-modal consistency alignment objectives are based on our reasoning that well-separated audio sources should be semantically consistent with the visual concepts that guide their separation as well as their text labels. Given an input video V of T frames $V = \{V_1, \cdots V_T\}$ and an audio spectrogram $A^S \in \mathbb{R}^{F\text{x}N}$ with F frequency bins and N STFT time frames, our model predicts a separated audio spectrogram $\hat{A}^S \in \mathbb{R}^{F\text{x}N}$. We extract a latent representation for the predicted spectrogram: $\hat{f}^A = \mathcal{M}_\theta(\hat{A}^S) \in \mathbb{R}^D$ where $\mathcal{M}$ denotes our audio separation model. For each video, we use its encoded predicted spectrogram and latent caption to provide pseudo-language supervision in our alignment objectives detailed below.

Audio-language consistency loss. To encourage our audio encoder to learn audio representations that are effective for separating sound components when conditioned on either text or video queries, we aim to maximize the pairwise similarities between all three modalities. The key insight is that the audio sources that are well-separated by the visual concepts in a video should have a strong semantic relevance to its latent caption which express the same concepts in natural language (Figure 3(a)). Theoretically, this is similar to the self-supervised multimodal learning objective of maximizing the similarity between the representations of an image and its corresponding caption as well as the dissimilarity of non-corresponding captions. In lieu of ground-truth object labels, we can maximize the alignment between the predicted audio representations and the latent captions over the entire vocabulary of captions $\mathcal{X}$:

$$\mathcal{L}_{\text{Audio-language}} = -\log\left(\frac{\exp(\hat{f}^A \cdot C^*/\tau)}{\sum_{x\in\mathcal{X}} \exp(\hat{f}^A \cdot x/\tau)}\right) \tag{2}$$

where τ is the temperature.

However, the problem of false negatives has been well-documented in image-text datasets, where captions for some images may be relevant to other images but are treated erroneously as negatives. Since we are training on unlabeled videos, we account for the high likelihood that some videos may contain similar sounding objects by using a lower weight for this objective. We theorize that maximizing the dissimilarity between the predicted audio and latent captions with similar sounding objects may make the effect of false negatives more pronounced and introduce a lot of noise during training if a high weight is used.

Tri-modal consistency loss While the audio-language consistency objective facilitates improving the alignment between audio sources and their corresponding latent captions, it does not provide a straightforward solution to disregard false negatives in its contrastive formulation. To address this, we further introduce a softer tri-modal alignment constraint which exploits the implicit localization capability of vision-language foundation models for supervision without requiring any negative samples. Specifically, we propose to use the intermediary visual modality to encourage well-separated audio components to be grounded correctly in the relevant spatiotemporal regions of a video. Prior work [27, 40] has demonstrated that the CLIP model can be engineered to extract a segmentation map based on natural language queries. Inspired by this, we use the latent captions to provide pseudo-target attention maps. Our key intuition is that enforcing a soft constraint on the predicted spectrograms such that they can be mapped to similar relevant regions as determined by the latent captions will encourage the model to implicitly learn the transitive alignment between the audio and language modalities.

To begin, we define $P_{Att}(A, b) = \sigma(A \cdot b)$ to be the attention operation where $\sigma(z)_i = z_i / \sum_j z_j$ is the softmax

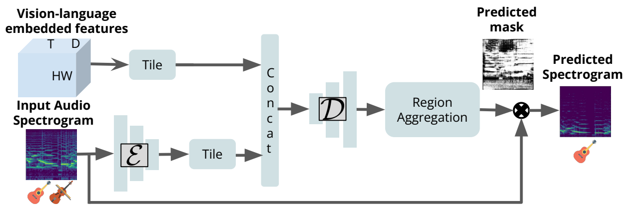


Figure 4. **Spectrogram mask prediction during inference.**

function. For a given video V, we extract a set of spatiotemporal region representations $f^V \in \mathbb{R}^{T\text{x}H\text{x}W\text{x}D}$ by encoding each frame separately, where H and W are the downsampled height and width of the input video. We encode the t-th frame as: $f_t^V = g_\theta^V(V_t)$, where $f_t^V \in \mathbb{R}^{HW\text{x}D}$. Then, we compute a probability distribution of similarity scores over the set of spatial region features with respect to its latent caption C^* for the t-th frame: $P_{VC}^t = P_{Att}(f_t^V, C^*) \in \mathbb{R}^{HW\times 1}$. Next, we compute a similar distribution over the region features of the t-th frame with respect to the encoded audio representation of the masked spectrogram: $P_{VA}^t = P_{Att}(f_t^V, \hat{f}^A) \in \mathbb{R}^{HW\times 1}$. We encourage the predicted audio representation to learn a similar mapping to that of the latent caption by minimizing the Kullback-Leibler (KL) divergence between the two attention distributions over the set of all time steps in a video $\mathcal{T}$:

$$\mathcal{L}_{\text{Trimodal}} = \mathbb{E}_{t\sim\mathcal{T}}\left[\sum_x P_{VC}^t(x)\log\frac{P_{VC}^t(x)}{P_{VA}^t(x)}\right], \quad (3)$$

where $P_{VA}^t(x)$ denotes the audio-video attention distribution over the t-frame of the x-th sample.

3.2. Audio Visual Separation through Text

Our Audio Visual Separation through Text (AV-SeT) model comprises an audio model and visual-language encoders. We use the pretrained CLIP model as our visual-language encoders to exploit its learnt visual-semantic alignment. Given a natural language query $L \in \mathbb{R}^{N_w}$ where N_w is the number of words, we extract its representation $f^L \in \mathbb{R}^{1\text{x}D}$ using the text encoder: $f^L = g^L(L)$. We stress that language queries are only used during inference.

Additionally, we adopt the U-Net [29] model, which has been demonstrated by prior work [12, 39] to be effective at audio source separation. The audio U-Net model comprises an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. Given an input audio spectrogram A^S, the encoder $\mathcal{E}$ passes it through a series of downsampling convolutional layers to output a set of bottleneck features: $f^A = \mathcal{E}(A^S)$, $f^A \in \mathbb{R}^{H^A\times W^A\times D}$, where H^A and W^A denote the downsampled height and width of the downsampled spectrogram, respectively. We enforce that the U-Net encoder learns a mapping of its audio representations directly into CLIP visual-semantic embedding space (Figure 4) so that it will be able to use both visual and language queries for audio separation. Given a visual representation f^V as query , we tile the visual representation by the factor $H^A W^A$ and concatenate them with the audio bottleneck representations along the channel dimension: $f^{AV} = concat(f^A, tile(f^V))$, where f^{AV} has the dimensions $\mathbb{R}^{H^A\times W^A\times 2D}$. We pass the concatenated representations into the decoder $\mathcal{D}$ consisting of a series of upsampling convolutional layers to generate a real-valued ratio mask: $\hat{M} = \mathcal{D}(f^{AV}) \in \mathbb{R}^{F\times N}$. The mask is then applied onto the input spectrogram to compute the predicted audio component that corresponds to the input video: $\hat{A}^S = \hat{M} \odot A^S$.

We adopt the self-supervised 'mix-and-separate' learning objective [12, 39] to train the audio decoder $\mathcal{D}$ for predicting spectrogram masks. In this formulation, audio from two or more videos are mixed and the goal is to use the visual information of a video to separate its audio component. This allows us to generate pseudo-target masks for training supervision. We compute L_{mask} as the L1 loss between the pseudo-target and predicted ratio marks [39]. We refer readers to the supplemental for details on L_{mask}.

Multiple Instance Learning for mask prediction. Prior work [12] has demonstrated the advantages of performing audio-visual separation at the object level for videos with multiple objects. In the absence of object detectors and labels, we propose an MIL formulation, where all spatiotemporal regions in the input video are treated as positive candidates. Specifically, given a spatiotemporal grid of region representations extracted from the visual encoder as defined in Section 3.2, we predict a spectrogram mask $\hat{M}_{t,r}$ for the r-th region in the t-th frame. Finally, we compute an aggregated spectrogram mask for the entire video:

$$\hat{M} = \frac{1}{T}\sum_{t=1}^{T}\sum_{r=1}^{HW'}\hat{M}_{t,r}. \quad (4)$$

Intuitively, this formulation encourages the model to learn to identify the salient regions with sounding objects for audio separation. We enforce this constraint within frames and average the predicted masks over the temporal dimension since we reason that a sounding object should have similar prediction masks across frames. The final objective function is computed as the sum of all three loss components:

$$\mathcal{L} = \lambda_1\mathcal{L}_{mask} + \lambda_2\mathcal{L}_{\text{Trimodal}} + \lambda_3\mathcal{L}_{\text{Audio-language}} \quad (5)$$

where λ_1, λ_2 and λ_3 denote the weights for the mask, trimodal and audio-language consistency losses, respectively.

4. Experiments

We evaluate the effectiveness of our proposed AV-SeT approach on the tasks of audio-text and audio-visual source separation in videos. To perform fair comparisons against prior work, we conduct experiments on the SOLOS [21],

Method	Object detectors	# Region proposals	Solos NSDR ↑	Solos SIR ↑	Solos SAR ↑	Duets NSDR ↑	Duets SIR ↑	Duets SAR ↑
Co-Separation [12]	Yes	2	7.38	13.70	10.82	7.42	13.81	10.60
AVSGS [5]	Yes	42	**9.04**	**14.45**	**12.24**	**10.25**	**15.60**	**12.82**
Sound-of-Pixels [39]	No	None	7.30	11.90	11.90	6.05	9.81	10.61
AV-Mix-and-Separate [12]	No	None	3.16	6.74	8.89	3.23	7.01	9.14
NMF-MFCC [32]	No	None	0.92	5.68	6.84	0.92	5.68	6.84
AV-SeT (Ours)	No	None	**7.98**	**13.92**	**12.35**	**8.08**	**13.97**	**11.33**

Table 1. **Source separation results on the MUSIC dataset.** We report the performance of the various approaches under 2 training settings - videos that contain a single instrument (solos) and two instruments (duets).

MUSIC [39] and AudioSet [13] datasets. We refer readers to the supplementary material for more details of these datasets and implementation details. Given a language query that is composed of a prompt template and an object, the goal of **audio-text source separation** is to extract its corresponding sound component from an audio input, which usually comprises a mixture of sounds emitted by various objects except for those devoid of sounds. Similarly, in **audio-visual source separation**, a model has to leverage the visual information in the input video to extract its corresponding sound component and align it with the relevant video pixels.

Evaluation metrics. We adopt the Normalized Signal-to-Distortion Ratio (NSDR), Signal-to-Interference Ratio (SIR) and Signal-to-Artifact Ratio (SAR) [25, 28]. The NSDR is computed as the difference between the SDR of the disentangled and mixed audio with respect to the ground-truth waveforms. This metric measures how well the separated audios sound in general. The SIR measures the amount of other audio sources in the predicted audio. Finally, the SAR metric measures the amount of artifacts in the predicted audio as compared to the ground-truth audio.

4.1. Quantitative results

We compare our approach to the following baselines. Similar to our approach, **Sound of Pixels (SOP)** [39] does not require object detectors but learns to perform video-level separation. We also implement the **AV-Mix-and-separate** baseline from [12] with the same audio-visual separation model as ours but is trained to perform video-level separation and without our alignment objectives. The off-the-shelf **NMF-MFCC** [32] baseline performs audio separation using non-negative matrix factorization (NMF) and Mel frequency cepstrum coefficients (MFCC). Next, we also compare to **Co-Separation** [12] that learns to perform audio separation at the object-level by relying on object detectors. Last but not least, we include **AVSGS** [5] that also relies on a high number of object proposals to construct a spatiotemporal scene graph to reason about context.

Audio-visual source separation on MUSIC. We report the results of our evaluation on the MUSIC dataset under two settings in Table 1. In the first setting, each video contains only a single instrument while the 'duet' setting includes videos that contain two different instruments. To begin, our AV-SeT approach outperforms SOP by a large margin despite not relying on object detectors too. As shown in Table 4, we also demonstrate that the performance gain is not due to just replacing ImageNet-pretrained Resnet18 [16] visual representations with those of the CLIP model since it leads to a large drop in performance. We hypothesize that SOP does not learn a direct projection of the audio inputs used in source separation into the CLIP embedding space. In contrast, our approach facilitates a better adaptation of the learnt audio representations to the CLIP embedding. We also found it important to modify CLIP's self-attention (see supp.). We observe that AV-SeT outperforms AV-Mix-and-Separate by a significant margin, particularly under the duet setting. This suggests that performing region-level audio separation provides more capacity for the model to reason about multiple sounding objects in videos.

More significantly, we observe that our approach outperforms the Co-Separation approach [12], which relies on object labels in the videos. Since Co-Separation uses a high confidence region proposal for each instrument to localize relevant regions, our improvements over it suggest that our latent captions are able to express multiple visual concepts that are present in the video. Last but not least, our approach is also comparable to AVSGS even without scene graph reasoning modules. The latter constructs a spatiotemporal visual scene graph over a large number of region proposals to reason about context between detected objects. We note that their audio-visual scene graph component can be combined with our AV-SeT approach to possibly improve performance but is beyond the scope of this work.

Audio-visual source separation on SOLOS. While the videos in the SOLOS dataset generally have less background noise than those of the MUSIC dataset, we see in Table 2 that AV-Mix-and-Separate and NMF-MFCC are still unable to generalize to the cleaner audio signals. Similar to the reported performance on the MUSIC dataset, we also observe that AV-SeT is comparable to the strongly supervised AVSGS approach on the SOLOS dataset. One pos-

Method	Object detectors	NSDR	SIR	SAR
Co-Separation [12]	Yes	7.11	12.09	10.05
AVSGS [5]	Yes	**9.20**	**14.05**	**12.17**
Sound-of-Pixels [39]	No	6.28	10.84	10.13
AV-Mix-and-Sep [12]	No	2.94	5.81	8.33
NMF-MFCC [32]	No	0.68	4.75	5.12
AV-SeT (Ours)	No	**8.58**	**14.16**	**12.35**

Table 2. **Source separation results on the SOLOS dataset.**

Method	Object detectors	NSDR	SIR	SAR
Co-Separation [12]	Yes	4.26	7.07	13.00
AVSGS [5]	Yes	**5.28**	**8.27**	**13.04**
Sound-of-Pixels [39]	No	1.66	3.58	11.50
AV-Mix-and-Sep [12]	No	1.68	3.30	12.20
NMF-MFCC [32]	No	0.25	4.19	5.78
AV-SeT (Ours)	No	**4.15**	**7.62**	**13.20**

Table 3. **Source separation results on the AudioSet dataset.**

Model	Visual encoder	NSDR	SIR	SAR
SOP [39]	ImageNet	6.28	10.84	10.13
SOP [39]	CLIP	4.42	8.36	8.21
Ours	CLIP	**8.58**	**14.16**	**12.35**

Table 4. **Using CLIP visual features in Sound of Pixels model on the SOLOS dataset.**

$\mathcal{L}_{\text{Audio-language}}$	$\mathcal{L}_{\text{Trimodal}}$	NSDR	SIR	SAR
✗	✗	5.47	10.55	10.95
✓	✗	8.08	13.74	12.18
✗	✓	8.10	13.84	11.79
✓	✓	**8.58**	**14.16**	**12.35**

Table 5. **Ablation of our audio-language and tri-modal consistency alignment objectives on the SOLOs dataset.**

# tokens	NSDR	SIR	SAR
1	7.31	11.34	11.71
2	7.67	13.07	11.45
3	**8.58**	**14.16**	**12.35**
4	8.02	13.39	11.53

Table 6. **Ablation over number of learnable tokens on SOLOS.**

sible reason behind the lower performance of AVSGS on videos with single instrument is that it is less critical to reason about context for videos with a single sounding object. Our better performance as compared to Co-Separation suggests that learning to perform audio separation at the coarse video region level under an MIL formulation can serve as an effective replacement for training object detectors. These results also indicate the great promise of latent captions in replacing ground-truth object annotations for supervision.

Audio-visual source separation on AudioSet. Finally, we report the results of our evaluation on the AudioSet dataset in Table 3. AudioSet has been documented by prior work [12] to be much noisier than the videos in the other two datasets, which explains the lower performance across all approaches. Unlike the SOP model, we observe that our training approach is more robust to noise in the dataset where the sounding object may not always be visible.

Ablation of $\mathcal{L}_{\text{Trimodal}}$ and $\mathcal{L}_{\text{Audio-language}}$. We conduct an ablation over our audio-language and tri-modal consistency losses in Table 5 on the SOLOS dataset to determine their contributions. Despite being trained on a larger and more diverse set of visual concepts, CLIP visual representations alone does not encourage the audio encoder to learn a strong alignment between the audio and vision modalities for audio-visual source separation. This finding is evidenced by the low performance when we only use the mask prediction loss (row 1). We observe the importance of our proposed alignment objectives where adding one or the other leads to a significant improvement in audio separation performance. Due to multiple training videos containing similar instruments, it is likely that we are treating some latent captions as false negative samples for each video in the consistency objective. However, we see that the consistency objective is still beneficial towards audio separation.

Number of learnable token embeddings in latent captions. We report the results of our ablation over the number of learnable token parameters used in the extraction of latent captions in Table 6. We observe that increasing the number of learnable tokens generally helps to improve the performance, although using 4 tokens hurts performance slightly. We hypothesize that a higher number of learnable tokens provide more capacity for the CLIP language model to express multiple visual concepts that are present in the videos. This finding suggests that using more tokens may be beneficial for videos with more complex visual scenes.

Audio-text source separation with descriptive natural language queries. Finally, we evaluate the capability of our trained model to separate audio components based on natural language queries in Table 7. We construct each query using the template 'person playing a {instrument}'. We stress that we only train on unlabeled videos without text supervision. To begin, we compare our AV-SeT model to a variant that is trained without our alignment objectives. As evidenced by the significant performance gap between the first two rows, our alignment losses are integral to learning a strong transitive alignment between the audio and language modalities. This suggests that just learning an alignment between the audio and visual modalities does not result in a *transitive* relationship between the former and natural language. Finally, we observe in the last two rows that our alignment objectives are still insufficient for closing the per-

Query Modality	Alignment objectives	SOLOS			MUSIC			Audioset		
		NSDR	SIR	SAR	NSDR	SIR	SAR	NSDR	SIR	SAR
Language	No	-3.05	2.79	3.77	-3.67	2.51	3.41	-5.02	1.98	14.94
Language	Yes	**6.92**	**12.07**	**10.41**	**6.45**	**11.18**	**10.77**	**2.36**	**4.71**	**10.28**

Table 7. **Audio-text separation with natural language queries.** We evaluate our model, that is trained only on unlabeled videos, on the task of audio-text separation. Note that we do not compare to other audio-visual separation baselines since there is no straightforward way to adapt them for language queries. We compare our full AV-SeT model to a variant that is only trained with the mask prediction loss.

formance gap between audio-text and audio-visual separation completely. This indicates that future work is needed for further improving the latent alignment between all three modalities.

4.2. Qualitative results

Visualizations of latent captions. To understand what our extracted latent captions encode, we provide some example attention visualizations of the latent captions with respect to the video frames in Figure 5. Interestingly, we observe that a latent caption is capable of describing multiple instances of the same object in the middle visualization, where it is focusing on all three clarinets.

Predicted audio-to-video attention visualizations. We also evaluate the semantic consistency between the predicted audio components and the visual cues that guide the separation process. To this end, we plot the attention map computed from the encoded predicted spectrograms and their corresponding video frames (Section 3.1) in Figure 6. We observe that our training approach helps the model to learn to ground the audio source in the video pixels despite not requiring bounding box supervision.

Predicted outputs of language-based audio separation. Finally, we present an example of how different language queries affect the separation performance for the same audio input in Figure 7. Interestingly, despite not relying on text annotations for training supervision, our model can separate audio components based on free-form natural language queries. We observe that providing more context in the language query (*e.g.*, "violin" versus "person playing a violin") can lead to improved separations.

Figure 5. **Attention map visualizations of latent captions.**

Limitations. While we have demonstrated that our proposed AV-SeT approach is able to generalize well to free-

Figure 6. **Predicted audio attention in videos.**

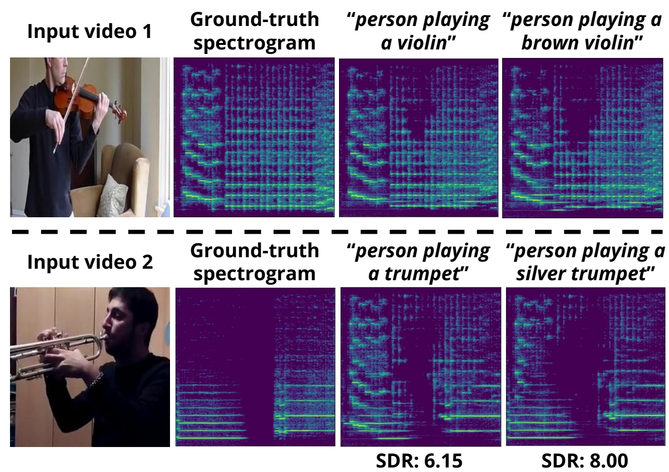


Figure 7. **Predicted audio spectrograms and masks with natural language queries.**

form natural language queries for source separation, we observe that it is only able to handle visually descriptive adjectives such as *person playing a small trumpet* instead of *a loud trumpet*. We hypothesize that this limitation is due to a higher likelihood of visually descriptive adjectives appearing in the alt text of the pretraining dataset used by CLIP.

5. Conclusion

In this work, we introduce a fully self-supervised approach for learning to separate audio based on a natural language query, or in combination with a video. In the absence of object labels, we propose to extract latent captions to provide pseudo-language supervision. Additionally, we introduce the novel tri-modal and audio-language alignment objectives that use the latent captions to improve the alignment of the audio, natural language and video modalities. By reducing the need for object labels during training, our work opens up the possibility of large-scale pretraining on unlabeled videos with diverse visual concepts.

Acknowledgements: This material is based upon work supported, in part, by DARPA under agreement number HR00112020054 and a gift from Adobe Research.

References

[1] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021. 2, 3

[2] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 2

[3] Francisco Jesus Canadas-Quesada, Pedro Vera-Candeas, Nicolas Ruiz-Reyes, Julio Carabias-Orti, and Pablo Cabanas-Molero. Percussive/harmonic sound separation by non-negative matrix factorization with smoothness/sparseness constraints. *EURASIP Journal on Audio, Speech, and Music Processing*, 2014(1):1–17, 2014. 2

[4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2

[5] Moitreya Chatterjee, Jonathan Le Roux, Narendra Ahuja, and Anoop Cherian. Visual scene graphs for audio source separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1204–1213, 2021. 1, 2, 6, 7, 12

[6] Changan Chen, Unnat Jain, Carl Schissler, Sebastia Vicenc Amengual Gari, Ziad Al-Halah, Vamsi Krishna Ithapu, Philip Robinson, and Kristen Grauman. Audio-visual embodied navigation. *environment*, 97:103, 2019. 2

[7] Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. " this is my unicorn, fluffy": Personalizing frozen vision-language representations. *arXiv preprint arXiv:2204.01694*, 2022. 3

[8] Derry Fitzgerald, Matt Cranitch, and Eugene Coyle. Shifted non-negative matrix factorisation for sound source separation. In *IEEE/SP 13th Workshop on Statistical Signal Processing, 2005*, pages 1132–1137. IEEE, 2005. 2

[9] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022. 3

[11] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)*, 41(4):1–13, 2022. 2

[12] Ruohan Gao and Kristen Grauman. Co-separating sounds of visual objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3879–3888, 2019. 1, 2, 5, 6, 7, 12, 13

[13] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 2, 6, 13

[14] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2

[15] Tengda Han, Weidi Xie, and Andrew Zisserman. Memory-augmented dense predictive coding for video representation learning. In *European conference on computer vision*, pages 312–329. Springer, 2020. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 2

[18] Kevin Kilgour, Beat Gfeller, Qingqing Huang, Aren Jansen, Scott Wisdom, and Marco Tagliasacchi. Text-driven separation of arbitrary sounds. *arXiv preprint arXiv:2204.05738*, 2022. 2

[19] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019. 2

[20] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2, 3

[21] Juan F. Montesinos, Olga Slizovskaia, and Gloria Haro. Solos: A dataset for audio-visual music analysis. In *22st IEEE International Workshop on Multimedia Signal Processing, MMSP 2020, Tampere, Finland, September 21-24, 2020*. IEEE, 2020. 2, 5, 13

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 13

[23] AJ Piergiovanni, Anelia Angelova, and Michael S Ryoo. Evolving losses for unsupervised video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 133–142, 2020. 2

[24] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 13

[25] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *In Proceedings of the 15th International Society for Music Information Retrieval Conference, ISMIR*. Citeseer, 2014. 6

[26] Akam Rahimi, Triantafyllos Afouras, and Andrew Zisserman. Reading to listen at the cocktail party: Multi-modal speech separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10493–10502, 2022. 2

[27] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 4

[28] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, volume 2, pages 749–752. IEEE, 2001. 6

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[30] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 2

[31] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17959–17968, 2022. 2

[32] Martin Spiertz and Volker Gnann. Source-filter based clustering for monaural blind source separation. In *Proceedings of the 12th International Conference on Digital Audio Effects*, volume 4, 2009. 6, 7

[33] Reuben Tan, Bryan Plummer, Kate Saenko, Hailin Jin, and Bryan Russell. Look at what i'm doing: Self-supervised spatial grounding of narrations in instructional videos. *Advances in Neural Information Processing Systems*, 34:14476–14487, 2021. 2

[34] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *arXiv preprint arXiv:2203.12602*, 2022. 2

[35] Beiming Wang and Mark D Plumbley. Musical audio stream separation by non-negative matrix factorization. In *Proc. DMRN summer conf*, pages 23–24, 2005. 2

[36] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 2

[37] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 2

[38] Hang Zhao, Chuang Gan, Wei-Chiu Ma, and Antonio Torralba. The sound of motions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1735–1744, 2019. 2

[39] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 1, 2, 3, 5, 6, 7, 12, 13

[40] Chong Zhou, Chen Change Loy, and Bo Dai. Denseclip: Extract free dense labels from clip. *arXiv preprint arXiv:2112.01071*, 2021. 4

A. Appendix

In this supplemental, we provide the following additional material to the main submission:

A.1. Latent caption extraction details

A.2. Extraction of CLIP region representations

A.3. Mix-and-separate training strategy and $\mathcal{L}_{\text{mask}}$

A.4. Dataset details

   (a) MUSIC
   (b) SOLOS
   (c) AudioSet

A.5. Implementation details

A.6. Additional ablation experiments

   (a) weights for $\mathcal{L}_{\text{Audio-language}}$ and $\mathcal{L}_{\text{Tri-modal}}$
   (b) shared parameters for audio U-Net encoder $\mathcal{E}$

A.7. Predicted separated audio samples

A.1. Latent caption extraction

We provide an illustration of our latent caption extraction operation (Section 3.1) in Figure 8 and a more detailed description of the entire operation. As mentioned earlier, we extract a latent caption from each unlabeled video to provide pseudo-language supervision. Given a video V, we begin by encoding its center frame using the CLIP visual encoder: $f^V_{\text{center}} = g^V(V_{\text{center}})$. Symmetrically, we seek to extract a language representation that corresponds to the encoded center frame semantically, described next.

The encoding function of the CLIP language transformer encoder g^L provides a mechanism that is amenable to searching for latent captions that already exist in its learnt vocabulary, which allows us to freeze its parameters and leverage its strong visual-semantic alignment with the vision modality. Instead of using the trained token embeddings, we introduce a learnable token parameter p and pass it into the language encoder g^L. We adopt the simple objective function of maximizing the cosine similarity between the center frame representation and the output of the language encoder, which allows us to update the weights of p through gradient back-propagation. We formulate the optimization operation mathematically as:

$$p^* = \arg\max_p sim\left(f^V_{center}, g^L(p)\right) \tag{6}$$

where $sim(x, y) = x^T y/(\|x\|\|y\|)$ and $||.||$ denotes the L_2 norm operator. We compute the final latent caption of the video as $C^* = g^L(p^*)$. The latent captions are used in our proposed alignment objectives to provide pseudo-language supervision.

A.2. Extraction of spatiotemporal region representations from CLIP in Section 3.1

We begin by providing an overview of the 2D attention pooling layer in the CLIP Resnet visual encoders. By default, the CLIP visual encoder outputs a global visual representation for each input image. While we use the Resnet variants instead of the transformer-based architectures in CLIP, the former differs from the standard Resnet architecture in two ways. First, the CLIP variant contains three convolutional stems instead of one. Second, and more importantly, the CLIP Resnet variant also replaces the global average pooling (GAP) layer with a 2D self-attention operation, which contains the key, query and value projections. Next, we describe in more detail this self-attention layer and how we modify it for our task.

CLIP 2D attention pooling. We begin by extracting a set of spatial region representations from an input image $\mathcal{I}$ as: $f^I = g^V(I) \in \mathbb{R}^{HW \times D}$, where H, W and D are the downsampled height, width and channel dimensions. Recall that a self-attention operation involves the use of keys, queries, and values. The CLIP model computes an average image representation as the query vector: $\overline{f}^I = \frac{1}{HW}\sum_{j=1}^{HW} f^I_j$, where f^I_j denotes the j-th row of f^I. Then, it computes a final representation for the entire image as follows:

$$\begin{aligned} K &= \overline{f}^I W_K \in \mathbb{R}^{1\times D} \\ Q &= f^I W_Q \in \mathbb{R}^{HW\times D} \\ V &= f^I W_V \in \mathbb{R}^{HW\times D} \end{aligned} \tag{7}$$

where W_K, W_Q and W_V are the key, query and value projection matrices, respectively and W_K, W_Q and $W_V \in \mathbb{R}^{D\times D}$. Lastly, we compute the final contextualized image representation as:

$$f^I_{\text{global}} = W_L\left(V^\top \text{softmax}\left(\frac{(QK^\top)}{\sqrt{D}}\right)\right) \tag{8}$$

where W_L is the final language projection layer that maps the visual representations into the joint visual-semantic embedding space and $W_L \in \mathbb{R}^{D\times D}$.

Modified attention operation. Our Multiple Instance Learning formulation necessitates the presence of region representations in each input frame since we are predicting a spectrogram mask for each region. Additionally, we require these region representations to be well-aligned with the language modality such that a region should have a high similarity with the language query if its visual concept is semantically consistent with that of the query. Consequently, we extract a set of spatiotemporal region representations f^V_{conv} for our input video V with T frames. We encode the t-frame as: $f^V_{t,\text{conv}} = g^V(V_t) \in \mathbb{R}^{HW\times D}$. Finally, we compute the set of language-aligned spatiotemporal region representations by projecting them through the value and language projection layers as follows:

$$\begin{aligned} f^V_{\text{val}} &= W_V f^V_{\text{conv}} \\ f^V &= W_L f^V_{\text{val}} \end{aligned} \tag{9}$$

We pass this set of spatiotemporal region representations into our audio separation model $\mathcal{M}$ along with an input audio spectrogram to predict a mask.

A.3. Mix-and-separate training objective in Section 3.2

To train the audio U-Net decoder $\mathcal{D}$ to predict spectrogram masks given fused audio-visual and audio-text representation inputs, we use the self-supervised "mix-and-separate" learning objective since we do not have ground-truth audio source annotations within each training video. Specifically, we synthetically combine the audio of multiple videos and the goal is to use the visual information within each video to separate its corresponding audio waveform. This objective allows us to compute ground-truth ratio spectrogram masks for training without annotations. Next, we describe the generation process of the ground-truth ratio masks for a pair of videos which is also commonly used in prior work [12,39]; the same process is generalizable to any number of input videos. Given a pair of ground-truth audio spectrograms A^S_1 and A^S_2, we compute their ratio masks as follows:

$$M_1 = \frac{A^S_1}{A^S_1 + A^S_2} \quad \text{and} \quad M_2 = \frac{A^S_2}{A^S_1 + A^S_2} \tag{10}$$

We adopt the mask prediction loss [5,12,39] to train the audio U-Net decoder $\mathcal{D}$ for audio separation. Given the pair of predicted masks $\hat{M}_1$ and $\hat{M}_2$, we compute the mask prediction loss as:

$$\mathcal{L}_{mask} = ||\hat{M}_1 - M_1||_1 + ||\hat{M}_2 - M_2||_1 \tag{11}$$

We note that it is also possible to compute the above-mentioned L1 regression loss using the ground-truth audio spectrograms but prior work [12,39] has demonstrated it is more numerically stable to use the ratio masks for supervision.

A.4. Ablation experiments

Ablation over region MIL mask prediction vs video-level prediction. We evaluate the effectiveness of learning to perform source separation at the region level as compared to the video level in Table 8. To perform video-level spectrogram mask prediction, we adopt the same video aggregation function in Sound of Pixels [39], where the region representations are maxpooled over the channel dimension to compute a final video representation that is passed into the audio U-Net decoder $\mathcal{D}$ (Section 4). We note that our proposed alignment objectives are used in the training of both model variants. We observe that training a model to perform region-level predictions under the MIL formulation results in a significant performance gain over performing video-level predictions, which validates our hypothesis that a model trained to perform video-level predictions may not be able to identify candidate objects that emit sound.

Effect of sharing parameters in U-Net encoder $\mathcal{E}$. Prior work [12] learns a separate audio encoder for encoding the predicted audio waveforms to classify them according to discrete audio category labels. Here, we aim to determine the benefit of using shared parameters for our audio encoder component of the U-Net model $\mathcal{E}$ in Table 9. In this case, unlike prior work [12], we observe that using a shared audio U-Net encoder to encode the input audio spectrogram for source separation and the predicted spectrogram for the two new losses is integral to improving the final performance of our trained model on audio-visual separation.

Ablation over weights of $\mathcal{L}_{\text{Audio-language}}$ and $\mathcal{L}_{\text{Tri-modal}}$. We report the results of our ablation over the weights of our proposed audio-language and tri-modal consistency alignment objectives in Table 10. The results of adding the audio-language consistency loss seem to validate our initial hypothesis that using a lower weight term for this loss is beneficial. As discussed earlier in Section 3.1, this is similar to the multimodal contrastive formulation used for training joint vision-language foundation models such as CLIP and ALIGN. Thus, there is a high probability that we are treating some latent captions as false negatives for each video even though they may contain similar sounding objects. Setting a low weight helps to alleviate this negative consequence. However, we observe that the audio-language consistency loss is still very helpful for improving audio-visual source separation as well as learning a strong transitive alignment between the audio and natural language modality. The reported results also suggest that adding the tri-modal consistency loss also helps to improve performance significantly. In this case, we note that this alignment objective is formulated as a KL divergence minimization problem and does not require negative samples. Consequently, it may not be as important to use a low weight for this term as compared to the audio-language consistency objective.

Prediction	NSDR	SIR	SAR
Video-level	6.72	11.47	10.58
Region-level	**8.58**	**14.16**	**12.35**

Table 8. **Comparison between video-level and region-level audio predictions with our trained model on the SOLOS dataset.**

Shared audio encoder params	SOLOS			MUSIC			Audioset		
	NSDR ↑	SIR ↑	SAR ↑	NSDR ↑	SIR ↑	SAR ↑	NSDR ↑	SIR ↑	SAR ↑
No	7.52	12.68	10.22	7.39	13.25	9.81	3.27	6.48	11.51
Yes	**8.58**	**14.16**	**12.35**	**8.08**	**13.97**	**11.33**	**11.33**	**7.62**	**13.20**

Table 9. **Ablation over using shared parameters for audio U-Net encoder.**

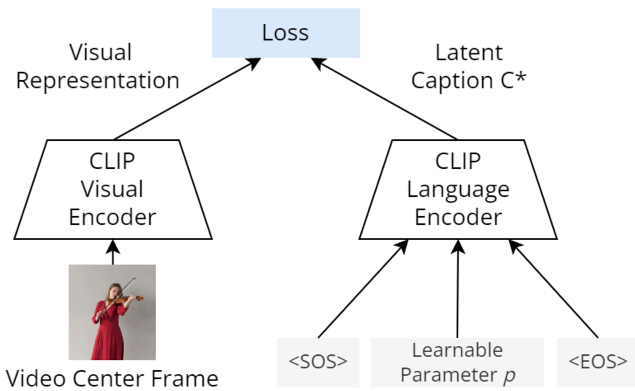


Figure 8. Extraction of latent captions for pseudo-supervision.

$\mathcal{L}_{\text{Audio-language}}$ weight	$\mathcal{L}_{\text{Trimodal}}$ weight	NSDR	SIR	SAR
0.0	0.0	5.47	10.55	10.95
1e-1	0.0	6.09	11.77	10.77
1e-2	0.0	8.08	13.74	12.18
1e-3	0.0	7.45	13.40	11.11
1.0	-	1.24	4.97	11.27
-	1e-1	8.02	13.82	11.76
0.0	1e-2	7.92	13.49	11.65
0.0	1e-3	8.10	13.84	11.79
0.0	1.0	6.81	12.61	11.00
1e-3	1e-2	8.58	14.16	12.35

Table 10. **Ablation results over the weights of the audio-language and tri-modal consistency alignment objectives on the SOLOs dataset.**

A.5. Datasets

We train and evaluate our proposed AV-SeT approach as well as other baselines on the widely-used SOLOS, MUSIC and AudioSet datasets which we describe below.

MUSIC [39]. The MUSIC dataset consists of videos that are downloaded from YouTube using queries about various musical instruments. It contains approximately 536 and 149 solo and duet videos, respectively. The entire set is comprised of videos containing 11 instrument categories: accordion, acoustic guitar, cello, clarinet, erhu, flute, saxophone, trumpet, tuba, violin and xylophone. Since the original splits of the dataset are not released, we adopt the same splits as [12], where the first and second videos in each instrument category are used as validation / test data and the rest are used for training.

SOLOS [21]. Similar to the MUSIC dataset, the SOLOS dataset contains 755 videos of musical videos that span 13 instrument categories. These videos are obtained from YouTube where the authors use queries of instruments as well as the 'solo' or 'auditions' tag. Unlike the MUSIC dataset, the SOLOS dataset does not contain duet videos.

AudioSet-Unlabeled [13]. AudioSet is a dataset that contains over two million 10 second video clips spanning 632 audio event classes that are sourced from YouTube. Compared to the MUSIC and SOLOS datasets, the audio clips in AudioSet are generally much noisier due to the presence of background sounds. Following prior work [12], we filter the video clips according to 15 musical instrument categories and select those from the 'unbalanced' split for training and the 'balanced' split for validation and testing.

A.6. Implementation details

We implement our proposed approach using the Pytorch deep learning library [22]. Consistent with prior work [12, 39], we downsample the audio clips to 11 kHz and use a Hann window size of 1022 samples[1] and a hop length of 256 samples in the STFT operation. This step results in an audio spectrogram of dimensions 512 x 256, which is re-sampled on a log-frequency scale to compute a final spectrogram of dimensions 256 x 256. We use the CLIP Resnet50 model [24] and its language encoder to extract a latent caption for each video as well as encode visual and language representations for audio separation. We set the dimension of the audio U-Net bottleneck features D to be the same as that of CLIP embedding space, which is 1024. We freeze the CLIP encoders during training and train the audio U-Net from scratch using a base learning rate of 4e-3. We train all models for 100 epochs with the SGD optimizer as well as using a linear warmup of 1000 steps and anneal the learning rate using a cosine decay schedule. We train our full model using 4 Quadro 6000 GPUs for approximately 8 days.

[1]While it is common to use powers of 2 as FFT size, we use 1022 as opposed to 1024 to be consistent with previous literature.

A.7. Demo video with predicted audio component generations

We provide sample predicted audio generated by our trained model on the SOLOS dataset in the attached video. The video contains 10 evaluation samples on the tasks of audio-language source separation and audio grounding in the input videos. For the first task, our objective is to separate an audio input based on a natural language query and the goal of the second task is to localize the predicted separated audio in its corresponding video. We compare the predictions of our full AV-SeT model to a baseline variant that is trained only with the mask prediction loss to determine the contributions of our proposed audio-language and tri-modal consistency alignment objectives. For each evaluation sample, we provide the following in order:

1. Input video with mixed audio input (composed of two different instruments)
2. Separated audio predicted by the baseline model
3. Separated audio predicted by the full AV-SeT model
4. Ground-truth audio
5. Attention heatmap between the separated audio in (3) and the center frame

We generally observe that our full AV-SeT model, that is trained with our proposed alignment objectives, is better able to separate the audio inputs based on natural language queries than the baseline model. In particular, we hear that the separated audio predicted by the baseline model is often very similar to the mixed audio input.