Language-Guided Audio-Visual Source Separation via Trimodal Consistency

Reuben Tan\textsuperscript{1} Arijit Ray\textsuperscript{1} Andrea Burns\textsuperscript{1} Bryan A. Plummer\textsuperscript{1} Justin Salamon\textsuperscript{2} Oriol Nieto\textsuperscript{2} Bryan Russell\textsuperscript{2} Kate Saenko\textsuperscript{1,3}
\textsuperscript{1}Boston University, \textsuperscript{2}Adobe Research, \textsuperscript{3}MIT-IBM Watson AI Lab, IBM Research
\{rxtan, aburns4, array, bplum, saenko\}@bu.edu, \{salamon, onieto, 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. Our project page including publicly available code can be found at https://cs-people.bu.edu/rxtan/projects/VAST.

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 object to its visual features and the corresponding components of the audio waveform without access to annotations during training. Existing audio-visual methods \cite{5, 14, 41} 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 \cite{5, 14}, these approaches require strong supervision including object label supervision.
and bounding box annotations (see Fig. 1 top). Overcoming these challenges would enable important downstream applications including holistic video understanding [33], embodied AI [6], and bidirectional audio-to-video retrieval [38].

To address these challenges, we make the following contributions. First, we propose Video-Audio Separation through Text (VAST), a self-supervised approach that leverages large vision-language “foundation” models [20, 26] 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).

Our second contribution introduces 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 can 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 VAST approach through extensive evaluations on the audio source separation task on the SOLOS [24], MUSIC [41], and AudioSet [15] 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 VAST 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. Conventional audio signal processing techniques rely on strong assumptions such as the number of sources in the audio waveforms to compute non-negative matrix factorization [3, 10, 37] of audio spectrograms. Deep learning methods commonly adopt the self-supervised ‘mix-and-separate’ strategy where multiple audio sources are combined into a synthetic mixture and then predict a spectrogram mask to retrieve queried audio components [5, 14, 41].

Multimodal source separation. Recent work in audio-text [21, 22] and audio-visual [5, 7, 14, 40, 41] 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, 14, 41] or free-form natural language queries [21, 22]. 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 VAST approach is also similar to the Voiceformer [28] 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. Finally, the task is also similar in nature to that of sound localization in videos [9, 19].

Multimodal representation learning. One key challenge is grounding information from the text and/or audio modalities in the video frames [2, 5, 14, 35]. Since annotating video datasets is a costly process, prior work has investigated self-supervised pretraining methods on unlabeled videos [17, 36]. This often involves utilizing complementary information from the audio, text and visual modalities [39] to learn robust representations that can generalize to downstream tasks such as action recognition [1, 25, 32] and text-to-video retrieval [2, 23]. These approaches often use large-scale pretraining datasets such as Kinetics-400 / 600 [4] and HowTo100M [23] before finetuning linear probes for target tasks and datasets. Recent work has also focused on prompting multimodal foundation models such as CLIP [26] and ALIGN [20] to adapt them to new tasks ranging from open-vocabulary object detection [16] to domain adaptation [13] without modifying their parameters.

3. Video-Audio Separation through Text

Given a set of unlabeled videos and their audio tracks, we strive to learn a self-supervised model that separates an audio source given a natural language query and, optionally, 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 (VAST) approach that learns an effective representation for audio-visual source separation, as well as a strong alignment between the audio and language modalities for audio-text source separation (see Figure 2 for an overview). To learn this representation, we exploit the strong joint visual-semantic alignment embodied in pre-trained 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 the visual and language modalities are already aligned, learning a strong correspondence between audio and vision should also provide an alignment between audio and language through videos via transitivity.

Despite the intuitiveness of this idea, our experiments will show that using the visual representations of existing audio-visual foundation models does not help them learn the above-mentioned transitive alignment. To address this limitation, we introduce our novel tri-modal and audio-language consistency alignment objectives (Figure 2). 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. We leverage vision-and-language foundation models to infer latent captions that correspond to the audio components without text annotations, thus providing pseudo-language supervision (Section 3.1). In addition, we adopt the “mix-and-separate” strategy [41] to train our audio separation model, where the audio from multiple videos are mixed and then separated using only the video’s visual information. For an input audio waveform $A$, we apply a Short-Time Fourier Transform (STFT) to obtain its magnitude spectrogram $A^S$ and phase $A^{\text{phase}}$. Our learning objective only makes use of $A^S$, which encodes how the magnitude of the audio frequencies changes over time. We reconstruct the predicted waveform by applying the inverse STFT on the predicted spectrogram using the original phase $A^{\text{phase}}$. We describe our VAST audio separation model $\mathcal{M}$ in Section 3.2.

### 3.1. Alignment between audio, language and video

A standard 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, 23]. However, learning this alignment 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 [8,11] 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.

#### Latent captions.

We perform “textual inversion” [8,11,12] 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 $V_{\text{center}}$ as: $f^V_{\text{center}} = g^V(V_{\text{center}})$, 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 similarity between $f^V_{\text{center}}$ and $g^L(p)$:

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

(1)

where $\text{sim}(x, y) = x^T y / (\|x\| \|y\|)$ and $\|\cdot\|$ 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 representation of a description of the frame and may replace ground-truth annotations effectively.
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 \( A^S \in \mathbb{R}^{F \times N} \) with \( F \) frequency bins and \( N \) STFT time frames, our model predicts a separated audio spectrogram \( \hat{A}^S \in \mathbb{R}^{F \times N} \). We extract a latent representation for the predicted spectrogram:

\[
\hat{f}^A = \mathcal{M}_\theta(\hat{A}^S) \in \mathbb{R}^D
\]

where \( \mathcal{M}_\theta \) 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 3a). 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 \( \mathcal{X} \):

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

where \( \tau \) 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. Intuitively, this weighting helps to alleviate the effect of false negatives and prevent it from introducing a lot of noise during training.

**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 issue, 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 [29, 42] has demonstrated that the CLIP model can be engineered to extract a segmentation map based on natural language queries. Inspired by this finding, we use the latent captions to provide pseudo-target attention maps.

Our 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. Let \( P_{\mathcal{M}}(A, b) = \sigma(A \cdot b) \) to be the attention operation where \( \sigma(z)_i = z_i / \sum_j z_j \) is the softmax function. For a given video \( V \), we extract a set of spatiotemporal region representations by encoding each frame separately. Specifically, we encode the \( t \)-th frame as:

\[
f_{t}^{V} = g_{k}^{Y}(V_t),
\]

where \( f_{t}^{V} \in \mathbb{R}^{H \times W \times D} \), where \( H \) and \( W \) are the downsampled height and width of the input video. Then, we compute a probability distribution of similarity scores over the set of

![Tri-modal Consistency](image-url)
Audio-Language Separation Model \( \mathcal{M}_{AL} \)

Input Audio Spectrogram

Predicted Spectrogram

Concatenated audio-language features

Predicted mask

Language feature

Audio-Language Separation Model \( \mathcal{M}_{AL} \)

Figure 4. Spectrogram mask prediction. During inference, our model is able to separate audio sources using video or language queries despite training without ground-truth text annotations. We note that we only use unlabeled videos during training. We illustrate the audio-visual separation mode \( \mathcal{M}_{AV} \) in the supplemental.

Spatial region features with respect to its latent caption \( C^* \) for the \( t \)-th frame: \( P_{VC}^{t} = P_{\text{Att}}(f^{t}_V, C^*) \in \mathbb{R}^{HW \times 1} \). 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: \( P_{VA}^{t} = P_{\text{Att}}(f^{t}_V, f^{A}) \in \mathbb{R}^{HW \times 1} \). We encourage the predicted audio representation to be similar to that of the latent caption by minimizing the Kullback-Leibler (KL) divergence between both attention distributions over the set of all time steps in a video \( T \):

\[
E_{\text{trimodal}} = \sum_{t=1}^{T} \sum_{x=1}^{HW} P_{VC}^{t}(x) \log \frac{P_{VC}^{t}(x)}{P_{VA}^{t}(x)},
\]

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

3.2. Audio Separation Model

We introduce two audio separation models \( \mathcal{M}_{AL} \) and \( \mathcal{M}_{AV} \) which share the same audio model parameters for separating audio sources. \( \mathcal{M}_{AL} \) (Figure 4) infers a spectrogram mask based on a language query while \( \mathcal{M}_{AV} \) (shown in supplemental) conditions its prediction on the input video. We adopt the U-Net [31] model as the audio component in \( \mathcal{M}_{AL} \) and \( \mathcal{M}_{AV} \), which has been demonstrated by prior work [14, 41] to be effective at audio source separation, and the CLIP visual and language encoders to exploit its learnt visual-semantic alignment. The U-Net model comprises an encoder \( \mathcal{E} \) and a decoder \( \mathcal{D} \). Given an input audio spectrogram \( A^{S} \), the encoder \( \mathcal{E} \) applies 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 height and width of the downsampled spectrogram, respectively.

Given a natural language query \( L \in \mathbb{R}^{N_{w}} \) where \( N_{w} \) is the number of words, we extract its representation \( f^{L} \in \mathbb{R}^{1 \times D} \) using the text encoder: \( f^{L} = g^{L}(L) \). We stress that language queries with ground-truth object labels are only used during inference. Then, we tile the language representation by the factor \( H^{A}W^{A} \) and concatenate with the audio bottleneck representations along the channel dimension: \( f^{AL} = \text{concat}(f^{A}, \text{tile}(f^{L})) \), where \( f^{AL} \) 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: \( M = \mathcal{D}(f^{AL}) \in \mathbb{R}^{F \times N} \). To predict the separated audio source, each element of the mask is multiplied with the corresponding location in the input spectrogram: \( \hat{A}^{S} = M \odot \check{A}^{S} \), where \( \odot \) denotes the Hadamard product.

Video and Multiple Instance Learning for mask prediction. Since language annotations are not used during training, we train only with unlabeled videos. During training, as well as inference on audio-visual separation, we use the CLIP visual encoder to extract a set of language-grounded spatiotemporal region representations \( f^{V} \in \mathbb{R}^{T \times H \times W \times D} \). Similarly, we can perform the same mask inference operation using videos by using the representation of each spatiotemporal region representation interchangeably in place of the language representation. This interchangeability is due to the fact that the CLIP representations project the video to the joint vision-language embedding space.

Prior work [14] 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 this section, we predict a spectrogram mask \( \hat{M}_{j} \) for the \( j \)-th region. Finally, we aggregate them over all regions to obtain a spectrogram mask for the entire video:

\[
\hat{M} = \frac{1}{T} \sum_{j=1}^{THW} \hat{M}_{j}.
\]

Intuitively, this formulation encourages the model to learn to identify the salient regions with sounding objects for audio separation. We obtain the predicted mask for each frame by computing the sum of the predicted masks over all spatial regions and average the predicted masks over the temporal dimension since we reason that a sounding object should have similar prediction masks across frames. We adopt the self-supervised ‘mix-and-separate’ learning objective [14, 41] to train our model 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 setting allows us to generate pseudo-target masks for training supervision. We compute \( L_{\text{mask}} \) as the L1 loss between the pseudo-target and predicted ratio masks [41]. Please refer to the supplemental for details of our spatiotemporal region representation, audio-visual separation model and \( L_{\text{mask}} \).

The final objective function is computed as:

\[
L = \lambda_{1}L_{\text{mask}} + \lambda_{2}E_{\text{trimodal}} + \lambda_{3}E_{\text{audio-language}}
\]
where $\lambda_1$, $\lambda_2$ and $\lambda_3$ denote the weights for the mask, trimodal and audio-language consistency losses, respectively.

### 4. Experiments

We evaluate the effectiveness of our proposed VAST approach on audio-text and audio-visual source separation in videos. To perform fair comparison to prior work, we conduct experiments on the SOLOS [24], MUSIC [41] and AudioSet [15] datasets. We refer readers to the supplemental for additional implementation and dataset details. Given a language query containing 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 is often comprised of a mixture of sounds emitted by various objects. 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) [27, 30]. NSDR computes the difference between the SDR of the predicted and mixed audio with respect to the ground-truth waveforms, which measures the sound quality of the separated audios. SIR measures the amount of contamination from other audio sources in the predicted audio. Finally, SAR measures the amount of artifacts in the predicted audio.

#### 4.1. Quantitative results

<table>
<thead>
<tr>
<th>Method</th>
<th>Object detectors</th>
<th># Region proposals</th>
<th>Solos NSDR ↑</th>
<th>Solos SIR ↑</th>
<th>Solos SAR ↑</th>
<th>Duets NSDR ↑</th>
<th>Duets SIR ↑</th>
<th>Duets SAR ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-Separation [14]</td>
<td>Yes</td>
<td>2</td>
<td>7.38</td>
<td>13.70</td>
<td>10.82</td>
<td>7.42</td>
<td>13.81</td>
<td>10.60</td>
</tr>
<tr>
<td>Sound-of-Pixels [41]</td>
<td>No</td>
<td>None</td>
<td>7.30</td>
<td>11.90</td>
<td>11.90</td>
<td>6.05</td>
<td>9.81</td>
<td>10.61</td>
</tr>
<tr>
<td>AV-Mix-and-Separate</td>
<td>No</td>
<td>None</td>
<td>3.16</td>
<td>6.74</td>
<td>8.89</td>
<td>3.23</td>
<td>7.01</td>
<td>9.14</td>
</tr>
<tr>
<td>NMF-MFCC [34]</td>
<td>No</td>
<td>None</td>
<td>0.92</td>
<td>5.68</td>
<td>6.84</td>
<td>0.92</td>
<td>5.68</td>
<td>6.84</td>
</tr>
<tr>
<td>VAST (Ours)</td>
<td>No</td>
<td>None</td>
<td>7.98</td>
<td>13.92</td>
<td>12.35</td>
<td>8.08</td>
<td>13.97</td>
<td>11.33</td>
</tr>
</tbody>
</table>

Table 1. Audio-visual source separation on MUSIC. We report results on videos that contain one instrument (solos) and two instruments (duets). Despite not training with object bounding boxes, VAST performs competitively with state-of-the-art detection-based approaches.

---

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ denote the weights for the mask, trimodal and audio-language consistency losses, respectively.

### 4. Experiments

We evaluate the effectiveness of our proposed VAST approach on audio-text and audio-visual source separation in videos. To perform fair comparison to prior work, we conduct experiments on the SOLOS [24], MUSIC [41] and AudioSet [15] datasets. We refer readers to the supplemental for additional implementation and dataset details. Given a language query containing 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 is often comprised of a mixture of sounds emitted by various objects. 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) [27, 30]. NSDR computes the difference between the SDR of the predicted and mixed audio with respect to the ground-truth waveforms, which measures the sound quality of the separated audios. SIR measures the amount of contamination from other audio sources in the predicted audio. Finally, SAR measures the amount of artifacts in the predicted audio.

#### 4.1. Quantitative results

<table>
<thead>
<tr>
<th>Method</th>
<th>Object detectors</th>
<th># Region proposals</th>
<th>Solos NSDR ↑</th>
<th>Solos SIR ↑</th>
<th>Solos SAR ↑</th>
<th>Duets NSDR ↑</th>
<th>Duets SIR ↑</th>
<th>Duets SAR ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Co-Separation [14]</td>
<td>Yes</td>
<td>2</td>
<td>7.38</td>
<td>13.70</td>
<td>10.82</td>
<td>7.42</td>
<td>13.81</td>
<td>10.60</td>
</tr>
<tr>
<td>Sound-of-Pixels [41]</td>
<td>No</td>
<td>None</td>
<td>7.30</td>
<td>11.90</td>
<td>11.90</td>
<td>6.05</td>
<td>9.81</td>
<td>10.61</td>
</tr>
<tr>
<td>AV-Mix-and-Separate</td>
<td>No</td>
<td>None</td>
<td>3.16</td>
<td>6.74</td>
<td>8.89</td>
<td>3.23</td>
<td>7.01</td>
<td>9.14</td>
</tr>
<tr>
<td>NMF-MFCC [34]</td>
<td>No</td>
<td>None</td>
<td>0.92</td>
<td>5.68</td>
<td>6.84</td>
<td>0.92</td>
<td>5.68</td>
<td>6.84</td>
</tr>
<tr>
<td>VAST (Ours)</td>
<td>No</td>
<td>None</td>
<td>7.98</td>
<td>13.92</td>
<td>12.35</td>
<td>8.08</td>
<td>13.97</td>
<td>11.33</td>
</tr>
</tbody>
</table>

Table 1. Audio-visual source separation on MUSIC. We report results on videos that contain one instrument (solos) and two instruments (duets). Despite not training with object bounding boxes, VAST performs competitively with state-of-the-art detection-based approaches.
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 are unable to generalize to the cleaner audio signals. Similar to the reported performance on the MUSIC dataset, we also observe that VAST is comparable to the strongly supervised AVSGS approach on the SOLOS dataset. One possible 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.

### 4.2. Audio-language source separation

Finally, we report the results of our evaluation on the AudioSet dataset in Table 3. AudioSet has been documented by prior work 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.

#### 4.1.1 Model Analysis

**Ablation of \( \mathcal{L}_{\text{Trimodal}} \) and \( \mathcal{L}_{\text{Audio-language}} \)**. Table 5 provides an ablation over our audio-language and tri-modal consistency losses. Despite being trained on a larger and more diverse set of visual concepts, we find the CLIP visual representations alone do not encourage the audio encoder to learn a strong alignment between the audio and vision modalities (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 audio-language 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.

### 4.1.2 SOLOS

**Using CLIP visual features in SOP model on the SOLOS dataset**. Using CLIP visual features naively in existing methods results in a drop in performance.

**Using CLIP visual features in SOP model on the SOLOS dataset**. Similar to the results in Table 1, we see that VAST significantly closes the gap with object detection-based approaches.
only train on unlabeled videos without text supervision. To begin, we compare our VAST model to a variant that is trained without our alignment objectives. Note that we do not compare to other audio-visual approaches since there is no straightforward way to adapt them for language queries. We compare our full VAST model to a variant that is only trained with the mask prediction loss.

As evidenced by the significant performance gap between both variants, our alignment losses are integral to learning a strong transitive alignment between the audio and language modalities. This finding suggests that just learning an alignment between the audio and visual modalities does not result in a transitive relationship between the audio and natural language modalities.

Additionally, for effective sound source separation given either video or text inputs, we hypothesize that it is critical to learn an audio encoder $E$ that is well-aligned with CLIP’s vision-language embedding. Note that the same shared audio encoder $E$ encodes both mixed and predicted audio sources. Thus, it can also be used as a general audio encoder and the decoder $D$ can condition its separation process on either language or video queries. Lastly, we observe that our alignment objectives are still insufficient for closing the performance gap between audio-text and audio-visual separation (Tables 1, 2 and 3) completely. This observation indicates that future work is needed for further improving the latent alignment between all three modalities.

### 4.3. Qualitative results

**Predicted audio-to-video attention visualizations.** To evaluate the semantic consistency between the predicted audio sources and the visual cues that guide the separation process, we plot the attention map between the encoded predicted spectrograms and their corresponding video frames (Section 3.1) in Figure 5. We observe that VAST 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 6. Significantly, despite not relying on text annotations for training supervision, our model can separate audio components based on free-form language queries. This demonstrates the effectiveness of our proposed idea to substitute ground-truth text annotations with latent captions during training. Additionally, we also observe that providing more context in the language query (e.g., “violin” versus “person playing a violin”) can lead to improved separations.

### 5. Conclusion

In summary, we introduce a self-supervised approach for learning to separate audio based on a language query, or a video. In the absence of object labels, we propose to extract latent captions to provide pseudo-language supervision. Furthermore, we introduce the novel tri-modal and audio-language alignment objectives that use the latent captions to improve the alignment of the audio, language and video modalities. By reducing the training need for object labels, our work opens up the possibility of large-scale pre-training 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.

<table>
<thead>
<tr>
<th>Modality</th>
<th>Alignment objectives</th>
<th>SOLOS NSDR ↑ SIR ↑ SAR ↑</th>
<th>MUSIC NSDR ↑ SIR ↑ SAR ↑</th>
<th>Audioset NSDR ↑ SIR ↑ SAR ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Language</td>
<td>No</td>
<td>-3.05 2.79 3.77</td>
<td>-3.67 2.51 3.41</td>
<td>-5.02 1.98 14.94</td>
</tr>
<tr>
<td>Language</td>
<td>Yes</td>
<td>6.92 12.07 10.41</td>
<td>6.45 11.18 10.77</td>
<td>2.36 4.71 10.28</td>
</tr>
</tbody>
</table>

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 VAST model to a variant that is only trained with the mask prediction loss.

Figure 5. Predicted audio attention in videos. Our consistency objectives help the model to localize the relevant regions more accurately without prior knowledge of sounding object locations.

Figure 6. Predicted audio spectrograms with language queries. Interestingly, we see that including visually grounded adjectives in the queries affects the quality of the separated audio sources.
References


[20] 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


[26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,


