Audio-Visual Instance Segmentation

Supplementary Material

A. Overview

B. Evaluation Metrics

In the era of competitive benchmarks, much research is evaluated on its ability to improve the scores. If benchmarks are using metrics to evaluate these scores which are skewed towards only certain aspects of a task, this will also steer research and models to focus more on these aspects.

B.1. mAP: mean Average Precision

mAP (mean Average Precision) is a standard evaluation metric in image instance segmentation. It is the area under the precision-recall curve across multiple intersection-over-union (IoU) thresholds. The mAP metric has been extended to video instance segmentation, as proposed in [55], where IoU computation differs from image instance segmentation because each instance contains a sequence of masks:

$$\text{IoU}(\mathbf{G}, \mathbf{P}) = \frac{\sum_{t=1}^{T} \left| \mathbf{m_t^G} \cap \mathbf{m_t^P} \right|}{\sum_{t=1}^{T} \left| \mathbf{m_t^G} \cup \mathbf{m_t^P} \right|} \tag{2}$$

The proposed IoU computes the spatio-temporal consistency of ground-truth and predicted segmentation results. If the algorithm detects object masks but fails to track the objects across frames, the IoU score will be reduced.

However, mAP is not perfectly suited to our AVIS task, because it can be increased by producing many different predictions with low confidence scores and does not decrease even if non-sounding objects are predicted. Moreover, the threshold required for an instance to be considered a positive match is set high, resulting in lots of improvements in detection, association, and localization being overlooked by the evaluation metric. In addition, mAP mixes association, detection and localisation in a manner that does not allow for differentiation among error types.

B.2. HOTA: Higher Order Tracking Accuracy

HOTA (Higher Order Tracking Accuracy) [43] performs a bijective (one-to-one) matching at a detection level while scoring association globally over trajectories, which is designed for multi-object tracking task. This makes HOTA a balanced metric for measuring both detection and association. When applied to the AVIS task, it can penalize those models that predict non-sounding objects.

A true positive (TP) refers to a matched pair of a ground-truth track set (gtDet) and a predicted detection set (prDet), for which the localisation similarity is greater than or equal to the threshold α. A false negative (FN) is a gtDet that is not matched to any prDet. A false positive (FP) is a prDet that is not matched to any gtDet. The matching between gtDets and prDets is bijective within each frame. For a given TP, denoted as c, the set of TPAs is the set of True Positive Associates (TPs) which have both the same ground-truth id set (gtID) and the same predicted id set (prID) as c. For a given TP, c, the set of False Negative Associates (FNAs) refers to the set of gtDets with the same gtID as c, but that were either assigned a different prID as c, or no prID if they were missed. For a given TP, c, the set of False Positive Associates (FPAs) denotes the set of prDets with the same prID as c, but that were either assigned a different gtID as c, or no gtID if they did not actually correspond to an object. Having defined the concepts for measuring successes and errors in detection (TPs, FPs, FNs) and association (TPAs, FPAs, FNAs), the HOTA score can be defined as:

$$\begin{aligned} \text{HOTA} &= \sqrt{\frac{\sum_{c \in \{\text{TP}\}} \mathcal{A}(c)}{|\text{TP}| + |\text{FN}| + |\text{FP}|}} \\ \mathcal{A}(c) &= \frac{|\text{TPA}(c)|}{|\text{TPA}(c)| + |\text{FNA}(c)| + |\text{FPA}(c)|} \end{aligned} \tag{3}$$

The HOTA can decompose into a separate detection accuracy score (DetA) and an association accuracy score (AssA) as follows:

$$\begin{aligned} \text{HOTA} &= \sqrt{\text{DetA} \cdot \text{AssA}} \\ \text{DetA} &= \frac{|\text{TP}|}{|\text{TP}| + |\text{FN}| + |\text{FP}|} \\ \text{AssA} &= \frac{1}{|\text{TP}|} \sum_{c \in \{\text{TP}\}} \mathcal{A}(c) \end{aligned} \tag{4}$$

B.3. FSLA: Frame-level Sound Location Accuracy

Besides considering the above object-based metrics, we propose a novel measure, namely frame-level sound localization accuracy (FSLA), tailored to measure the proportion of frames that are correctly predicted by the model out of the total number of frames. Specifically, we first use the Hungarian algorithm [32] to determine a one-to-one matching between ground-truth and predicted detections. For each frame, it can be treated as correct frame if it satisfies the

Algorithm 1: The FSLA Evaluation Metric

function FSLA $(\mathbf{M^P}, \mathbf{M^G}, \mathbf{C^P}, \mathbf{C^G}, \mathbf{ID^P}, \mathbf{ID^G})$;

Input : A predicted mask set $\mathbf{M^P} = \{\mathbf{m^P_{i,l}}\}^{x,L}_{i=1,l=1}$.
A labeled mask set $\mathbf{M^G} = \{\mathbf{m^G_{j,l}}\}^{y,L}_{j=1,l=1}$.
A predicted class set $\mathbf{C^P} = \{\mathbf{c^P_i}\}^{x}_{i=1}$.
A labeled class set $\mathbf{C^G} = \{\mathbf{c^G_j}\}^{y}_{j=1}$.
A predicted id set $\mathbf{ID^P} = \{\mathbf{id^P_i}\}^{x}_{i=1}$.
A labeled id set $\mathbf{ID^G} = \{\mathbf{id^G_j}\}^{y}_{j=1}$.
The video frames $\mathbf{F}$. The video length L.
N_{fna}, N_{fsa}, and N_{fma} are the number of silent, single- and multi-sound-source frames.

Output: FSLA, FSLAn, FSLAs, FSLAm

$N_{fnt}, N_{fst}, N_{fmt} \leftarrow 0$
$S(x, y) = \text{HungarianMatch}(\mathbf{M^P}, \mathbf{M^G}, \mathbf{ID^P}, \mathbf{ID^G})$
for $\alpha \leftarrow 0.05$ *to* 0.95 *step* 0.05 **do**
  for $l \leftarrow 1$ *to* L *step* 1 **do**
    if F_l *is a silent frame* **then**
      $N_{fnt} \leftarrow N_{fnt} + 1$
    end
    if F_l *is a single-sound-source frame* **then**
      if $\mathbf{c^P_{i,l}} == \mathbf{c^G_{j,l}}$ **then**
        if $S(i,j)$ *and* $IoU(\mathbf{m^P_{i,l}}, \mathbf{m^G_{j,l}}) > \alpha$ **then**
          $N_{fst} \leftarrow N_{fst} + 1$
        end
      end
    end
    if F_m *is a multi-sound-source frame* **then**
      if $\mathbf{c^P_{i,l}} == \mathbf{c^G_{j,l}}$ **then**
        if $S(i,j)$ *and* $IoU(\mathbf{m^P_{i,l}}, \mathbf{m^G_{j,l}}) > \alpha$ **then**
          $N_{fmt} \leftarrow N_{fmt} + 1$
        end
      end
    end
  end
  FSLAn(α) $\leftarrow N_{fnt}/N_{fna}$, FSLAs($\alpha$) $\leftarrow N_{fst}/N_{fsa}$
  FSLAm(α) $\leftarrow N_{fmt}/N_{fma}$
  FSLA(α) $\leftarrow (N_{fnt} + N_{fst} + N_{fmt})/L$
end
FSLA $\leftarrow \overline{\text{FSLA}(\alpha)}$
FSLAn, FSLAs, FSLAm $\leftarrow \overline{\text{FSLAn}(\alpha)}$, $\overline{\text{FSLAs}(\alpha)}$, $\overline{\text{FSLAm}(\alpha)}$

following conditions: 1) The number of sounding objects is correct; 2) The category of the sounding objects is correct; 3) The IoU (Intersection over Union) between the ground truth and the predicted sounding objects is greater than threshold α. The final score is computed by averaging over all classes before averaging different α thresholds (0.05 to 0.95 in 0.05 intervals). The pseudo code of the FSLA metric is shown in Algorithm 1. Compared to other metrics, our FSLA allows for easier localization of incorrect frames and offers a more intuitive explanation of the model's performance across different time periods. Additionally, it can be decomposed into a set of sub-metrics (FSLAn, FSLAs and FSLAm) which can be used for model evaluation in scenarios with no sound source, a single sound source, and multiple sound sources. This results in FSLA being able to guide how models can be improved, or understand where they are likely to fail when used.

C. Implementation Details

The audio and video frames are sampled at rates of 16 kHz and 1 FPS, respectively. For the image encoder, we attempt two different backbones, ResNet-50/101 [24] and Swin-L [42]. For the audio encoder, we adopt VGGish [18] pre-trained on AudioSet, with its parameters frozen during the training phase. Unless specified, the window size W is set to 6, and both the number of frame queries and video queries are set to 100. Our model is implemented on top of the detectron2[1] and trained on the proposed AVISeg dataset for 48,000 iterations with a batch size of 1. We use the AdamW optimizer and the step learning rate schedule. The initial learning rate is set to 1e-4 and reduced by a factor of 0.1 at 32,000 iterations. By default, the shorter side of frames are resized to 360 and 448 pixels during inference. The mask predictions are obtained without any post-processing, such as NMS. We keep predictions with a confidence threshold greater than 0.3. The experiments are conducted on 2 NVIDIA Quadro 6000 GPUs.

D. More Ablation Studies

Impact of similarity loss and hyper-parameter setup. As shown in Table 7, removing similarity loss yields a significant decrease across all metrics. This is because the model struggles to learn correct associations between object tokens and video queries, leading to feature misalignment, identity switches and tracking failures, especially for different instances of the same category. Additionally, we test several hyper-parameters and set $\lambda_{\text{sim}} = 0.5$ as default, which achieves the best performance.

Table 7. Impact of similarity loss and hyper-parameter setup.

similarity loss	$\lambda_{\text{sim}} = 0.1$	$\lambda_{\text{sim}} = 0.5$	$\lambda_{\text{sim}} = 1.0$	FSLA	HOTA	mAP
✘				32.71	52.45	35.77
✔	✔			38.97	59.92	38.22
✔		✔		42.78	61.73	40.57
✔			✔	42.08	61.63	40.49

E. Details of Multi-modal Large Models

E.1. Sam4AVS

Model. As shown in Figure 6 (a), Sam4AVS [59] leverages the large-scale audio-language model CLAP [53] to classify the input audio. For a single-source audio, the class

[1] https://github.com/facebookresearch/detectron2

name with the highest score is selected, while for a multi-source audio, the two highest-scoring class names are chosen. The predicted class names are input into Grounding DINO [41] to generate box predictions, and these boxes are then utilized to query SAM [30] for mask generation.

Experiment. We reproduce Sam4AVS and make it suitable for the AVIS task. Specifically, we divide each audio into multiple 1-second segments and feed them into CLAP separately. Then, we select the class name with top-1 score to generate masks for each video frame. Furthermore, masks of the same category throughout the entire video are considered to belong to the same object.

Problem. Only using audio information to predict the category of sounding objects proves insufficient and unreliable in complex scenarios. For instance, humans can imitate the sound of a cat meowing, and both cars and airplanes may generate similar engine sounds. Sam4AVS neglects visual cues, potentially leading to inaccurate classification of sounding objects. When provided with a class name, Sam4AVS tends to segment all objects belonging to the predicted class, rather than those sounding ones. Additionally, Sam4AVS processes images individually, which prevents it from establishing temporal correlations or tracking instances of sounding objects.

E.2. BuboGPT

Model. As shown in Figure 6 (b), BuboGPT [62] aligns audio-vision-language modalities while leveraging a large language model to generate description of sounding objects. It employs an existing visual grounding pipeline to find the above sounding objects described above in an image and output their final masks. More specifically, BuboGPT uses ImageBind [19] as the audio encoder, BLIP-2 [34] as the vision encoder and Vicuna [14] as the large language model. BuboGPT first aligns audio or visual features with language by training the modality Q-Former [34] and linear projection layer on audio or image caption datasets, respectively. Subsequently, it conducts multi-modal instruct tuning on a large instruction-following dataset, prompting Vicuna to generate description of sound source. The prompt template, i.e., prompt1 depicted in Figure 6 (b), is defined as follows:

```
<Vision><ModalityHere></Vision> <Audio><
    ModalityHere></Audio> Please find the source
    that emits the given sound in this image.
```

To explore the relationships between different visual objects and descriptions of sound source, BuboGPT adopts an off-the-shelf visual grounding pipeline based on SAM [30]. This pipeline consists of four modules: 1) a tagging module RAM [61] to produce multiple text tags/labels that are relevant to the input image; 2) a grounding module Grounding DINO [41] responsible for localizing a bounding box in the image corresponding to each tag/label; 3) an entity-matching module GPT-4 [1] that leverages the reasoning capabilities of the large language model to retrieve matched entities from tags and image descriptions; 4) a segmentation module SAM [30] designed to get fine-grained masks. The prompt template of the entity-matching module, i.e., prompt2 depicted in Figure 6 (b), is defined as follows:

```
You are a helpful assistant. Now I will give you
    a list of entities and give you a paragraph
    or sentence. You need to first extract the
    entity given in the text and then find the
    corresponding entity having similar or
    identical meanings in the given list. Find
    all the pairs. Are you clear? let us think
    step by step. The extracted entities must
    come from the given text and the
    corresponding entity must come from the given
     list. If multiple entities can be linked to
    the same span of text or vice versa, just
    keep one and do not merge them. Here is an
    example: <List>['dog','sheepdog','grass','
    chase sheepdog','field','field park','grassy
    ','corgi','brown dog','brown','park']</List>
    <Text>A brown dog running in the grassy field
    </Text> The answer is: brown dog - brown dog
    \n grassy field - field
```

Experiment. We reproduce BuboGPT and make it suitable for the AVIS task. Specifically, we split each video into multiple non-overlapping visual and audio snippet pairs, where each snippet spans 1 second. BuboGPT takes an image and the corresponding 1-second audio as input, and generate masks for each video frame. Furthermore, masks of the same tag/category throughout the entire video are considered to belong to the same object.

Problem. Compared to Sam4AVS, BuboGPT integrates both audio and visual information to classify and localize sounding object instances, resulting in more accurate sound source localization. However, it still only process one image at a time, which prevents it from establishing temporal correlations or tracking instances of sounding objects. Moreover, RAM predicts tags/categories rather than providing detailed descriptions of the objects. Therefore, the entity-matching module struggles to differentiate between different object instances of the same category.

E.3. PG-Video-LLaVA

Model. As shown in Figure 6 (c), PG-Video-LLaVA [44] transcribes audio cues into texts and extracts spatio-temporal features from videos. Then they are input into a large language model to generate description of sounding objects. Finally, PG-Video-LLaVA uses an off-the-shelf tracker along with a visual grounding module, allowing it to spatially segment sounding objects in videos according to the generated descriptions. Specifically, PG-Video-LLaVA takes video frames as input and employs the CLIP [46] visual encoder to extract video features by averaging frame-level features across temporal and spatial dimensions. For

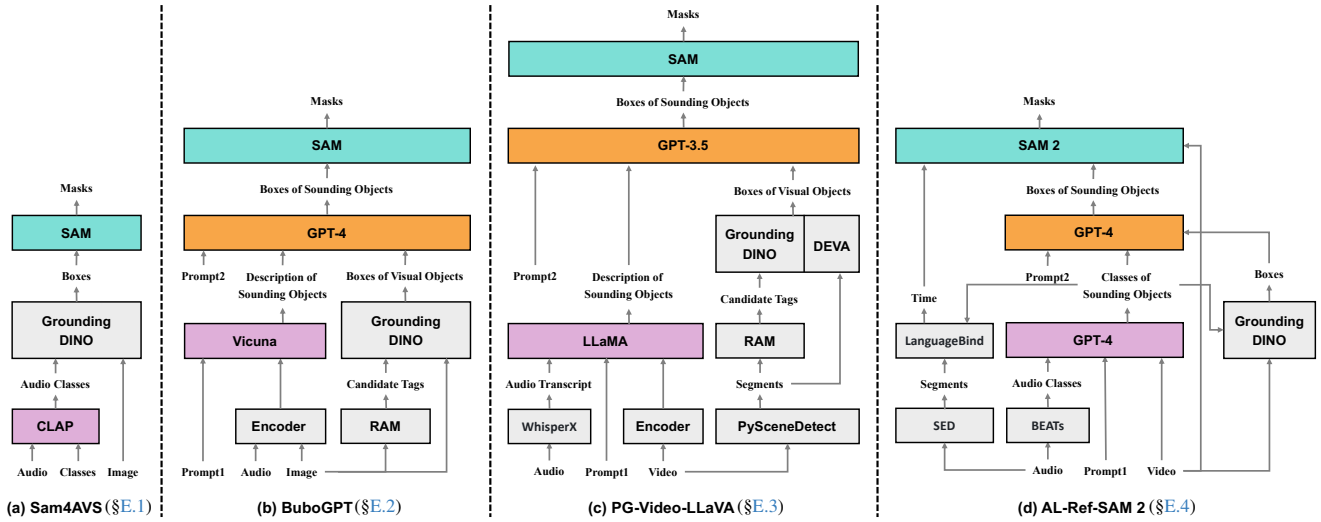

Figure 6. Pipeline comparison of multi-modal large models, including (a) Sam4AVS [59], (b) BuboGPT [62], (c) PG-Video-LLaVA [44], and (d) AL-Ref-SAM 2 [27]. **Multi-modal fusion module** aligns audio-X modalities and outputs classes or descriptions of sounding objects. **Assistant module** leverages the reasoning capabilities of large language models to retrieve matched sounding objects. **Segmentation module** adopts an off-the-shelf visual grounding pipeline to localize sounding objects and generate corresponding fine-grained masks.

the audio modality, PG-Video-LLaVA utilizes WhisperX [3], a speech recognition system, to detect voice activity and generate audio transcripts. The integration of the audio transcript with the video features is executed in the large language model LLaMA [50] through a carefully designed prompt template, i.e., prompt1 depicted in Figure 6 (c):

```
You are PG-Video-LLaVA, a large vision-language
    assistant. You are able to understand the
    video content that the user provides, and
    assist the user with a variety of tasks using
     natural language. Your task is to find the
    source that emits the given sound in this
    video. <Video-Tokens> The noisy audio
    transcript of this video is: <Audio-
    Transcript>
```

After obtaining descriptions of sounding objects from LLaMA, these are employed for grounding within the corresponding video frames. Key noun phrases are extracted from the generated text via GPT-3.5, focusing on the category of sounding objects. The prompt template of GPT-3.5, i.e., prompt2 depicted in Figure 6 (c), is similar to BuboGPT. Simultaneously, an image tagging model, RAM [61], tags visual elements in each frame, constructing a detailed map of the video content. The video is segmented into smaller parts using PySceneDetect, based on changes in scene composition. In each segment, a grounding ensemble, composed of GroundingDINO [41], DEVA [13], and SAM [30], employs the image tags to generate segmentation masks and tracking IDs for the identified visual elements. The visual cues from these segmentation masks are subsequently matched with the textual noun phrases through CLIP [46]. This matching process links the text to the corresponding visual elements in the video.

Experiment. We reproduce PG-Video-LLaVA and make it suitable for the AVIS task. Specifically, each noun phrase from GPT-3.5 serves as an instance and is then input into the grounding module to generate segmentation masks throughout the entire video.

Problem. PG-Video-LLaVA extends image-based large multi-modal models to the video domain, and provides a more accurate understanding of video content compared to Sam4AVS and BuboGP. Nevertheless, it can only describe what the sounding object in the video is but cannot pinpoint the exact time intervals for each sounding object. Moreover, for each video, its feature are obtained by simply averaging image features, which may result in the loss of some valuable information. For each audio, PG-Video-LLaVA only identifies speech segments, filtering out non-speech audio components (e.g., music, machine or animal sounds), and transcribes the speech into text. In addition, RAM predicts tags/categories rather than providing detailed descriptions of the objects. Therefore, GPT-3.5 struggles to differentiate between different object instances of the same category.

E.4. AL-Ref-SAM 2

Model. As shown in Figure 6 (d), AL-Ref-SAM 2 [27] employs an intuitive three-stage pipeline for the audio-visual segmentation task: 1) extract reference information from the multi-modal input, 2) identify the sounding object in the initial frame based on the extracted reference, and 3) segment the identified sounding object throughout the entire video. Specifically, AL-Ref-SAM 2 applies an audio

classifier, BEATs [8], to categorize the audio clip.

```
The image is composed of multiple frames from a
    video spliced from left to right, and the
    frame number is marked with a circle in the
    upper left corner of each frame. Using an
    audio classification model, we obtained the
    audio labels with the highest confidence in
    the video: {$OBJ_1$,$OBJ_2$,...,$OBJ_k$}.
    Please process these audio labels based on
    the content of the image, filtering out audio
     labels that do not exist in the video or are
     abstract labels that cannot be associated
    with specific objects. Additionally, merge
    audio labels that represent the same object.
    Then, according to the retained audio labels,
     output the category of one or more objects
    in the video that may be making sounds in a
    list surrounded by [].
```

```
I have input an image stitched together from
    frames of a video, each frame is marked with
    an ID in the upper left corner. Please first
    describe in detail the events happening in
    the video and then help me select the single
    frame that best demonstrates the \"{reference
    }\" and may result in a good segmentation
    result of the object previously described,
    and return their IDs in the upper left corner
     to me in a list surrounded by [].

The above content is an image that contains
    sampled frames of a video, with the frame
    numbers labeled in the top-left corner. In
    the {$p_f$} frame, three objects are marked
    with colored boxes: {$bbox_1$,$bbox_2$,
    $bbox_3$}. Please follow these steps:
1. Describe the Scene: Describe the video and
    each frame. Describe each object in the frame
    .
2. Describe the Objects within Each Box: Describe
     the objects in the above boxes and their
    relationships.
3. Analyze the Provided Description: Given the
    description \"{reference}\" and analyze its
    syntax, identifying the main object described
     in the sentence. Adhere to syntax analysis
    principles, and do not assume that an object
    is the main subject simply because it an has
    extensive description. This analysis will
    help you distinguish the box that needs to be
     selected from the image.
4. Identify the Object that Best Matches the
    Description:
Ensure you select the precise bounding box of the
     referring object by following these tips:
    Include only the main object described,
    excluding other objects. Include the whole
    main object. Do not include other objects
    mentioned in the description that are not the
     main object.
5. Output the Result: Output the single number in
     list [] format.
```

To avoid the disturbance presence of background noise and the ambiguity of audio information, it integrates visual context and leverages the vision-language understanding capabilities of existing large multi-modal model, GPT-4 [1], to accurately identify the categories of the actual sounding objects present in the video. The prompt template, i.e., prompt1 depicted in Figure 6 (d), is defined as mentioned above. Since the selected referent may be silent in certain frames, AL-Ref-SAM 2 further utilizes sound event detection (SED) to segment the whole audio clip and filter out silent frames from the generated mask sequence. Then, GPT-4 processes the identified categories and video clip to identify a high-quality box of the referent in a specific frame where the referent clearly appears. The prompt template, i.e., prompt2 depicted in Figure 6 (d), is defined as above. Finally, the selected bounding box serves as the pivot box to prompt SAM 2 [47] to segment the referent and propagate its mask forward and backward through the entire video.

Experiment. We reproduce AL-Ref-SAM 2 and make it suitable for the AVIS task. Specifically, each category is considered as a individual object instance.

Problem. Compared to PG-Video-LLaVA, AL-Ref-SAM 2 is capable of determining the exact time intervals during which objects emit sound. However, it cannot distinguish between different object instances of the same category, because BEATs and GPT-4 only output the category of the audio rather than a description of sounding objects.

F. More Qualitative Results

As shown in Figure 7 and Figure 8, we provide some qualitative comparisons with other methods on 4 scenarios. **1)** Video instance segmentation methods (e.g., VITA [26]) can accurately segment and track objects, but fails to determine when these objects are producing sound, e.g., "person" in Figure 7 and "lion" in Figure 8, due to the absence of audio input. **2)** With the help of audio information, audio-visual semantic segmentation methods (e.g., COMBO [56]) are capable of correctly localizing the sound source in most cases. However, such methods show difficulties in processing long sequences, which may result in multiple identity switches in tracking, e.g., "person" in Figure 7. **3)** Multi-modal large models (e.g., PG-Video-LLaVA [44] and AL-Ref-SAM 2 [27]) serve audio as a form of language and leverage foundation models to achieve audio-referred visual grounding. As discussed in Section E, these methods not only fail to distinguish between different object instances of the same category, e.g., "person" in Figure 7, but also struggle to determine the exact time intervals for each sounding object, e.g., "lion" in Figure 8.

In addition, we show more visual results of our baseline model in Figure 9. Our model accurately localizes sound sources, segments sounding objects, and determines when they are emitting sound.

G. Failure Cases

Figure 10 displays additional failure cases of our model on the AVISeg dataset. We observe that inaccurate sound source localization tends to occur in complex multi-source scenarios, especially when multiple objects within the same category emit sound, e.g., two "girls", three "tubas", two "dogs" and three "men" in Figure 10. This because audio signals from homogeneous sounding objects often exhibit similarity and indistinguishable, making them complicating the alignment with visual content. It motivates us to explore how to more effectively disentangle high-density audio signals and establish robust correspondences between audio and visual contents in complex multi-source scenarios and long video sequences.

H. Future Works

As a pioneering work, the current approach is not perfect and thus leaves much room for improvement, which we summarize below:

1) Long-range temporal modeling. Recent work by StreamingLLM [54] introduces the concept of "attention sinks", additional initial tokens that consistently participate in attention computations during sliding window processing. This enables models trained with finite attention windows and generalize to infinite-length sequences without requiring further fine-tuning. Adopting this technique could potentially enhance long-range consistency and improve performance across extended audio-visual sequences.

2) Audio decoupling and audio-visual fusion. As discussed in Section G, our model's performance may be limited in scenarios where multiple objects of the same category are producing sound. To better associate mixed-source audio with visual objects, product quantization [29, 56] can be considered to decompose the mixed audio semantics into several disentangled single-source semantics with noise suppression. This approach has the potential to provide a more compact and robust audio representation for audio-visual interaction, especially in complex scenarios.

3) Online audio-visual segmentation. Many recently introduced methods have demonstrated promising performance for audio-visual segmentation tasks. However, they are restricted in real-time applications as they operate offline, requiring the entire video to be processed before the predictions. Therefore, developing online methods that processes video frames sequentially, without access to future frames, will be an important topic.

4) Prompt engineering and instruction tuning. With the help of large language models, existing multi-modal large models (MMLMs) exhibits impressive audio-visual understanding abilities. Nevertheless, they are far from satisfactory in fine-grained audio-referred visual grounding tasks, especially in instance-aware sound source localization and long videos. By carefully designing the text prompts or fine-tuning on the AVISeg-based instruction-tuning dataset, MMLMs can produce more accurate responses and detailed descriptions of sounding objects.

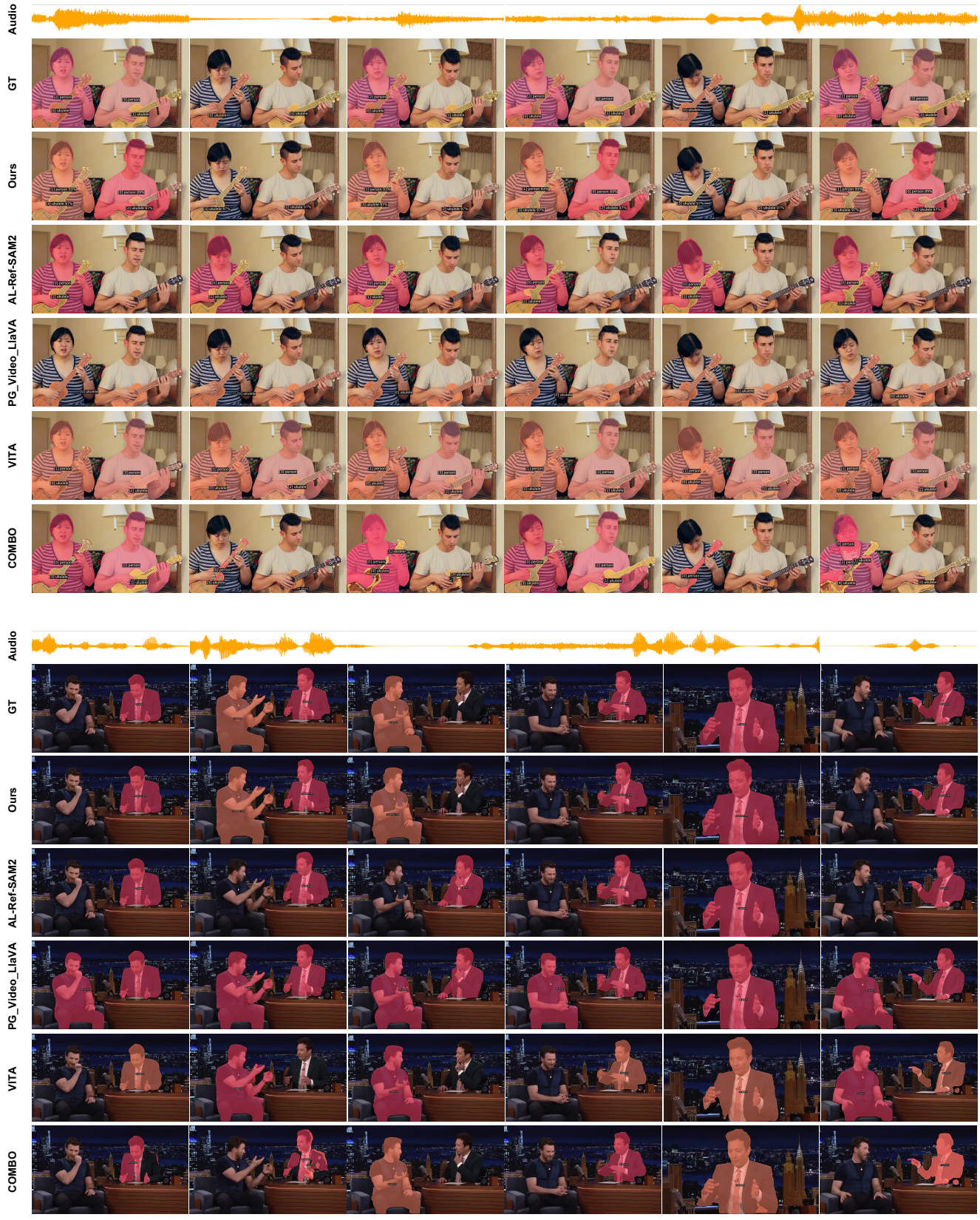

Figure 7. Qualitative comparison of our model with VIS (VITA), AVSS (COMBO) and multi-modal large models (PG-Video-LLaVA and AL-Ref-SAM 2) on Music (Top) and Speaking (Bottom) scenarios.

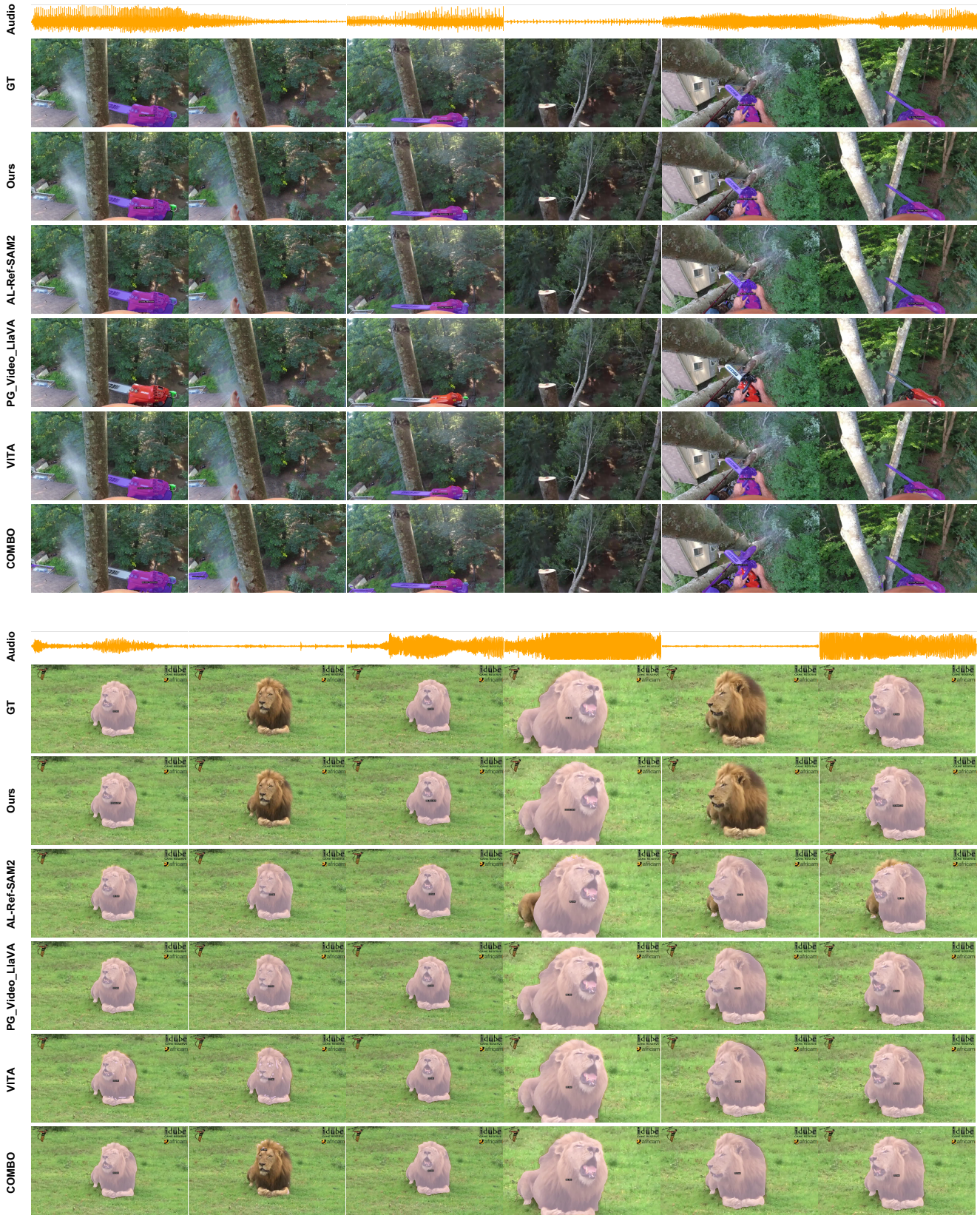

Figure 8. Qualitative comparison of our model with VIS (VITA), AVSS (COMBO) and multi-modal large models (PG-Video-LLaVA and AL-Ref-SAM 2) on Machine (Top) and Animal (Bottom) scenarios.

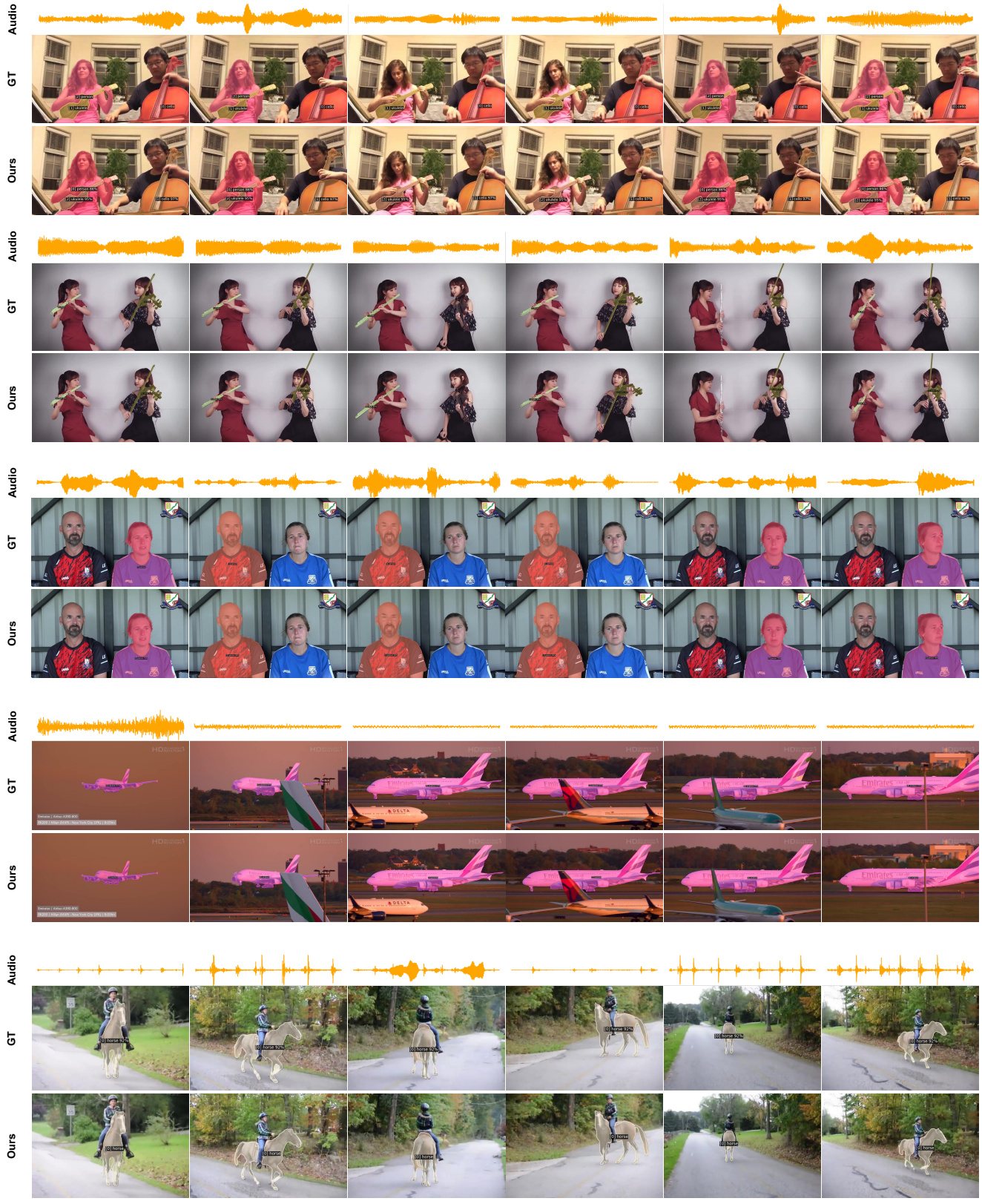

Figure 9. More visual results of our baseline model on AVISeg dataset from four scenarios. Each row have six sampled frames from a video sequence. Zoom in to see details.

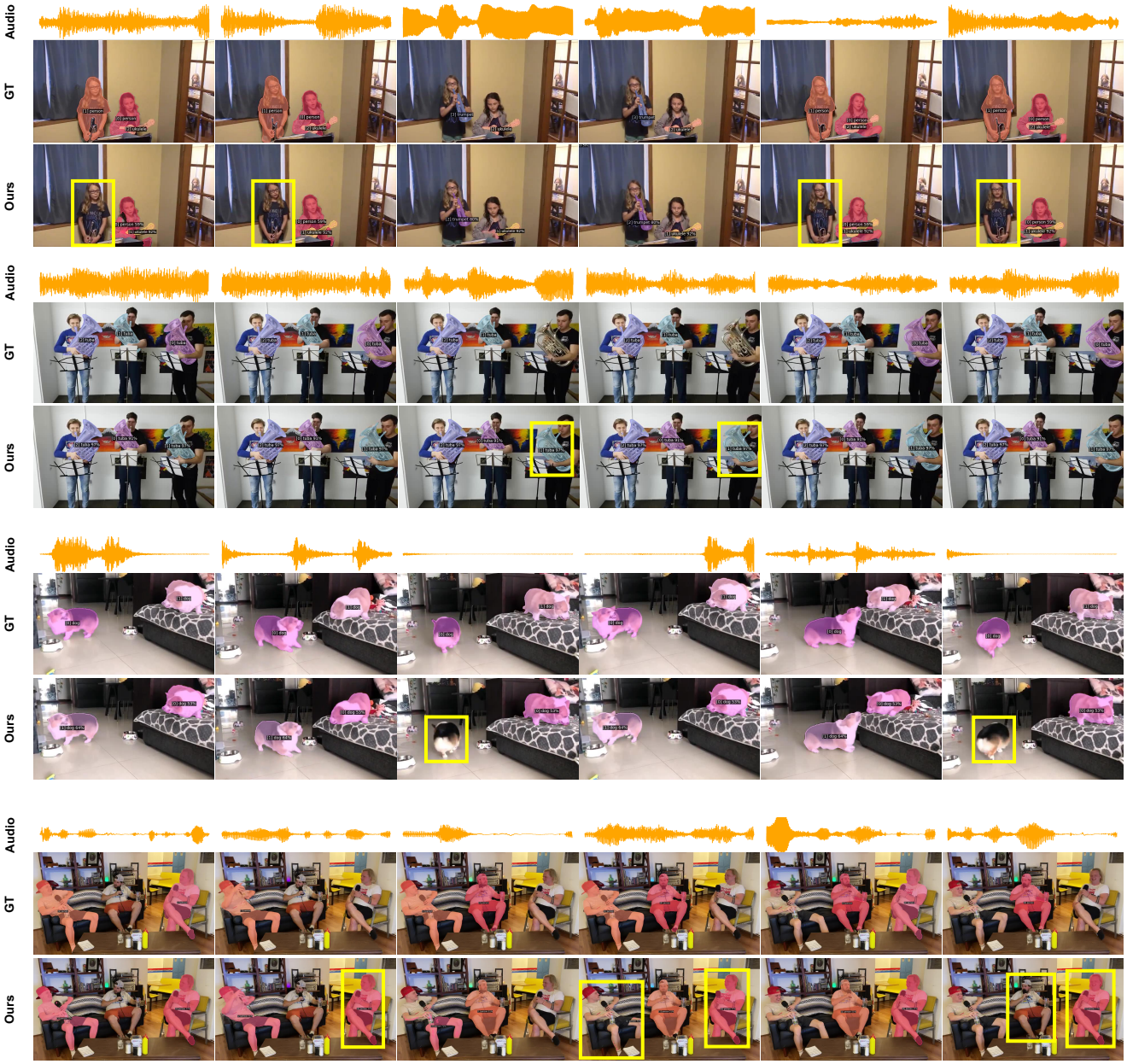

Figure 10. Failure cases of our baseline model on AVISeg dataset. Each row has six sampled frames from a video sequence. The yellow boxes indicate the incorrect segmentation regions. Zoom in to see more details.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 7, 3, 5

[2] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *European Conference on Computer Vision*, pages 158–177. Springer, 2020. 2

[3] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. In *Interspeech*, pages 4489–4493, 2023. 4

[4] Jiale Cao, Rao Muhammad Anwer, Hisham Cholakkal, Fahad Shahbaz Khan, Yanwei Pang, and Ling Shao. Sipmask: Spatial information preservation for fast image and video instance segmentation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2020. 2

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020. 1, 2, 3, 5, 6

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 721–725. IEEE, 2020. 3

[7] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 4

[8] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In *International Conference on Machine Learning*, pages 5178–5193. PMLR, 2023. 5

[9] Yaru Chen, Ruohao Guo, Xubo Liu, Peipei Wu, Guangyao Li, Zhenbo Li, and Wenwu Wang. Cm-pie: Cross-modal perception for interactive-enhanced audio-visual video parsing. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8421–8425. IEEE, 2024. 1

[10] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. Unraveling instance associations: A closer look for audio-visual segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26497–26507, 2024. 3

[11] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021. 2, 3, 4, 6, 7

[12] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 1, 2, 3, 4, 6, 8

[13] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 1316–1326, 2023. 4

[14] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org*, 2(3):6, 2023. 7, 3

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee, 2009. 6

[16] Hao Fang, Tong Zhang, Xiaofei Zhou, and Xinxin Zhang. Learning better video query with sam for video instance segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024. 6, 7

[17] Shengyi Gao, Zhe Chen, Guo Chen, Wenhai Wang, and Tong Lu. Avsegformer: Audio-visual segmentation with transformer. In *AAAI Conference on Artificial Intelligence*, pages 12155–12163, 2024. 1, 3, 6, 7

[18] 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 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 776–780, 2017. 5, 2

[19] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 3

[20] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. Sotr: Segmenting objects with transformers. In *IEEE/CVF International Conference on Computer Vision*, pages 7157–7166, 2021. 2

[21] Ruohao Guo, Dantong Niu, Liao Qu, Yanyu Qi, Ji Shi, Wenzhen Yue, Bowei Xing, Taiyan Chen, and Xianghua Ying. Instance-level panoramic audio-visual saliency detection and ranking. In *ACM International Conference on Multimedia*, pages 9426–9434, 2024. 3

[22] Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. Open-vocabulary audio-visual semantic segmentation. In *ACM International Conference on Multimedia*, pages 7533–7541, 2024. 1, 3

[23] Ruohao Guo, Xianghua Ying, Yanyu Qi, and Liao Qu. Unitr: A unified transformer-based framework for co-object and multi-modal saliency detection. *IEEE Transactions on Multimedia*, 2024. 1

[24] 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. 5, 2

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE/CVF International Conference on Computer Vision*, pages 2961–2969, 2017. 2

[26] Miran Heo, Sukjun Hwang, Seoung Wug Oh, Joon-Young Lee, and Seon Joo Kim. Vita: Video instance segmentation via object token association. *Advances in Neural Information Processing Systems*, 35:23109–23120, 2022. 2, 3, 4, 5, 6, 7

[27] Shaofei Huang, Rui Ling, Hongyu Li, Tianrui Hui, Zongheng Tang, Xiaoming Wei, Jizhong Han, and Si Liu. Unleashing the temporal-spatial reasoning capacity of gpt for training-free audio and language referenced video object segmentation. *arXiv preprint arXiv:2408.15876*, 2024. 7, 4, 5

[28] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. *Advances in Neural Information Processing Systems*, 34:13352–13363, 2021. 2, 3, 6

[29] Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2010. 6

[30] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 4, 7, 3

[31] Rajat Koner, Tanveer Hannan, Suprosanna Shit, Sahand Sharifzadeh, Matthias Schubert, Thomas Seidl, and Volker Tresp. Instanceformer: An online video instance segmentation framework. In *AAAI Conference on Artificial Intelligence*, pages 1188–1195, 2023. 2, 6

[32] Harold W Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. 3, 6, 1

[33] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19108–19118, 2022. 4

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, pages 19730–19742. PMLR, 2023. 3

[35] Kexin Li, Zongxin Yang, Lei Chen, Yi Yang, and Jun Xiao. Catr: Combinatorial-dependence audio-queried transformer for audio-visual video segmentation. In *ACM International Conference on Multimedia*, pages 1485–1494, 2023. 1, 3

[36] Zhangbin Li, Dan Guo, Jinxing Zhou, Jing Zhang, and Meng Wang. Object-aware adaptive-positivity learning for audio-visual question answering. In *AAAI Conference on Artificial Intelligence*, pages 3306–3314, 2024. 4

[37] Zhangbin Li, Jinxing Zhou, Jing Zhang, Shengeng Tang, Kun Li, and Dan Guo. Patch-level sounding object tracking for audio-visual question answering. *arXiv preprint arXiv:2412.10749*, 2024. 4

[38] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014. 3

[39] Dongfang Liu, Yiming Cui, Wenbo Tan, and Yingjie Chen. Sg-net: Spatial granularity network for one-stage video instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9816–9825, 2021. 2

[40] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 3, 4

[42] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 5, 2

[43] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129: 548–578, 2021. 3, 1

[44] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 7, 3, 4, 5

[45] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip HS Torr, and Song Bai. Occluded video instance segmentation: A benchmark. *International Journal of Computer Vision*, 130(8): 2022–2039, 2022. 4, 8

[46] 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. 3, 4

[47] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5

[48] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018. 4

[49] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 9627–9636, 2019. 2

[50] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 4

[51] Yuqing Wang, Zhaoliang Xu, Xinlong Wang, Chunhua Shen, Baoshan Cheng, Hao Shen, and Huaxia Xia. End-to-end video instance segmentation with transformers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8741–8750, 2021. 2

[52] Junfeng Wu, Yi Jiang, Song Bai, Wenqing Zhang, and Xiang Bai. Seqformer: Sequential transformer for video instance segmentation. In *European Conference on Computer Vision*, pages 553–569. Springer, 2022. 2, 3, 6, 7

[53] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1–5. IEEE, 2023. 2

[54] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations*, 2023. 6

[55] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 5188–5197, 2019. 2, 3, 4, 6, 1

[56] Qi Yang, Xing Nie, Tong Li, Pengfei Gao, Ying Guo, Cheng Zhen, Pengfei Yan, and Shiming Xiang. Cooperation does matter: Exploring multi-order bilateral relations for audio-visual segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27134–27143, 2024. 1, 3, 6, 7, 5

[57] Shusheng Yang, Yuxin Fang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Crossover learning for fast online video instance segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 8043–8052, 2021. 2

[58] Shusheng Yang, Xinggang Wang, Yu Li, Yuxin Fang, Jiemin Fang, Wenyu Liu, Xun Zhao, and Ying Shan. Temporally efficient vision transformer for video instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2885–2895, 2022. 2, 6, 7

[59] Jiarui Yu, Haoran Li, Yanbin Hao, Jinmeng Wu, Tong Xu, Shuo Wang, and Xiangnan He. How can contrastive pre-training benefit audio-visual segmentation? a study from supervised and zero-shot perspectives. In *British Machine Vision Association*, pages 367–374, 2023. 7, 2, 4

[60] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. Dvis: Decoupled video instance segmentation framework. In *IEEE/CVF International Conference on Computer Vision*, pages 1282–1291, 2023. 2, 6, 7

[61] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize anything: A strong image tagging model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 3, 4

[62] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023. 7, 3, 4

[63] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021. 1

[64] Jinxing Zhou, Dan Guo, and Meng Wang. Contrastive positive sample propagation along the audio-visual event line. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7239–7257, 2022. 1

[65] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *European Conference on Computer Vision*, pages 386–403. Springer, 2022. 1, 3, 4

[66] Jinxing Zhou, Dan Guo, Ruohao Guo, Yuxin Mao, Jingjing Hu, Yiran Zhong, Xiaojun Chang, and Meng Wang. Towards open-vocabulary audio-visual event localization. *arXiv preprint arXiv:2411.11278*, 2024. 1

[67] Jinxing Zhou, Dan Guo, Yuxin Mao, Yiran Zhong, Xiaojun Chang, and Meng Wang. Label-anticipated event disentanglement for audio-visual video parsing. In *European Conference on Computer Vision*, pages 35–51. Springer, 2024.

[68] Jinxing Zhou, Dan Guo, Yiran Zhong, and Meng Wang. Advancing weakly-supervised audio-visual video parsing via segment-wise pseudo labeling. *International Journal of Computer Vision*, 132(11):5308–5329, 2024. 1

[69] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *International Journal of Computer Vision*, pages 1–21, 2024. 1, 3, 4

[70] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. 5