What’s Making That Sound Right Now? Video-centric Audio-Visual Localization

Supplementary Material

A. The AVATAR Details

A.1. Data Quantity Analysis

	Videos
Total	5,000

(a) Videos

	Frames
Total	24,266
Off-screen	670

(b) Frames

	Instances
Total	28,516
Single-sound	15,372
Multi-entity	9,322
Mixed-sound	3,822

(c) Instances

Table A. The Statistics of AVATAR

The detailed quantitative statistics of the AVATAR benchmark are summarized in Tab. A. AVATAR consists of 5,000 videos and 24,266 labeled frames, with approximately five frames sampled per video during the frame sampling stage. However, the total number of labeled frames is not an exact multiple of five due to two key factors. First, to ensure event diversity, frame sampling was designed to prevent overlapping events within a ±0.7-second window, resulting in some videos containing fewer than five selected events. Second, during human verification, blurred frames, those with artificial scene transitions, and frames with post-processed visual effects or background music were excluded from labeling. Consequently, a total of 24,266 valid frames were included in AVATAR.

Notably, AVATAR contains 670 frames corresponding to Off-screen scenarios. Among existing audio-visual localization benchmarks, [20] is the only work that explicitly addresses audio-visual negative samples. Specifically, it provides 42 frames from the Flickr-SoundNet test set and 379 frames from the VGG-Sound test set, each containing either non-audible events or non-visible sound sources. In addition, it includes synthetic negative samples generated by mismatching audio and visual frames from different videos. In contrast, AVATAR offers 670 non-synthetic Off-screen frames—approximately 1.8 to 16 times more than prior benchmarks—facilitating a more comprehensive evaluation of model performance under Off-screen conditions.

As shown in Tab. A, AVATAR includes a total of 28,516 audio-visual instances, categorized into Single-sound (15,372), Multi-entity (9,322), and Mixed-sound (3,822) instances.

A.2. Diversity of Audio-Visual Categories

The AVATAR dataset comprises 80 audio-visual categories, designed to cover a broad range of domains encountered in everyday life. These categories span a diverse spectrum, including human activities, musical performances, animal sounds, vehicle sounds, and tool sounds. By encompassing a wide variety of real-world scenarios, AVATAR provides a comprehensive evaluation framework for assessing model performance. A detailed list of categories and the distribution of instances per category are presented in Fig. T.

A.3. Visualization Examples

In this section, we present qualitative examples from the AVATAR benchmark across different scenarios, highlighting the diversity and complexity of the dataset. Each scenario introduces unique challenges in audio-visual localization, requiring models to effectively process spatial and temporal information. We provide representative examples for the following scenarios: Single-sound, Mixed-sound, Multi-entity, Off-screen, and a Cross-event subset.

A.3.1. Single-sound

Single-sound scenario represents the most fundamental case, where a single instance within the frame generates a distinct audio event. This setting serves as a baseline for evaluating a model’s core audio-visual localization capabilities, requiring accurate alignment between the visual source and the corresponding audio without interference from complex acoustic environments.

Fig. L illustrates various examples of this scenario. Row 1 features a chainsaw cutting a tree, while row 5 presents a moving car that needs to be localized. Additional examples include a woman speaking (row 4), a motorcycle (row 6), keyboard typing (row 11), and an electric shaver (row 12). These cases assess the model’s ability to associate a single sound source with its visual counterpart across diverse contexts.

As the foundation for audio-visual localization evaluation, this scenario is crucial for determining how well a model learns fundamental audio-visual associations before progressing to more complex cases.

A.3.2. Mixed-sound

Mixed-sound scenario involves multiple overlapping audio sources, challenging the model’s ability to correctly associate sound with the corresponding visual objects. This setting requires advanced audio perception, as models must filter relevant sources amidst competing sounds. It includes:

multiple instances of the same category, instances from different categories, and partially off-screen sound sources.

Fig. M presents diverse cases within this scenario. Row 9 depicts six horn players (same category) performing simultaneously, requiring the model to discern multiple co-occurring sources. Row 4 showcases a banjo, violin, guitar, and double bass being played together (different categories), requiring category-wise differentiation. In row 3, a woman plays the violin while accompanied by a piano, where only the violin is visible, posing a challenge in distinguishing on-screen and off-screen sound sources.

This scenario evaluates how well models can handle real-world acoustic complexities by correctly localizing and matching relevant sound-emitting objects under challenging conditions.

A.3.3. Multi-entity

Multi-entity scenario requires models to identify the specific instance producing sound among multiple visually similar objects. Unlike simple category-level matching, this setting demands fine-grained differentiation between active and inactive instances within the same visual category.

Fig. N illustrates examples where multiple instances of the same category appear within the frame, such as a group of men (row 1), a collection of singing bowls (row 3), dogs (row 4), timpani (row 9), birds (row 10), and harps (row 12). The key challenge is to precisely localize the active source rather than merely associating the entire category with the sound.

This scenario extends beyond standard category-based localization, assessing a model's ability to perform precise instance-level audio-visual differentiation. As a novel challenge introduced in AVATAR, it enables rigorous evaluation of fine-grained audio-visual perception in real-world conditions.

A.3.4. Off-screen

Off-screen scenario addresses cases where the sound source is not visible within the frame. This setting evaluates a model's ability to suppress false positives by distinguishing between on-screen and off-screen audio events, testing its robustness in handling occluded or unseen sources.

Fig. O presents various examples. Row 8 features an interview where the interviewer's voice originates from off-screen, requiring the model to avoid mistakenly associating the sound with the visible person. Row 11 depicts a flutist performing alongside an off-screen piano, where the piano sound must not be attributed to the on-screen individual. Row 12 involves a police siren blaring before the police car enters the frame, necessitating the model to delay localization until the source appears.

This scenario is critical for assessing robustness and preventing false positive localization, making it a unique aspect of the AVATAR benchmark.

A.3.5. Cross-event

Cross-event is a subset of AVATAR, comprising 450 videos where distinct audio-visual events occur sequentially over time. This subset evaluates a model's temporal understanding by requiring it to adapt to event transitions and accurately localize different sound sources at different time steps. Videos in this subset were selected based on changes in the dominant audio-visual category. The selection process, detailed in Algorithm 1, identifies frames where a new audio-visual category emerges, signaling a transition to a different sound-producing entity.

Fig. P showcases examples from this conditions. Row 2 features a conversation where an off-screen man and an on-screen child take turns speaking. Row 7 presents a case where a child's voice alternates with a barking dog. In row 9, a man speaks first, followed by a woman beside him.

Cross-event is designed to evaluate a model's temporal reasoning, ensuring it can dynamically adjust to shifts in audio-visual sources while maintaining accurate localization.

A.4. License

The benchmark was constructed using videos available under the Creative Commons Attribution 4.0 International License (CC BY 4.0) on YouTube. **The AVATAR permits usage for all purposes, including commercial applications.**

B. Labeling Details

This section describes the annotation interface and human workflow used in constructing the AVATAR benchmark. As detailed in Sec. 3.1.3, human involvement in the labeling process occurs in two key stages: bounding-box annotation and instance segmentation.

B.1. Bounding-box Annotation

In the bounding-box annotation stage, human annotators review a short clip of ± 0.05 seconds around the target frame to determine whether the frame belongs to the Off-screen scenario. If the sound source is not visible, annotators provide a textual label for the corresponding audio-visual category. The interface used for this process is shown in Fig. Q.

If the sound source is visible (non-Off-screen scenario), annotators follow a sequential process. First, they identify the sound-emitting object by reviewing bounding boxes generated through Audio-guided Bounding Box Filtering, based on the clip surrounding the frame, and removing those that do not correspond to the actual sound source. Next, they assign the appropriate audio-visual category, using a predefined category list derived from CAV-MAE [9] audio classification results. If the correct category is not available, they manually input metadata. Finally, annotators assign a scenario label to each instance based on predefined

Algorithm 1 Cross-event Video Sampling Pseudocode

```
Input:
$X_{Total}$ := All videos in the AVATAR benchmark (Total video set)
Output:
$X_{CE}$ := List of Cross-event videos
Initialize:
$X_{CE} \leftarrow$ An empty list to store Cross-event videos
$*$ $instance$ in I has attributes segmentation, audio_visual_category

  for each $video$ $X \in X_{Total}$ do
      $VideoInstances \leftarrow$ An empty set to store instances' audio-visual categories in video
      for each labeled $frame$ $V^t \in X$ do
          $FrameInstances \leftarrow$ An empty set to store instances' audio-visual categories in frame
          for each $instance$ $I \in$ I do
              $FrameInstances$.add($I$.audio_visual_category)
          end for

          $new\_category \leftarrow (FrameInstances - VideoInstances)$
          if $new\_category$ then
              $X_{CE}$.append($X$)
              break
          end if
          $VideoInstances \leftarrow VideoInstances \cup FrameInstances$
      end for
  end for
```

classification rules. In ambiguous cases, such as a scene with both a guitarist and a singer, guitar playing is classified under the Mixed-sound scenario, while singing could belong to both Mixed-sound and Multi-entity scenarios. In such cases, AVATAR prioritizes classification under Multi-entity. The bounding-box annotation process was conducted using the open-source tool Label Studio[1], and the corresponding interface is shown in Fig. R.

B.2. Instance Segmentation

Following the bounding-box annotation, the instance segmentation stage generates segmentation masks for the annotated bounding boxes. Annotators verify the segmentation masks generated by the Segment Anything Model (SAM) [16] and refine them if necessary using SAM guided annotation. If the automatically generated masks are inaccurate, annotators manually adjust them. The interface for this process is shown in Fig. S. Instance segmentation was performed using the open-source tool SALT[2].

C. Experiment Details

C.1. Cross-event Sampling Strategy

The Cross-event subset is designed to evaluate the model's adaptability in dynamic environments. This subset consists of videos where the sound-emitting instances change over time, enabling an assessment of whether the model can accurately localize sound sources even as they shift throughout the video.

To construct the Cross-event subset, we apply a selection process to all videos in the AVATAR benchmark based on specific criteria. The detailed procedure is outlined in Algorithm 1.

Algorithm 1 operates as follows. First, we iterate through all videos in the AVATAR benchmark, extracting the categories of instances present in each frame. Specifically, we examine the audio-visual categories of instances occurring in each frame of a video and determine whether a new, previously unseen category emerges within the video. If a frame contains a newly introduced category that was not present in earlier frames, the video is added to the Cross-event subset, and no further frames from that video need to be examined. This strategy efficiently identifies videos in which sound sources change over time.

By employing this sampling strategy, the resulting Cross-event subset includes videos with diverse sound-emitting instances that dynamically change over time, rather than being restricted to a single sound source. Consequently, this subset enables a more rigorous evaluation of a model's ability to localize sound sources accurately in complex and dynamic environments.

D. Additional Experimental Results

D.1. Effect of Audio Window Length

We set SLAVC [20], EZ-VSL [21], and SSL-TIE [18] as our baselines, all of which are trained on VGGSound [5]. These models perform inference by taking an audio-image pair as input and localizing sound sources at the frame level. While we follow the same inference setup for evaluation, this approach does not directly align with our evaluation protocol, which considers temporal evolution.

To mitigate this discrepancy, we conducted additional experiments by segmenting the audio into windows of 10 seconds, 5 seconds, 2 seconds, and 1 second, as shown in Fig. A. The results demonstrate a sharp decline in CIoU and AUC performance across the entire video as the audio window size decreases. Notably, for certain models, performance remained stable despite shorter audio windows. This trend becomes more apparent when analyzing the difference between cross-event performance and overall video performance. Specifically, even as the audio window size decreases, the reduction in CIoU and AUC performance in the cross-event setting remains similar to that observed with the 10-second window.

In conclusion, this study maintains the inference approach of the baseline models while leveraging 10-second global audio features for experimentation.

[1] https://labelstud.io/

[2] https://github.com/anuragxel/salt

	Total	
Method	CIoU(%)↑	AUC(%)↑
Spatio-Temporal + Max	5.06	6.76
Spatio-Temporal + Min	11.36	12.16
Temporal-Spatio + Mean	13.24	13.86
Spatio-Temporal + Mean	**13.37**	**13.98**

Table B. Ablation Study on Our Approach. Comparison of different optimization objectives for negative bags and computation orders in the Audio-Spatio-Temporal Attention block. Spatio-Temporal + Mean achieves the best performance, demonstrating the effectiveness of optimizing negative bags for improved localization.

D.2. Ablation Studies

In this study, we analyze the impact of the optimization objective for negative bags and the computation order in the Audio-Spatio-Temporal Attention Block. Tab. B presents a comparison of different optimization strategies—Min, Max, and Mean—as well as two processing orders: Spatio-Temporal and Temporal-Spatio.

The results indicate that the choice of optimization objective for negative bags significantly affects performance, whereas the computation order has a relatively minor impact. Notably, the Spatio-Temporal + Mean approach achieves the highest CIoU and AUC scores, demonstrating its effectiveness in optimizing negative bags for improved sound source localization. These findings highlight the critical role of negative bag optimization in enhancing audio-visual localization performance.

D.3. Qualitative Result

In this section, we provide a qualitative analysis of model performance across various scenarios in the AVATAR benchmark. Through heatmap visualizations, we examine how effectively each model localizes sound-emitting objects and compare their performance under different conditions, including Single-sound, Mixed-sound, Multi-entity, Off-screen scenarios and Cross-event subset. The evaluated models include TAVLO 10k (Ours), SLAVC [20], EZ-VSL [21], and SSL-TIE [18].

D.3.1. Single-sound

Fig. B, Fig. C, and Fig. D present the heatmap results for each model in the Single-sound scenario. This scenario serves as a fundamental benchmark for evaluating audio-visual localization ability. We provide heatmap visualizations for diverse audio-visual categories, including an ambulance (Fig. B), baby laughter (Fig. C), and a horn (Fig. D).

TAVLO 10k (Ours) demonstrates consistent localization performance across various categories, with the size and position of the heatmap remaining aligned with the ground truth across frames.

D.3.2. Mixed-sound

Fig. E, Fig. F, and Fig. G present the heatmap results for each model in the Mixed-sound scenario. This scenario evaluates the ability to accurately detect sound-emitting objects in the presence of multiple overlapping audio sources. We consider three variations of the Mixed-sound scenario: (1) multiple instances from the same audio-visual category (Accordion) producing sound simultaneously (Fig. E), (2) instances from different audio-visual categories (Piano and Trombone) appearing together (Fig. F), and (3) a combination of an on-screen instance (Violin) and an off-screen instance (Piano), both emitting sound concurrently (Fig. G).

TAVLO 10k (Ours) consistently generates heatmaps that accurately capture the corresponding instances across all Mixed-sound variations. This demonstrates the model's robustness in complex audio environments, effectively localizing multiple sound sources even under challenging conditions.

D.3.3. Multi-entity

Fig. H and Fig. I present the heatmap results for each model in the Multi-entity scenario. This scenario evaluates a model's ability to accurately localize the sound-emitting instance among multiple instances belonging to the same visual category. We analyze model performance using two cases: Vibraphone (Fig. H) and Mandolin (Fig. I).

In Fig. H, two Vibraphones are visible in the scene, but only the left one is being actively played, requiring precise localization of the sound-emitting instance. Traditional audio-image mapping approaches struggle to differentiate between the two, as they rely heavily on global audio-visual correspondence. This limitation is evident in the experimental results, where baseline models either generate heatmaps covering both Vibraphones or incorrectly localize the right Vibraphone. In contrast, TAVLO 10k (Ours) consistently localizes the correct Vibraphone across all four frames.

Similarly, in Fig. I, the task is to localize the sound-emitting Mandolin among two instances—one being actively played and the other resting on a sofa. TAVLO 10k (Ours) successfully localizes the correct Mandolin across all five frames, while baseline models frequently mislocalize the non-playing Mandolin on the sofa.

These qualitative results highlight the limitations of baseline models, which rely heavily on audio-visual mapping, and demonstrate the fine-grained object differentiation capability of TAVLO 10k in Multi-entity scenarios.

D.3.4. Off-screen

Fig. J presents the qualitative results for the Off-screen scenario. This scenario evaluates a model's robustness by assessing its ability to avoid localizing sound sources that are not visible within the frame.

In Fig. J, the sound of a motorcycle is heard during the video, but the source remains off-screen. TAVLO 10k (Ours) appropriately handles this case by not generating a localization response for the corresponding frames.

D.3.5. Cross-event

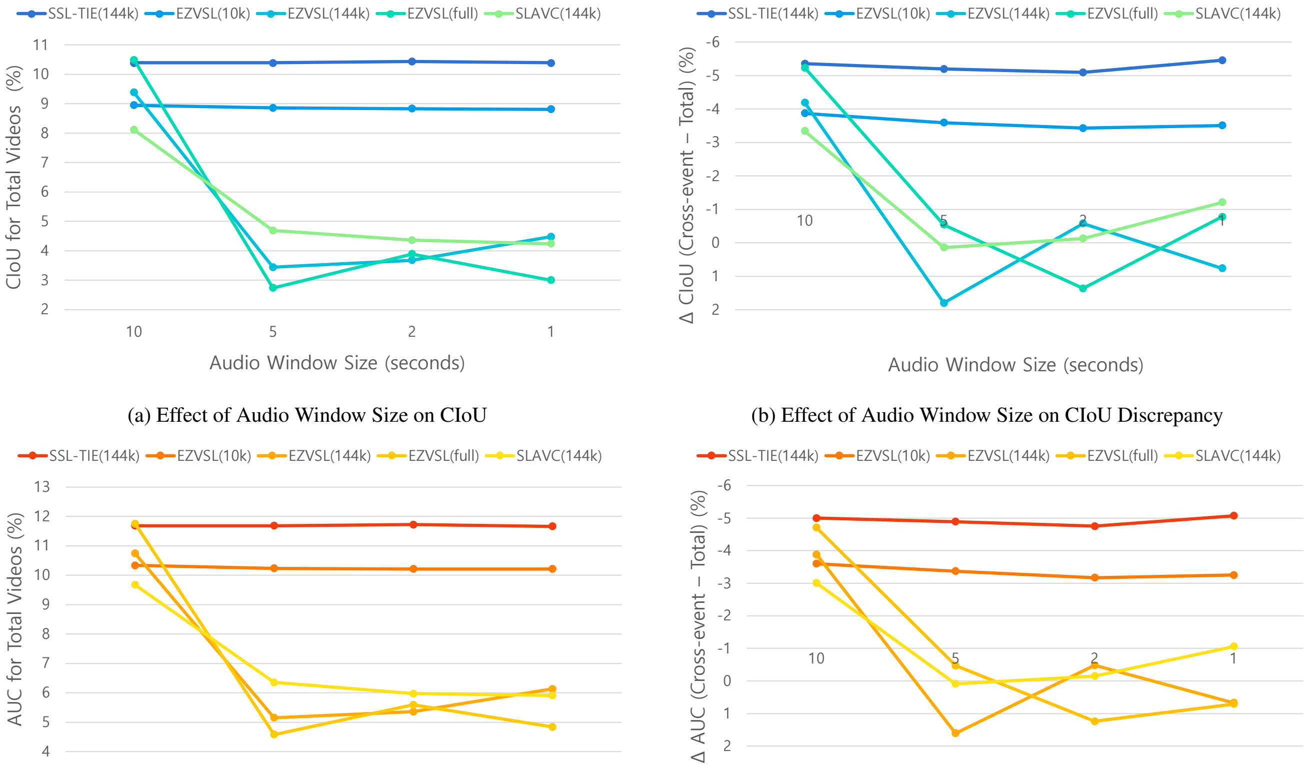
Fig. K illustrates the qualitative results for the Cross-event scenario, which evaluates a model's adaptability when the sound source changes during the video. In this scenario, the initial frame features a female narrator's voice, followed by a transition to the sound of a Glockenspiel being played on-screen.

TAVLO 10k (Ours) successfully avoids generating a false positive during the off-screen phase in the first frame and then accurately localizes the Glockenspiel over the subsequent four frames. In contrast, all baseline models incorrectly localize the Glockenspiel in the first frame, likely due to their reliance on global audio features and a lack of temporal awareness in assessing the current sound context. Moreover, the baselines exhibit unstable localization performance across the following frames, further highlighting their limitations in handling dynamic audio transitions.

E. Emoji

The icons used in Fig. 1 were sourced from Flaticon[3]. The emojis are license-free.

[3] https://www.flaticon.com



(a) Effect of Audio Window Size on CIoU

(b) Effect of Audio Window Size on CIoU Discrepancy

(c) Effect of Audio Window Size on AUC

(d) Effect of Audio Window Size on AUC Discrepancy

Figure A. Effect of Audio Window Size on Baselines Performance. As the audio window size decreases, the performance (CIoU and AUC) for total videos significantly drops. Additionally, even when performance is maintained, the difference between total and cross-event CIoU and AUC remains comparable to that of the full 10s window, suggesting that shorter windows do not mitigate the performance gap between cross-event and total video localization.

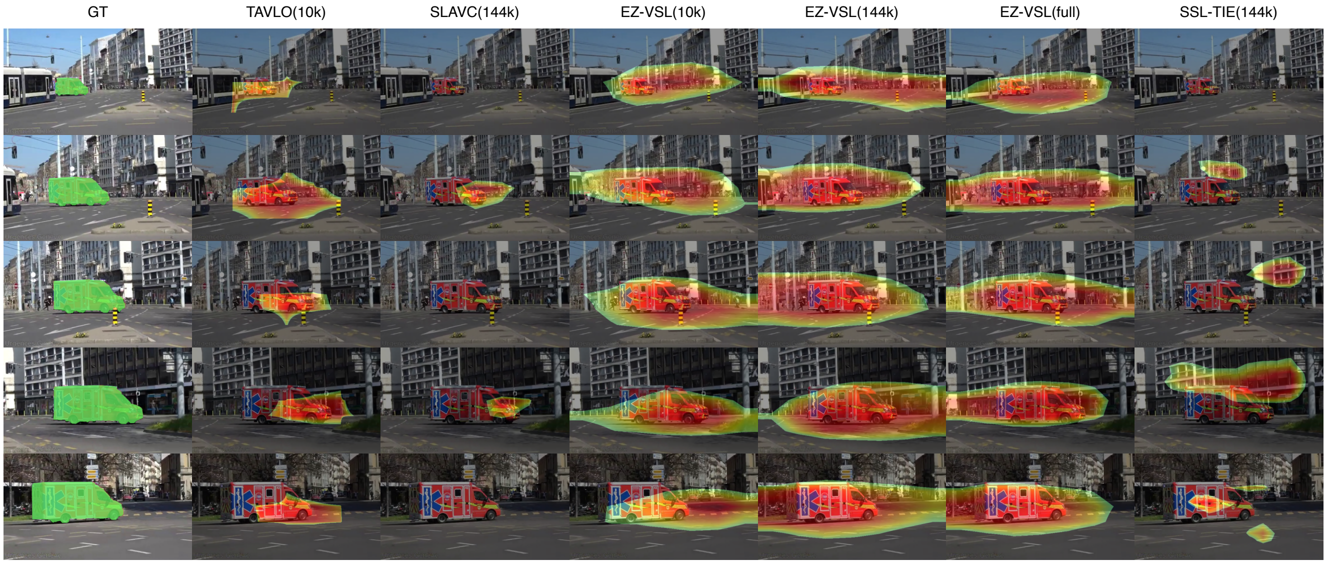


Figure B. Single-sound Scenario Visualization Result 1, Ambulance (frame: 9, 29, 49, 85, 156)

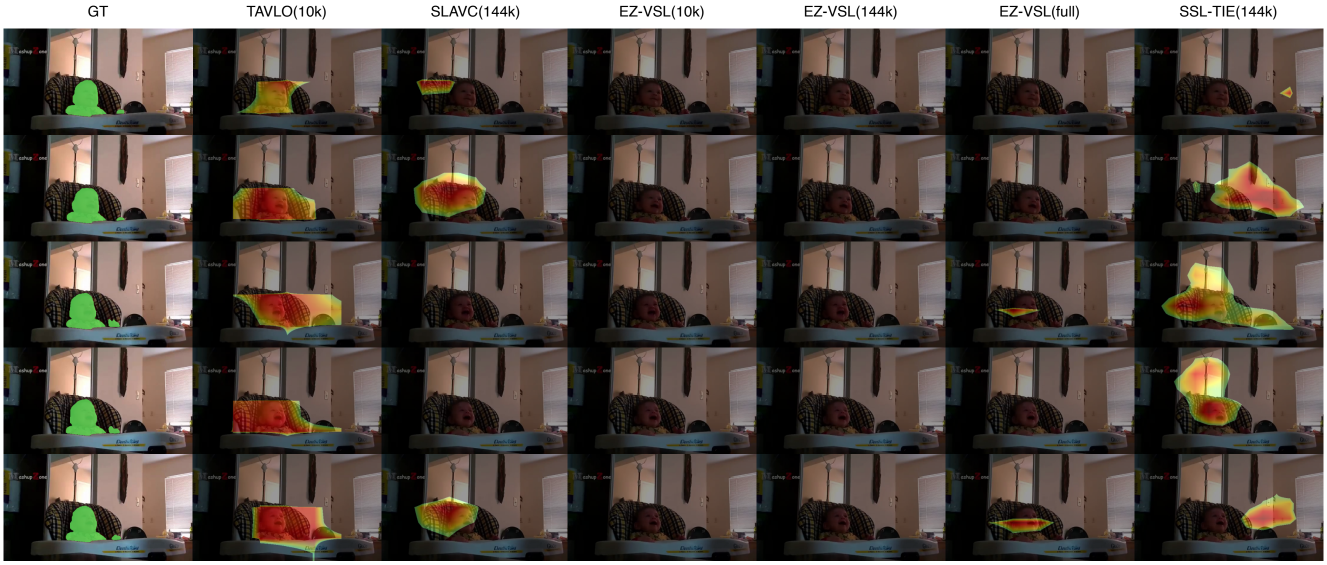


Figure C. Single-sound Scenario Visualization Result 2 - Baby laughter (frame: 10, 53, 230, 256, 293)

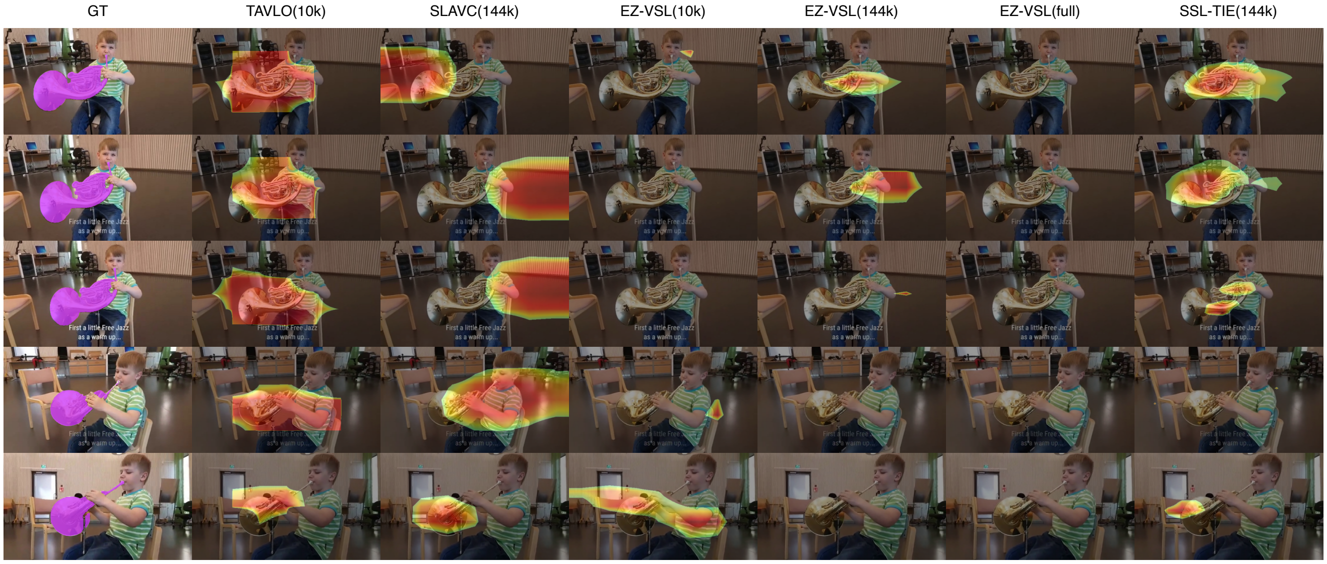


Figure D. Single-sound Scenario Visualization Result 3 - Horn (frame: 24, 52, 88, 171, 286)

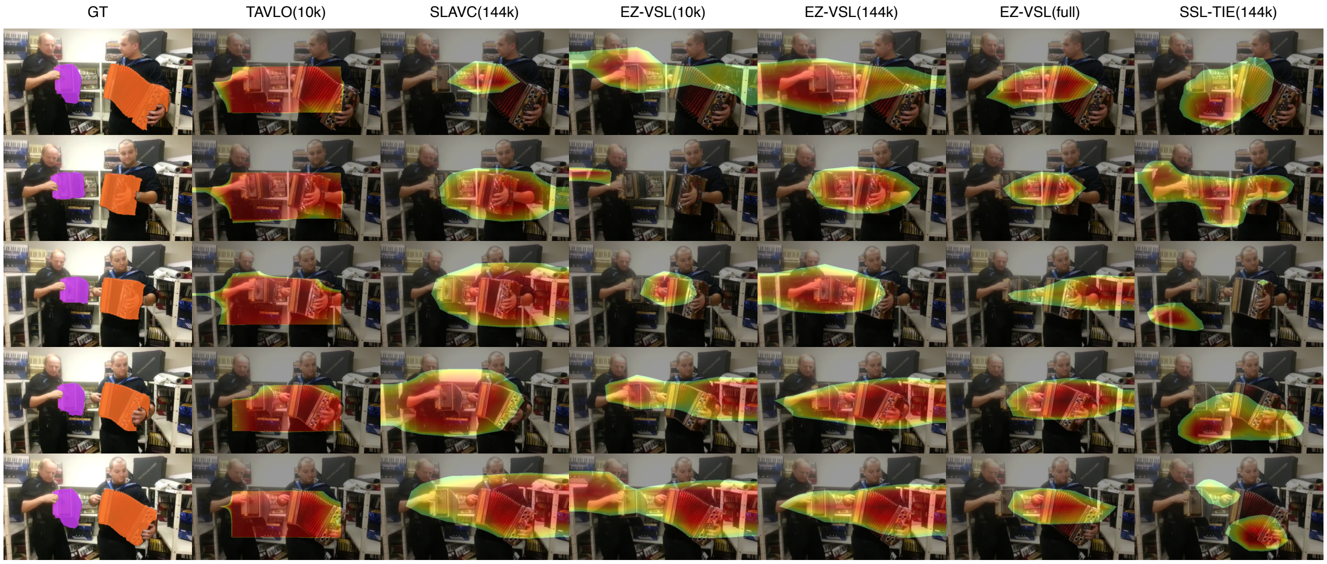


Figure E. Mixed-sound Scenario Visualization Result 1 - Accordion (frame: 44, 131, 167, 203, 258)

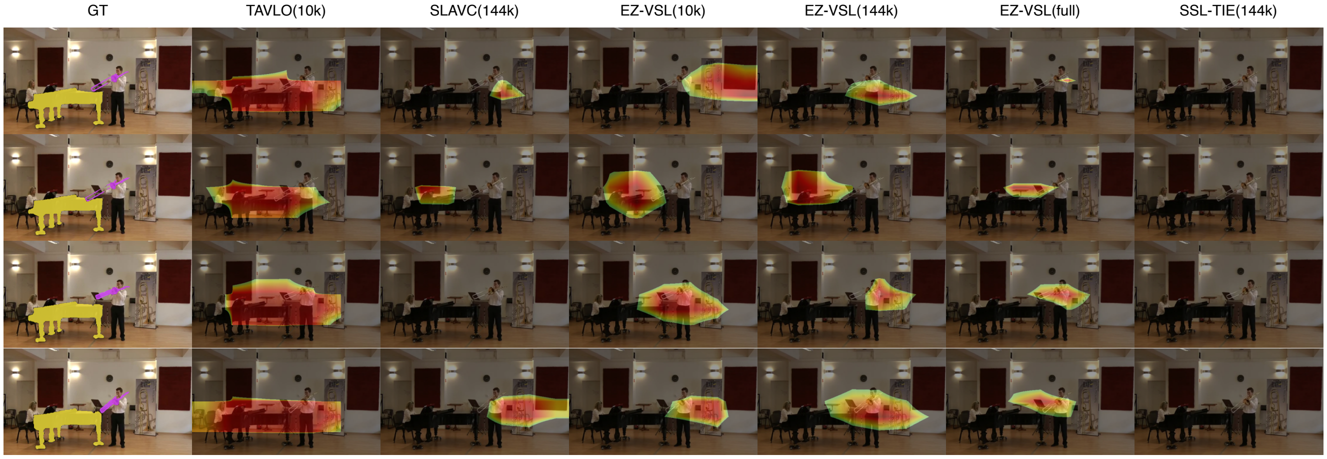


Figure F. Mixed-sound Scenario Visualization Result 2 - Piano and Trombone (frame: 28, 90, 202, 234)

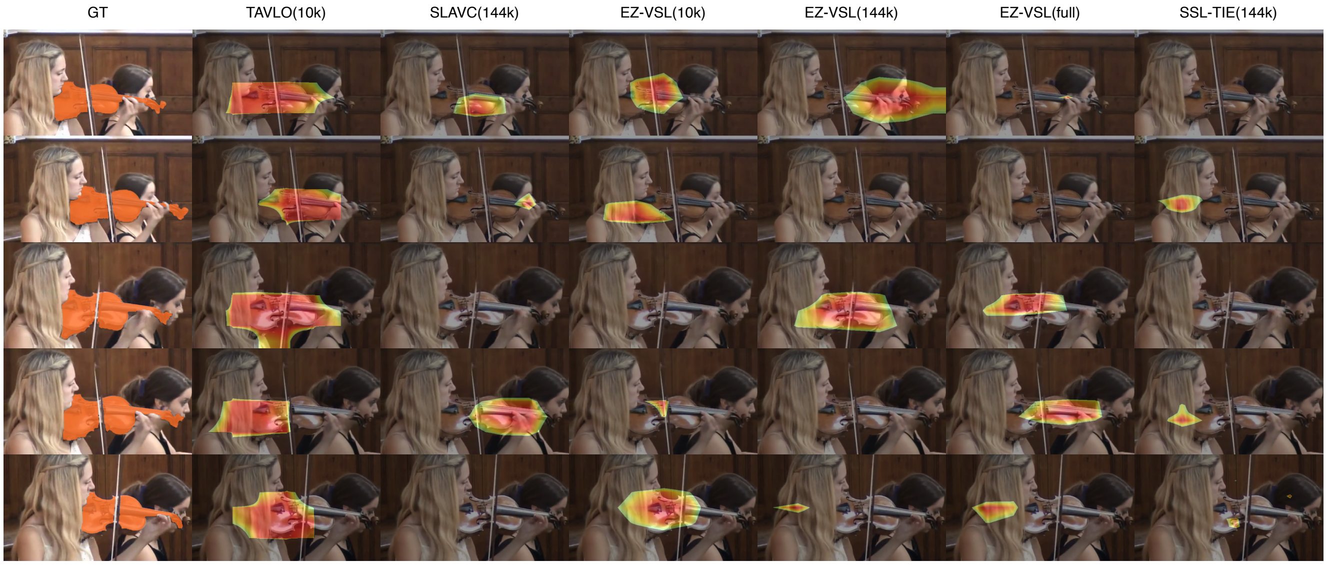


Figure G. Mixed-sound Scenario Visualization Result 3 - Violin and Off-screen Piano (frame: 22, 59, 139, 168, 230)

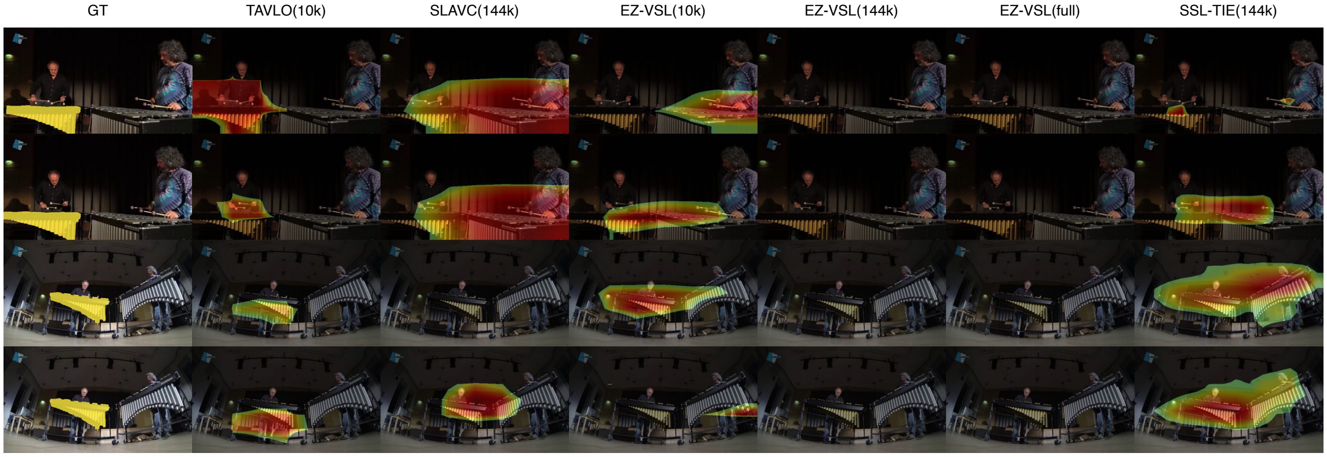


Figure H. Multi-entity Scenario Visualization Result 1 - Vibraphone (frame: 8, 54, 143, 229)

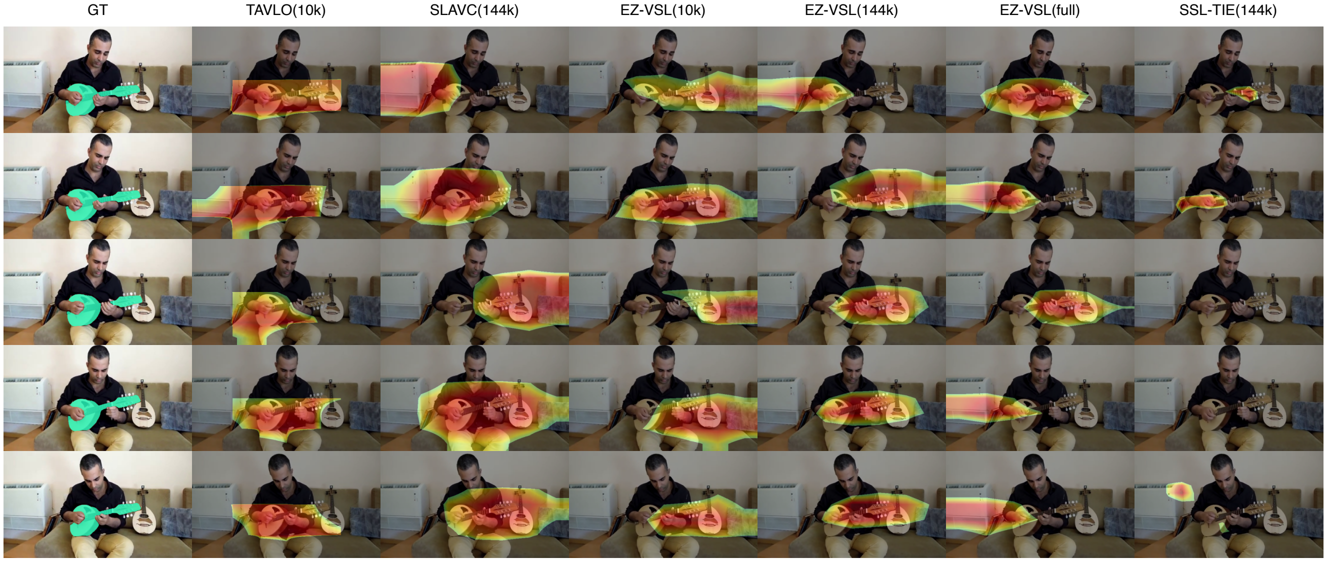


Figure I. Multi-entity Scenario Visualization Result 2 - Mandolin (frame: 40, 82, 174, 225, 290)

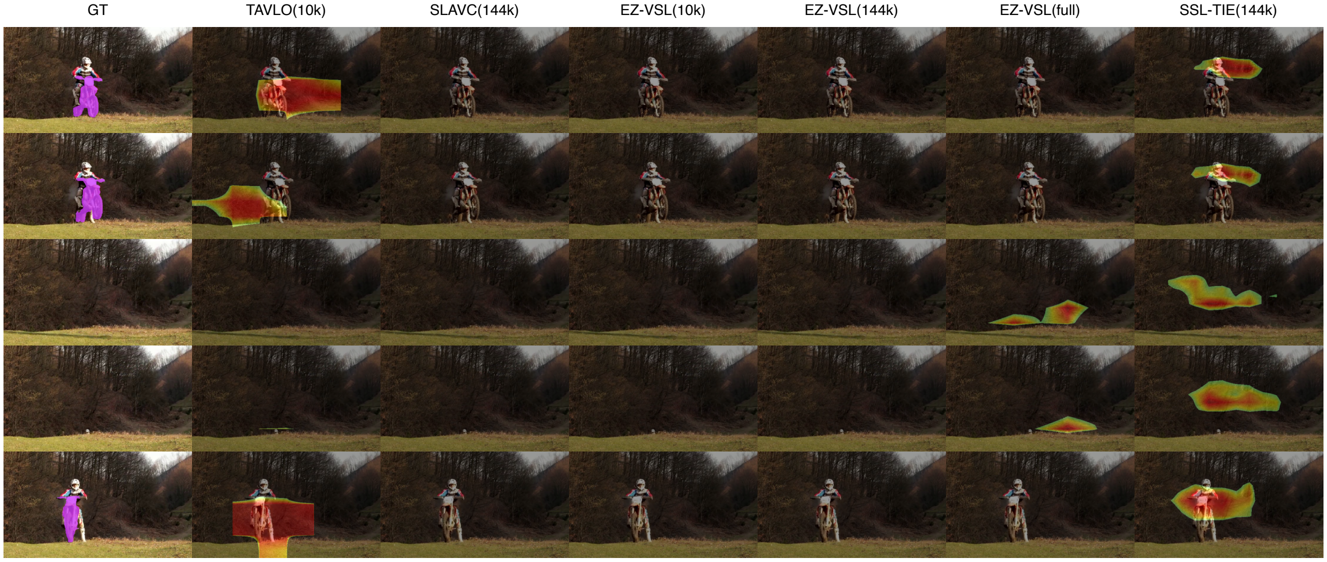


Figure J. Off-screen Scenario Visualization Result 1 - Motorcycle (frame: 40, 64, 120, 166, 276)

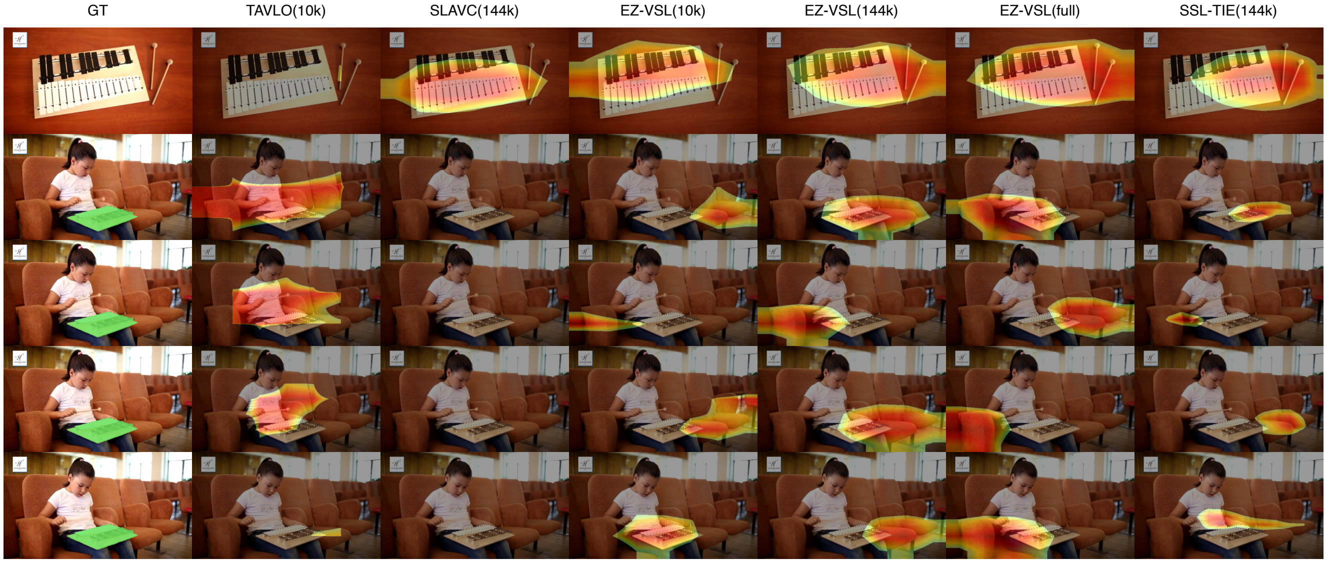


Figure K. Cross-event Subset Visualization Result 1 - Female speech and Glockenspiel (frame: 3, 104, 130, 153, 236)

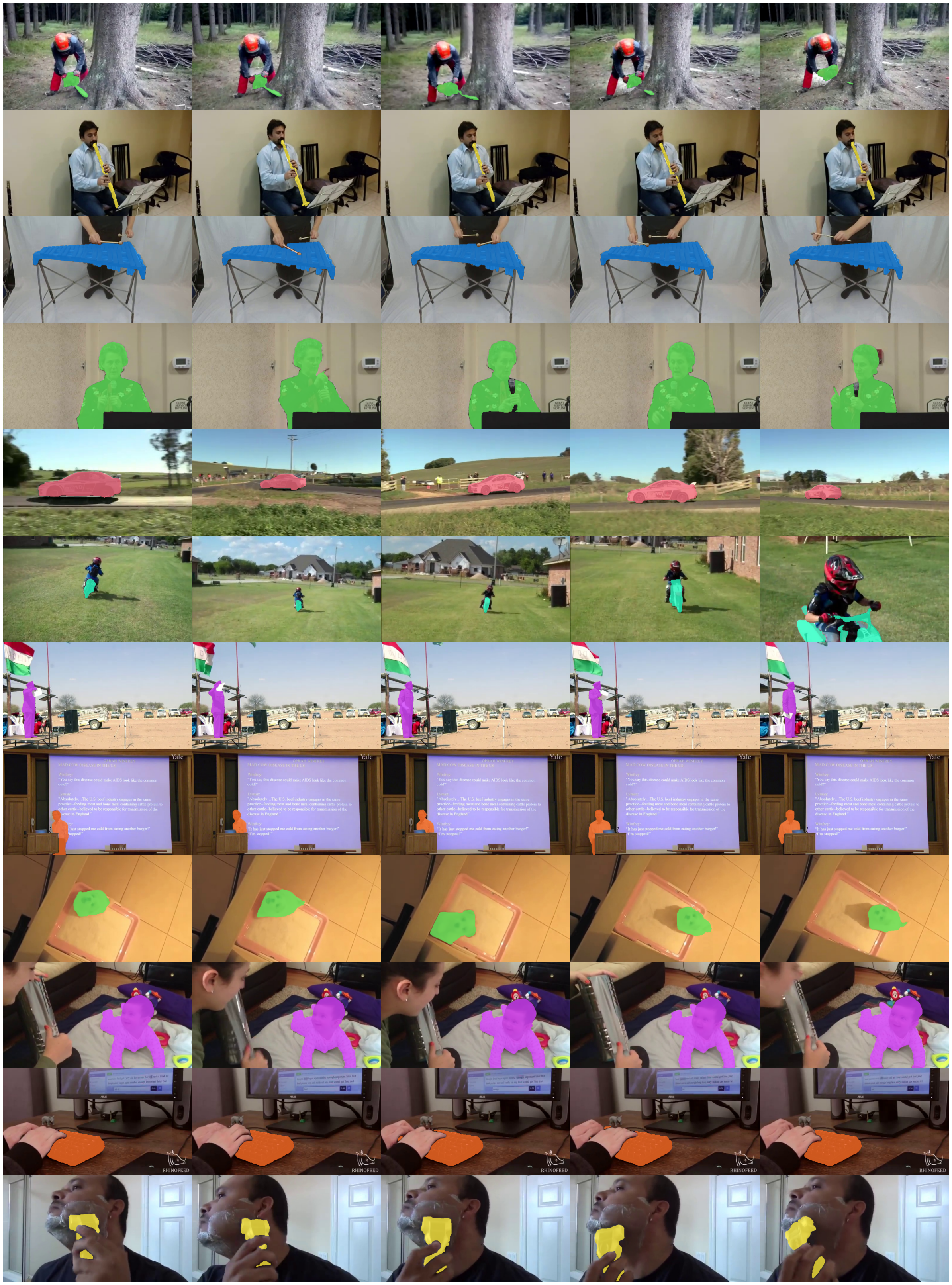


Figure L. Single-sound Scenario Examples

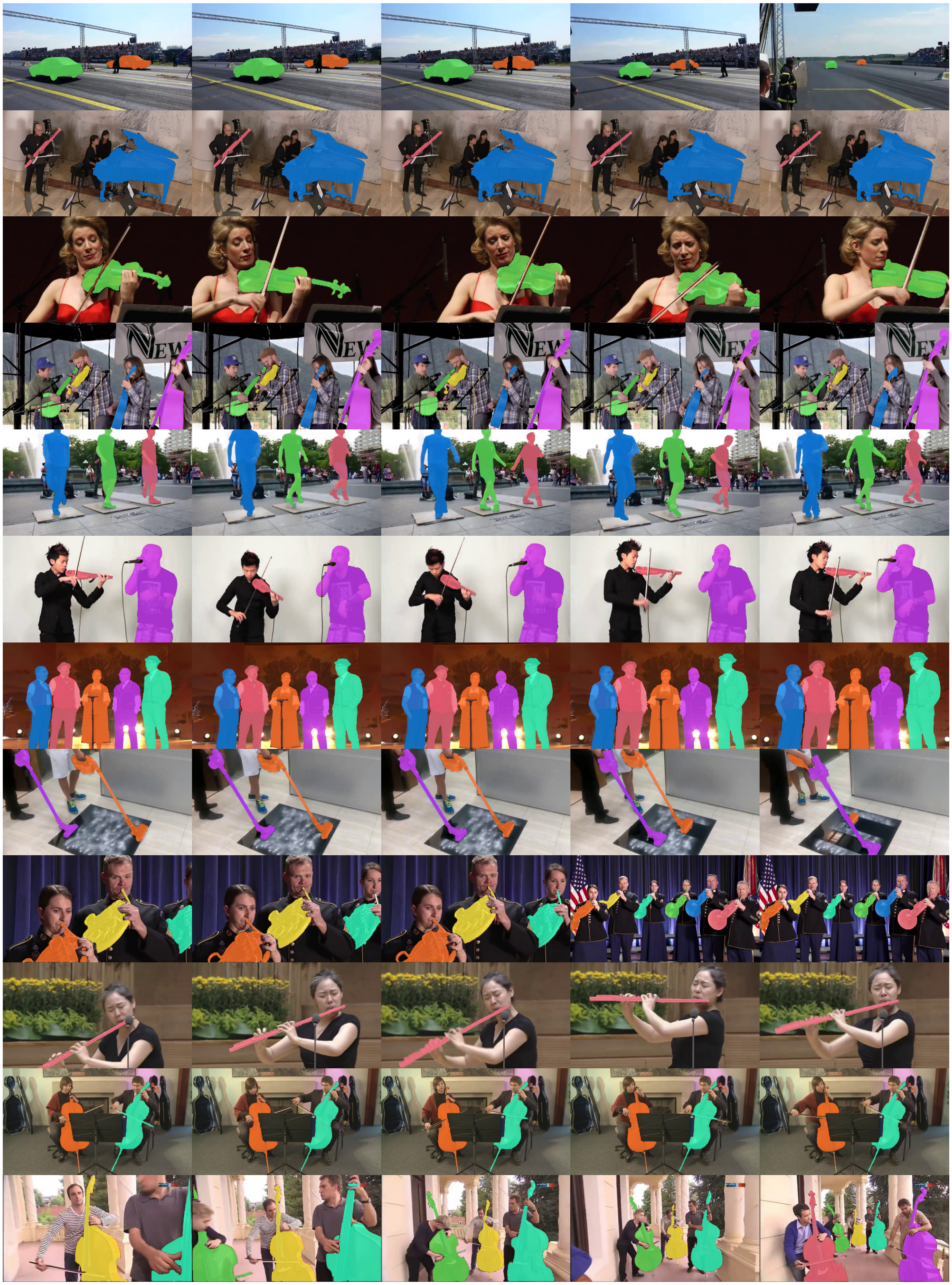


Figure M. Mixed-sound Scenario Examples

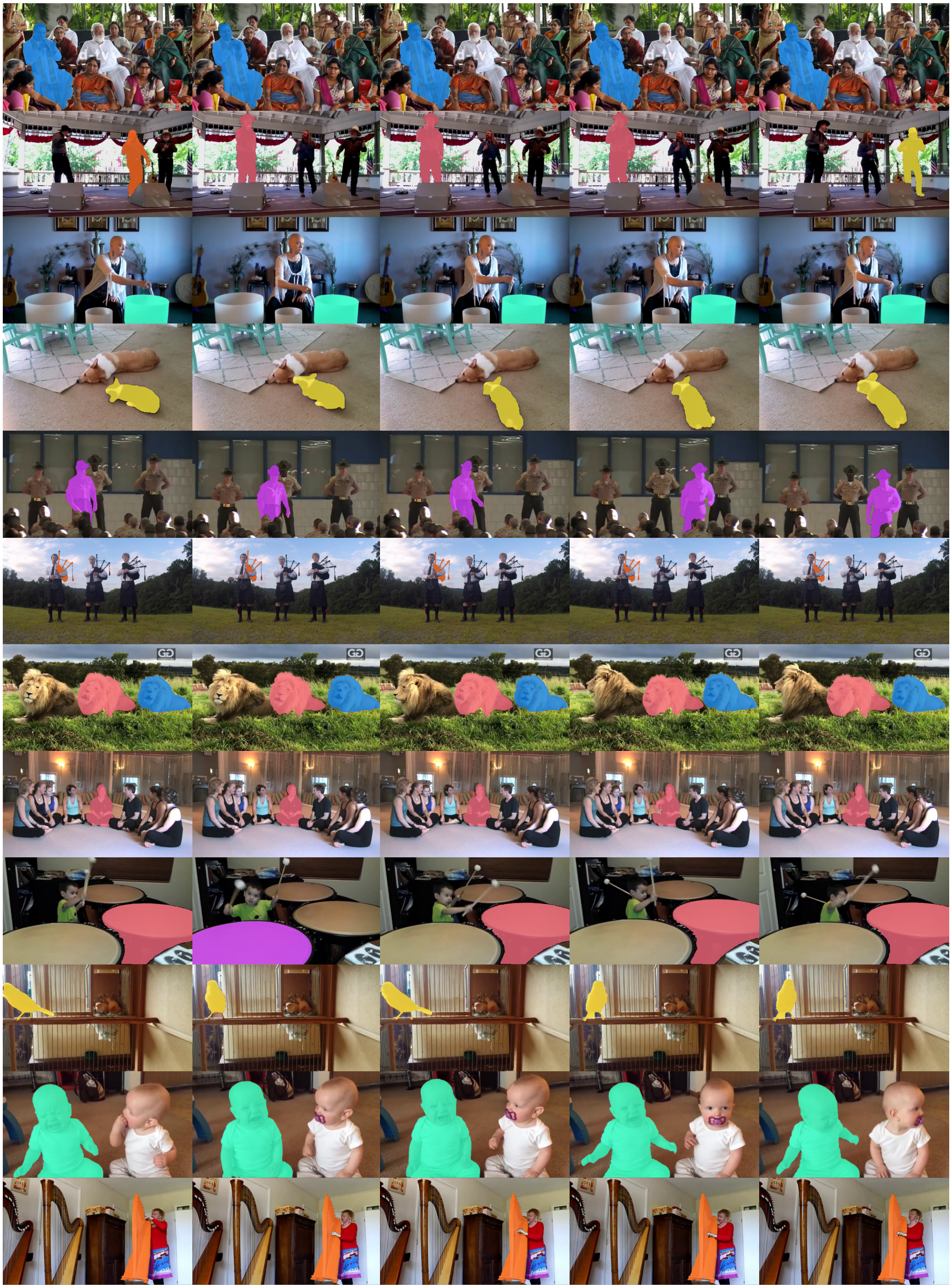


Figure N. Multi-entity Scenario Examples

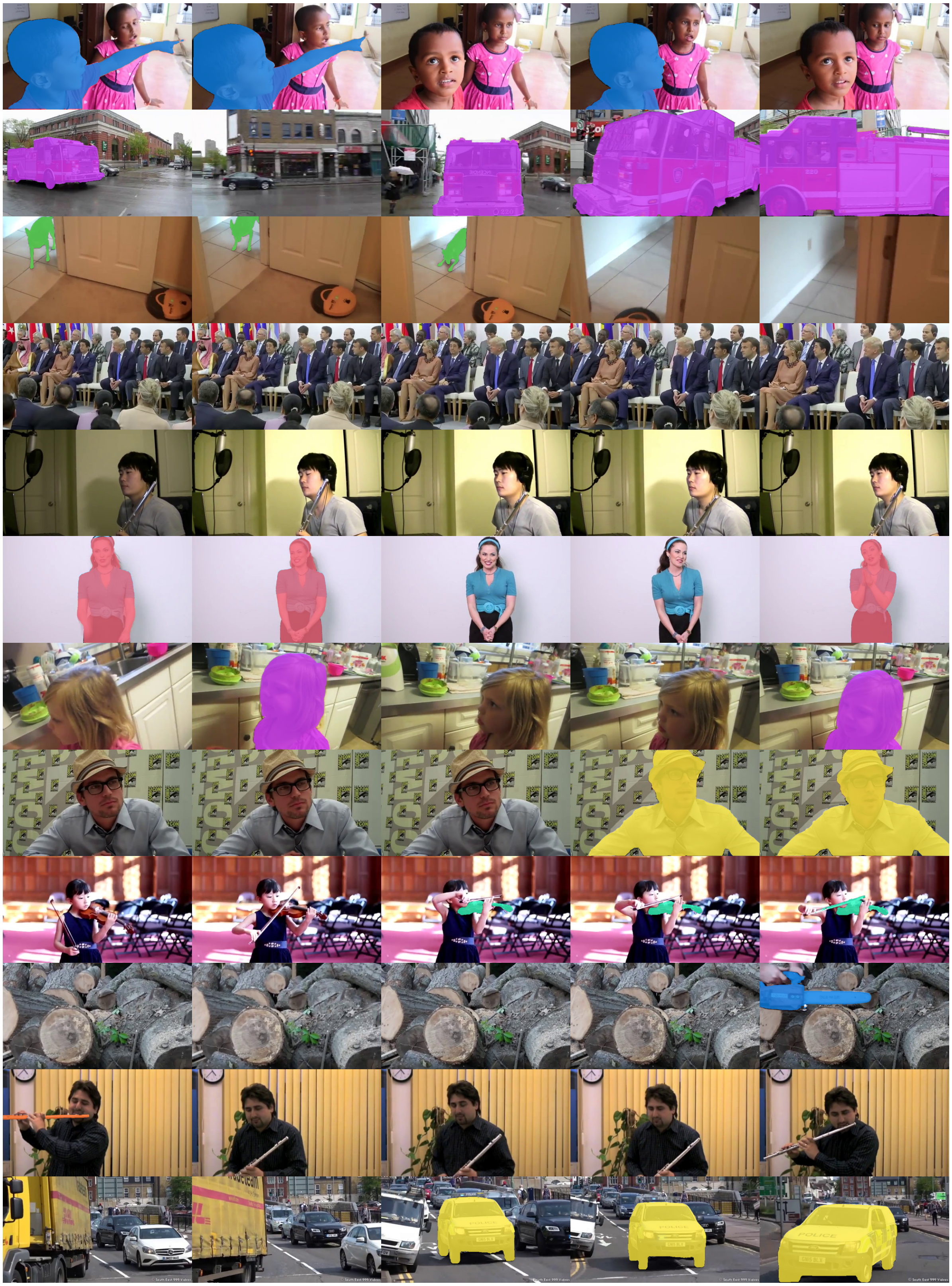


Figure O. Off-screen Scenario Examples

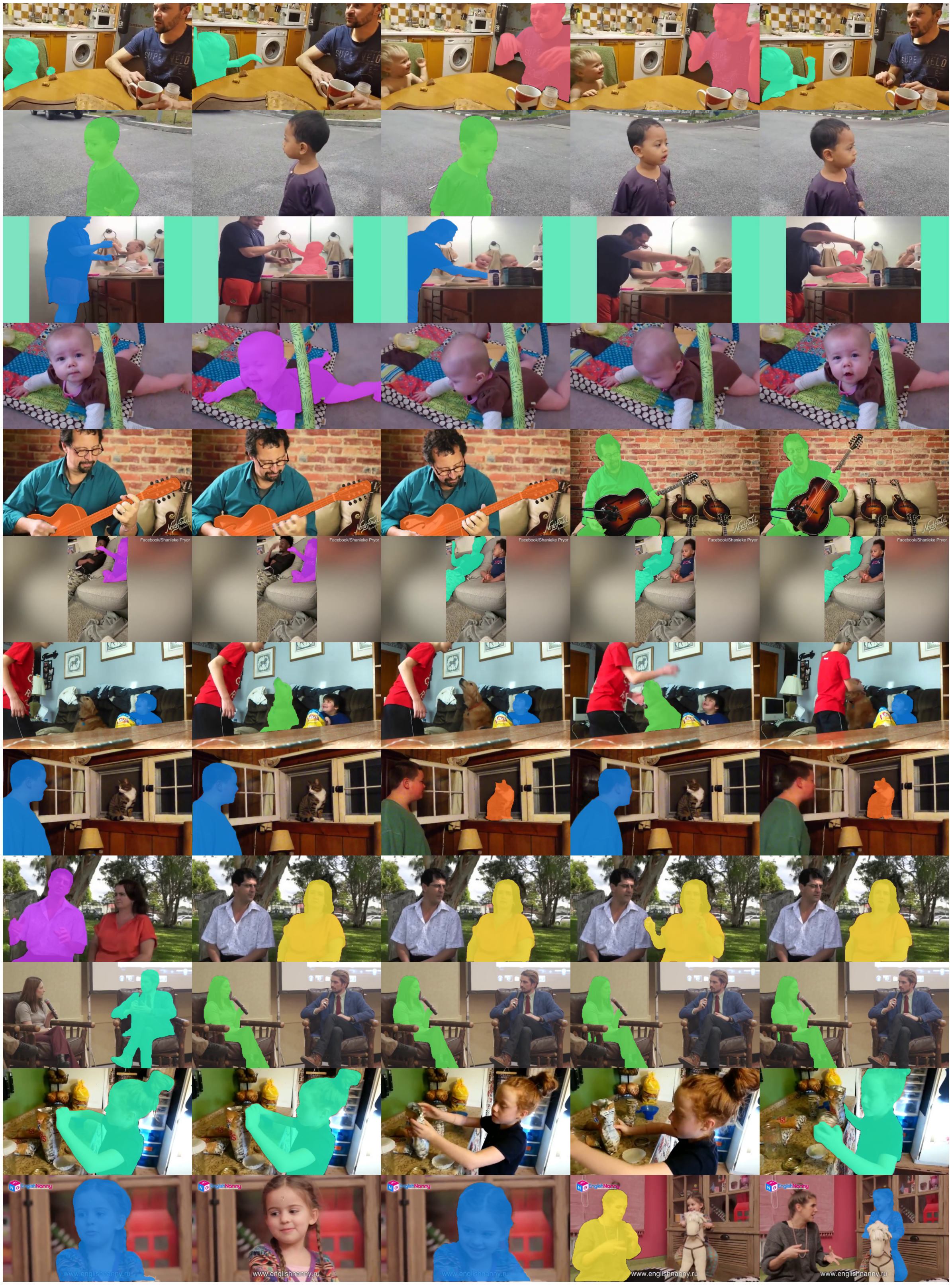
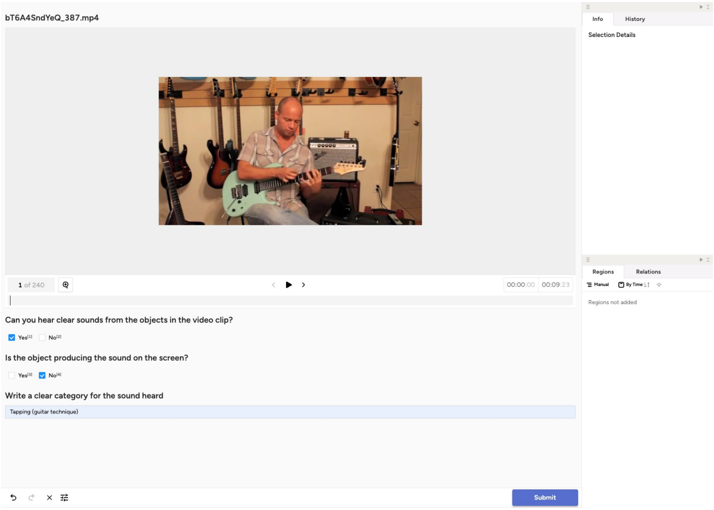


Figure P. Cross-event Subset Examples



bT6A4SndYeQ_387.mp4

Can you hear clear sounds from the objects in the video clip?

☑ Yes[1] □ No[2]

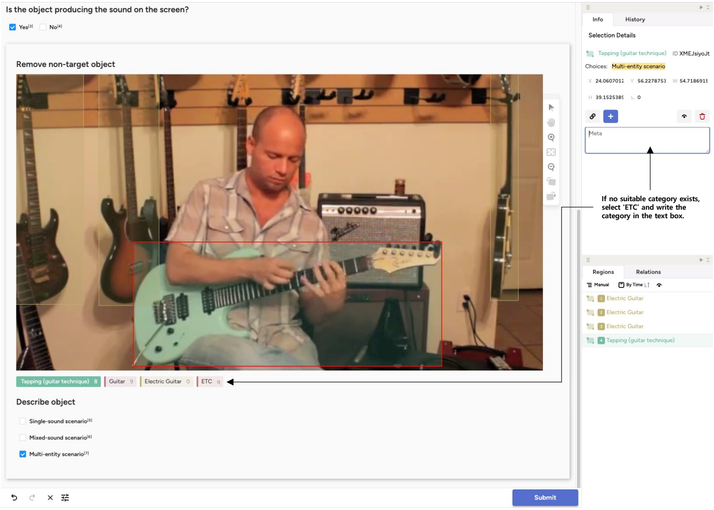
Is the object producing the sound on the screen?

□ Yes[3] ☑ No[4]

Write a clear category for the sound heard

Tapping (guitar technique)

Figure Q. Interface for annotating Off-screen scenarios in the Bounding-box Annotation stage



Is the object producing the sound on the screen?

☑ Yes[3] □ No[4]

Remove non-target object

Tapping (guitar technique) 8 | Guitar 9 | Electric Guitar 0 | ETC q

Describe object

□ Single-sound scenario[5]

□ Mixed-sound scenario[6]

☑ Multi-entity scenario[7]

Selection Details

Tapping (guitar technique) ID XMEJsiyoJt

Choices: Multi-entity scenario

X 24.0607012 Y 56.2278753 W 54.7186915

H 39.1525389 ∟ 0

Meta

If no suitable category exists, select 'ETC' and write the category in the text box.

Figure R. Interface for verifying and assigning scenarios to bounding boxes in the Bounding-box Annotation stage

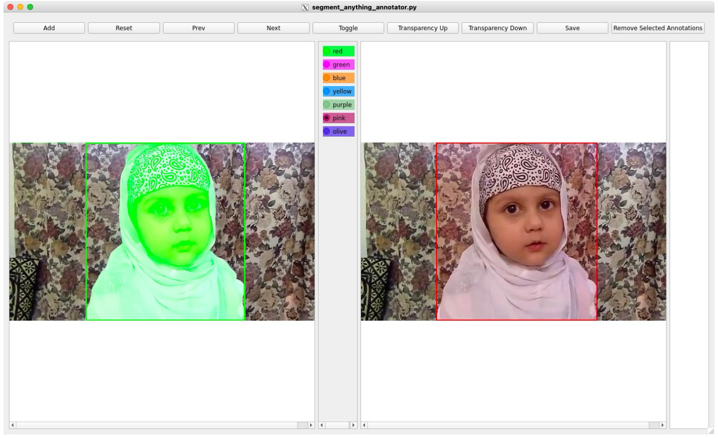


Figure S. Interface for reviewing and refining segmentation masks in the Instance Segmentation stage

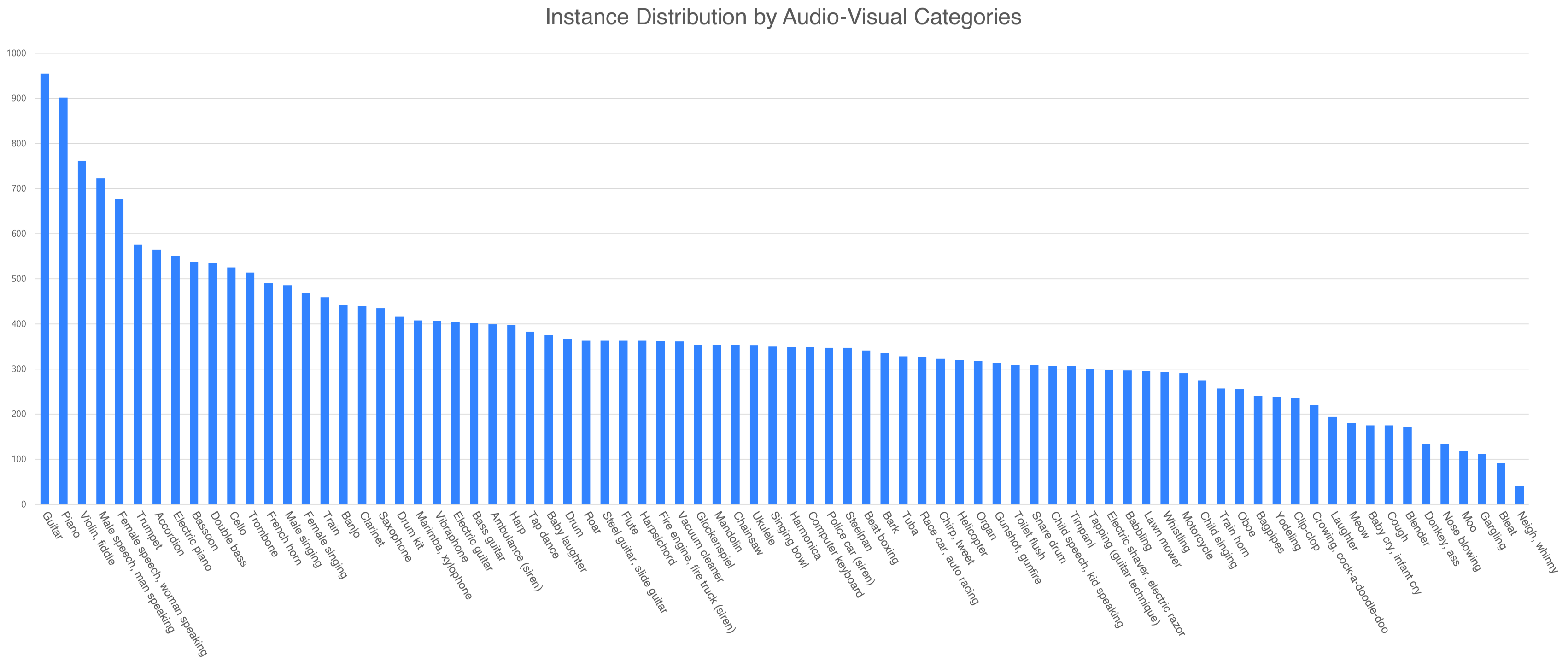


Figure T. Distribution of Instances across Audio-Visual Categories in the AVATAR Benchmark. Each bar represents the total number of instances associated with a specific audio-visual category.