Clink! Chop! Thud! – Learning Object Sounds from Real-World Interactions

Supplementary Material

8. Additional sounding object detection qualitative results

We provide additional qualitative results for sounding object detection on Epic Kitchens in Fig. 10 and Ego4D in Fig. 11.

9. Additional visual clustering results

Additional examples of visual embedding clusters from Ego4D are shown in Fig. 8. In each case, cluster examples correspond to scenarios with similar sounds.

10. Similarity map color legend

Fig. 9 shows the color scale used for visualizing the similarity maps for sounding object detection. Darker colors correspond to the extremes with blue being a low value and red being a high value. Since similarity is calculated using cosine similarity between the vision and audio embeddings, the scores are in the range $[0, 1]$ with a higher value depicting greater correspondence.

11. Sounding object detection annotation

We use Labelbox [23] to develop our annotation interface, shown in Fig. 7. First, annotators answer whether the action is sounding or not. If the action is not sounding, then no further labels are required. If the action is sounding, they then label the locations of the two objects involved by placing keypoints to track the objects across multiple frames. The objects are also labelled with either a noun present in the provided narration or a text input field is provided for the annotators to describe the objects and optionally provide a more descriptive narration.

12. Additional implementation details

We use a learning rate of 5e-5 and 4 video frames per clip during pretraining. Each frame is 224×224 and we use a patch size of 16×16. During training, the 4 frames are sampled randomly. During sounding action discovery evaluation, the 4 frames are sampled uniformly. Meanwhile, sounding object detection uses 1 frame sampled from the middle of the clip.

For the audio encoder, we use AST [17] pretrained on ImageNet [32]. The input to the audio encoder are fbank features calculated on the waveform using 128 Mel frequency bins, 10ms frame shift, and a Hanning window. We use a sample rate of 16kHz.

For the language encoder, we use the pretrained CLIP model from Huggingface, specifically "openai/clip-vit-base-patch32", which we keep frozen.

For our visual encoder, we initialize from a pretrained slot attention model trained on MS COCO 2017 [24] from [21] that uses 7 slots. We keep the encoder and slot embeddings of the slot attention encoder frozen and train the decoder weights.

We project all modalities into a common 256-dimensional embedding space. We use video clips that are 1.5s long, which was found to be the ideal length in [6]. Given the timestamp of the narration, we extract 0.5s from before and 1s after.

Finally, we use a confidence score threshold of 0.35 for OWLv2 [27] when detecting object candidates in a scene for sounding object detection.

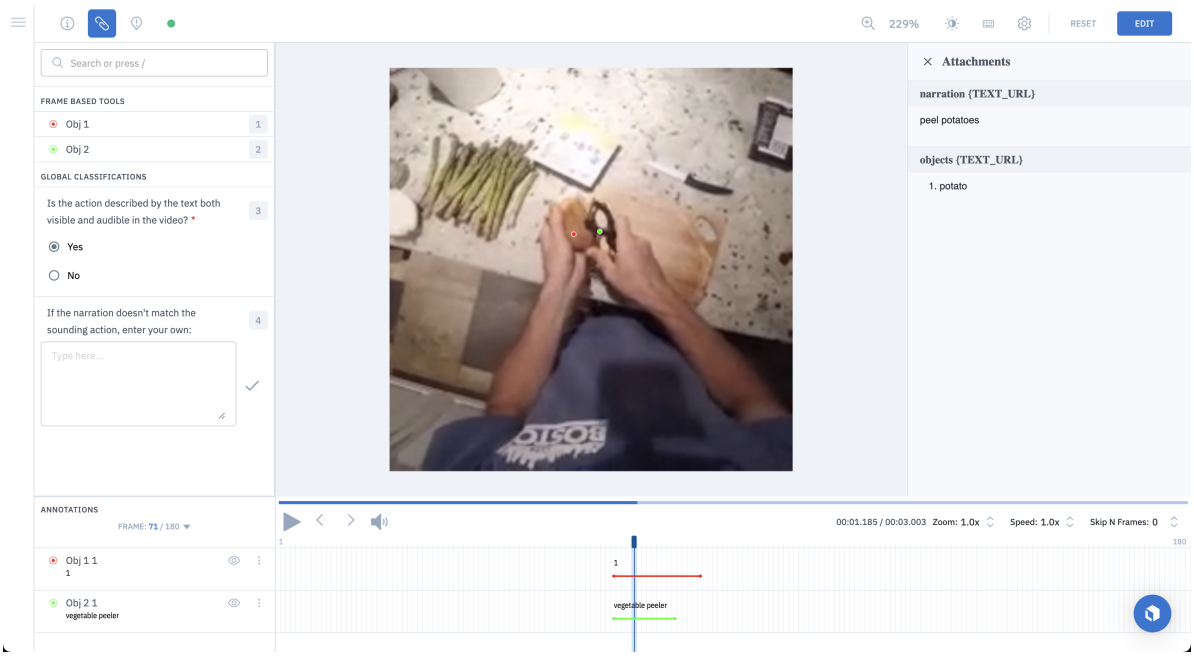

Figure 7. Screenshot of the Labelbox [23] interface used to annotate ground truth object masks for our sounding object detection benchmark. In addition to answering the questions in the left column, annotators can scrub through individual frames and apply keypoints to the objects involved in the action. These keypoints are then used with SAM 2 [31] to extract ground truth object masks.

(a) Visual embeddings that correspond to sounds of food sizzling.

(b) Visual embeddings that correspond to sounds of plants rustling.

Figure 8. Additional visual embedding clustering results. Each cluster shown corresponds to visual frames with diverse perspectives and backgrounds. But the common trait is all corresponding sounds belong to the same category.

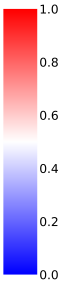

Figure 9. Colorbar legend used to visualize object region scores in sounding object detection.

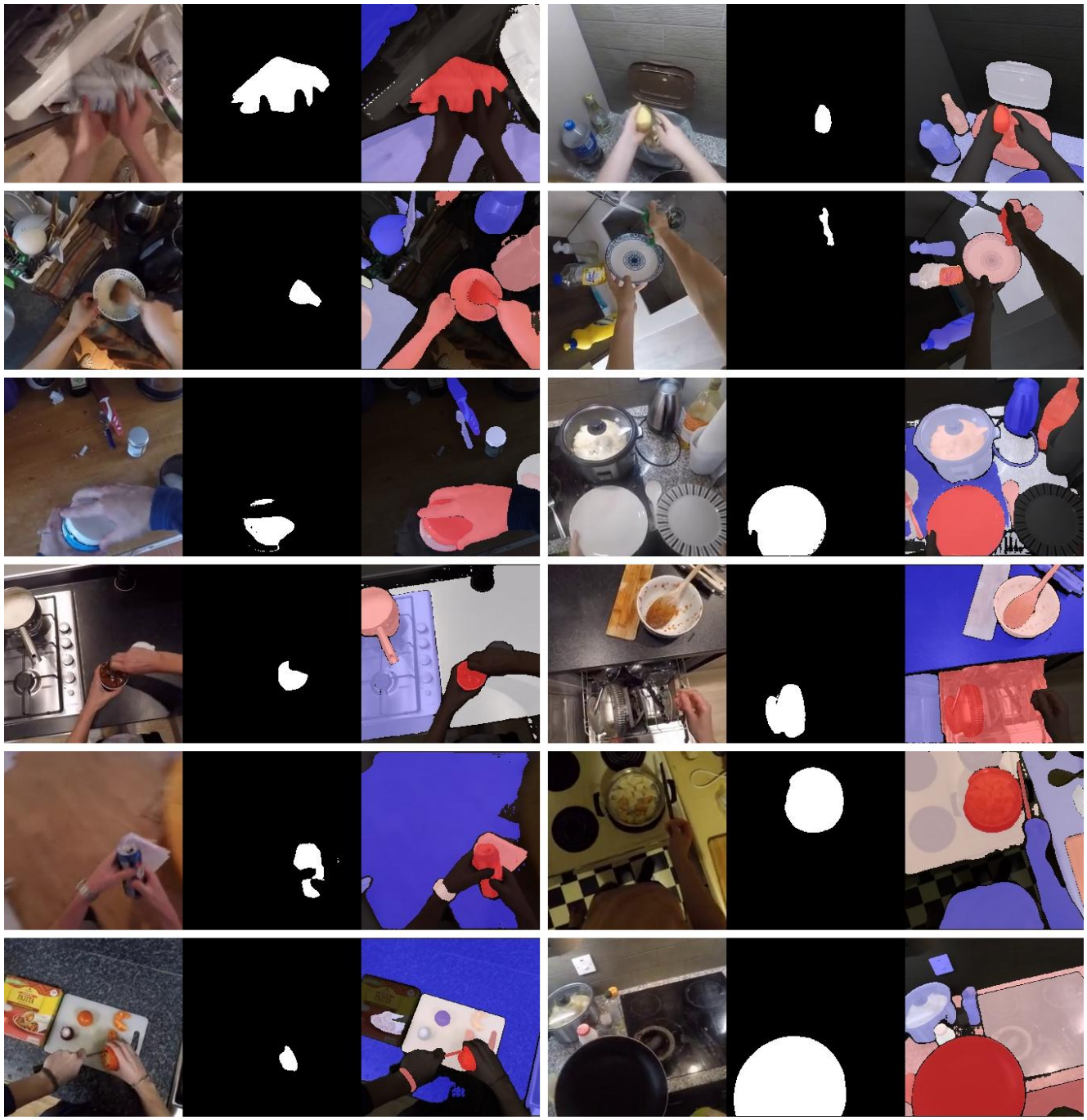

Figure 10. Additional qualitative results for sounding object detection on Epic Kitchens.

Figure 11. Additional qualitative results for sounding object detection on Ego4D.