SAM2-LOVE: Segment Anything Model 2 in Language-aided Audio-Visual Scenes

Supplementary Material

A. Case Study

In this section, we will show many cases to demonstrate the robustness of our model to different categories and expressions. As shown in Figure 6, the first case shows the model is robust to the text that contains a negative expression, which requires the model to understand the meaning of the visual element referred to by the text. The $[seg]$ token successfully prompts the SAM2 to locate the correct target in the former frames. Although the target is irrelevant to the sounding object and disappears in the last several frames, our model can still output empty masks, showcasing strong temporal consistency. The second case shows that the fusion transformer can effectively assist in locating the true object, which means that the $[seg]$ token contains not only the highly compressed semantics from three modalities but also the accurate positional information. This demonstrates that our proposed token propagation and accumulation strategies can retain spatial consistency despite multi-view motorcycle changes. In the third case, we can observe that the fusion transformer also shows a good understanding of the sequence order corresponding to the audio hidden in the text.

B. Explanation of Token behavior

First, we introduce two concepts: **forward knowledge transfer** and **backward knowledge transfer**. The goal of forward knowledge transfer is to leverage prior knowledge to facilitate the learning of a new task, whereas the target of backward knowledge transfer is to utilize the newly acquired knowledge to improve performance on prior tasks. Building on this concept, we conceptualize token propagation as a form of forward transfer, where knowledge from prior frames is utilized to facilitate the learning of subsequent frames, while token accumulation is interpreted as a form of backward knowledge transfer, where the acquired knowledge can be adapted to the previous frame by continuously accumulating and replaying prior tokens, preventing the forgetting of the model. Thus, we contend that while these two strategies are formally distinct, they are conceptually unified, as both aim to transfer knowledge to all frames.

C. Backbone Model Details

In this section, we present a detailed description of the model architecture. To balance computational efficiency and resource constraints, the text encoder is a distilled variant of the RoBERTa-base model, trained following the same procedure as DistilBERT. This model is case-sensitive, distinguishing between lowercase and uppercase words (e.g., english vs. English). It comprises 6 layers, a hidden dimension of 768, and 12 attention heads, resulting in a total of 82 million parameters, compared to 125 million in RoBERTa-base. Notably, DistilRoBERTa achieves approximately twice the inference speed of RoBERTa-base. The Vision Transformer (ViT) model utilized in our work is a base-sized version pre-trained on ImageNet-21k, which includes 14 million images across 21,843 classes, at a resolution of 224x224. Subsequently, it was fine-tuned on the ImageNet-2012 dataset, comprising 1 million images and 1,000 classes, at the same resolution. We directly use the pre-trained model provided on HuggingFace. The VGGish model employed in our work was pre-trained on AudioSet. Features are extracted from the pre-classification layer following activation, resulting in a 128-dimensional feature tensor corresponding to 0.96 seconds of the original video. The pre-trained model checkpoint was obtained directly from the $torchaudio$ library.

RoBERTa Size	Seen (%)			Unseen (%)		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
DitillRoBERTa (82M)	42.6	50.9	46.7	66.3	71.6	69.0
RoBERTa-base (125M)	**42.7**	**51.1**	**46.9**	**66.6**	**72.0**	**69.3**

Table 7. Ablations on the different versions of RoBERTa.

Comparison of RoBERTa size. As shown in Table 7, two versions of RoBERTa achieve comparable performance, and the improvement of RoBERTa-base can be attributed to the additional approximately 40M parameters. Since our main contributions do not focus on the backbone design and consider efficiency, we chose DistillRoBERTa as our text encoder.

More Implementation Details. Since the video length of Ref-AVS is 10 seconds corresponding to the audio length of 10 seconds, the total frame number is N is 10, with a sample rate of 1 frame per second. For the input of SAM2, We only use a single data augmentation to resize the input size to 1024×1024. For the input of the three modality backbone, we use their predefined transformation of the data to prepare their corresponding inputs. The image input will be downsampled 16 to obtain a latent space of 14×14 corresponding to $h \times w$. The maximum length of the tokenized expression is set to $P = 25$. The audio features with $d_A = 128$ are extracted from the mono-channel waveform.

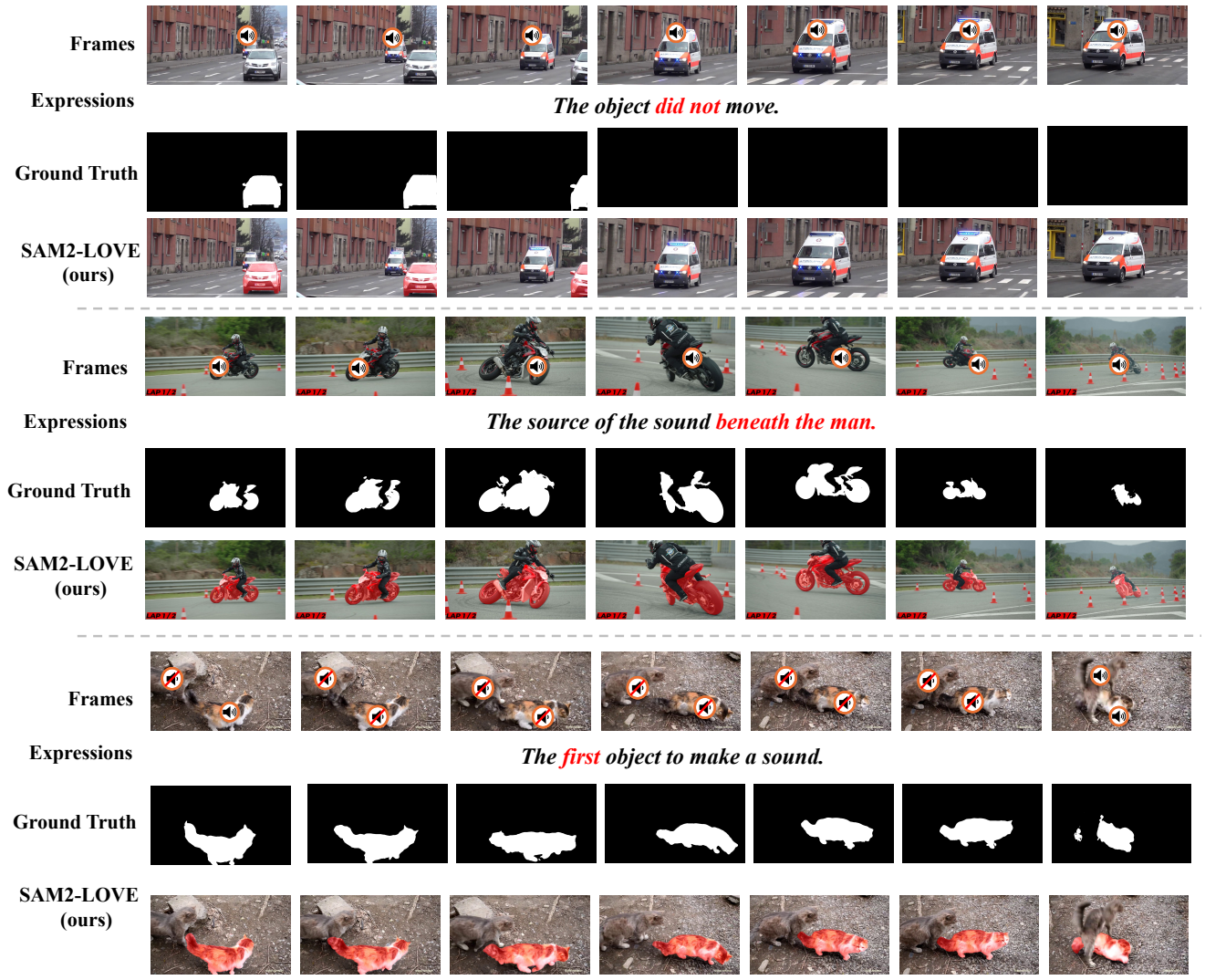

Figure 6. The visualization results of the referred objects in the Ref-AVS. Notably, we use the legends of the trumpet to represent the different sound volumes of the audio objects, from silent to increasing loudness.