Multi-modal Instruction Tuned LLMs with Fine-grained Visual Perception

Supplementary Material

6. Implementation Details

6.1. Network Architecture

The base MLLM of **AnyRef** is based on LLaVA-7B-v1-1, where we freeze the image encoder and visual-language projection layer, and train the LLM using LoRA[13]. For audio input, we adopt the pre-trained audio encoder from ImageBind-H [9], which is also frozen during fine-tuning. The audio-language projection layer, akin to the visual-language projection layer, comprises a single trainable MLP layer. The adopted segmentation model is SAM-H [15], featuring a frozen image encoder and a trainable mask decoder. In line with [16], we utilize a 2-layer MLP to map the mask embedding to the input of the mask decoder.

6.2. Instruction Formulation

In contrast to LISA[16], **AnyRef** is required to generate textual descriptions along with masks. Hence, we must create supplementary question-answer templates distinct from LISA's, which primarily uses answer templates such as "`It is <SEG>`" or "`Sure, it is <SEG>`". Additionally, owing to the introduced *Refocusing Mechanism* (as detailed in Sec. 3.1.2), tokens preceding the `<obj>` token are intended to offer textual guidance, while the phrase "`it is`" does not cover any meaningful information.

For textural referring segmentation datasets (including general semantic/instance segmentation and referring expression segmentation), the question template is "`Can you segment {exp} in this image?`", where `{exp}` represents either the textural class name or the description. And the answer is "`{exp} <obj>.`", indicating that **AnyRef** needs to repeat the grounding description to offer textural guidance for the `<obj>` token, as part of the *Refocusing Mechanism*. We apply the same approach for image-level referring and audio-visual samples, replacing `{exp}` with the textural target class name."

6.3. Human Evaluation on Referring Expression Generation

We follow previous work [50] to conduct human evaluation on generated referring expressions. We randomly sample 100 images with corresponded bounding boxes for each test split, and generate textural expressions using different models. The images, along with the bounding boxes and the generated referring expressions, are presented to five different evaluators. They assess whether the target object could be unambiguously identified based on the provided referring expression; if so, the referring expression is labelled as correct.

7. Qualitative Results

We provide additional visualizations for all feasible tasks within the **AnyRef**, including referring expression segmentation, image-referring segmentation, audio-visual segmentation and region-level referring expression generation. The results are produced from the same general model, without fine-tuning on each single task.

In Fig. 6, we show the referring segmentation results from textural descriptions, comparing the ground-truth, the concurrent model LISA [16] and **AnyRef**. In comparison to LISA, our **AnyRef** excels in identifying the correct referring object and generated more accurate masks corresponding to the ground-truth (*e.g.*, in the 2nd and 5th rows, accurate mask shapes of the urn and giraffe were detected without including the background).

In Fig. 7 and Fig. 8, we present the image-level referring segmentation results without and with corresponding mask annotations. When lacking explicit mask annotations, **AnyRef** demonstrates the capability to comprehend image references and identify similar objects or regions within the queried image. With the inclusion of mask annotations, background noise is eliminated, resulting in clearer identification of referring objects (*e.g.*, in the last row of Fig. 8, **AnyRef** accurately identifies trains solely from the train door), which enhances object recognition and reduces ambiguity. Unlike [27], which utilizes visual prompt engineering (*e.g.*, adjusting background intensity, blurring, or adding outlines) to emphasize objects, we found that using a black background performs adequately in our scenarios. Therefore, we opt not to include additional visual prompting in the image references to maintain simplicity.

In Fig. 9, we show the audio-visual segmentation results on AVSBench [60]. Compared with ground-truth mask annotations, **AnyRef** is able to produce more precises masks.

In Fig. 10, we show the region-level referring expression generation results between **AnyRef** and two recent region-aware LMMs, Shikra [4] and KOSMOS-2 [31]. Thank to the powerful LLM pre-trained from vision instructions (where **AnyRef** is initialized from [23]), we suprisingly found that it still keeps the OCR ability (*e.g.*, it correctly recognizes "Happy halloween" on the shirt in the last row of Fig. 10). Moreover, **AnyRef** has the capability to generate grounding masks corresponding to bounding boxes simultaneously, towards more fine-grained object perceptions.

8. Discussions

We further investigate whether using multiple references from different modalities improves performances, *e.g.*, providing textural descriptions and audio inputs or image-level references simultaneously. The prompt is built like: "`Can you segment {exp} with sound of <aud_ref><aud_feat></aud_ref> in this image?`" However, we obverse no performance improvement. We speculate that due to the LLM's robust understanding and reasoning abilities, alongside its capacity for multi-modal feature alignments, it has the capability to transfer prompts from various modalities into the textual space. Therefore, repeatedly prompting the same concepts may not enhance its performance but could potentially confuse the LLM." However, this area remains relatively unexplored and requires further investigation, which goes beyond the scope of this paper. We hope to delve deeper into it in future research.

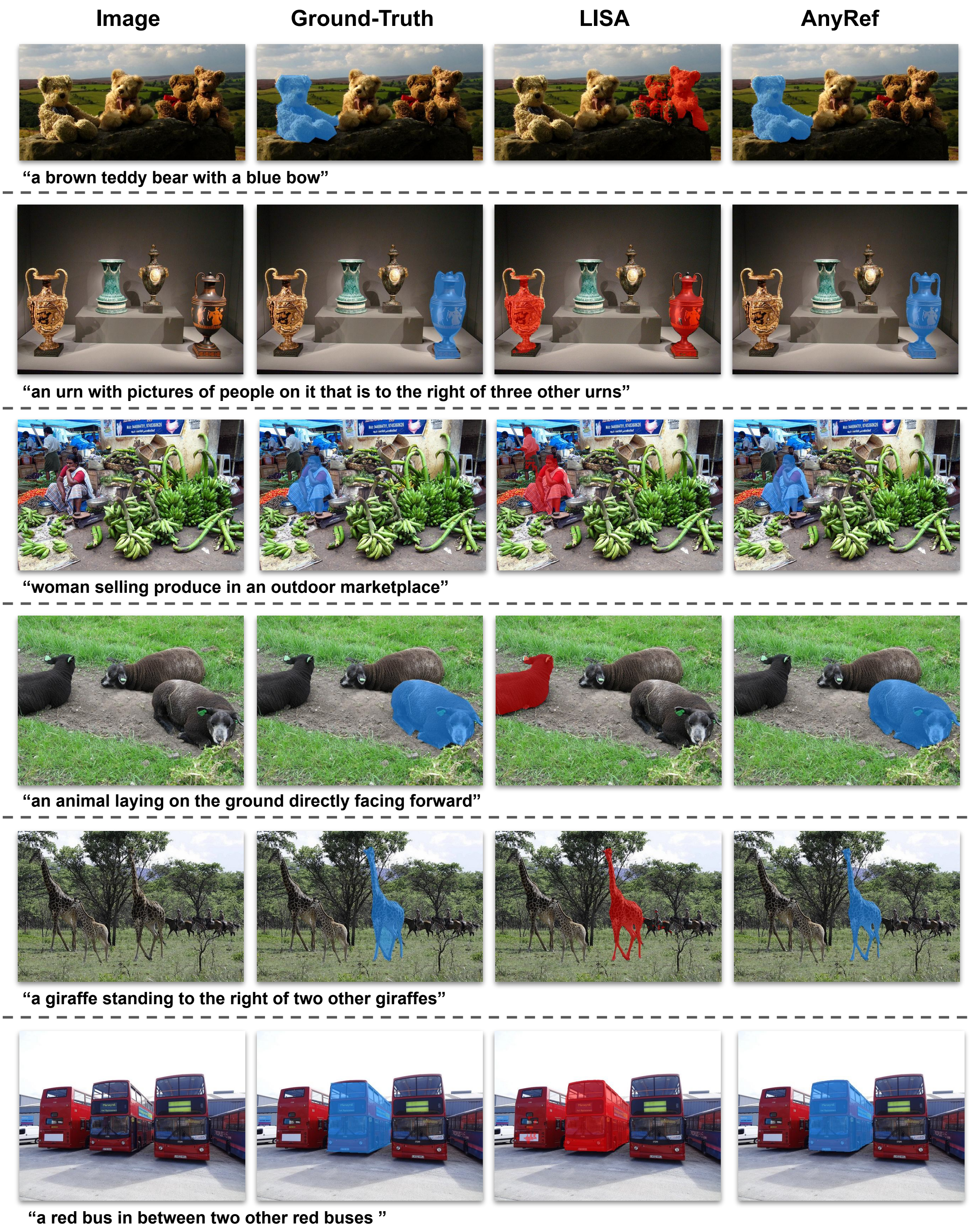


Image
Ground-Truth
LISA
AnyRef
“a brown teddy bear with a blue bow”
“an urn with pictures of people on it that is to the right of three other urns”
“woman selling produce in an outdoor marketplace”
“an animal laying on the ground directly facing forward”
“a giraffe standing to the right of two other giraffes”
“a red bus in between two other red buses ”

Figure 6. Qualitative results of **Referring Expression Segmentation**. The text below indicates the textural referring expression. LISA [16] occasionally fails to identify the correct referring object, whereas **AnyRef** can produce more precise masks compared to the ground-truth (*e.g.*, urn in the 2nd row and giraffe in the 5th row).

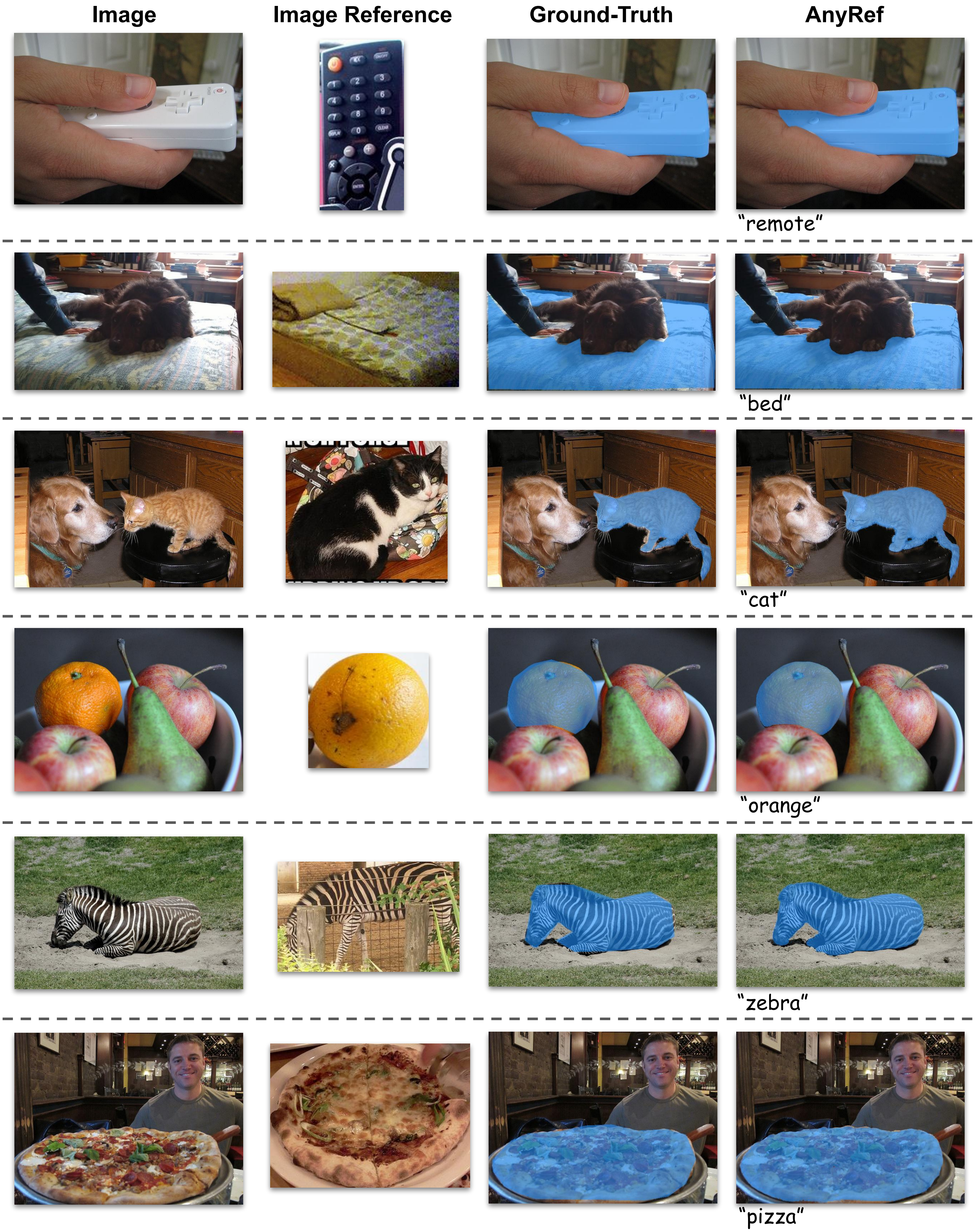


Figure 7. Qualitative results of **Image Referring Segmentation**.

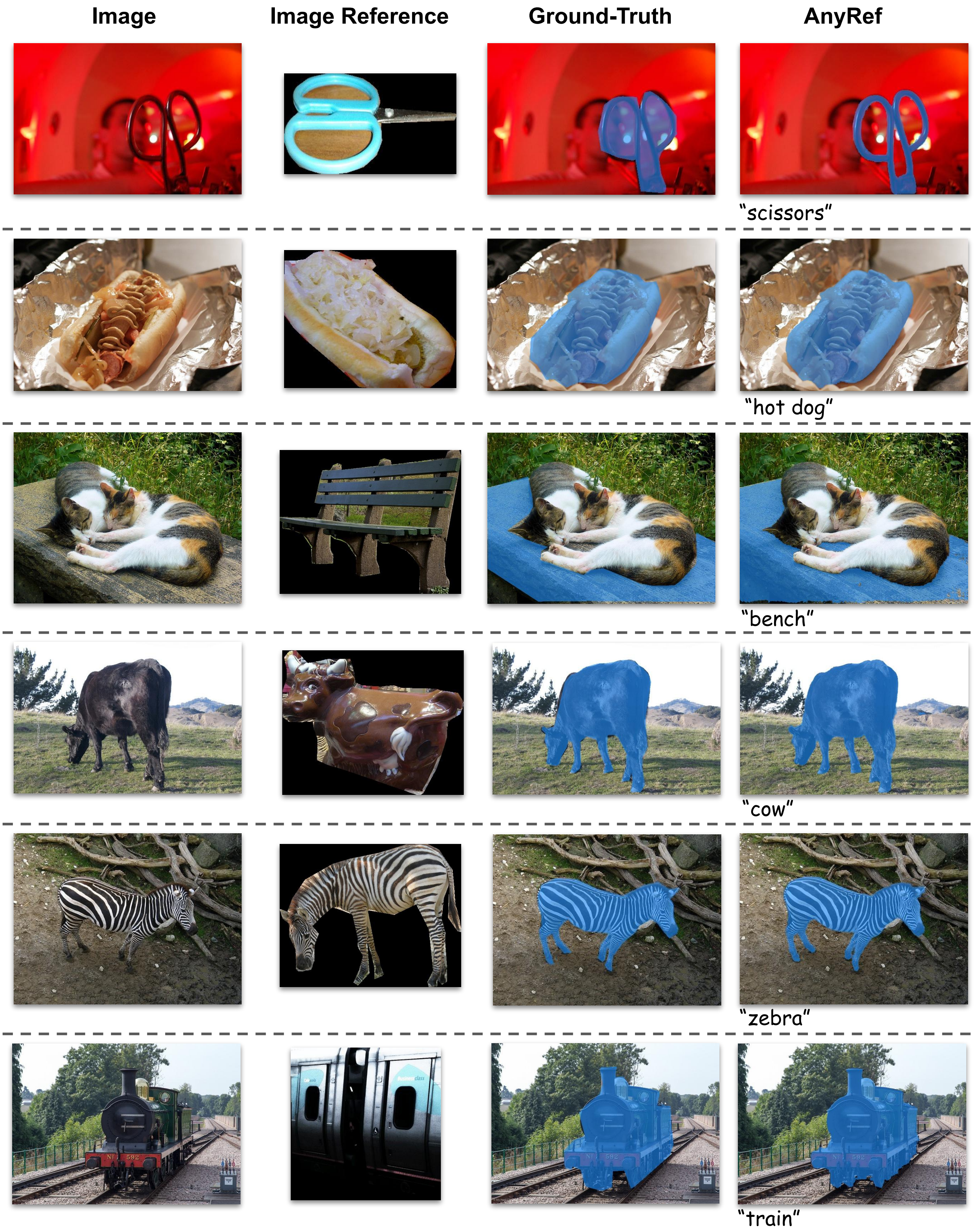


Figure 8. Qualitative results of **Image Referring Segmentation with masks**.

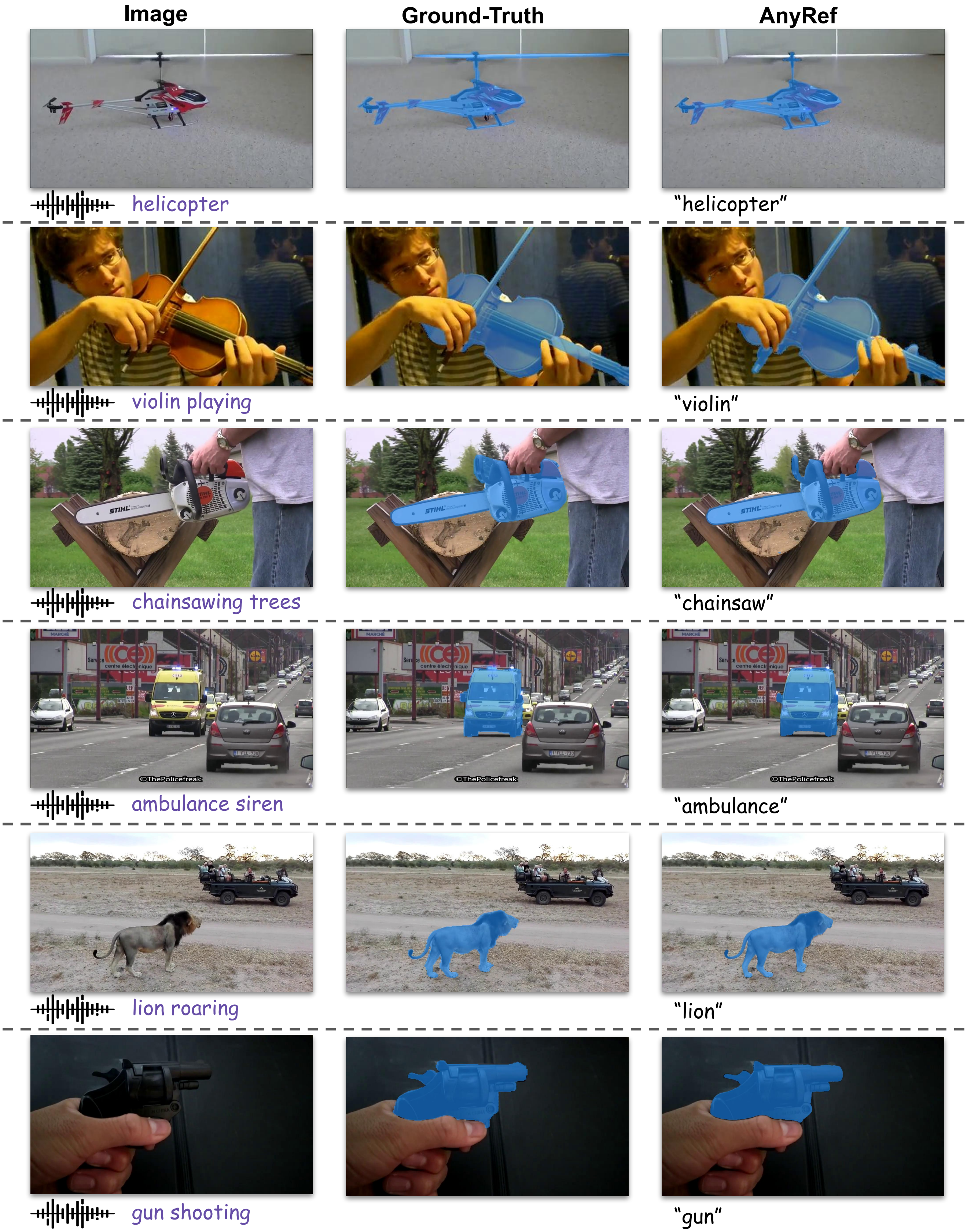
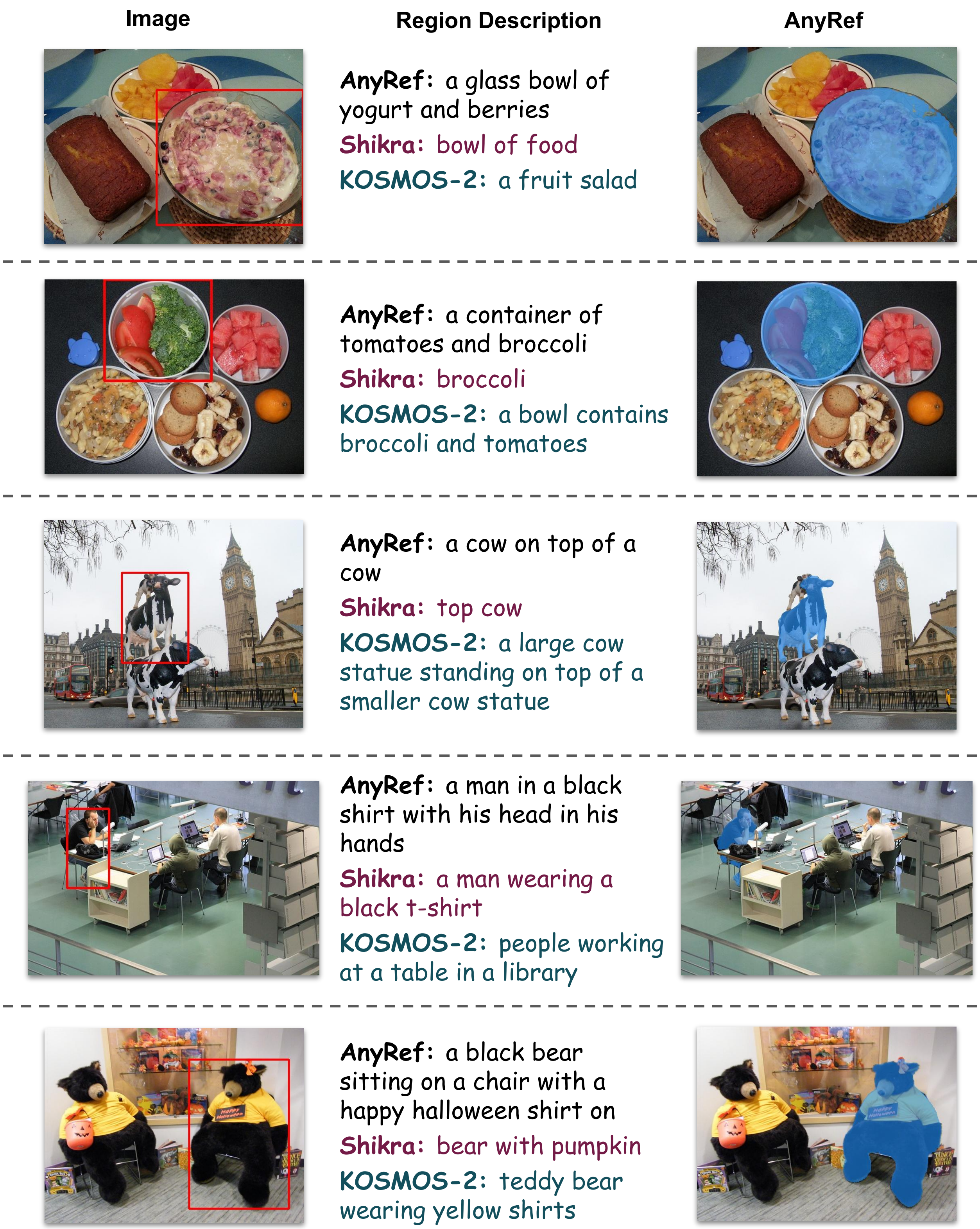


Figure 9. Qualitative results of **Audio-Visual Segmentation**. Furthermore, **AnyRef** can also generate textual class name of the target sounding object based on the audio input.



AnyRef: a glass bowl of yogurt and berries

Shikra: bowl of food

KOSMOS-2: a fruit salad

AnyRef: a container of tomatoes and broccoli

Shikra: broccoli

KOSMOS-2: a bowl contains broccoli and tomatoes

AnyRef: a cow on top of a cow

Shikra: top cow

KOSMOS-2: a large cow statue standing on top of a smaller cow statue

AnyRef: a man in a black shirt with his head in his hands

Shikra: a man wearing a black t-shirt

KOSMOS-2: people working at a table in a library

AnyRef: a black bear sitting on a chair with a happy halloween shirt on

Shikra: bear with pumpkin

KOSMOS-2: teddy bear wearing yellow shirts

Figure 10. Qualitative results of **Region-level Referring Expression Generation** compared with Shikra [4] and KOSMOS-2 [31]. **AnyRef** is initialized from LLaVA [23], so it still keeps the OCR ability (*e.g.*, it recognizes "happy halloween" on the shirt). Additionally, **AnyRef** can also output the grounding mask corresponding to the bounding-box.