arXiv:2501.04001v4 [cs.CV] 24 Aug 2026

ByteDance

# Sa2VA: Marrying SAM2 with MLLM for Dense Grounded Understanding of Images and Videos

**Haobo Yuan**[1*], **Xiangtai Li**[2*†‡], **Tao Zhang**[2,3*], **Yueyi Sun**[4], **Zilong Huang**[2], **Shilin Xu**[4], **Shunping Ji**[3], **Yunhai Tong**[4], **Lu Qi**[3‡], **Jiashi Feng**[2], **Ming-Hsuan Yang**[1]

University of California, Merced[1] Bytedance Seed[2] Wuhan University[3] Peking University[4]

[†]Project lead [*]Equal technical contribution [‡]Corresponding authors

Code: https://github.com/Bytedance/Sa2VA

Model: https://hf.co/collections/ByteDance/sa2va-model-zoo

## Abstract

This work presents Sa2VA, the first comprehensive, unified model for dense grounded understanding of both images and videos. Unlike existing multi-modal large language models, which are often limited to specific modalities and tasks, Sa2VA supports a wide range of image and video tasks, including referring segmentation and conversation, with minimal single-stage instruction tuning. Sa2VA combines SAM-2, a foundation video segmentation model, with MLLM, advanced vision-language models, and unifies text, image, and video into a shared LLM token space. Using the LLM, Sa2VA generates instruction tokens that guide SAM-2 in producing precise masks, enabling a grounded, multi-modal understanding of both static and dynamic visual content. Additionally, we introduce Ref-SAV, an auto-labeled dataset containing over 72k object expressions in complex video scenes, designed to boost model performance. We also manually validate 2k video objects in the Ref-SAV datasets to benchmark referring video object segmentation in complex environments. Experiments show that Sa2VA achieves strong performance across multiple tasks, particularly in referring video object segmentation, highlighting its potential for complex real-world applications. In addition, Sa2VA can be easily extended into various MLLMs, including Qwen-VL and Intern-VL, which can be updated with rapid process in current open-sourced MLLMs. Code and models have been provided to the community.

**Date:** August 25, 2026

**Correspondence:** Xiangtai Li at xiangtai94@gmail.com, Lu Qi at luqi360@whu.edu.cn

**Index Terms---** Multi-modal Large Language Models, Grounded Understanding, Referring Segmentation.

## 1 Introduction

Multi-modal Large Language Models (MLLMs) have made significant progress, fueled by the rapid development of Large Language Models (LLMs) [38, 81, 99]. Numerous MLLMs have been applied to image- and video-level tasks such as visual question answering (VQA) [1, 78], narrative story generation [32, 100, 120], and interactive editing [35, 44, 83]. One important direction is to understand video content in a fine-grained manner, including segmenting and tracking pixels with language descriptions, and performing VQA on visual prompts in the video. In particular, we aim to achieve promptable fine-grained analysis of video, enabling the user to be in the loop when




**Figure 1 Illustration of capabilities of our proposed Sa2VA.** (a). Given a video, Sa2VA can segment the referred object and understand the whole scene. (b). Sa2VA supports image chat, video chat, image referring segmentation, video referring segmentation, and grounded caption generation with single-stage instruction-tuning. (c). Sa2VA achieves strong results on multiple image, video referring segmentation, and chat benchmarks compared to existing MLLMs [74, 98, 112].

playing the video in an interactive mode, as shown in Fig. 1 (a). This capability will enable various applications, such as short video editing [8, 62, 82], robot navigation [26, 28, 29], and surveillance analysis [70].

However, neither state-of-the-art video perception models [16, 54, 75, 111] nor video MLLMs [31, 48, 55, 61, 108] can achieve this. The former are limited by constrained semantic concepts and lack open-ended capabilities, such as video question answering (Video-QA) [51, 55]. While the state-of-the-art SAM-2 model [75] can perform promptable segmentation and tracking, it cannot handle text-aware tasks, such as understanding language expressions or video conversation. On the other hand, video MLLMs can understand videos and perform Video-QA. For example, the latest LLaVA [46] can achieve strong results on Video-QA. However, it can neither carry out perception tasks nor understand visual prompts Several works [23, 73, 92, 98, 112] have explored the combination of perception models and MLLMs. However, existing works mainly focus on exploring image tasks [23, 105, 112] or trying to solve one specific video task [4, 40, 68, 98]. To the best of our knowledge, *no work* has successfully integrated the strengths of both sides to build a unified model that supports both image and video inputs across a diverse range of tasks (Tab. 1).

Based on the above analysis, we set out to unify two visual foundation models, SAM-2 [75] and MLLMs [59] into one framework. As such, our model not only inherits the spatial-temporal perception capabilities of SAM-2 and the open-ended abilities of MLLMs but also benefits from learning additional knowledge derived from the additional training data. To achieve this goal, three key challenges must be addressed: (1) *Task formulation*: How to effectively formulate a range of tasks into a unified training setting, particularly for multi-modal inputs. (2) *Performance*

*balance*: How to resolve conflicts between tasks, such as ensuring strong referring visual understanding abilities without sacrificing the language proficiency of MLLMs. Note that existing methods [93, 112] find that conversation tasks are drastically degraded when grounding tasks are performed. The models are thus degraded to specialists. (3) *Knowledge Inheritance*: How to leverage the pre-trained knowledge from SAM-2 and MLLMs to build a unified model that is both robust and flexible with different foundation models. We note that SAM-2 is trained on more than 1B masks, and most MLLMs are trained with massive instruction-following data pairs. Furthermore, both model families are evolving rapidly, and a flexible framework is crucial to more easily leverage their future progress.

In this work, we present Sa2VA, the first model that integrates SAM-2 with MLLMs, creating a unified, grounded understanding of both images and videos. We first formulate various tasks into a single-stage visual instruction tuning process, including referring segmentation, visual question answering (VQA), and grounded conversation generation (GCG) for both image and video. We leverage the token-length flexibility of LLMs, treating all input images, videos, and visual prompts as tokens without additional specific design. Through joint training, we demonstrate that conflicts between grounding and conversation tasks can be effectively resolved. Unlike existing approaches that use MLLMs as agents or connectors to call on visual experts, our model is trained end-to-end, showcasing model and dataset scalability. Moreover, we adopt a decoupled design in which SAM-2's encoder and memory module are frozen, allowing us to retain its tracking capabilities. This design also makes our method a flexible framework, enabling our model to evolve with increasingly powerful MLLMs. To be more specific, our framework is compatible with advanced MLLMs, such as InternVL [122] and Qwen3-VL [95].

In addition, we empirically observe that existing video segmentation datasets are limited to small-scale collections of short clips with a limited number of occlusions. Thus, to bridge the gaps between current datasets and the real-world requirements of interactive applications, we introduce Ref-SAV: a new, challenging reference video segmentation dataset curated through an automated pipeline based on the recent SA-V datasets [75]. Finally, we benchmark several existing models on this dataset and find significant room for further exploration.

Sa2VA is co-trained with multiple datasets, including image and video modalities and various tasks. Without any bells and whistles, Sa2VA achieves strong performance on six referring image and video segmentation datasets while maintaining strong image and video QA capabilities compared to previous MLLMs, as shown in Fig. 1(c). On the proposed Ref-SAV dataset, our method achieves over 15% performance gain over existing approaches under the zero-shot test setting, and obtains even stronger results using a large-scale Ref-SAV training set (about 37k videos). As shown in Fig. 1(c), our work builds a new strong baseline for a unified, dense, grounded understanding of images and video. Our main contributions are:

- We develop Sa2VA, the first simple framework that combines SAM-2 and MLLM into one model. Sa2VA achieves strong spatial-temporal grounded understanding performance across various benchmarks. Sa2VA can be built upon with various modern MLLMs, including InternVL [122] and Qwen-VL [3].
- We introduce a challenging human-checked video benchmark, Ref-SAV. The benchmark introduces heavier occlusions, long text inputs, and motion blurs.
- We develop a simple data annotation pipeline to build the Ref-SAV training dataset, where we find that the training dataset improves model performance on Ref-SAV.
- Extensive experiments show the effectiveness of Sa2VA on various open-sourced benchmarks (over 15 benchmarks, including image, video VQA, visual prompt, referring image/video segmentation). To the best of our knowledge, few existing MLLMs have shown comparable performance across such a diverse set of domains.

## 2 Related Work

**Multi-modal Large Language Models.** Earlier works [36, 49, 50] explore better multi-modal fusion methods and feature extractors and design various fusion architectures, particularly for vision language tasks. With the advances of LLMs [5, 38, 81], multi-modal instruction tuning on LLMs [2, 11, 59, 79] have recently attracted much attention. Notably, various benchmarks [24, 37, 56, 64, 80] have been established to enhance model performance, reinforcing the notion that data is crucial in developing state-of-the-art MLLMs. One representative work, LLaVA [59], considers the visual features as visual tokens. Furthermore, LLaVA provides a unified data format to consolidate extensive VQA tasks. Several works [105]

**Table 1 Capabilities of different representative models.** We compare Sa2VA against other representative works. As shown, while there are several groups of models that share similar capability sets (e.g., focusing on video conversation or dense grounding), no single existing method covers the full spectrum of tasks. In contrast, Sa2VA is designed to comprehensively support all listed modalities and tasks (also see Fig. 1 (a) and (b)).

| Method | Support Inputs | | | Dense Grounding | | | | Conversation | | | | End to End |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Image | Video | Visual Prompts | RES | Ref-VOS | Inter-VOS | GCG | Image-Chat | Video-Chat | Video Caption | Interactive Caption | Training |
| *Image-Chat Only Models* | | | | | | | | | | | | |
| LLAVA [59] | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| *Image & Video Chat Models* | | | | | | | | | | | | |
| LLaVA-OneVision [46] | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ |
| InternVL [14] | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ |
| LLaMA-VID [55] | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ |
| P-LLaVA [96] | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ | ✓ | ✓ | ✗ | ✓ |
| *Visual Prompt & Dense Grounding Models* | | | | | | | | | | | | |
| Osprey [105] | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| VIP-LLaVA [6] | ✓ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| GLAMM [74] | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✓ |
| LISA [45] | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✓ |
| OMG-LLaVA [112] | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ |
| *Dense Grounding Models* | | | | | | | | | | | | |
| F-LLM [93] | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✗ |
| VISA [98] | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ | ✗ |
| HyperSeg [86] | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| InstructSeg [85] | ✓ | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ | ✗ | ✗ | ✓ |
| Sa2VA (Ours) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

subsequently explore stronger visual cues to improve the visual inputs of LLaVA. Meanwhile, numerous methods [19, 20, 34, 45, 57, 72, 76, 107–110, 113] incorporate additional components to adapt LLaVA for visual grounding, detection, segmentation, and video VQA. Recently, a new trend has emerged to unify image, video, and multi-image analysis within a single framework [46, 61]. LLaVA-OneVision [46] designs a single model to handle four different input sources. In visual perception, several works [54, 75, 97, 114] also explore multi-dataset and multi-task co-training. SAM-2 [75] proposes a unified model for joint image and video interactive segmentation. Our model, Sa2VA, integrates SAM-2 into existing MLLMs to create a unified end-to-end model. It aims to combine image and video for dense, grounded understanding, encompassing tasks such as segmentation, chat, and captioning.

**Referring Segmentation.** This task aims to output specific masks (image) or tracked masks (video) driven by language description. Earlier studies [18, 43, 58, 91, 104, 117] examine various fusion modules to enhance performance. Then, several stronger models [97] adopt DETR-like methods [7, 12, 69, 119, 123] to achieve unified segmentation and tracking in video. Equipped with LLMs, several recent works [9, 45, 72, 87, 94, 112] have been developed to accomplish more complex referring tasks, including reasoning for referring or joint mask and caption generation. For example, LISA [45] involves reasoning-based segmentation. Then, GLaMM [74] annotates a new dataset and introduces region-level captioning and segmentation tasks. Meanwhile, several recent models exploit joint instruction tuning for referring segmentation and conversation [11, 98, 115]. Our method expands on these studies in the video domain by utilizing SAM-2 [75], while maintaining superior performance in both image/video referring tasks and conversation tasks.

**Video Segmentation and Grounding.** Existing video segmentation methods [39, 53, 121] mainly focus on segmenting and tracking pixels in closed sets, while a few approaches [30, 118] explore open-vocabulary settings. However, concepts are still limited compared to the knowledge space of LLMs. For video grounding, a recent method [33] explores LLM in the joint understanding of video and audio. On the other hand, VISA [98] and VideoLISA [4] also explore referring video segmentation. However, they lack comprehensive training, which restricts their capability in other tasks and scaling up. In contrast, Sa2VA enables fine-grained spatial-temporal modeling of both static (image) and dynamic (video) visual content, achieving strong performance across various tasks.

# 3 Method

## 3.1 Unifying Multi-task Representations

Developing a unified model to tackle various image and video understanding tasks is challenging due to their differences in spatial and temporal information.

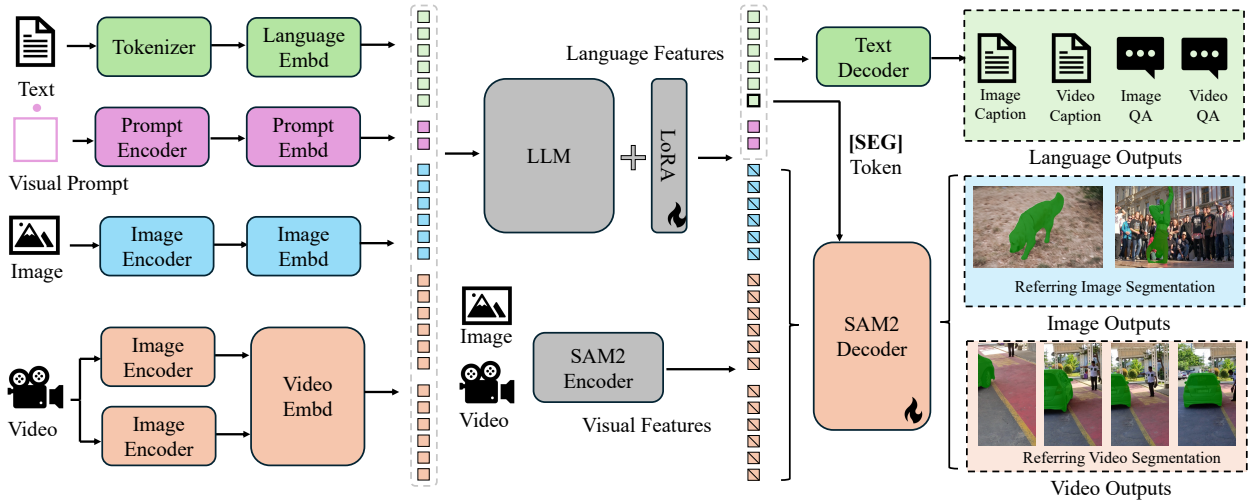



**Figure 2 Proposed Sa2VA model.** The model first encodes the input texts, visual prompts, images, and videos into token embeddings. These tokens are then processed through a large language model (LLM). The output text tokens are used to generate the "[SEG]" token and associated language outputs. The SAM-2 decoder receives the image and video features from the SAM-2 encoder, along with the "[SEG]" token, to generate corresponding image and video masks. Modules with a fire icon are trained during the single-stage instruction-tuning.

To address this, we begin by re-examining the task formulations and proposing a unified representation, which lays the groundwork for the development of Sa2VA.

**Referring Image/Video Object Segmentation.** For image-referring segmentation, given input text tokens $T_i \in \mathbb{R}^{N\times D}$, the model processes input images $I_i \in \mathbb{R}^{H\times W\times 3}$ and produces corresponding binary masks $M_o \in \mathbb{R}^{H\times W}$ that align with the text description, where $N$ and $D$ are the number and dimension of text tokens. For video object referring segmentation, the model takes input videos $V_i \in \mathbb{R}^{T\times H\times W\times 3}$ and outputs binary spatio-temporal masks (masklets) $M_o \in \mathbb{R}^{T\times H\times W}$ where $T$, $H$, and $W$ are video frame number, height, and width.

**Image/Video Chat and Grounded Caption Generation.** For image and video chat tasks, given input text tokens $T_i$ and the corresponding images $I_i$ or videos $V_i$, the model produces the answer text $T_o$. For grounded caption generation tasks, the model simultaneously outputs the masks $M_o$ and the aligned text $T_o$.

**Visual Prompt Understanding Tasks.** For visual prompt understanding tasks, in addition to text tokens $T_i$ and image $I_i$, the model also takes additional visual prompt tokens $VP_i$ (boxes or points on the image) as inputs, outputting corresponding masks $M_o$ and aligned text answers $T_o$.

**Unified Task Representation.** Existing works address the above tasks through specific design models or partially unified models (different tasks utilize different model weights). In this work, we argue that all the aforementioned tasks can be unified into a single-stage instruction-tuning process, as we can leverage LLMs to manage various visual tokens. The overall process can be formulated as:

$$T_o, M_o = LLM(\{I_i, V_i, VP_i\}, T_i). \quad (1)$$

The input tokens vary for different tasks. For chat-only tasks, the model only outputs text tokens $T_o$. For referring segmentation tasks, the model outputs masks (image) or masklets (video) $M_o$. For grounded caption generation tasks, the model outputs both $T_o$ and $M_o$. Since the model primarily processes images/videos, text, and masks as inputs and outputs, it is reasonable to unify these tasks within a single framework and train the model end-to-end. This approach enables a single model to support all the aforementioned tasks effectively.

## 3.2 Sa2VA Framework

The overall framework of Sa2VA is illustrated in Fig. 2. It consists of two components: the MLLM model and SAM-2.

**Pre-trained MLLMs.** We adopt pre-trained MLLMs, which contain one visual encoder, one visual projection layer, and one LLM. The visual encoder takes

**Algorithm 1:** Ref-VOS Inference Pipeline

**Input:** Video length $N$; Number of key frames $M$; Video frames $S_N$ ($X_1$, $X_2$, $X_3$,..., $X_N$); Language description $T$;
**Output:** Sequence of masks $M_1$, $M_2$, $M_3$,..., $M_N$;
**Run:** Sa2VA Model for Ref-VOS;
Extract key frames: $S_M \leftarrow S_N$;
Visual embeddings: $E_v \leftarrow \text{Encoder}(S_M)$;
Language embeddings: $E_l \leftarrow \text{Encoder}(T)$;
Answers: $A \leftarrow \text{LLM}(\{E_v, E_l\})$;
Prompt embedding: $P_l \leftarrow \text{Linear}(\text{Find}(A,$ '[SEG]'));
**for** $i = 1, 2, \ldots, M$ **do**
    SAM-2 feature: $F_i \leftarrow \text{Encoder}(X_0)$;
    Mask: $M_i \leftarrow \text{Decoder}(\{P_l, F_i\})$;
    Update Memory: $Mem \leftarrow$ Cross-Attention($\{Mem, M_i\}$);
**for** $i = M+1, M+2, \ldots, N$ **do**
    SAM-2 feature: $F_i \leftarrow \text{Encoder}(X_0)$;
    Mask: $M_i \leftarrow \text{Decoder}(\{Mem, F_i\})$;
    Update Memory: $Mem \leftarrow$ Cross-Attention($\{Mem, M_i\}$);
**emit** $M_1$, $M_2$, $M_3$,..., $M_N$;

input images, videos, and sub-images and outputs visual features. The visual projection layer maps the features into tokens. These tokens, along with the input text tokens, serve as the input for LLMs, which generate text token predictions based on them. Note that we utilize pre-trained MLLMs, as seen in prior research [45, 74, 98, 112], to harness their powerful capabilities. For both image and video conversation tasks, we use the same pipeline as the original pre-trained MLLMs (e.g., InternVL2 [14], Qwen-VL [2]).

**Decoupled Design.** We append SAM-2 [75] alongside the pre-trained MLLMs. Note that we do not take the SAM-2's output tokens (visual features or decoder outputs) into LLM. There are three reasons behind this design choice. First, we aim to simplify the combination without incurring additional computation costs. Second, adding extra tokens will impede knowledge inheritance from MLLMs because MLLMs and SAM-2 are misaligned. Third, with this design, we can fully convert our work into a flexible framework that utilizes pre-trained, evolving MLLMs, as the MLLM community progresses rapidly.

**Tuning SAM-2 Decoder via "[SEG]" Tokens.** Similar to previous works [45, 98] which utilize SAM, we employ the special token "[SEG]" to connect the decoder of SAM-2 with the MLLM. The hidden states of the "[SEG]" token serve as a new kind of prompt, which are fed into SAM-2's Decoder, where they are decoded into segmentation masks. The hidden states of "[SEG]" can be viewed as a novel spatial-temporal prompt for SAM-2. SAM-2 segments the relevant object mask in images and videos based on this spatial-temporal prompt. During training, the SAM-2 decoder can be adjusted to comprehend the spatial-temporal prompt, and gradients can be backpropagated through the "[SEG]" token to the MLLM, enabling it to gain knowledge from the training datasets.

**Utilizing SAM-2 knowledge for mask tracking.** For Ref-VOS tasks, we develop a straightforward framework to attain strong results on public benchmarks. Specifically, we initially extract key frames (the first several frames) from the video. These key frames are then processed through MLLM. This MLLM then generates a "[SEG]" token as the SAM-2 prompt to create the masks for the key frames. Next, we utilize the memory encoded by the key frame features to generate the mask for the remaining frames. We present the default inference algorithm in Algorithm 1.

**Benefits of a unified framework.** Compared with training separate task-specific models, a single shared set of weights provides four concrete benefits. (1) Positive cross-task transfer: joint co-training across image/video × chat/segmentation produces gains that cannot be recovered by single-task training alone (Tab. 9; e.g., removing image-segmentation data reduces RefCOCO performance from 77.4 to 20.2 cIoU). (2) Preserved conversational ability: Sa2VA retains the chat capability of its base MLLM while achieving state-of-the-art grounding performance (Tab. 7). (3) Lower maintenance cost than per-dataset fine-tuning: dataset-specific fine-tuning brings only marginal improvements (Tab. 11) but requires maintaining a separate checkpoint for each dataset. (4) Deployment efficiency: a single model handles all tasks without per-task switching, and the appended SAM-2 module adds only about 220M parameters while running at 39.5 FPS; thus, the overall cost is dominated by the shared MLLM (Tab. 19). We provide further discussion of these benefits in the Appendix.

## 3.3 Ref-SAV Dataset and Benchmark

**Data Annotation Pipeline.** We create an automatic annotation pipeline to generate referring object text expressions for the SA-V dataset [75]. As illustrated in Fig. 3, the pipeline consists of three stages:

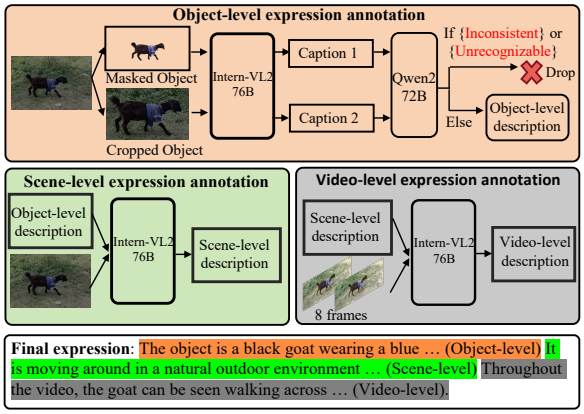



**Figure 3 Data annotation pipeline**. Our proposed automatic data annotation pipeline consists of three stages: object-level, scene-level, and video-level text expression annotation.

(1) *Object-level annotation.* We first select the frame with the largest object area from the video and mask out the non-object pixels. The cropped (without masking) and full (with masking) images are then fed separately into InternVL2-76B [14] to generate detailed descriptions. The descriptions are processed by Qwen2-72B [99] for consistency checking, and conflicting descriptions are discarded. This method allows us to filter out error-prone cases. (2) *Scene-level annotation.* Both the image and the object-level description from the previous stage are sent into InternVL2-76B [14] to generate a detailed object description that includes relationships to the scene and surrounding objects. (3) *Video-level annotation.* We sample 8 frames uniformly from the video, applying yellow borders to emphasize the object in each frame. These frames, paired with the scene-level description, are processed by InternVL2-76B [14] to generate a video-level description that captures the object's movement and actions.

**Ref-SAV Dataset.** Using the above pipeline, we have automatically annotated detailed object expressions for the SA-V dataset, resulting in the creation of Ref-SAV. The training set of Ref-SAV contains 37,311 videos and 72,509 object expressions, which do not have any human labeling. For the Ref-SAV validation set, we select a subset of videos from the training set of the SA-V dataset to create the Ref-SAV evaluation benchmark, as the validation and test sets contain only a limited number of videos. These videos are completely separate from the Ref-SAV training dataset. The evaluation benchmark has two parts: 1) *The long-expression set*, generated using the automatic annotation pipeline mentioned above and carefully filtered by human annotators. 2) *The short-expression set*, which is manually annotated. This evaluation benchmark includes 1,147 videos and 1,945 object expressions, comprising 1,694 long expressions and 251 short expressions. We note that the annotation pipeline uses a two-model agreement filter: captions generated by InternVL2-76B are cross-checked by Qwen2-72B. All reported test results are evaluated against human ground truth; moreover, the short-expression set is fully human-annotated, whereas the long-expression set is human-filtered. In the Appendix, we discuss the choice of annotation model, explain why it does not introduce an unfair family-internal advantage, and provide external validation of the automatically generated labels.

### 3.4 Sa2VA Training and Testing.

**One For All Co-Training.** Sa2VA is co-trained on multiple datasets. For VQA tasks, we utilize text regression loss $L_{text}$, similar to common MLLMs. For segmentation tasks $L_{mask}$, we employ pixel-wise cross-entropy loss $L_{CE}$ and dice loss $L_{DICE}$. It is important to note that there is no pre-training stage as in the base MLLMs [14, 59]; instead, we conduct supervised fine-tuning in a single-stage training with the following loss:

$$\mathcal{L}_{instruction} = \mathcal{L}_{text} + \mathcal{L}_{mask}, \quad \mathcal{L}_{mask} = \mathcal{L}_{CE} + \mathcal{L}_{DICE}. \tag{2}$$

**One For All Testing.** All tasks can be encompassed within the Eq. 1 paradigm. During the inference stage, we encode the necessary task requirements, such as text prompts, visual prompts, image features, and video features, into tokens to input into the LLM. The output tokens from the LLM are then decoded into text responses (LLM prediction head), segmentation masks (SAM-2 decoder), and responses from the SAM-2 mask tracking module according to the task definition.

## 4 Experiments

**Baseline.** We construct the baseline by combining the SOTA public MLLM model InternVL2 [14] with SAM-2 [75]. Similar to previous works [45, 74, 112], the segmentation mask is obtained by decoding the hidden state of the "[SEG]" token through SAM2's decoder. Inspired by Mask2Former-VIS [15], an object shares the same "[SEG]" token throughout the video, enabling our model to handle both image and video referring segmentation tasks in a unified manner. In addition, we further use more advanced MLLM models, InternVL2.5 [13], to demonstrate the model

**Table 2** Sa2VA training datasets.

| Type | Datasets |
|---|---|
| Image QA | LLaVA 1.5 (665K) |
| Image Seg | RefCOCO (17K), RefCOCO+ (17K), RefCOCOg (22K), Grand-f (214K) |
| Video QA | Video-ChatGPT (100K) |
| Video Seg | Ref-YTVOS (3.5K), MeVIS (0.6K), ReVOS (1.7K), **Ref-SAV (37K)** |

**Table 3** Comparison with previous Ref-VOS benchmarks.

| Property | DAVIS17-RVOS | ReVOS | Ref-YT-VOS | MeVIS | Ours |
|---|---|---|---|---|---|
| Short Text | ✓ | ✓ | ✓ | ✓ | ✓ |
| Long Text | ✗ | ✓ | ✗ | ✗ | ✓ |
| Large Object Motion | ✗ | ✗ | ✓ | ✗ | ✓ |
| Large Camera Motion | ✓ | ✓ | ✗ | ✗ | ✓ |
| Heavy Occlusion | ✗ | ✗ | ✗ | ✗ | ✓ |

scaling effect. Our paradigm treats the MLLM as a freely interchangeable plug-in, and it is our unified paradigm that endows an arbitrary conversational MLLM with strong referring segmentation while preserving its dialogue ability. Sa2VA thus does not stem from a stronger backbone, and even a relatively weak backbone (LLaVA-1.5-7B) plugged into our framework already surpasses the prior specialist (Appendix). Varying the backbone under our paradigm (Tab. 8, Tab. 14) yields consistent gains, confirming the improvement comes from the framework, not any particular backbone; the same modularity let us swap SAM-2 for SAM-3 (Appendix).

**Datasets and Metrics.** We use four types of datasets to train Sa2VA, which include image QA, video QA, image segmentation, and video segmentation datasets. As shown in Tab. 2, Sa2VA's training data includes approximately 1.1 million image-text or video-text pairs. Since InternVL2 has been trained with a large amount of image QA and video QA data, we only used 665K LLaVA 1.5 [63] and 100K Video-ChatGPT [66] data to prevent the MLLM from forgetting its image and video QA capabilities. We use 56K referring expression data [41, 103] and 214K grounding conversation generation data [74] for image-level text-driven segmentation. For video-level referring expression segmentation, we used 5.8K existing referring VOS data from Ref-YouTubeVOS [77], MeVIS [18], and ReVOS [98]. Additionally, we used 37K long text referring VOS data (Ref-SAV), generated by our proposed automatic annotation pipeline, to enhance Sa2VA's understanding of long referring text and its object grounding capabilities for complex videos. For image referring segmentation, we adopt cIoU since it balances the large and small objects. For referring video object segmentation, we adopt J&F. For image and video chat tasks, we follow previous works [55, 59] and report performance accordingly.

**Implementation Details.** We adopt the xtuner [17] framework for training. Our model inherits from MLLM models and does not require pre-training stages, which allows it to effectively leverage the information from the different base models. Unless specified otherwise, the following details describe the training of our default InternVL2 models. During the instruction tuning phase, we use the AdamW optimizer with an initial learning rate of 4e-5 and a weight decay of 0.05. We employ a 5% warmup ratio and apply gradient clipping with max gradient norm as 1.0. For LoRA parameters, we set the rank (r) to 256. The MLLM input image size is 448×448, and the SAM-2 input image size is 1024×1024. The maximum sequence length for the LLM is set to 8,192. We train for one epoch with a total batch size of 256 (16x2x8), achieved using a per-device batch size of 2 and 8 gradient accumulation steps. The training is conducted using bfloat16 (bf16) mixed precision on 16 NVIDIA H100 80GB GPUs. The instruction tuning stage lasts approximately 24 hours. The training strategy is listed in Tab. 5 We adopt VLMEvalKit [21] for the evaluation of chat tasks, and we have provided our scripts for segmentation tasks. For Ref-SAV testing, we adopt the original open-sourced codebase [90, 97, 98] and model weights to infer the video results. By default, we use single-stage instruction tuning and *do not fine-tune on any specific task* to show the effectiveness of single-stage instruction tuning.

**Table 4** Statistics comparison with other Ref-VOS datasets.

| Datasets | Video | Object | Expression | Mask | Avg. Len. |
|---|---|---|---|---|---|
| Ref-YTVOS | 3,978 | 7,451 | 15,009 | 131k | 9.68 |
| MeViS | 2,006 | 8,171 | 28,570 | 443k | 7.07 |
| ReVOS | 1,042 | 9,084 | 35,074 | 273k | 10.5 |
| Ref-SAV | 37,311 | 72,509 | 72,509 | 6.0m | 83.6 |

**Table 5** Training strategies for model components.

| | MLLM | SAM-2 Encoder | SAM-2 Decoder | SAM-2 Memory |
|---|---|---|---|---|
| Training Strategy | LoRA | Frozen | Finetune | Frozen |

**Ref-SAV Dataset.** In Tab. 3, we further compare the existing Ref-VOS benchmarks from five different aspects: short text, long text, large object motion, large camera motion, and heavy occlusion. The previous benchmarks only contain partial aspects of these five challenging cases. On the other hand, our benchmark is built from SAM-2 [75] and labeled with long text, which is more challenging than previous benchmarks. This is why the previous Ref-VOS models cannot achieve good results on our benchmark. Furthermore, as shown in Tab. 4, our dataset has more detailed captions (Avg. Len. refers to the average

**Table 6** Experiment results on various image/video benchmarks. We report cIoU for the image segmentation sets, J&F for the video segmentation sets, and AP50 for the GCG benchmark. For MME dataset [24], A / B denotes the perception (A) and cognition (B) scores, respectively, while C(+) represents the total score (C = A + B). Although it is more common to report perception and cognition separately, we report the sum as in their original paper if the individual scores are missing.

| Method | Image Segmentation | | | Video Segmentation | | | | Image Chat | | | Video Chat | | GCG |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RefCOCO [41] | RefCOCO+ [41] | RefCOCOg [103] | MeViS [18] | Ref-DAVIS17 [43] | Ref-YTVOS [77] | ReVOS [88] | MME [24] | MMBench [64] | SEED-Bench [47] | Video-MME [25] | MMBench-Video [22] | GCG [74] |
| LLAVA-1.5-13B [60] | - | - | - | - | - | - | - | 1531(+) | 68.8 | 70.1 | - | - | - |
| Video-LLaVA-7B [57] | - | - | - | - | - | - | - | - | 60.9 | - | 39.9 | 1.03 | - |
| LLaMA-VID-7B [55] | - | - | - | - | - | - | - | 1521(+) | 65.1 | 59.9 | - | 1.08 | - |
| mPLUG-Owl3-8B [102] | - | - | - | - | - | - | - | - | 77.6 | - | 53.5 | 1.35 | - |
| InternVL2-8B [14] | - | - | - | - | - | - | - | - | 81.7 | **76.2** | **54.0** | 1.28 | - |
| PixelLM-7B [76] | 73.0 | 66.3 | 69.3 | - | - | - | - | 309/135 | 17.4 | - | - | - | - |
| LaSagnA [84] | 76.8 | 66.4 | 70.6 | - | - | - | - | 0/0 | 0.0 | - | - | - | - |
| LISA-7B [45] | 74.9 | 65.1 | 67.9 | - | - | - | - | 1/1 | 0.4 | - | - | - | - |
| GLaMM-7B [74] | 79.5 | 72.6 | 74.2 | - | - | - | - | 14/9 | 36.8 | - | - | - | 28.9 |
| LLaVA-G-7B [109] | 77.1 | 68.8 | 71.5 | - | - | - | - | - | - | - | - | - | - |
| GSVA-13B [94] | 79.2 | 70.3 | 75.7 | - | - | - | - | - | - | - | - | - | - |
| OMG-LLaVA-7B [112] | 78.0 | 69.1 | 72.9 | - | - | - | - | 1177/235 | 47.9 | 56.5 | - | - | 29.9 |
| VideoLISA-3.8B [4] | 73.8 | 63.4 | 68.3 | 44.4 | 68.8 | 63.7 | - | - | - | - | - | - | - |
| VISA-13B [98] | 72.4 | 59.8 | 65.5 | 44.5 | 70.4 | 63.0 | 50.9 | - | - | - | - | - | - |
| Sa2VA-1B (Ours) | 77.4 | 69.9 | 72.3 | 41.7 | 72.3 | 65.3 | 47.6 | 1381/405 | 68.3 | 64.8 | 39.9 | 1.07 | 23.8 |
| Sa2VA-4B (Ours) | 80.4 | 74.3 | 76.7 | 46.2 | 73.8 | 70.0 | 53.2 | 1553/540 | 76.8 | 72.6 | 50.4 | 1.23 | 28.2 |
| Sa2VA-8B (Ours) | **81.9** | **76.5** | **78.9** | **46.9** | **75.2** | **70.7** | **57.6** | **1651/578** | **82.4** | 75.5 | 52.1 | **1.34** | **31.0** |
| Sa2VA-26B (Ours) | 82.5 | 78.8 | 79.7 | 46.2 | 77.0 | 70.1 | 58.4 | 1691/538 | 83.7 | 76.8 | 52.6 | 1.45 | 33.5 |

**Table 7** Performance on image-level benchmarks with MLLMs that have segmentation capability. We report cIoU for the image segmentation sets. A / B denotes the perception (A) and cognition (B) scores, respectively, while C(+) represents the total score (C = A + B).

| Method | MME [24] | MMBench [64] | SEED-Bench [47] | AI2D [42] | MMStar [10] | MMMU [106] | $SQA^{test}$ [65] | RefCOCO | RefCOCO+ | RefCOCOg |
|---|---|---|---|---|---|---|---|---|---|---|
| LLAVA-1.5-13B [60] | 1531(+) | 68.8 | 70.1 | - | - | - | - | 0.0 | 0.0 | 0.0 |
| LISA-7B [45] | 1/1 | 0.4 | - | 0.0 | - | - | - | 74.9 | 65.1 | 67.9 |
| PixelLM-7B [76] | 309/135 | 17.4 | - | 0.0 | - | - | - | 73.0 | 66.3 | 69.3 |
| LaSagnA-7B [84] | 0/0 | 0.0 | - | 0.0 | - | - | - | 76.8 | 66.4 | 70.6 |
| GLaMM-7B [74] | 14/9 | 36.8 | - | 28.2 | - | - | - | 79.5 | 72.6 | 74.2 |
| OMG-LLaVA-7B [112] | 1177/235 | 47.9 | 56.5 | 42.9 | - | - | - | 78.0 | 69.1 | 72.9 |
| Sa2VA-4B (ours) | 1553/540 | 76.8 | 72.6 | 79.9 | 53.7 | 46.2 | 95.8 | 80.4 | 74.3 | 76.7 |
| Sa2VA-8B (ours) | 1651/578 | 82.4 | 75.5 | 82.1 | 60.3 | 44.7 | 96.8 | 81.9 | 76.5 | 78.9 |

**Table 8** Experiment results using stronger InternVL2.5 in our Sa2VA.

| Base MLLM | Image Segmentation | | | Video Segmentation | | | Image Chat | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RefCOCO | RefCOCO+ | RefCOCOg | MeViS (val_u) | Ref-YTVOS | Ref-DAVIS17 | MME | MMBench | SEED-Bench | AI2D | MMStar | $SQA^{test}$ |
| InternVL2.0-4B | 80.4 | 74.3 | 76.7 | 52.1 | 70.0 | 73.8 | 1553/540 | 76.8 | 72.6 | 79.9 | 53.7 | 95.8 |
| InternVL2.0-8B | 81.9 | 76.5 | 78.9 | 57.0 | 70.7 | 75.2 | 1651/578 | 82.4 | 75.5 | 82.1 | 60.3 | 96.8 |
| InternVL2.5-1B | 79.6 | 73.6 | 77.7 | 53.4 | 68.0 | 69.5 | 1504/434 | 71.9 | 71.0 | 69.2 | 48.6 | 89.9 |
| InternVL2.5-4B | 82.4 | 77.6 | 79.7 | 55.9 | 71.3 | 73.7 | 1691/610 | 81.8 | 74.9 | 81.4 | 57.9 | 96.8 |
| InternVL2.5-8B | 82.6 | 78.0 | 80.3 | 58.9 | 72.3 | 75.9 | 1690/610 | 84.4 | 76.5 | 82.7 | 62.4 | 97.4 |
| InternVL2.5-26B | 82.9 | 79.3 | 81.2 | 61.8 | 75.1 | 78.6 | 1698/653 | 85.8 | 78.3 | 85.7 | 67.0 | 98.4 |

length of the referring description) and many more masks (about 10x). The proposed data can lead to better instruction-following capability for complex descriptions.

## 4.1 Main Results

**Comparison With SOTA MLLMs.** As shown in Tab. 6, Sa2V-8B achieves cIoU scores of 81.9, 76.5, and 78.9 on RefCOCO, RefCOCO+, and RefCOCOg, respectively, surpassing GLaMM-7B by 2.4, 3.9, and 4.7 cIoU. Sa2VA performs favorably against the state-of-the-art methods on RefCOCO+ and RefCOCOg, significantly outperforming existing grounding MLLMs, including LISA, GLaMM, PixelLLM, PSALM, and OMG-LLaVA. Additionally, Sa2VA demonstrates strong conversational capabilities, achieving scores of 2229 (1651/578), 82.4, and 75.5 on MME, MMbench, and SEED-Bench, while existing grounding MLLMs perform poorly in conversation due to catastrophic forgetting. Sa2VA achieves performance comparable to that of InternVL2 in image QA benchmarks, indicating that Sa2VA largely maintains the chat performance of the base MLLM.

Sa2VA also performs well on video benchmarks. It achieves scores of 46.9, 75.2, and 57.6 J&F on MeVIS, Ref-DAVIS17, and ReVOS, respectively, surpassing the previous SOTA VISA-13B by 2.4, 4.8, and 6.7 J&F. Additionally, Sa2VA-8B earns a score of 1.34

on the video QA benchmark MMBench-Video, outperforming InternVL2-8B's 1.28. Sa2VA does not achieve dominant results on video chat tasks, as it has both video segmentation and understanding capabilities, which current methods are not equipped to handle simultaneously. The main experimental results indicate that our Sa2VA is a versatile and powerful MLLM.

**Detailed Results on Image Benchmarks.** As indicated by [45, 93, 112], performing instruction tuning negatively impacts performance drastically on chat tasks. We utilize co-training to address this issue. In Tab. 7, we also present additional results on image chat datasets. Our model maintains good results across multiple image chat benchmarks while achieving strong referring segmentation performance, demonstrating its versatility.

**Results on Ref-SAV validation set.** In Tab. 10, we benchmark several state-of-the-art Ref-VOS models using our proposed Ref-SAV benchmark. We find that even the foundation model [97] and the recent video MLLM model [98] fail to achieve strong results on our benchmarks. This is due to our benchmark featuring more occlusions, longer text descriptions, and diverse annotations from SAM-2 [75]. Conversely, our method, Sa2VA, whether with or without the Ref-SAV training set, can yield strong results. Our method can be further enhanced with our training set, indicating that our Ref-SAV training set serves as a valuable supplement to the video understanding community.

**Comparison with specific fine-tuned methods.** The original Sa2VA setting is more challenging because it requires the use of the same model weights to complete all tasks. For fair comparison, similar to prior works [74, 112], we evaluate with MLLMs on specific fine-tuned datasets. As shown in Tab. 11, we find 1.5% improvement across three datasets compared to the co-training model, with Sa2VA performing best among recent methods. From another perspective, fine-tuning on a single dataset does not yield much benefit (0.4 on RefCOCO). However, this requires fine-tuning each dataset and producing a new model. Therefore, we advocate for more general models like Sa2VA for their versatility and convenience.

**InternVL-2.5 Results.** We further conduct ablation experiments to evaluate the scalability of Sa2VA with stronger MLLMs and additional datasets. As shown in Tab. 8, upgrading the MLLM from InternVL2 [14] to InternVL2.5 [13] consistently improves performance across image, video, and multimodal benchmarks.

## 4.2 Ablation Studies

**Effectiveness of Joint Co-training.** The performance of Sa2VA on image and video QA and segmentation benchmarks can be attributed to unified instruction tuning and joint co-training. We perform ablation studies on the training datasets, with the results presented in Tab. 9. When trained without image QA datasets, Sa2VA's scores on MME and MMBench drop by 129 and 4.9, respectively. Without image segmentation datasets, Sa2VA achieves only 20.2, 20.6, and 23.2 cIoU on RefCOCO, RefCOCO+, and RefCOCOg, while performance on the video segmentation benchmark drops significantly. Sa2VA on MMBench-Video decreases by 34% when trained without video QA datasets. When training without video segmentation datasets, Sa2VA's performance on MeVIS and Ref-DAVIS17 drops by 6.4 and 3.3 J&F, respectively. The ablation results indicate that joint co-training with these four types of datasets is critical to Sa2VA's performance across the various tasks. In Tab. 13, adding 3M image-QA samples from Infinity-MM [27] yields a notable +2.1 gain on MMBench, with negligible changes in segmentation accuracy. When incorporating our Ref-SAV dataset, Sa2VA-1B achieves a +1.7 J&F improvement on MeViS while maintaining image-level performance. These results demonstrate that Sa2VA benefits synergistically from both model scaling and data expansion, indicating substantial headroom for further improvement.

**Ablation study on the Segmentation Token Design.** Tab. 12 shows the results of our ablation on the segmentation token design. We analyze two primary alternative strategies based on whether to use the same special token or different special tokens for different frames. First, we test the **repetitive** strategy, which requires the LLM to output the same "[SEG]" token $N$ times for $N$ frames. This approach slightly reduces video segmentation performance. Second, we analyze using **unique**, frame-specific tokens (e.g., "[SEG 1]" for frame 1, "[SEG 2]" for frame 2, and so on). This design caused a larger performance drop. The reason is that it prevents knowledge sharing with the image segmentation task. The model is typically pre-trained to associate a single, generic "[SEG]" token with the concept of segmentation. By introducing new tokens ("[SEG 1]", ..., "[SEG n]"), there is a mismatch between images and videos that blocks this knowledge transfer, forcing the model to learn the meaning of each new segmentation token independently. In other words, carrying richer or more specialized information through multiple dedicated

**Table 9** Ablation study on co-training effect on multiple datasets. We use Sa2VA-1B to test the performance.

| Data | Image Segmentation | | | Video Segmentation | | Image Chat | | | Video Chat | |
|---|---|---|---|---|---|---|---|---|---|---|
| | RefCOCO | RefCOCO+ | RefCOCOg | MeViS (val_u) | Ref-DAVIS17 | MME | MMBench | SEED-Bench | Video-MME | MMBench-Video |
| All Data | 77.4 | 69.9 | 72.3 | 50.8 | 72.3 | 1381/405 | 68.3 | 64.8 | 39.9 | 1.07 |
| w/o Image QA | 78.0 | 70.1 | 72.2 | 48.3 | 73.0 | 1298/359 | 63.4 | 63.8 | 39.7 | 0.39 |
| w/o Image Segmentation | 20.2 | 20.6 | 23.2 | 38.0 | 48.8 | 1393/408 | 70.1 | 65.7 | 41.2 | 1.08 |
| w/o Video QA | 78.0 | 70.4 | 72.6 | 50.7 | 74.3 | 1370/402 | 69.1 | 65.0 | 41.3 | 0.71 |
| w/o Video Segmentation | 77.4 | 69.1 | 72.4 | 44.4 | 69.0 | 1403/398 | 67.8 | 64.9 | 40.4 | 1.04 |

**Table 10** Ref-SAV validation sets. zs: zero-shot testing.

| Method | Long | | | Short | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|
| | J | F | J&F | J | F | J&F | J | F | J&F |
| UniRef++ [90] (zs) | 14.1 | 10.8 | 12.5 | 9.0 | 8.2 | 8.6 | 11.6 | 9.5 | 10.5 |
| UNINEXT [97] (zs) | 11.7 | 8.3 | 10.0 | 5.8 | 4.4 | 5.1 | 8.8 | 6.4 | 7.6 |
| MeVIS [18] (zs) | 12.1 | 7.1 | 11.3 | 6.2 | 5.3 | 5.5 | 12.2 | 9.8 | 10.3 |
| VISA [98] (zs) | 16.1 | 12.2 | 14.1 | 12.3 | 9.6 | 9.2 | 13.2 | 11.3 | 11.8 |
| Sa2VA-8b (zs) | 47.7 | 50.9 | 49.3 | 31.5 | 35.0 | 33.3 | 39.6 | 43.0 | 41.3 |
| Sa2VA-8b (Ours) | 57.0 | 60.4 | 58.7 | 39.5 | 42.9 | 41.2 | 48.3 | 51.7 | 50.0 |

**Table 11** Comparison with Fine-tuned Models.

| Model Type | RefCOCO | RefCOCO+ | RefCOCOg |
|---|---|---|---|
| LAVT [101] | 72.7 | 62.1 | 61.2 |
| GlaMM-7B [74] | 79.5 | 72.6 | 74.2 |
| OMG-LLaVA-7B [112] | 78.0 | 69.1 | 72.9 |
| F-LLM-7B [93] | 76.1 | 65.2 | 68.5 |
| Sa2VA-8B (ours) | 81.9 | 76.5 | 78.9 |
| Sa2VA-8B (ft) | 82.3 | 77.3 | 79.3 |

tokens is a promising direction. However, it appears to require dedicated pre-training to re-establish the association between these tokens and segmentation outputs. This is beyond our single-stage instruction-tuning budget and is therefore left to future work.

## 4.3 More Comparison Results

**Comparison with recent video MLLM models on referral video segmentation** Recently, several works have also combined MLLMs into video referring segmentation tasks. As shown in Tab. 15, we list and compare our models against several recent methods on three datasets. These competitors include VideoGLaMM [68], VideoLISA [4], HyperSeg [86], InstructSeg [85], and etc. We also include the combined strong baseline (SAM2 [75] + GlaMM [74]) reported in [68]. As shown in Tab. 15, our Sa2VA-8B model achieves strong performance across all three datasets. Our Sa2VA-4B model also demonstrates strong competitive performance, achieving the highest score on

**Table 12** Ablation study on "[SEG]" token design.

| Type | RefCOCO | RefCOCO+ | RefCOCOg | DAVIS | MeVIS(val_u) |
|---|---|---|---|---|---|
| Single | 77.4 | 69.9 | 72.3 | 72.3 | 50.8 |
| Repetitive | 77.3 | 70.2 | 72.5 | 71.1 | 49.6 |
| Unique | 77.6 | 70.3 | 72.4 | 68.6 | 46.3 |

**Table 13** Ablation study on using more datasets.

| Dataset | Size | RefCOCO | RefCOCOg | MMBench | MME | MeVIS(val_u) |
|---|---|---|---|---|---|---|
| baseline | 1.2M | 77.4 | 72.3 | 68.3 | 1381/405 | 50.8 |
| Inifinity-MM [27] | 1.2M+3M | 77.1(-0.3) | 72.6(+0.3) | 70.4(+2.1) | 1396/346(-44) | 51.2(+0.4) |
| Ref-SAV | 1.2M+37K | 77.2(-0.2) | 72.6(+0.3) | 68.2(-0.1) | 1384/418(+16) | 52.5(+1.7) |

MeViS, while maintaining the ability for image/video chat, GCG, and visual prompt understanding. At the same time, other referring video segmentation methods only focus on this task, and due to the lack of training data for other tasks, the obtained model will perform poorly on other tasks such as video or image chat.

**Comparison with vision expert models.** As shown in Tab. 16, we compare our model with recent vision expert models designed specifically for referring segmentation. Although these specialist models are typically more lightweight, our general-purpose 26B model (Sa2VA-26B) still achieves the strongest results, outperforming them across all five datasets. This is significant, as these expert models lack conversational capabilities and are mostly limited to a single modality (video or image), whereas our model is a generalist.

**Effectiveness of training dataset on more methods.** We further demonstrate the effectiveness of our proposed Ref-SAV training dataset by evaluating a representative method, UniRef++ [90]. As shown in Tab. 17, we compare two distinct settings to isolate the impact of our data. The zero-shot setting involves directly testing the pre-trained UniRef++ model on the Ref-SAV validation dataset without any new training. In contrast, the fine-tuning setting first trains the model on our proposed Ref-SAV training dataset before evaluation on the same validation set. The results clearly show that the model fine-tuned on our dataset achieves a significant improvement across all metrics. Most notably, the Overall J&F score increases substantially from 10.5 to 14.6. This considerable gain, which is consistent across both "Long" (12.5 to 17.2 J&F) and "Short" (8.6 to 12.0 J&F) validation splits, indicates that our automatic data engine has great potential for boosting the per-

**Table 14** Experiment results of Sa2VA with recent base MLLMs.

| Base MLLM | Image Segmentation | | | Video Segmentation | | | Image Chat | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RefCOCO | refCOCO+ | refCOCOg | MeViS (val_u) | ReVOS | Ref-DAVIS17 | MME | MMBench | SEED-Bench | AI2D | MMStar | $SQA^{test}$ |
| Sa2VA-8B | 81.9 | 76.5 | 78.9 | 57.0 | 57.6 | 75.2 | 1651/578 | 82.4 | 75.5 | 82.1 | 60.3 | 96.8 |
| Sa2VA-InternVL3-2B | 81.4 | 75.7 | 80.3 | 53.9 | 56.2 | 74.5 | 1631/559 | 79.8 | 73.9 | 77.1 | 59.2 | 93.7 |
| Sa2VA-InternVL3-8B | 83.3 | 78.9 | 81.8 | 56.4 | 60.8 | 76.3 | 1743/633 | 83.0 | 76.2 | 84.3 | 65.9 | 97.5 |
| Sa2VA-InternVL3-14B | 83.6 | 79.9 | 83.6 | 59.2 | 60.7 | 76.6 | 1746/724 | 84.3 | 76.6 | 85.2 | 67.4 | 98.7 |
| Sa2VA-Qwen2.5VL-3B | 79.6 | 74.0 | 77.1 | 51.6 | 52.0 | 73.4 | 1533/572 | 78.4 | 73.9 | 81.1 | 57.7 | 80.3 |
| Sa2VA-Qwen2.5VL-7B | 82.4 | 77.5 | 81.5 | 56.4 | 58.3 | 79.4 | 1552/676 | 84.5 | 75.0 | 84.5 | 62.3 | 87.9 |
| Sa2VA-Qwen3VL-4B | 81.7 | 77.4 | 80.0 | 57.1 | 56.7 | 75.9 | 1660/655 | 86.3 | 77.3 | 85.4 | 66.3 | 91.6 |

**Table 15** Comparison with Recent Video MLLMs.

| Model Type | MeViS | ReVOS | Ref-DAVIS17 |
|---|---|---|---|
| PG-Video-LLaVA [67] | 18.9 | - | - |
| GLaMM + SAM2 [68] | 38.7 | - | - |
| VideoGLaMM (3.8B) [68] | 45.2 | - | - |
| VISA-13B [98] | 44.5 | 50.9 | 70.4 |
| VideoLISA-3.8B [4] | 44.4 | - | 68.8 |
| HyperSeg-3B [86] | - | 55.7 | 71.2 |
| InstructSeg [85] | - | 54.5 | 71.1 |
| Sa2VA-4B (ours) | 46.2 | 53.2 | 73.8 |
| Sa2VA-8B (ours) | 46.9 | 57.6 | 75.2 |

**Table 16** Comparison with Vision Expert Models.

| Model Type | RefCOCO | RefCOCO+ | RefCOCOg | MeVIS | Ref-DAVIS17 |
|---|---|---|---|---|---|
| LAVT [101] | 72.7 | 62.1 | 61.2 | - | - |
| ReferFormer [89] | - | - | - | 31.0 | 61.1 |
| UniRef++-L [90] | 81.4 | 74.0 | 76.0 | - | 67.2 |
| EVF-SAM [116] | 82.4 | 76.5 | 78.2 | - | - |
| LMPM [18] | - | - | - | 37.2 | - |
| UniVS [52] | - | - | - | - | 59.4 |
| Sa2VA-26B (ours) | 82.5 | 78.8 | 79.7 | 46.2 | 77.0 |

formance of Ref-VOS models.

**Results on visual prompts understanding tasks.** We also report on visual prompt understanding tasks, following previous works [105, 112]. Specifically, we evaluate region caption performance on the RefCOCOg dataset. As shown in Tab. 18, our method, Sa2VA-4B, also achieves the best results among recent visual prompt understanding models. It scores 17.3 on the METEOR metric, significantly outperforming the next-best method, Osprey [105], which scored 16.6. This indicates that Sa2VA can also generate strong region-aware captions.

## 4.4 Discussions

**Sa2VA with recent MLLM models.** Table 14 presents a comprehensive evaluation of our Sa2VA framework integrated with different recent base Multimodal Large Language Models (MLLMs), including Qwen2.5-VL [3], Qwen3-VL [95], and InternVL-3 [122]. The results clearly demonstrate that the choice of the base MLLM significantly influences performance, and this effect varies across different tasks.

**Table 17** Vision Expert on Ref-SAV validation sets.

| Method | Long | | | Short | | | Overall | | |
|---|---|---|---|---|---|---|---|---|---|
| | J | F | J&F | J | F | J&F | J | F | J&F |
| UniRef++ [90] (zero-shot) | 14.1 | 10.8 | 12.5 | 9.0 | 8.2 | 8.6 | 11.6 | 9.5 | 10.5 |
| UniRef++ (fine-tuning) | 19.2 | 15.1 | 17.2 | 12.3 | 11.7 | 12.0 | 15.8 | 13.4 | 14.6 |

**Table 18** Region caption performance on RefCOCOg dataset.

| Method | Sa2VA-4B | OMG-LLaVA [112] | Osprey [105] | GLaMM [74] | GRIT [88] | Kosmos-2 [71] |
|---|---|---|---|---|---|---|
| METEOR | 17.3 | 15.3 | 16.6 | 16.2 | 15.2 | 14.1 |

For instance, while the Sa2VA-InternVL3-14B model achieves the strongest overall performance, particularly on image and video segmentation benchmarks like RefCOCO (83.6) and MeViS (59.2), other models exhibit competitive strengths in specific domains. Notably, the Sa2VA-Qwen3VL-4B model records the highest score on MMBench (86.3) and shows highly competitive results on other image chat datasets. This variation highlights that different base MLLM architectures possess distinct capabilities, making model selection a critical factor for task-specific optimization. To facilitate further research and allow the community to build upon these findings, we have publicly released all trained model checkpoints.

**Discussion on the inference speed.** The computational overhead in Sa2VA is dominated by the MLLM component due to its auto-regressive nature, where each inference step generates tokens sequentially. In contrast, the primary additional module, SAM-2, is relatively lightweight at 220M parameters (compared to the MLLMs, which range from 1B to 26B) and efficient, achieving 39.5 FPS on vision tasks with only a single forward pass per image. While the variable output length of the MLLM makes precise latency estimation challenging, to provide a more concrete comparison, we also construct a controlled benchmark with fixed frame and language lengths ("Please segment the man." as the text prompt with 5 frames and totaling 1280 tokens) to estimate computational cost more systematically. The results of this benchmark

**Table 19** Computational cost comparision.

| Model | Params | Base MLLM | Base Language Model | GPU | Inference Time |
|---|---|---|---|---|---|
| Sa2VA-1B | 1B | InternVL2.5-1B | Qwen/Qwen2.5-0.5B-Instruct | NVIDIA-H100 | 0.123s |
| Sa2VA-4B | 4B | InternVL2.5-4B | Qwen/Qwen2.5-3B-Instruct | NVIDIA-H100 | 0.282s |
| Sa2VA-8B | 8B | InternVL2.5-8B | internlm/internlm2_5-7b-chat | NVIDIA-H100 | 0.201s |
| Sa2VA-26B | 26B | InternVL2.5-26B | internlm/internlm2_5-20b-chat | NVIDIA-H100 | 0.463s |

are detailed in Table 19. The results show inference times of 0.123s for Sa2VA-1B, 0.282s for Sa2VA-4B, 0.201s for Sa2VA-8B, and 0.463s for Sa2VA-26B. We note that inference speed is largely dependent on optimization, and the discrepancy between the 4B model (0.282s) and the 8B model (0.201s) may be due to the different base LLMs (Qwen vs. InternLM) employed. Nonetheless, these results confirm that the majority of the model's computational cost stems from the auto-regressive MLLM component rather than the efficient, single-pass SAM-2 module.

**Inference Strategy.** We also analyzed the impact of our keyframe sampling strategy. Our default approach of using the first five frames was chosen for simplicity and consistency, but we investigated whether this method limits the model's ability to capture long-range temporal dependencies or objects that appear later in the video. To test this, we performed an ablation study on the MeVIS dataset using the Sa2VA-8B model. We compared the performance of selecting the first 1, 3, 4, and 5 consecutive frames against a uniform sampling strategy that selects 5 frames spread across the entire video. The results of this comparison are detailed in Table 20. As the table shows, performance generally improves as the number of consecutive frames increases, from 55.1 J&F for a single frame to 59.5 J&F for four frames. However, the "First 5 Frames" strategy (58.9 J&F) is significantly outperformed by the "Uniform 5 Frames" strategy, which achieves 62.9 J&F. Notably, both 5-frame methods have an identical inference time (0.207s) and process the same number of image tokens (1280). This finding suggests that more sophisticated, interval-based sampling strategies improve performance. However, we would like to keep Sa2VA as simple as possible and leave the design of advanced sampling techniques to future work.

## 4.5 Visualizations

**Results on image referring segmentation.** As shown in Fig. 5, we visualize the image referring segmentation task. With different language descriptions, our Sa2VA can segment different masks with the clues from the description. Additional qualitative results on video referring segmentation, visual-prompt understanding, and failure cases are provided in the

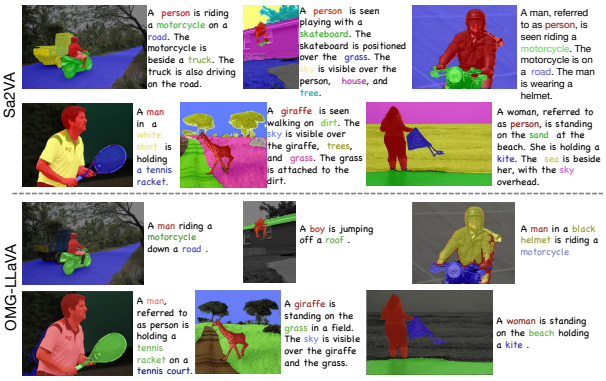



**Figure 4** Visualization results on GCG tasks. Top: our method. Bottom: OMG-LLaVA [112]. Note that our method has stronger, fine-grained grounding and text alignment than OMG-LLaVA [112], a strong previous baseline.

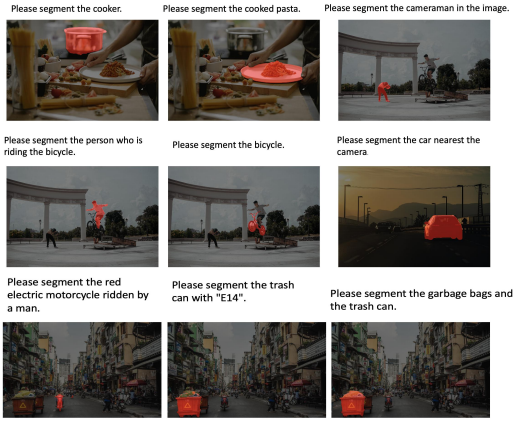



**Figure 5** Visualization results on the image referring segmentation.

Appendix.

**Results on GCG task.** In Fig. 4, we further show several visual examples in grounded caption generation tasks. Compared with the previous SOTA model, OMG-LLaVA, our model achieves better results in both mask quality and text-mask alignments. The former indicates better-aligned segmentation outputs, while the latter shows the latter indicates good text-to-region alignments.

**Video demo.** Since dynamic visual content such as videos cannot be effectively presented within the static format of the manuscript, we have provided additional video demos in the supplementary material. These videos are intended to offer a clearer and more intuitive understanding of our method's performance. Please refer to the ZIP file for the demos.

**Table 20** Ablation study on keyframe sampling strategies using Sa2VA-8B on the MeVIS (val_u) dataset.

| Sampling Strategy | MeVIS J&F (val_u) | Inference Time (per sample) | # Image Tokens |
|---|---|---|---|
| First 1 Frame | 55.1 | 0.151s | 256 |
| First 3 Frames | 58.7 | 0.167s | 768 |
| First 4 Frames | 59.5 | 0.181s | 1024 |
| First 5 Frames | 58.9 | 0.207s | 1280 |
| **Uniform 5 Frames** | **62.9** | **0.207s** | **1280** |

## 5 Failure Cases and Future Work

**Failure cases of Sa2VA and Future work.** Although Sa2VA achieves strong results across various referring segmentation datasets, there are still several failure cases that need to be explored. We identify two shortcomings of Sa2VA in our future work. One is a long video with hard, distinguished referring examples (Appendix). This is because our Sa2VA works in online mode without knowing the entire video content to align the long, complex text. Tab. 20 shows that uniformly sampling 5 frames substantially outperforms using the first 5 frames (62.9 vs. 58.9 J&F on MeViS val_u) at the same inference cost, indicating that broader temporal coverage is beneficial. We therefore view interval-based or adaptive temporal sampling, together with stronger long-range memory, as a principled remedy for long-video alignment. The other open challenge is how to improve VQA performance on both images and videos without degrading referring segmentation. As shown in the ablation table in the main paper, increasing the scale of VQA data leads to a performance drop on referring segmentation tasks. Thus, how to balance these two types of supervision while preserving grounded prediction knowledge remains an unsolved problem.

**Hypothesis on the VQA--grounding conflict.** We hypothesize that this residual imbalance reflects a representation–objective conflict rather than a capacity limitation. VQA is optimized with a next-token language-modeling loss, which favors holistic, sequence-level semantic representations. In contrast, referring segmentation requires the hidden state of the "[SEG]" token to preserve fine-grained, spatially discriminative information for the SAM-2 decoder. As the amount of VQA data increases, the LLM representations may drift toward a global-semantic attractor, thereby weakening the spatial signal carried by "[SEG]". We further conjecture that visual generative objectives, such as image/video synthesis or masked reconstruction, may interfere less with dense grounding because they also require spatially grounded representations, and may even serve as a mutual regularizer. We provide extended analysis and discuss possible mitigations in the Appendix.

**Reinforcement learning as a future direction.** The language-modeling objective used in this work establishes a strong single-stage baseline. A natural next step is reinforcement learning with verifiable rewards, such as mask IoU or boundary F-score, optionally combined with text-side preference signals as in RLHF/GRPO. Such task-specific rewards could inject grounding supervision only where it is needed, thereby mitigating the VQA–grounding interference discussed above and improving robustness on challenging, distinguishing referring expressions in long videos. We leave RL-based post-training to future work.

**Future directions on Ref-SAV benchmarks.** As discussed in Tab. 3, long text, occlusion inputs, and longer videos with camera motion are critical problems in our benchmarks. Thus, the new models need to design a more robust Ref-VOS system to handle these challenges with several specific designs, including more robust memory designs. In addition, understanding long text and improving long text grounding ability is also important to explore.

## 6 Conclusion

In this work, we present Sa2VA, a versatile framework that integrates SAM-2 with MLLMs to achieve a dense, grounded understanding of both images and video. Our method can handle various image and video understanding tasks, including referring image/video segmentation and image/video conversation, with just single-stage instruction tuning. By leveraging the knowledge from both MLLM and SAM-2, our model has strong capabilities in both mask and language generation. To demonstrate the effectiveness of our proposed method, we propose a challenging referring video object segmentation benchmark, Ref-SAV. Extensive experimental results show that Sa2VA achieves strong performance on various tasks across benchmarks. In addition, Sa2VA can also be extended with various modern MLLMs, with potential for building stronger baselines for pixel multimodal systems.

## Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior / Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and

conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government. This work was also supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project).

## References

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In ICCV, 2015.

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966, 2023.

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025.

[4] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. NeurIPS, 2024.

[5] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In NeurIPS, 2020.

[6] Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. Making large multimodal models understand arbitrary visual prompts. In CVPR, 2024.

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In ECCV, 2020.

[8] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. Pix2video: Video editing using image diffusion. In ICCV, 2023.

[9] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. arXiv preprint arXiv:2406.16866, 2024.

[10] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? arXiv preprint arXiv:2403.20330, 2024.

[11] Wei-Ge Chen, Irina Spiridonova, Jianwei Yang, Jianfeng Gao, and Chunyuan Li. Llava-interactive: An all-in-one demo for image chat, segmentation, generation and editing. arXiv preprint arXiv:2311.00571, 2023.

[12] Xinghao Chen, Siwei Li, Yijing Yang, and Yunhe Wang. Deco: Unleashing the potential of convnets for query-based detection and segmentation. In ICLR, 2025.

[13] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271, 2024.

[14] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. arXiv preprint arXiv:2404.16821, 2024.

[15] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In CVPR, 2022.

[16] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Joon-Young Lee, and Alexander Schwing. Putting the object back into video object segmentation. In CVPR, 2024.

[17] XTuner Contributors. Xtuner: A toolkit for efficiently fine-tuning llm. https://github.com/InternLM/xtuner, 2023.

[18] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In ICCV, 2023.

[19] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. arXiv preprint arXiv:2401.16420, 2024.

[20] Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Songyang Zhang, Haodong Duan, Wenwei Zhang, Yining Li, Hang Yan, Yang Gao, Zhe Chen, Xinyue Zhang, Wei Li, Jingwen Li, Wenhai Wang, Kai Chen, Conghui He, Xingcheng Zhang, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer2-4khd: A pioneering large vision-language model handling resolutions from 336 pixels to 4k hd. In NeurIPS, 2024.

[21] Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In ACMMM, 2024.

[22] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. arXiv preprint arXiv:2406.14515, 2024.

[23] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In NeurIPS, 2024.

[24] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023.

[25] Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. arXiv preprint arXiv:2405.21075, 2024.

[26] Brent Griffin, Victoria Florence, and Jason Corso. Video object segmentation-based visual servo control and object depth estimation on a mobile robot. In WACV, 2020.

[27] Shuhao Gu, Jialing Zhang, Siyuan Zhou, Kevin Yu, Zhaohu Xing, Liangdong Wang, Zhou Cao, Jintao Jia, Zhuoyi Zhang, Yixuan Wang, et al. Infinity-mm: Scaling multimodal performance with large-scale and high-quality instruction data. arXiv preprint arXiv:2410.18558, 2024.

[28] Tianrui Guan, Divya Kothandaraman, Rohan Chandra, Adarsh Jagan Sathyamoorthy, Kasun Weerakoon, and Dinesh Manocha. Ga-nav: Efficient terrain segmentation for robot navigation in unstructured outdoor environments. RA-L, 2022.

[29] Tianrui Guan, Ruitao Song, Zhixian Ye, and Liangjun Zhang. Vinet: Visual and inertial-based terrain classification and adaptive navigation over unknown terrain. In ICRA, 2023.

[30] Pinxue Guo, Tony Huang, Peiyang He, Xuefeng Liu, Tianjun Xiao, Zhaoyu Chen, and Wenqiang Zhang. Openvis: Open-vocabulary video instance segmentation. arXiv preprint arXiv:2305.16835, 2023.

[31] Kai Han, Jianyuan Guo, Yehui Tang, Wei He, Enhua Wu, and Yunhe Wang. Free video-llm: Prompt-guided visual perception for efficient training-free video llms. arXiv preprint arXiv:2410.10441, 2024.

[32] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. Animate-a-story: Storytelling with retrieval-augmented video generation. arXiv preprint arXiv:2307.06940, 2023.

[33] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In CVPR, 2024.

[34] Kuan-Chih Huang, Xiangtai Li, Lu Qi, Shuicheng Yan, and Ming-Hsuan Yang. Reason3d: Searching and reasoning 3d segmentation via large language model. In 3DV, 2025.

[35] Nisha Huang, Yuxin Zhang, and Weiming Dong. Style-a-video: Agile diffusion for arbitrary text-based video style transfer. SPL, 2024.

[36] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. arXiv preprint arXiv:2004.00849, 2020.

[37] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In CVPR, 2019.

[38] Louis Martin Hugo Touvron, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and et al. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288, 2023.

[39] Sukjun Hwang, Miran Heo, Seoung Wug Oh, and Seon Joo Kim. Video instance segmentation using inter-frame communication transformers. In NeurIPS, 2021.

[40] Shibo Jie, Yehui Tang, Ning Ding, Zhi-Hong Deng, Kai Han, and Yunhe Wang. Memory-space visual prompting for efficient vision-language fine-tuning. arXiv preprint arXiv:2405.05615, 2024.

[41] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In EMNLP, 2014.

[42] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In ECCV, 2016.

[43] Anna Khoreva, Anna Rohrbach, and Bernt Schiele. Video object segmentation with language referring expressions. In ACCV, 2018.

[44] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. In ICML, 2024.

[45] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In CVPR, 2024.

[46] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326, 2024.

[47] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In CVPR, 2024.

[48] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. arXiv preprint arXiv:2410.05993, 2024.

[49] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In ICML, 2022.

[50] 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 ICML, 2023.

[51] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355, 2023.

[52] Minghan Li, Shuai Li, Xindong Zhang, and Lei Zhang. Univs: Unified and universal video segmentation with prompts as queries. In CVPR, 2024.

[53] Xiangtai Li, Haobo Yuan, Wenwei Zhang, Guangliang Cheng, Jiangmiao Pang, and Chen Change Loy. Tube-link: A flexible cross tube baseline for universal video segmentation. In ICCV, 2023.

[54] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In CVPR, 2024.

[55] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In ECCV, 2024.

[56] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In EMNLP, 2023.

[57] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In EMNLP, 2024.

[58] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In CVPR, 2023.

[59] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In NeurIPS, 2023.

[60] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In CVPR, 2024.

[61] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[62] Shaoteng Liu, Yuechen Zhang, Wenbo Li, Zhe Lin, and Jiaya Jia. Video-p2p: Video editing with cross-attention control. In CVPR, 2024.

[63] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun Zhu, et al. Llava-plus: Learning to use tools for creating multimodal agents. In ECCV, 2024.

[64] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In ECCV, 2024.

[65] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In NeurIPS, 2022.

[66] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In ACL, 2024.

[67] 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.

[68] Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. arXiv preprint arXiv:2411.04923, 2024.

[69] Zhenliang Ni, Xinghao Chen, Yingjie Zhai, Yehui Tang, and Yunhe Wang. Context-guided spatial feature reconstruction for efficient semantic segmentation. In ECCV, 2024.

[70] Prashant W Patil, Akshay Dudhane, Ashutosh Kulkarni, Subrahmanyam Murala, Anil Balaji Gonde, and Sunil Gupta. An unified recurrent video object segmentation framework for various surveillance environments. IEEE TIP, 2021.

[71] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824, 2023.

[72] Lu Qi, Yi-Wen Chen, Lehan Yang, Tiancheng Shen, Xiangtai Li, Weidong Guo, Yu Xu, and Ming-Hsuan Yang. Generalizable entity grounding via assistance of large language model. arXiv preprint arXiv:2402.02555, 2024.

[73] Miao Rang, Zhenni Bi, Chuanjian Liu, Yehui Tang, Kai Han, and Yunhe Wang. Eve: Efficient multimodal vision language models with elastic visual experts. arXiv preprint arXiv:2501.04322, 2025.

[74] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S. Khan. Glamm: Pixel grounding large multimodal model. In CVPR, 2024.

[75] 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.

[76] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In CVPR, 2024.

[77] Seonguk Seo, Joon-Young Lee, and Bohyung Han. URVOS: Unified referring video object segmentation network with a large-scale benchmark. In ECCV, 2020.

[78] Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. Prompting large language models with answer heuristics for knowledge-based visual question answering. In CVPR, 2023.

[79] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023.

[80] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Austin Wang, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In NeurIPS, 2024.

[81] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models. arXiv:2302.13971, 2023.

[82] Wen Wang, Yan Jiang, Kangyang Xie, Zide Liu, Hao Chen, Yue Cao, Xinlong Wang, and Chunhua Shen. Zero-shot video editing using off-the-shelf image diffusion models. arXiv preprint, 2023.

[83] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. In arXiv, 2024.

[84] Cong Wei, Haoxian Tan, Yujie Zhong, Yujiu Yang, and Lin Ma. LaSagnA: Language-based segmentation assistant for complex queries. arXiv preprint arXiv:2404.08506, 2024.

[85] Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Zheng Zhao, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. arXiv preprint, 2024.

[86] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model. In CVPR, 2025.

[87] Fangyun Wei, Jinjing Zhao, Kun Yan, Hongyang Zhang, and Chang Xu. A large-scale human-centric benchmark for referring expression comprehension in the LMM era. In NeurIPS, 2024.

[88] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In ECCV, 2024.

[89] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In CVPR, 2022.

[90] Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, and Ping Luo. Uniref++: Segment every reference object in spatial and temporal spaces. In ICCV, 2023.

[91] Jianzong Wu, Xiangtai Li, Chenyang Si, Shangchen Zhou, Jingkang Yang, Jiangning Zhang, Yining Li, Kai Chen, Yunhai Tong, Ziwei Liu, et al. Towards language-driven video inpainting via multimodal large language models. In CVPR, 2024.

[92] Qiong Wu, Xiangcong Yang, Yiyi Zhou, Chenxin Fang, Baiyang Song, Xiaoshuai Sun, and Rongrong Ji. Grounded chain-of-thought for multimodal large language models. arXiv preprint arXiv:2503.12799, 2025.

[93] Size Wu, Sheng Jin, Wenwei Zhang, Lumin Xu, Wentao Liu, Wei Li, and Chen Change Loy. F-lmm: Grounding frozen large multimodal models. arXiv preprint arXiv:2406.05821, 2024.

[94] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In CVPR, 2024.

[95] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. arXiv preprint arXiv:2509.17765, 2025.

[96] Lin Xu, Yilin Zhao, Daquan Zhou, Zhijie Lin, See Kiong Ng, and Jiashi Feng. Pllava: Parameter-free llava extension from images to videos for video dense captioning. arXiv preprint arXiv:2404.16994, 2024.

[97] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In CVPR, 2023.

[98] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In ECCV, 2024.

[99] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. Qwen2 technical report. arXiv preprint arXiv:2407.10671, 2024.

[100] Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seed-story: Multimodal long story generation with large language model. arXiv preprint arXiv:2407.08683, 2024.

[101] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In CVPR, 2022.

[102] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl3: Towards long image-sequence understanding in multi-modal large language models. arXiv preprint, 2024.

[103] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In ECCV, 2016.

[104] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In CVPR, 2018.

[105] Yuqian Yuan, Wentong Li, Jian Liu, Dongqi Tang, Xinjie Luo, Chi Qin, Lei Zhang, and Jianke Zhu. Osprey: Pixel understanding with visual instruction tuning. In CVPR, 2024.

[106] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In CVPR, 2024.

[107] Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models. IJCV, 2024.

[108] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In EMNLP, 2023.

[109] Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Chunyuan Li, Jainwei Yang, et al. Llava-grounding: Grounded visual chat with large multimodal models. In ECCV, 2024.

[110] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei

Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. arXiv preprint arXiv:2309.15112, 2023.

[111] Tao Zhang, Xingye Tian, Yu Wu, Shunping Ji, Xuebo Wang, Yuan Zhang, and Pengfei Wan. DVIS: Decoupled video instance segmentation framework. In ICCV, 2023.

[112] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Change Loy Chen, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In NeurIPS, 2024.

[113] Tao Zhang, Xiangtai Li, Zilong Huang, Yanwei Li, Weixian Lei, Xueqing Deng, Shihao Chen, Shunping Ji, and Jiashi Feng. Pixel-sail: Single transformer for pixel-grounded understanding. arXiv preprint arXiv:2504.10465, 2025.

[114] Tao Zhang, Xingye Tian, Yikang Zhou, Shunping Ji, Xuebo Wang, Xin Tao, Yuan Zhang, Pengfei Wan, Zhongyuan Wang, and Yu Wu. Dvis++: Improved decoupled framework for universal video segmentation. IEEE TPAMI, 2025.

[115] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In CVPR, 2024.

[116] Yuxuan Zhang, Tianheng Cheng, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. arXiv preprint arXiv:2406.20076, 2024.

[117] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. In ECCV, 2024.

[118] Hao Zhou, Tiancheng Shen, Xu Yang, Hai Huang, Xiangtai Li, Lu Qi, and Ming-Hsuan Yang. Rethinking evaluation metrics of open-vocabulary segmentaion. arXiv preprint arXiv:2311.03352, 2023.

[119] Yikang Zhou, Tao Zhang, Shunping Ji, Shuicheng Yan, and Xiangtai Li. Improving video segmentation via dynamic anchor queries. In ECCV, 2024.

[120] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. arXiv preprint arXiv:2405.01434, 2024.

[121] Feng Zhu, Zongxin Yang, Xin Yu, Yi Yang, and Yunchao Wei. Instance as identity: A generic online paradigm for video instance segmentation. In ECCV, 2022.

[122] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479, 2025.

[123] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. arXiv preprint arXiv:2010.04159, 2020.

# Appendix

## A Concrete Benefits of the Unified Framework

Sa2VA employs a single set of weights for all tasks, rather than training a separate model for each input type and task. Beyond the summary in Sec 3.2, we detail four benefits of this unified design, each supported by the experiments in the main paper.

**Positive cross-task transfer.** Joint co-training across the image/video × chat/segmentation grid yields gains that single-task training cannot recover. As shown in the co-training ablation (Tab. 9), removing any single modality of training data causes large drops; for instance, removing image-segmentation data collapses RefCOCO from 77.4 to 20.2 cIoU, and removing video-segmentation data degrades MeViS and Ref-DAVIS17 by 6.4 and 3.3 J&F. The unified paradigm is thus essential, independently of any newly introduced data.

**No catastrophic forgetting of conversational ability.** A recurring failure mode of grounding MLLMs is that instruction tuning for segmentation destroys their chat ability. Sa2VA instead retains the conversational performance of its base MLLM across MME, MMBench, SEED-Bench, and AI2D while simultaneously reaching state-of-the-art grounding (Tab. 6, Tab. 7).

**Lower maintenance than per-dataset fine-tuning.** Fine-tuning the model on each individual dataset yields only marginal improvements (Tab. 11, 1.5% on the referring sets and only 0.4 on RefCOCO for single-dataset fine-tuning), yet it requires a separately stored and maintained checkpoint per dataset. The unified model attains comparable or better accuracy with a single checkpoint.

**Deployment and inference efficiency.** A single set of weights serves all tasks with no per-task model switching cost. The appended SAM-2 adds only ≈220M parameters and runs at 39.5 FPS, so the overall computational cost is dominated by the shared autoregressive MLLM (Tab. 19: 0.123 s for Sa2VA-1B, 0.282 s for 4B, 0.201 s for 8B, and 0.463 s for 26B). The visual-prompt / "[SEG]" token budget is identical across tasks, so unification does not introduce additional token overhead.

## B Annotation Model Choice and Validation

**Choice of InternVL2-76B.** We adopted InternVL2-76B as the primary captioner because it was one of the strongest open-source MLLMs available when the Ref-SAV training set was built. Importantly, the annotation pipeline (Fig. 3, Sec. 3.3) is not a single-model pseudo-labeler: object-level captions generated by InternVL2-76B are cross-checked by Qwen2-72B, and inconsistent or unrecognizable cases are discarded. This two-model agreement filter reduces the risk of systematic single-model bias.

**Decoupled annotation model and backbones.** Using an InternVL-family model for both annotation and the backbone does not introduce a family-internal advantage. First, the annotation model is InternVL2-76B, whereas the main experiments adopt backbones spanning InternVL2-1B/4B/8B/26B, InternVL2.5, InternVL3, and the unrelated Qwen2.5-VL and Qwen3-VL families (Tab. 8 and Tab. 14), which have distinct weights and even distinct model families. Second, Ref-SAV transfers cleanly to backbones outside the InternVL family: UniRef++, which is unrelated to InternVL, improves from 10.5 to 14.6 overall J&F after fine-tuning on Ref-SAV (Tab. 17), and the Qwen-family backbones obtain similarly large gains (Tab. 14). Such cross-family transfer would not arise from family-internal label leakage.

**Validation of annotation quality.** The Ref-SAV evaluation is anchored to human ground truth: the short-expression evaluation set is fully human-annotated, and the long-expression evaluation set is human-filtered (Sec. 3.3). Therefore all reported test numbers measure agreement with human labels rather than with the automatic captioner. As external validation, Tab. 17 shows that an independent method consistently improves after training on our automatic labels, across both the long (12.5→17.2 J&F) and short (8.6→12.0 J&F) splits, indicating that the automatic data engine produces useful supervision.

## C Generalization to the Newer Segmenter: SAM-3

Generalization to evolving segmentation models is an explicit design goal of Sa2VA. The decoupled design (Sec. 3.2) keeps the segmenter's encoder and memory frozen and bridges to the MLLM only through the spatial-temporal "[SEG]" prompt token, so any prompt-able segmentation decoder can be plugged in without modifying the MLLM side. To verify this, we replace SAM-2 with the more recent SAM-

**Table A1** Ablation study on the segmenter choice under a fixed Sa2VA pipeline (Qwen3-VL-4B).

| Segmenter | RefCOCO (cIoU) | RefCOCO+ (cIoU) | RefCOCOg (cIoU) | MeViS (J&F) | ReVOS (J&F) | Ref-DAVIS17 (J&F) |
|---|---|---|---|---|---|---|
| SAM-2 | 83.6 | 79.2 | 82.4 | 61.4 | 61.4 | 76.2 |
| SAM-3 | **83.7** | **79.4** | **83.0** | **65.3** | **66.3** | **77.1** |

**Table A2** Controlled same-backbone comparison against LISA-7B. All Sa2VA variants share an identical training recipe.

| Model (backbone) | RefCOCO (cIoU) | RefCOCO+ (cIoU) | RefCOCOg (cIoU) | MeViS (J&F) | ReVOS (J&F) | Ref-DAVIS17 (J&F) |
|---|---|---|---|---|---|---|
| LISA-7B (LLaVA-1.5-7B)† | 74.9 | 65.1 | 67.9 | – | – | – |
| Sa2VA (LLaVA-1.5-7B) | 80.3 | 73.1 | 79.2 | 54.8 | 54.0 | 74.6 |
| Sa2VA (Qwen2.5-VL-7B) | 82.3 | 77.1 | 80.5 | 58.1 | 58.6 | **77.4** |
| Sa2VA (Qwen3-VL-4B) | **83.6** | **79.2** | **82.4** | **61.4** | **61.4** | 76.2 |

3 while keeping the rest of the pipeline unchanged, using a Qwen3VL-4B backbone. Since SAM-3 retains a promptable visual segmentation interface, the "[SEG]" token transfers directly through the existing prompt projection. To isolate the effect of the segmenter, both rows are trained under an identical recipe without the Ref-SAV data, so that the comparison reflects the segmenter alone rather than any additional training data. We train Sa2VA with publicly released codebase and have released the model and training config for reproduce.

As reported in Tab. A1, upgrading the segmenter consistently improves performance, with the largest gains on the video benchmarks (e.g., MeViS 61.4 → 65.3 and ReVOS 61.4 → 66.3 J&F). This result supports two conclusions: (i) the framework is modular with respect to the segmenter, as the decoupled design and the "[SEG]" bridge enable a drop-in replacement; and (ii) the framework benefits from stronger segmenters without retraining the MLLM. We note that the contribution of Sa2VA lies not in adopting a stronger segmenter, but in unifying diverse tasks within a single architecture. The SAM-3 integration is also released in our public repository to facilitate further research.

## D Backbone Fairness: Same-Backbone Comparison

To isolate the contribution of our unified paradigm from that of a stronger backbone, we drop a LISA-comparable language backbone—the original LLaVA-1.5-7B—into our framework and compare against LISA-7B under the same backbone. All Sa2VA rows in Tab. A2 are trained from scratch under an identical recipe (1 epoch, LoRA $r = 128$, frozen LLM and visual encoder; image scores are val cIoU and video scores are J&F), without the Ref-SAV data, so that the comparison reflects the backbone alone rather than any additional training data. On the same LLaVA-1.5-7B backbone, our paradigm lifts referring segmentation well above LISA-7B (+5.4 RefCOCO, +8.0 RefCOCO+, +11.3 RefCOCOg) and additionally endows it with referring video segmentation that LISA does not support at all. The gain therefore stems from the unified paradigm rather than a stronger backbone; swapping in stronger MLLM backbones (Qwen2.5-VL, Qwen3-VL) improves the numbers further, but even the deliberately weak LLaVA-1.5 backbone already surpasses the prior specialist. We train Sa2VA with our publicly released codebase and have released the trained models and training configs for reproducibility.

## E Disentangling Data and Paradigm

A key question is whether the performance of Sa2VA stems from the newly constructed data or from the unified training paradigm. We first note that Sa2VA is efficient to train: it requires no pre-training stage and only a single epoch of supervised instruction tuning with LoRA ($r = 256$), taking ≈24 hours on 16 NVIDIA H100 80GB GPUs (bf16, total batch size 256), as detailed in Sec. 4. We then disentangle the two factors from both directions.

**The paradigm matters with data held fixed.** The co-training ablation (Tab. 9) is conducted without the Ref-SAV data, and removing any single modality of training data already causes large drops. Hence

the joint paradigm is necessary independently of the new dataset.

**The same data cannot compensate for a weaker paradigm.** Fine-tuning a non-unified baseline (UniRef++) on the same Ref-SAV data improves its overall J&F only from 10.5 to 14.6 (Tab. 17), far below Sa2VA. Under an identical data-construction cost, a conventional paradigm cannot approach the unified framework, indicating an inherent advantage of the framework rather than a mere data advantage.

**Targeted data helps, but only through the paradigm.** With the paradigm fixed, adding 3M generic VQA samples barely affects segmentation, whereas adding only 37K targeted Ref-SAV samples yields +1.7 J&F on MeViS (Tab. 13). The paradigm requires data targeted to the task, and the data is effective only within the paradigm. In summary, data and paradigm are complementary, and neither factor alone accounts for the performance of Sa2VA.

## F The VQA--Grounding Representation Conflict

We expand on the hypothesis introduced in Sec. 4.4 regarding why scaling up VQA data can hurt referring segmentation.

**Why VQA and grounding compete in the shared token space.** VQA is optimized with a next-token language-modeling loss whose gradient pressure favors holistic, sequence-level semantic representations: visual tokens are encouraged to be compressed into a compact summary suitable for autoregressive answer generation. Referring segmentation, in contrast, requires the hidden state of the "[SEG]" token to retain fine-grained, spatially discriminative information, so that the SAM-2 decoder can localize the correct pixels. As VQA data is scaled up, the representations of the LLM drift toward the global-semantic objective, weakening the spatial signal carried by the "[SEG]" token and degrading segmentation. This representation-objective conflict is most visible in Tab. 9: removing image-segmentation supervision collapses RefCOCO from 77.4 to 20.2 cIoU, showing how fragile the spatial signal carried by the "[SEG]" token is once the training objective shifts away from dense grounding. Consistent with this, scaling generic VQA data (Tab. 13) yields no segmentation improvement even as it raises chat scores.

**Visual generative tasks may align better with dense grounding.** Generative objectives such as image/video synthesis or masked reconstruction require spatially grounded representations for pixel-level reconstruction, which are structurally closer to dense grounding than the holistic-semantic objective of VQA. We thus conjecture that jointly training with generative objectives may interfere less with grounding than VQA does, and could even act as a mutual regularizer. Validating this empirically, for example by adding a masked image/video reconstruction head or a diffusion-style auxiliary loss, is a promising direction for future work.

**Possible mitigations within the current paradigm.** Beyond generative co-training, we are exploring (i) task-specific token subspaces with lightweight gating, (ii) representation-level regularization that constrains "[SEG]" features to a separable subspace, and (iii) reinforcement learning with verifiable segmentation rewards (Sec. 5).

## G Additional Qualitative Results

Owing to the page limit, the following qualitative results on video referring segmentation, visual prompt understanding, and failure cases are presented here rather than in the main text.

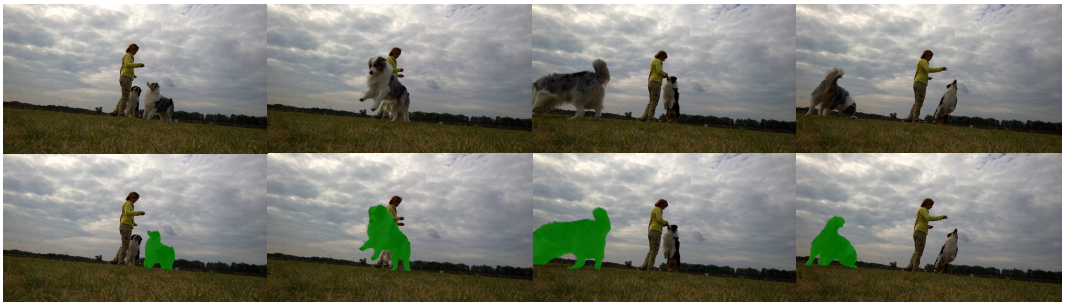
**Results on video referring segmentation.** As shown in Fig. A1, our method performs well in handling diverse and challenging conditions, showcasing a remarkable ability to adjust and perform effectively even in occlusion scenes or highly dynamic environments.

**Results on visual prompt understanding.** As shown in Fig. A2, our method can generate descriptions based on the visual prompts. The descriptions generated by our method demonstrate a high degree of contextual awareness, effectively capturing details within the visual cues.

**Failure cases.** As shown in Fig. A3, we present representative failure cases, which mainly arise from long videos with hard, fine-grained referring expressions, as discussed in Sec. 5.



Please segment the object is a dark-colored backpack with light-colored accents, featuring multiple compartments and pockets, securely fastened to an individual's back. The person is dressed in dark clothing and ascending an escalator in a public setting, likely a mall or transportation hub. The backpack has adjustable straps and a top handle, appearing functional for carrying various items. The individual moves steadily up the escalator, indicating a purposeful journey.



Please segment a white dog with black patches.

**Figure A1** Visualization results on video referring segmentation.




**Figure A2** Visualization results on the visual prompt understanding task. We use the masks predicted by our model under the GCG task as visual prompts, and generate region-level descriptions for these masks. The object masks and their corresponding region captions are highlighted in the same color.



The object is a silver car, captured from the rear, driving on a busy road at night. Its brake lights are illuminated, suggesting it is either slowing down or stopped. The car's make is visible on the back, and the license plate is clearly displayed. It is among various vehicles, including a white SUV directly in front and a truck to the left. Throughout the scene, the car remains in motion, navigating through traffic alongside other vehicles.



The object is a laptop. It starts inside a black bag on a desk and is taken by a person wearing a red outfit, who is sitting on a black rolling office chair. The person places the laptop on the desk, opens it, and interacts with it by typing on the keyboard and using the trackpad. The laptop stays on the desk in front of the person with the screen facing them. The setting appears to be an office or classroom, as indicated by a whiteboard in the background.

**Figure A3** Visualization failure cases.