X2SAM: Any Segmentation in Images and Videos

Hao Wang[1,2], Limeng Qiao[3], Chi Zhang[3], Lin Ma[3], Guanglu Wan[3], Xiangyuan Lan[2⋆], and Xiaodan Liang[1,2⋆]

[1] Shenzhen Campus of Sun Yat-sen University, China
[2] Peng Cheng Laboratory, China
[3] Meituan Inc., China

Abstract. Multimodal Large Language Models (MLLMs) have demonstrated strong image-level visual understanding and reasoning, yet their pixel-level perception across both images and videos remains limited. Foundation segmentation models such as the SAM series produce high-quality masks, but they rely on low-level visual prompts and cannot natively interpret complex conversational instructions. Existing segmentation MLLMs narrow this gap, but are usually specialized for either images or videos and rarely support both textual and visual prompts in one interface. We introduce X2SAM, a unified segmentation MLLM that extends any-segmentation capabilities from images to videos. Given conversational instructions and visual prompts, X2SAM couples an LLM with a Mask Memory module that stores guided vision features for temporally consistent video mask generation. The same formulation supports generic, open-vocabulary, referring, reasoning, grounded conversation generation, interactive, and visual grounded segmentation across image and video inputs. We further introduce the Video Visual Grounded (V-VGD) segmentation benchmark, which evaluates whether a model can segment object tracks in videos from interactive visual prompts. With a unified joint training strategy over heterogeneous image and video datasets, X2SAM delivers strong video segmentation performance, remains competitive on image segmentation benchmarks, and preserves general image and video chat ability. Code is available at `https://github.com/wanghao9610/X2SAM`.

Keywords: Any Segmentation · Multimodal Large Language Model

1 Introduction

Multi-modal Large Language Models (MLLMs) have exhibited substantial advancements alongside the rapid development of Large Language Models (LLMs) [3, 38] and multi-modal pre-training methods [12, 33]. These models have shown remarkable effectiveness in a wide range of applications, including image captioning [45], VQA [1], and visual editing [7]. However, while current MLLMs excel at global visual understanding, their capability to generate dense, pixel-level outputs for precise spatial and temporal comprehension remains limited.

⋆ Corresponding authors: `lanxy@pcl.ac.cn`, `xdliang328@gmail.com`.

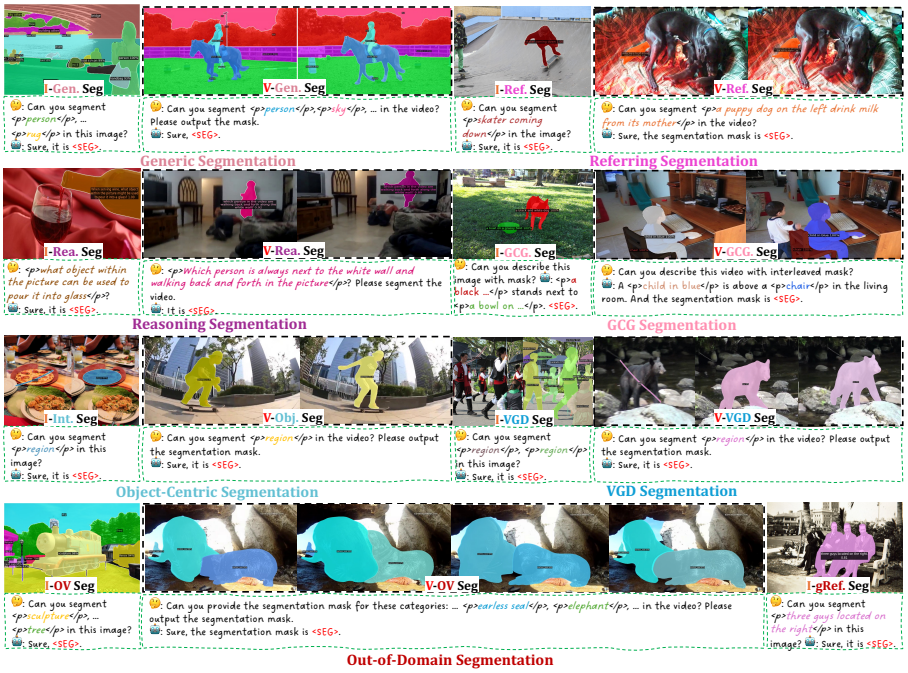

Fig. 1: Comprehensive capabilities of X2SAM. Guided by conversational instructions, X2SAM performs diverse Image and Video segmentation tasks. X2SAM provides a unified interface for a wide range of image and video segmentation tasks, including generic, referring, reasoning, grounded conversation generation, interactive, visual grounded, object-centric, and out-of-domain segmentation. This design highlights how image-level any-segmentation capabilities can be extended to video inputs with textual and visual prompts.

This limitation poses a considerable challenge in directly addressing fine-grained tasks across both static images and dynamic video sequences.

Foundation segmentation models, such as SAM [13] and its video-extended successor SAM2 [36], generate dense masks across spatial and temporal domains. Nevertheless, they depend on explicit low-level visual prompts (e.g., points or boxes) and cannot natively interpret complex conversational text instructions. Conversely, as illustrated in Fig. 2, recent segmentation MLLMs have attempted to bridge language understanding and mask generation, but they remain structurally fragmented. Image segmentation MLLMs (e.g., LISA [14]) process textual instructions but are restricted to static images and usually lack visual prompting support. Video segmentation MLLMs (e.g., VISA [47], VideoLISA [6]) support temporal text-to-mask generation but do not provide a unified architecture for both static images and visual prompts. Achieving a single framework that interprets complex multi-modal instructions, including both text and visual prompts, for segmentation across images and videos remains a critical challenge.

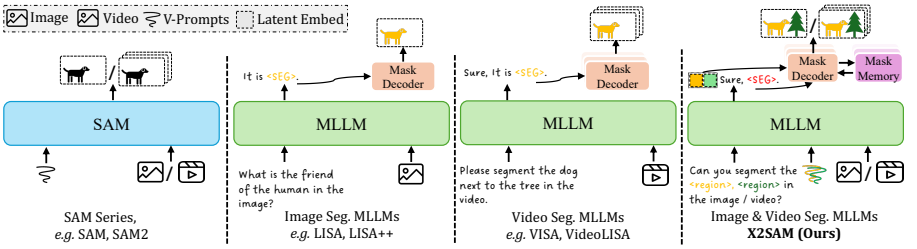

Fig. 2: Comparison of X2SAM with existing methods, including SAM Series models, Image Segmentation MLLMs, and Video Segmentation MLLMs. X2SAM processes both textual and visual prompts in a shared image-video segmentation framework, improving coverage over prior image-only or video-only segmentation MLLMs.

In this work, we introduce X2SAM, a framework that unifies diverse image and video segmentation tasks and extends the image-centric any-segmentation paradigm toward a unified image-and-video setting. As illustrated in Fig. 1, X2SAM provides a conversational interface for text-driven and visually prompted segmentation across static images and dynamic videos. To realize this capability and overcome limitations of prior paradigms (Fig. 2), our approach addresses three technical challenges: (1) *Comprehensive Prompt Integration*: augmenting LLMs to process interleaved textual instructions and visual prompts (V-Prompts) for both image and video inputs. (2) *Spatio-Temporal Task Formulation*: casting diverse image segmentation paradigms into a shared formulation that can represent video targets over time. (3) *Temporal Coherence via Mask Memory*: replacing independent frame-by-frame decoding with a Mask Memory module that interacts with the Mask Decoder and stores guided vision features to maintain mask consistency across video sequences.

As illustrated in Fig. 3, we develop a unified MLLM architecture that processes global visual representations and fine-grained visual features. Guided by latent condition embeddings from the LLM, the Mask Decoder works with the newly introduced Mask Memory module to generate temporally consistent segmentation masks. Moreover, we expand the visual prompting capabilities of MLLMs by introducing the Video Visual Grounded (V-VGD) segmentation task. This task equips the model to segment any instance object in a video using interactive visual prompts, grounding targets across frames.

As shown in Tab. 1, we compare X2SAM with existing methods across inputs, outputs, and tasks. X2SAM is the first to natively support seven segmentation tasks, e.g., generic, open-vocabulary, referring, reasoning, grounded conversation generation, object-centric, and visual grounded segmentation, for image and videos.

Supported by a unified joint training strategy that accelerates learning across multi-modalities, X2SAM undergoes co-training with a diverse range of image and video datasets. Experimental results show that X2SAM achieves strong performance across image and video benchmarks, with particularly consistent gains

on video segmentation tasks, establishing a practical baseline for unified pixel-level spatio-temporal understanding. In summary, our contributions are as follows:

- We introduce X2SAM, a unified framework that extends the any segmentation paradigm from images to videos. By integrating an MLLM with a Mask Memory module, X2SAM formulates diverse image and video segmentation tasks into a standardized, temporally consistent format.
- We propose a new benchmark, Video Visual Grounded (V-VGD) segmentation, which provides interactive visual prompts for MLLMs to ground and segment instance objects consistently across video frames.
- We present a unified joint training strategy to co-train X2SAM on both image and video data. Extensive evaluations show that X2SAM supports a broad set of segmentation tasks, remains competitive on image benchmarks, and achieves strong results on video and out-of-domain evaluations.

2 Related Work

Multi-modal Large Language Model. Multi-modal learning has witnessed progressive developments alongside the rapid evolution of Large Language Models (LLMs) [3, 38] and multi-modal pre-training methods [12, 33]. The field has evolved from early models focused on task-specific fusion and feature extraction [16], to generalized, instruction-tuned frameworks leveraging visual feature tokenization [21, 23]. While current MLLMs demonstrate remarkable effectiveness in global visual understanding tasks such as image captioning [45] and VQA [1], their capability to generate dense, pixel-level outputs for precise spatial and temporal comprehension remains highly limited. This poses a considerable challenge when directly addressing fine-grained tasks across static images and dynamic video sequences.

Image Segmentation MLLMs. Foundation models like SAM [13] and its extensions [36] have profoundly impacted the segmentation landscape by introducing visual grounding signals, vastly improving mask generation performance. Building upon this, researchers have explored combining MLLMs with segmentation models to handle open-world challenges, unified task architectures [2, 11], and language-guided tasks [17, 51]. Image segmentation MLLMs, such as LISA [14], successfully process complex textual instructions to output segmentation masks. However, these models are structurally restricted to static images and frequently lack comprehensive support for interactive visual prompting (V-Prompts), limiting their ability to treat grounded visual inputs as freely as textual inputs.

Video Segmentation MLLMs. Extending dense segmentation capabilities to dynamic video sequences introduces significant temporal complexities [18, 40]. Recent video segmentation MLLMs, including VISA [47] and VideoLISA [6], have attempted to bridge this gap by enabling temporal text-to-mask generation. Despite these advancements, the current landscape remains structurally

Table 1: Comparison of Chat-based and Segmentation-based MLLMs across inputs, outputs, and tasks.

Method	Inputs				Outputs		Tasks			
	Image	*Video*	*T-Prompts*	*V-Prompts*	*Text*	*Mask*	*Image-Chat*	*Video-Chat*	*#Image-Seg.*	*#Video-Seg.*
Chat-based MLLMs										
LLaVA [24]	✓	✗	✗	✗	✓	✗	✓	✗	0	0
LLaVA-Next [22]	✓	✓	✓	✗	✓	✗	✓	✓	0	0
LLaVA-OV [15]	✓	✓	✓	✗	✓	✗	✓	✓	0	0
Intern-VL [8]	✓	✓	✓	✗	✓	✗	✓	✓	0	0
Qwen-VL [4]	✓	✓	✓	✗	✓	✗	✓	✓	0	0
Seg.-based MLLMs										
LISA [14]	✓	✗	✓	✗	✓	✓	✓	✗	2	2
VISA [47]	✓	✓	✓	✗	✓	✓	✓	✓	2	2
GLaMM [35]	✓	✗	✓	✗	✓	✓	✗	✓	2	0
VideoGLaMM [31]	✗	✓	✓	✗	✓	✓	✗	✓	0	2
PSALM [52]	✓	✓	✓	✓	✓	✓	✓	✗	5	1
HyperSeg [41]	✓	✓	✓	✓	✓	✓	✓	✗	5	4
Sa2VA [50]	✓	✓	✓	✓	✓	✓	✓	✓	3	3
X-SAM [39]	✓	✓	✓	✓	✓	✓	✓	✗	**7**	0
X2SAM	✓	✓	✓	✓	✓	✓	✓	✓	**7**	**7**

fragmented. Existing video-centric MLLMs lack the unified architecture for both images and videos. Furthermore, standard frame-by-frame decoding approaches struggle to systematically store and track multi-modal guided features, failing to maintain robust mask consistency and temporal coherence across continuous video frames.

Analysis against SAM2 and X-SAM. X2SAM is related to SAM2 [36] and X-SAM [39], but targets a distinct setting. SAM2 enables promptable image and video segmentation with memory-based propagation, yet it mainly relies on low-level visual prompts and lacks language-driven reasoning or grounded conversation. X-SAM supports MLLM-based segmentation with textual and visual prompts, but is image-centric and does not model temporal object identity. X2SAM is not a simple X-SAM+SAM2 cascade. It unifies image and video segmentation in an instruction-following framework, where textual prompts, visual prompts, and generated `<SEG>` tokens are converted into mask-aware conditions. Its language-conditioned Mask Memory stores guided visual features from the MLLM-conditioned decoder, coupling semantic grounding with temporal propagation. Thus, unlike frame-wise X-SAM or cascaded propagation, X2SAM jointly optimizes grounding, decoding, and memory for temporally consistent instruction-based mask generation.

3 Method

To extend segmentation capabilities seamlessly from static images to dynamic video sequences, we propose a novel segmentation-oriented MLLM, termed X2SAM. We first present the formal problem formulation of X2SAM, encompassing the definition of inputs, outputs, and task formulations. Subsequently, we elaborate on the architectural framework of X2SAM, detailing the input processing pipeline, the encoders and LLM, the redesigned mask decoder, and the mask

memory module. Finally, we discuss the training methodology of X2SAM, highlighting our unified joint training strategy and the associated training objectives.

3.1 Formulation

Inputs. The inputs to X2SAM comprise a textual or visual prompt coupled with either a single image or a video sequence. The textual prompt constitutes a natural language instruction that delineates the target segmentation task, whereas the visual prompt represents an interactive visual cue (e.g., points or boxes) that designates the objects of interest. The image or video sequence serves as the primary visual input to be processed by the framework.
Outputs. The outputs of X2SAM comprise a contextual language response and a corresponding segmentation mask. The language response represents the natural language output generated by the LLM, while the segmentation mask provides a binary, pixel-level delineation of the target specified by the prompt.
Unified Formulation. To accommodate a comprehensive set of image and video segmentation tasks, we introduce a unified formulation for X2SAM. In this formulation, the objects of interest across all tasks are treated as conditional states, while the language instruction serves as the contextual input. Following X-SAM [39], we incorporate two special tokens, `<p>` and `</p>`, to demarcate the beginning and end of the object condition, respectively, along with a dedicated `<SEG>` token to indicate the corresponding segmentation mask. The LLM's output representation for the `<SEG>` token functions as a dedicated directive, guiding the mask decoder to segment the objects of interest. Furthermore, task-specific templates are devised to facilitate aligned language response generation by the LLM.

3.2 Framework

Overview. As illustrated in Fig. 3, X2SAM takes as input a language instruction $\mathrm{X_q}$ and a visual input $\mathrm{X_v} \in \mathbb{R}^{T\times H\times W\times C}$, where $T = 1$ for images and $T > 1$ for videos, and jointly outputs a language response $\mathrm{Y_q}$ and a segmentation mask $\mathrm{Y_m} \in \mathbb{R}^{T\times H\times W}$. The model adopts a dual-branch visual extraction architecture: a vision encoder f_v extracts global representations $\mathrm{Z_v}$, while a mask encoder g_m captures fine-grained features $\mathrm{Z_m}$ for dense prediction. The projected global features $\mathrm{H_v} = W_\mathrm{v}(\mathrm{Z_v})$, region features $\mathrm{H_r}$ from the region sampler g_r, and tokenized textual embeddings $\mathrm{H_q}$ are fed into the LLM f_ϕ. The LLM auto-regressively generates $\mathrm{Y_q}$ together with a dedicated `SEG` latent embedding, serving as a semantic bridge between language understanding and mask prediction. This embedding is transformed by the MLLM projector into the prompt token embedding $\mathrm{Z_p}$. Finally, the mask decoder g_ψ synthesizes $\mathrm{Y_m}$ by integrating $\mathrm{Z_p}$, learnable mask queries $\mathrm{Q_m}$, and temporally refined visual features $\mathrm{Z_w}$. These features are produced by the mask memory module g_ω, which maintains a first-in-first-out (FIFO) cache of guided visual features from preceding frames for temporally consistent segmentation.

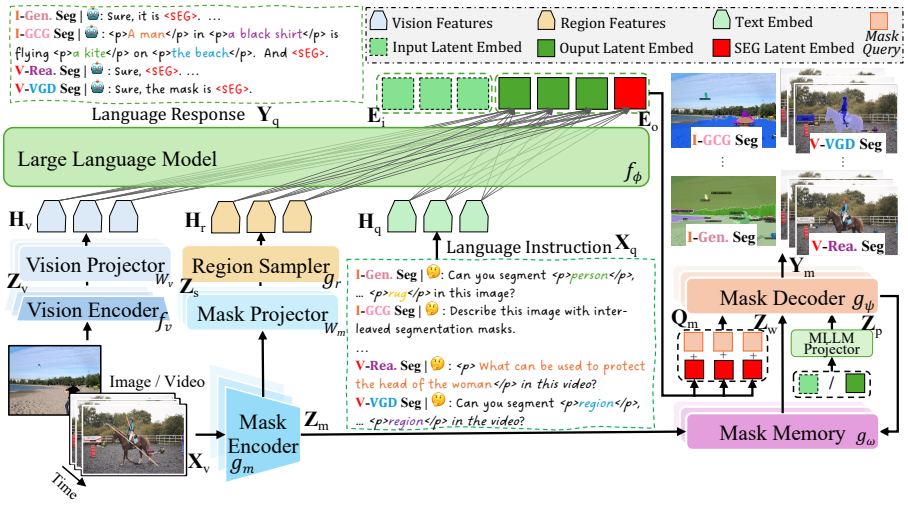

Fig. 3: Overview of X2SAM. The Vision Encoder extracts global visual representations, while the Mask Encoder captures fine-grained visual features. The Large Language Model generates the language response and produces the latent condition embedding, which guides the Mask Decoder in generating the segmentation mask. The Mask Memory module stores guided vision features for each video frame, and the Region Sampler extracts region-of-interest embeddings from both images and videos.

Input Processing. Given the visual input X_v and instruction X_q, X2SAM employs two complementary visual processing pipelines. For global understanding, we follow Qwen3-VL-4B [5], where visual inputs are augmented with timestamps, partitioned into spatial patches, and projected into latent embeddings Z_v. For high-resolution mask prediction, we adopt SAM2 [36], which processes videos frame-wise to extract fine-grained mask features Z_m. When region-specific information is required, the region sampler g_r extracts localized visual prompt embeddings from Z_m. In parallel, the textual instruction is formatted with task-specific templates, tokenized, and embedded into text latent representations H_q.

Region Sampler. We design a parameter-free region sampler to facilitate the injection of visual prompts into the LLM. Specifically, we conduct point-sampling [49] on regions of interest utilizing the mask encoder's high-resolution features Z_m. We then apply adaptive pooling to aggregate these point-sampled features into cohesive region-level representations H_r.

Vision Encoder and LLM. Large Vision-Language Models (LVLMs) inherently possess robust semantic understanding. We adopt the vision encoder, vision projector, and LLM backbone from Qwen3-VL [5], endowing X2SAM with state-of-the-art multimodal reasoning and broad visual comprehension capabilities.

Mask Encoder and Decoder. We utilize the robust and lightweight mask encoder from SAM2 [36]. However, to overcome limitations in parallel mask generation, we discard its original mask decoder and redesign a novel architecture inspired by X-SAM [39]. As illustrated in Fig. 4(b), we introduce structured

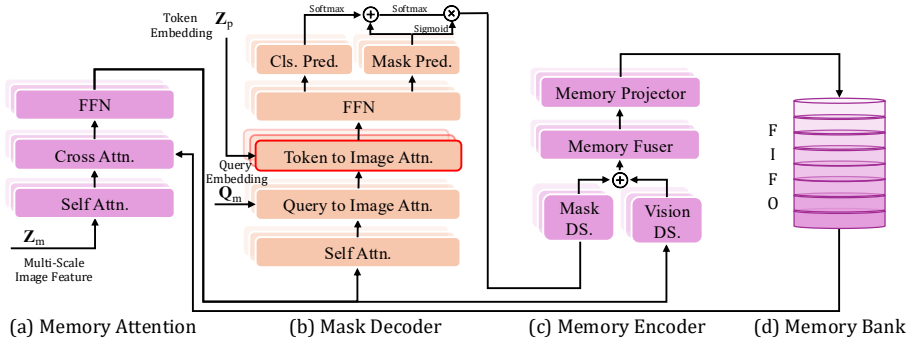

Fig. 4: Architecture of the mask memory and mask decoder. (a) *Memory Attention* attends the guided vision features of previous frames in the video. (b) *Mask Decoder* generates the segmentation mask of the current frame. (c) *Memory Encoder* encodes the downsampled vision features and mask logits of the current frame. (d) *Memory Bank* stores the guided vision features of each frame in the video, and updates the memory bank via FIFO (First In First Out) strategy.

attention modules, namely *Query-to-Image Attention* and *Token-to-Image Attention*, to inject token-level conditional information into the mask decoder. This allows the LLM's semantic token embedding Z_p to directly interact with spatial features. We employ zero-initialization for the Token-to-Image Attention parameters, ensuring smooth and stable integration of token-level conditional information during early training.

Mask Memory. To maintain temporal coherence across video frames, we propose a Mask Memory module (detailed in Fig. 4) that operates as a temporal cache. Its data flow follows the four parts in Fig. 4: 1) *Memory Attention* (Fig. 4a): attends to guided vision features from previous frames and produces temporally-refined vision features for the current frame. 2) *Mask Decoder* (Fig. 4b): generates the current-frame segmentation mask and mask logits from temporally-refined features and the LLM-derived segmentation token. 3) *Memory Encoder* (Fig. 4c): encodes downsampled vision features and current-frame mask logits into guided vision features. 4) *Memory Bank* (Fig. 4d): stores guided vision features of processed frames and updates the memory bank using a First-In-First-Out (FIFO) strategy.

3.3 Training

Agnostic Segmentor Training. We first perform category-agnostic segmentor training to provide the mask decoder with a stable initialization before multi-modal instruction tuning. Following X-SAM [39], the mask encoder is kept frozen and only the mask decoder is optimized with mask-level supervision. This stage encourages the decoder to learn class-independent shape and boundary priors from dense annotations, thereby reducing its dependence on semantic category labels during the subsequent joint training stage. Our mask loss $\mathcal{L}_{\text{mask}}$ combines

binary cross-entropy loss $\mathcal{L}_{\text{bce}}$ and dice loss $\mathcal{L}_{\text{dice}}$ [30]:

$$\mathcal{L}_{\text{mask}} = \lambda_{\text{bce}}\mathcal{L}_{\text{bce}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}}, \tag{1}$$

where $\lambda_{\text{bce}} = 5.0$ and $\lambda_{\text{dice}} = 5.0$ balance the relative weighting of each objective.

Unified Joint Training. We train X2SAM jointly on heterogeneous image and video datasets under a unified optimization framework. The main challenge is that image and video samples differ substantially in temporal length and memory footprint. To address this issue, we adopt a dimension-shifting pipeline together with modality-aware batching. Given a visual input tensor $\mathbf{X}_{\text{v}} \in \mathbb{R}^{B\times T\times H\times W\times 3}$, where $T = 1$ for images and $T > 1$ for videos, we first transpose it to $\mathbb{R}^{T\times B\times H\times W\times 3}$ and split it into T frame-level tensors of shape $\mathbb{R}^{B\times H\times W\times 3}$. Each frame is then processed by the mask encoder using the same image-level interface, while temporal dependencies are introduced through the mask memory module during sequential mask decoding. The predicted frame-level masks are finally concatenated along the temporal dimension to recover the sequence-level output $\mathbb{R}^{B\times T\times H\times W}$. To improve training efficiency under memory constraints, we further adapt the batch organization to the input modality. We set the base per-device batch size to $B = 1$ for video samples to avoid excessive memory consumption, while image-only batches are expanded with an image batch multiplier, yielding an effective image batch size on per device to better utilize GPU parallelism. We also use modality-specific gradient accumulation, updating image batches every step and accumulating video gradients over multi-steps to stabilize optimization under the same memory budget. In addition, a temporal-aware sampler groups video clips with the same temporal length into the same batch, reducing unnecessary padding and improving computational efficiency. Our joint training objective $\mathcal{L}_{\text{joint}}$ integrates the auto-regressive loss $\mathcal{L}_{\text{ar}}$ [34] for language generation, the mask loss $\mathcal{L}_{\text{mask}}$ for mask segmentation, and the focal loss $\mathcal{L}_{\text{cls}}$ [19] for mask classification:

$$\mathcal{L}_{\text{joint}} = \begin{cases} \mathcal{L}_{\text{ar}}, & \text{image \& video chat} \\ \mathcal{L}_{\text{ar}} + \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{cls}}, & \text{image \& video segmentation} \end{cases} \tag{2}$$

4 Experiments

4.1 Tasks, Datasets, and Metrics

Tasks. X2SAM is engineered to perform segmentation across both static images and video sequences, driven by textual or visual prompts. The framework spans a comprehensive suite of 14 segmentation tasks, stratified into image-based and video-based modalities. The seven image-based tasks include: generic segmentation (I-Gen.), open-vocabulary segmentation (I-OV), referring segmentation (I-Ref.), reasoning segmentation (I-Rea.), grounded conversation generation segmentation (I-GCG), image interactive segmentation (I-Int.), and visual grounded segmentation (I-VGD). Correspondingly, the seven video-based tasks comprise: generic segmentation (V-Gen.), open-vocabulary segmentation (V-OV), referring

segmentation (V-Ref.), reasoning segmentation (V-Rea.), grounded conversation generation segmentation (V-GCG), video object segmentation (V-Obj.), and visual grounded segmentation (V-VGD).

Datasets. Our training has two phases: class-agnostic segmentor training and unified joint training. For the agnostic segmentor phase, we use mask-only SA-1B [13] to train the mask decoder. The unified joint training phase integrates data from all 14 segmentation tasks, supplemented by image and video chat datasets. For image segmentation and chat tasks, we follow the mixed fine-tuning configuration of X-SAM [39]. The video segmentation corpus includes: VIPSeg [28], VSPW [29], and YT-VIS19 [48] for generic segmentation; YT-RefVOS [37] for referring segmentation; ReVOS [47] for reasoning segmentation; VideoGLaMM [31] for grounded conversation generation; and YT-VOS19 [46] and DAVIS17 [32] for video object segmentation. We introduce two datasets for video visual grounded segmentation, derived from YT-VIS19 and VIPSeg. For video chat training, we use VideoInstruct100K [27]. To assess generalization, we evaluate on standard validation or test splits and out-of-domain benchmarks: gRefCOCO [20] for image referring segmentation, ADE20K [53] for image open-vocabulary segmentation, and YT-VIS-21 [48] for video open-vocabulary segmentation.

Metrics. We evaluate X2SAM across image and video benchmarks, following established evaluation protocols. For panoptic, instance, and semantic segmentation tasks in both images and videos, we report (V)PQ, mAP, and mIoU, respectively. Image referring and reasoning segmentation are evaluated using cIoU and gIoU. Video-based referring, reasoning, and video object segmentation tasks are evaluated using $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$. For GCG segmentation, we report METEOR, CIDEr, AP50, and mIoU, and VGD segmentation performance is quantified using mAP and AP50. Image interactive segmentation utilizes mIoU and cIoU.

4.2 Implementation Details

Baseline Setup. We initialize the vision encoder, vision projector, and LLM with pre-trained weights from Qwen3-VL [5], while the mask encoder is initialized from SAM2 [36]. We employ LoRA [10] for LLM fine-tuning. The mask decoder is initialized using the pre-trained agnostic segmentor. Unless otherwise specified, the region sampler uses mask-encoder features with an adaptive pooling kernel size of 4. The ablation baseline omits the token-to-image attention layers in the mask decoder and excludes the mask memory module; the full X2SAM model adds both components. Both the mask and MLLM projectors are implemented as MLPs.

Training Setup. In the agnostic segmentor training phase, the mask encoder is frozen and the mask decoder is trained with an effective batch size of 128 and a learning rate of 1×10^{-4}. In the unified joint training phase, we optimize the projectors, LoRA parameters of the LLM, mask encoder, mask decoder, and mask memory. The learning rate is 1×10^{-5} for the mask encoder and 1×10^{-4} for the other trainable modules. The effective batch size is 32 for video data and 128 for image data. The memory capacity in the mask memory module is fixed

Table 2: Ablation on the mask decoder.

Method	I-Ref. RefCOCO/+/g cIoU	I-Rea. Val/Test	V-Ref. YT21 $\mathcal{J}\&\mathcal{F}$	V-Rea. Ref./Rea./All $\mathcal{J}\&\mathcal{F}$
Baseline	82.9/**78.0**/**79.5**	58.6/55.9	53.6	36.3/36.9/36.5
+ T2I(Rand)	82.5/77.2/79.4	62.4/56.0	58.2	44.9/45.7/44.9
+ T2I(Zero)	**83.3**/77.8/**79.5**	**63.0**/**56.8**	**60.8**	**47.6**/**47.7**/**47.3**

Table 3: Ablation on the joint training.

Method	Train Time Hours	I-Gen. Pan./Sem./Ins. PQ/mIoU/mAP	I-OV Pan./Sem./Ins. PQ/mIoU/mAP	V-Gen. Pan./Sem./Ins. VPQ/mIoU/mAP	V-OV Ins. mAP
Separate	/	52.6/64.0/43.8	25.3/32.5/19.5	42.9/61.1/66.3	57.1
Simple	~5.2K	**54.4**/65.0/**45.0**	29.3/36.3/18.9	46.8/64.6/68.1	58.7
Unified	**3.3K**	54.1/**65.3**/44.8	**29.4**/**37.5**/**19.3**	**47.1**/**64.7**/**68.3**	**59.1**

Table 4: Ablation on the mask memory.

Method	V-Gen. Pan./Sem./Ins. VPQ/mIoU/mAP	V-Ref. YT21/DV17 $\mathcal{J}\&\mathcal{F}$	V-Rea. Ref./Rea./All $\mathcal{J}\&\mathcal{F}$
Baseline	42.9/61.1/66.3	53.6/41.1	36.3/36.9/36.5
+Single Scale	42.7/60.8/66.1	52.5/41.9	38.0/37.5/37.6
+Mask Guide	44.5/62.3/66.7	63.3/49.4	51.4/51.4/51.1
+Class Guide	44.8/61.9/**66.8**	64.6/48.3	52.0/51.9/51.6
+Multi Scale	**45.0**/**62.5**/66.6	**65.0**/**49.6**	**53.7**/**53.9**/**53.5**

Table 5: Ablation on the memory size.

#Mem. Size	V-Gen. Pan./Sem./Ins. VPQ1/mIoU/mAP	V-OV Ins. mAP	V-Ref. YT21/DV17 $\mathcal{J}\&\mathcal{F}$	V-Rea. Ref./Rea./All $\mathcal{J}\&\mathcal{F}$
1	43.9/62.0/68.0	58.5	**66.8**/50.7	56.9/56.9/56.5
2	43.9/62.0/67.9	58.2	66.5/**50.8**	57.6/57.7/57.2
4	43.8/**63.2**/**68.4**	58.9	66.7/50.3	57.7/**58.0**/57.4
6	43.9/62.6/**68.4**	**60.2**	66.5/50.5	**58.0**/57.9/**57.5**
8	**45.0**/62.5/66.6	58.3	65.0/49.6	53.7/53.9/53.5

at 8 frames unless the memory-size ablation states otherwise. Optimization is performed using AdamW [26] with a weight decay of 0.05. To ensure balanced representation across heterogeneous sources, we apply a dataset-balanced resampling strategy [39] with a temperature parameter $t = 0.1$. All main experiments are trained for one epoch on 32 NVIDIA H800 GPUs, while ablation studies are conducted on 16 GPUs due to limited computational resources.

4.3 Ablation Studies

Mask Decoder. We ablate the impact of the Token-to-Image (T2I) attention module and its initialization strategies in Tab. 2. Compared to the baseline, employing a random initialization (+ T2I(Rand)) disrupts early training, occasionally yielding sub-optimal results and falling behind the baseline in static image tasks (e.g., I-Ref RefCOCO/+/g decreases from 82.9/78.0/79.5 to 82.5/77.2/79.4). Conversely, our proposed zero-initialization strategy (+ T2I(Zero)) facilitates a stable injection of the LLM's semantic token embeddings into the spatial features. This approach consistently outperforms the baseline, achieving peak performance across image metrics (e.g., 83.3/77.8/79.5 in I-Ref) and substantial gains in video tasks (e.g., increasing V-Ref YT21 from 53.6 to 60.8 $\mathcal{J}\&\mathcal{F}$).

Joint Training. Tab. 3 evaluates the efficacy of our unified joint training strategy against separate and simple joint training paradigms. The simple joint training method achieves strong performance (e.g., 54.4 I-Gen. PQ and 64.6 V-Gen. mIoU) but requires approximately 5.2K GPU hours. Our unified joint training strategy reduces the training cost to 3.3K GPU hours, a 36.5% reduction, while maintaining comparable image performance and improving video-level metrics such as V-Gen. mIoU (64.7) and V-OV Ins. mAP (59.1).

Mask Memory. As shown in Tab. 4, directly adopting single-scale memory features provides only marginal improvements on V-Rea. and slightly degrades V-Gen. and V-Ref., indicating that naive temporal memory is insufficient. In

Table 6: Comparison of state-of-the-art segmentation methods across image and video segmentation benchmarks, ranging from non-MLLM-based to MLLM-based, and from specialists to generalists. "✗" denotes unsupported. "–" indicates unreported. Best results are in **bold**, second-best are underlined.

Method	Textual Prompt					Visual Prompt	
	Image Segmentation						
	I-Gen. *Pan./Sem./Ins.* PQ/mIoU/mAP	***I-OV*** *Pan./Sem./Ins.* PQ/mIoU/mAP	***I-Ref.*** *RefCOCO/+/g* cIoU	***I-Rea.*** *Val/Test* gIoU	***I-GCG*** *Val/Test* mIoU	***I-Int.*** *Point/Box* mIoU	***I-VGD*** *Point/Box* mAP
Non-MLLM-based Image Specialists							
SAM-L [13]	✗	✗	✗	✗	✗	51.8/76.6	12.8/31.7
Mask2Former-L [9]	57.8/67.4/48.6	✗	✗	✗	✗	✗	✗
ODISE [44]	55.4/65.2/46.0	22.6/29.9/14.4	✗	✗	✗	✗	✗
MLLM-based Image Generalists							
LISA-7B [14]	✗	✗	74.9/65.1/67.9	52.9/47.3	✗	✗	✗
GLaMM [35]	✗	✗	79.5/72.6/74.2	✗	65.8/64.6	✗	✗
OMG-LLaVA [51]	53.8/–/–	✗	78.0/69.1/72.9	✗	65.5/64.7	✗	✗
Sa2VA-8B [50]	✗	✗	81.6/76.2/78.7	–	–	✗	✗
X-SAM [39]	**54.7/66.5/47.0**	20.9/28.8/16.2	**85.1/78.0/83.8**	56.6/57.8	**69.4/69.0**	65.4/69.6	**47.9/49.5**
X2SAM	54.1/64.8/45.8	**31.2/38.2/20.2**	84.0/**78.4**/81.9	**71.1/68.5**	67.1/65.2	**67.7/70.3**	45.9/48.5

Method	*Video Segmentation*						
	V-Gen. *Pan./Sem./Ins.* PQ/mIoU/mAP	***V-OV*** *YT21-Ins.* mAP	***V-Ref.*** *YT21/DV17* $\mathcal{J}\&\mathcal{F}$	***V-Rea.*** *Ref./Rea./All* $\mathcal{J}\&\mathcal{F}$	***V-GCG*** *V-GLaMM* mIoU	***V-Obj.*** *YT19-All* $\mathcal{J}\&\mathcal{F}$	***V-VGD*** *YT19/VIPSeg* mAP
Non-MLLM-based Video Specialists							
SAM2-H [36]	✗	✗	✗	✗	✗	88.8	39.2/25.6
OMG-Seg [17]	49.8/–/56.4	50.5	✗	✗	✗	✗	✗
UniRef++-L [43]	✗	✗	66.9/67.2	✗	✗	85.9	✗
MLLM-based Video Generalists							
VISA-7B [47]	✗	✗	61.5/69.4	50.9/43.0/46.9	✗	✗	✗
VideoLISA [6]	✗	✗	61.7/67.7	–	✗	✗	✗
UniPixel-7B [25]	✗	✗	71.0/76.4	65.8/61.5/63.7	✗	✗	✗
HyperSeg [41]	–	–	68.5/71.2	58.5/53.0/55.7	✗	✗	✗
VideoGLaMM [31]	✗	✗	66.8/–	–	54.3	✗	✗
X2SAM	**47.3/65.1/69.9**	**60.3**	**78.5/79.0**	**69.3/70.7/69.9**	**75.8**	**74.0**	**73.8/55.4**

contrast, introducing mask guidance brings consistent gains across tasks, especially improving V-Ref. YT21 from 53.6 to 63.3 $\mathcal{J}\&\mathcal{F}$, which demonstrates the importance of mask-level cues for temporal alignment. Class guidance further improves semantic discrimination, leading to stronger results on V-Rea. Finally, the full multi-scale design achieves the best overall performance, obtaining 45.0 VPQ, 62.5 mIoU, 65.0 $\mathcal{J}\&\mathcal{F}$ on V-Ref. *YT21*, and 53.5 $\mathcal{J}\&\mathcal{F}$ on V-Rea. *All*.

Memory Size. As reported in Tab. 5, increasing the memory size from 1 to 6 generally benefits tasks requiring long-term temporal context, particularly open-vocabulary and reasoning-oriented video understanding. The best V-OV mAP of 60.2 and V-Rea. All score of 57.5 $\mathcal{J}\&\mathcal{F}$ are achieved with memory size 6, suggesting that richer historical information helps resolve temporal ambiguity. However, the gains are not monotonic across all benchmarks. Some V-Gen. and V-Ref. metrics saturate earlier, and increasing the memory size to 8 leads to lower V-Ref. and V-Rea. performance despite improving V-Gen. VPQ. This indicates that excessive memory may introduce redundant or noisy temporal cues, while a moderate memory size provides a better trade-off, we adopt 6 frames as the final memory size.

Table 7: Comparison across image and video reasoning segmentation benchmarks.

Method	I-Rea. Seg.				V-Rea. Seg.		
	$\textbf{\textit{Val}}_{Overall}$ cIoU/gIoU	$\textbf{\textit{Test}}_{Short}$ cIoU/gIoU	$\textbf{\textit{Test}}_{Long}$ cIoU/gIoU	$\textbf{\textit{Test}}_{Overall}$ cIoU/gIoU	$\textbf{\textit{ReVOS}}_{Ref.}$ $\mathcal{J}/\mathcal{F}/\mathcal{J}\&\mathcal{F}$	$\textbf{\textit{ReVOS}}_{Rea.}$ $\mathcal{J}/\mathcal{F}/\mathcal{J}\&\mathcal{F}$	$\textbf{\textit{ReVOS}}_{Overall}$ $\mathcal{J}/\mathcal{F}/\mathcal{J}\&\mathcal{F}$
Non-MLLM-Based Specialists							
SEEM-L [54]	21.2/25.5	11.5/20.1	20.8/25.6	18.7/24.4	–	–	–
ReferFormer-B [42]	–	–	–	–	31.2/34.3/32.7	21.3/25.6/23.4	26.2/29.9/28.1
MLLM-Based Generalists							
LISA-7B [14]	54.0/52.9	40.6/40.6	51.0/49.4	48.4/47.3	44.3/47.1/45.7	33.8/38.4/36.1	39.1/42.7/40.9
VISA-7B [47]	57.8/52.7	–	–	–	49.2/52.6/50.9	40.6/45.4/43.0	44.9/49.0/46.9
HyperSeg [41]	56.7/59.2	–	–	–	56.0/60.9/58.5	50.2/55.8/53.0	53.1/58.4/55.7
X2SAM	**64.5/71.1**	**53.5/60.0**	**66.7/68.9**	**65.6/68.5**	**66.2/72.4/69.3**	**67.5/74.0/70.3**	**66.7/73.0/69.9**

Table 8: Comparison on out-of-domain tasks, including image generalized referring segmentation, image and video open-vocabulary segmentation benchmarks.

Method	I-Ref. Seg.			I-OV Seg.			V-OV Seg.
	$\textbf{\textit{gRefCOCO}}_{Val}$ cIoU/gIoU	$\textbf{\textit{gRefCOCO}}_{TestA}$ cIoU/gIoU	$\textbf{\textit{gRefCOCO}}_{TestB}$ cIoU/gIoU	$\textbf{\textit{A150}}_{Pan.}$ PQ	$\textbf{\textit{A150}}_{Sem.}$ mIoU	$\textbf{\textit{A150}}_{Ins.}$ mAP	$\textbf{\textit{YT-VIS-21}}_{Ins.}$ AP/AP50
Non-MLLM-Based Specialists							
ReLA [20]	62.4/63.6	69.3/70.0	59.9/61.0	✗	✗	✗	✗
ODISE [44]	✗	✗	✗	22.6	29.9	14.4	✗
OMG-Seg [17]	✗	✗	✗	27.9	–	–	50.5/–
MLLM-Based Generalists							
LISA-7B [14]	38.7/–	52.6/–	44.8/–	✗	✗	✗	✗
PSALM [52]	42.0/43.3	52.4/54.5	50.6/52.5	13.7	18.2	9.0	✗
HyperSeg [41]	47.5/–	57.3/–	52.5/–	16.1	22.3	–	53.8/–
X-SAM [39]	59.9/65.6	63.0/67.2	62.0/65.2	20.9	28.8	16.2	✗
X2SAM	**63.1/68.1**	**67.3/71.2**	**63.4/66.7**	**31.2**	**38.2**	**20.2**	**60.3/78.0**

4.4 Benchmark Results

Overall Performance. Tab. 6 presents a comparison of X2SAM against specialist models and MLLM-based generalists across image and video segmentation benchmarks. On image tasks, X2SAM remains competitive with the image-centric generalist X-SAM [39] while improving image open-vocabulary segmentation (I-OV) from 20.9 to 31.2 PQ. It also maintains strong performance on generic (I-Gen.) and referring (I-Ref.) segmentation, and obtains 67.1/65.2 mIoU on I-GCG validation/test splits. These results suggest that extending the architecture to video does not collapse its static image segmentation ability. On video tasks, X2SAM outperforms existing MLLM-based video generalists on most reported video segmentation benchmarks. Compared with UniPixel-7B [25], X2SAM improves V-Ref. on both Ref-YT21 and Ref-DV17, and achieves strong V-Gen. performance with 65.1 mIoU. On video grounded conversation generation (V-GCG), X2SAM improves over VideoGLaMM [31] by +21.5 mIoU (75.8 vs. 54.3). These gains support the role of the unified formulation and mask memory in video-level segmentation, while the remaining gap to task-specific video object segmentation (VOS) specialists highlights the cost of maintaining a general interface.

Reasoning Segmentation. Tab. 7 reports X2SAM's performance on reasoning segmentation, where targets require implicit knowledge and logical deduc-

Table 9: Comparison across image and video visual grounded segmentation benchmarks.

Method	I-VGD Seg.		V-VGD Seg.			
	$COCO_{Point}$	$COCO_{Box}$	$YT\text{-}VIS19_{Point}$	$YT\text{-}VIS19_{Box}$	$VIPSeg_{Point}$	$VIPSeg_{Box}$
	AP/AP50	AP/AP50	AP/AP50	AP/AP50	AP/AP50	AP/AP50
Non-MLLM-Based Specialists						
SAM-L [13]	12.8/22.8	31.7/50.1	–	–	–	–
SAM2-H [36]	–	–	39.2/53.5	54.0/73.3	25.6/36.3	40.4/54.7
MLLM-Based Specialists						
PSALM [52]	2.0/3.3	3.7/5.8	✗	✗	✗	✗
X-SAM [39]	**47.9/72.5**	**49.5/74.7**	✗	✗	✗	✗
X2SAM	45.9/68.2	48.5/71.6	**73.8/93.5**	**74.4/93.9**	**55.5/75.4**	**57.8/78.3**

tion. On I-Rea. Seg., X2SAM achieves 64.5 cIoU and 71.1 gIoU on validation, outperforming HyperSeg [41] by +7.8 cIoU and +11.9 gIoU. It also sets new state-of-the-art test results for both short and long queries, showing robust instruction understanding. On V-Rea. Seg., X2SAM reaches 69.9 $\mathcal{J}\&\mathcal{F}$ on ReVOS, improving over HyperSeg [41] by +14.2 points and surpassing the video-specialist ReferFormer-B [42]. Its consistent performance on both Ref. and Rea. subsets suggests effective integration of LLM-based target reasoning and spatio-temporal mask tracking.

Out-of-Domain Segmentation. As shown in Tab. 8, we evaluate X2SAM on out-of-domain tasks covering unseen datasets and novel categories. On gRefCOCO [20], which includes multi-target and no-target expressions, X2SAM improves over both the non-MLLM specialist ReLA [20] and recent MLLM-based generalists such as PSALM [52] and X-SAM [39]. On ADE20K (A150) [53], X2SAM achieves 31.2 PQ, 38.2 mIoU, and 20.2 mAP, outperforming compared open-vocabulary segmentation methods including ODISE [44] and X-SAM. In the video domain, X2SAM obtains 60.3 AP on YT-VIS-21, exceeding OMG-Seg [17] and HyperSeg [41]. These results indicate that the unified training setup transfers beyond the primary training distribution, although the evaluation remains bounded by the available OOD benchmarks.

Visual Grounded Segmentation. Tab. 9 presents a comparative analysis of visual grounded segmentation performance utilizing point and bounding box prompts. Within the image domain (I-VGD Seg.), X2SAM demonstrates highly competitive efficacy on the COCO benchmark, achieving an AP of 45.9 and 48.5 for point and box prompts, respectively, thereby performing comparably to the image-specialized X-SAM model. Crucially, the temporal modeling advantages of our approach become distinctly evident in the video domain (V-VGD Seg.), where X2SAM substantially improves over SAM2-H under the evaluated V-VGD protocol. Specifically, under box prompt evaluation, X2SAM attains an impressive 74.4 AP on YT-VIS19 and 57.8 AP on the more complex VIPSeg dataset, exhibiting significant improvements over SAM2-H (54.0 AP and 40.4 AP). These substantial performance gains validate the effectiveness of our region sampler and

mask memory modules in robustly propagating temporal visual prompts across highly dynamic video frames.

5 Discussion

Conclusion. We presented X2SAM, a unified segmentation-oriented MLLM for pixel-level understanding across images and videos. X2SAM casts diverse segmentation tasks into a shared instruction-following formulation, supporting both textual instructions and visual prompts within the same framework. By integrating a Mask Memory module, it extends image-centric any-segmentation capabilities to video sequences with improved temporal consistency. We also introduced the Video Visual Grounded (V-VGD) benchmark to evaluate video object grounding from interactive visual prompts. In addition, our adaptive joint training strategy enables efficient co-training over heterogeneous image and video datasets, reducing the training cost while maintaining balanced performance across modalities. Extensive experiments show that X2SAM achieves a strong balance between task coverage and accuracy: it remains competitive on image segmentation benchmarks, improves many video segmentation tasks, and preserves general image and video understanding abilities. More broadly, X2SAM demonstrates that a single instruction-following interface can unify text- and vision-prompted segmentation across both images and videos without task-specific designs.

Limitations and Outlook. X2SAM still has several limitations. First, unified training over heterogeneous image and video datasets remains computationally expensive, especially for video samples with high memory cost. Second, the fixed-size FIFO memory may be insufficient for long videos with prolonged occlusions, large appearance changes, or sparse target reappearance. Third, as a unified generalist model, X2SAM may still lag behind specialized models on narrowly focused tasks such as optimized video object segmentation or image-only segmentation. Future work will explore more efficient training, lightweight backbones, and adaptive long-range memory to improve scalability and robustness. In particular, an adaptive memory that allocates capacity on demand could better cope with long, occlusion-heavy videos than a fixed-size FIFO cache, while the unified instruction interface extends naturally to streaming and interactive segmentation.

Acknowledgements

This work is supported by National Key Research and Development Program of China (2024YFE0203100), Scientific Research Innovation Capability Support Project for Young Faculty (No. ZYGXQNJSKYCXNLZCXM-I28), Shenzhen Science and Technology Program (Grant No. QNXMA20250701093544048) and General Embodied AI Center of Sun Yat-sen University, National Natural Science Foundation of China (62536003) and (62402252), and Guangdong High-Level Talent Programme (2024TQ08X283).

References

1. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: ICCV. pp. 2425–2433 (2015)
2. Athar, A., Hermans, A., Luiten, J., Ramanan, D., Leibe, B.: Tarvis: A unified approach for target-based video segmentation. In: CVPR. pp. 18738–18748 (2023)
3. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al.: Qwen technical report. arXiv preprint arXiv:2309.16609 (2023)
4. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
5. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
6. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Liu, L., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. NeurIPS (2024)
7. Chen, J., Shen, Y., Gao, J., Liu, J., Liu, X.: Language-based image editing with recurrent attentive models. In: CVPR. pp. 8721–8729 (2018)
8. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. Science China Information Sciences **67**(12), 220101 (2024)
9. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR. pp. 1290–1299 (2022)
10. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. Iclr **1**(2), 3 (2022)
11. Jain, J., Li, J., Chiu, M.T., Hassani, A., Orlov, N., Shi, H.: Oneformer: One transformer to rule universal image segmentation. In: CVPR. pp. 2989–2998 (2023)
12. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916. PMLR (2021)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
14. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. In: CVPR. pp. 9579–9589 (2024)
15. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. arXiv preprint arXiv:2408.03326 (2024)
16. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML. pp. 12888–12900. PMLR (2022)

17. Li, X., Yuan, H., Li, W., Ding, H., Wu, S., Zhang, W., Li, Y., Chen, K., Loy, C.C.: Omg-seg: Is one model good enough for all segmentation? In: CVPR. pp. 27948–27959 (2024)
18. Li, X., Zhang, W., Pang, J., Chen, K., Cheng, G., Tong, Y., Loy, C.C.: Video k-net: A simple, strong, and unified baseline for video segmentation. In: CVPR (2022)
19. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: ICCV. pp. 2980–2988 (2017)
20. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: CVPR. pp. 23592–23601 (2023)
21. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR. pp. 26296–26306 (2024)
22. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llavanext: Improved reasoning, ocr, and world knowledge (2024)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. NeurIPS **36**, 34892–34916 (2023)
25. Liu, Y., Ma, Z., Pu, J., Qi, Z., Wu, Y., Ying, S., Chen, C.W.: Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. In: NeurIPS (2025)
26. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
27. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
28. Miao, J., Wang, X., Wu, Y., Li, W., Zhang, X., Wei, Y., Yang, Y.: Large-scale video panoptic segmentation in the wild: A benchmark. In: CVPR. pp. 21033–21043 (2022)
29. Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., Yang, Y.: Vspw: A large-scale dataset for video scene parsing in the wild. In: CVPR. pp. 4133–4143 (2021)
30. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 3DV. pp. 565–571. Ieee (2016)
31. Munasinghe, S., Gani, H., Zhu, W., Cao, J., Xing, E., Khan, F., Khan, S.: Videoglamm: A large multimodal model for pixel-level visual grounding in videos. CVPR (2024)
32. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR. pp. 724–732 (2016)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PmLR (2021)
34. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training. Tech. rep., OpenAI, San Francisco, CA, USA (2018)
35. Rasheed, H., Maaz, M., Shaji, S., Shaker, A., Khan, S., Cholakkal, H., Anwer, R.M., Xing, E., Yang, M.H., Khan, F.S.: Glamm: Pixel grounding large multimodal model. In: CVPR. pp. 13009–13018 (2024)
36. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)

37. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV. pp. 208–223. Springer (2020)
38. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
39. Wang, H., Qiao, L., Jie, Z., Huang, Z., Feng, C., Zheng, Q., Ma, L., Lan, X., Liang, X.: X-sam: From segment anything to any segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 26187–26196 (2026)
40. Wang, H., Wang, W., Liu, J.: Temporal memory attention for video semantic segmentation. In: ICIP. pp. 2254–2258. IEEE (2021)
41. Wei, C., Zhong, Y., Tan, H., Liu, Y., Zhao, Z., Hu, J., Yang, Y.: Hyperseg: Towards universal visual segmentation with large language model. arXiv preprint arXiv:2411.17606 (2024)
42. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR. pp. 4974–4984 (2022)
43. Wu, J., Jiang, Y., Yan, B., Lu, H., Yuan, Z., Luo, P.: Uniref++: Segment every reference object in spatial and temporal spaces. ICCV (2023)
44. Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., De Mello, S.: Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: CVPR. pp. 2955–2966 (2023)
45. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML. pp. 2048–2057. PMLR (2015)
46. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-vos: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018)
47. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Visa: Reasoning video object segmentation via large language models. ECCV (2024)
48. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV. pp. 5188–5197 (2019)
49. You, H., Zhang, H., Gan, Z., Du, X., Zhang, B., Wang, Z., Cao, L., Chang, S.F., Yang, Y.: Ferret: Refer and ground anything anywhere at any granularity. arXiv preprint arXiv:2310.07704 (2023)
50. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025)
51. Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., Loy, C.C., Yan, S.: Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. NeurIPS **37**, 71737–71767 (2024)
52. Zhang, Z., Ma, Y., Zhang, E., Bai, X.: Psalm: Pixelwise segmentation with large multi-modal model. In: ECCV. pp. 74–91. Springer (2024)
53. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. IJCV **127**(3), 302–321 (2019)
54. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. NeurIPS **36**, 19769–19782 (2023)