UniPixel: Unified Object Referring and Segmentation for Pixel-Level Visual Reasoning

Ye Liu[1,2], **Zongyang Ma**[2,3], **Junfu Pu**[2], **Zhongang Qi**[4], **Yang Wu**[5], **Ying Shan**[2], **Chang Wen Chen**[1]*

[1] The Hong Kong Polytechnic University [2] ARC Lab, Tencent PCG
[3] Chinese Academy of Sciences [4] vivo Mobile Communication Co. [5] Tencent AI Lab
coco.ye.liu@connect.polyu.hk
https://polyu-chenlab.github.io/unipixel/

Figure 1: **UniPixel** flexibly supports a large variety of fine-grained image and video understanding tasks, including referring/reasoning/interactive segmentation, motion-grounded video reasoning, and referred video description & question answering. It can also handle a novel **PixelQA** task that jointly requires object-centric referring, segmentation, and question answering in videos.

Abstract

Recent advances in Large Multi-modal Models (LMMs) have demonstrated their remarkable success as general-purpose multi-modal assistants, with particular focuses on holistic image- and video-language understanding. Conversely, less attention has been given to scaling fine-grained pixel-level understanding capabilities, where the models are expected to realize pixel-level alignment between visual

* *Corresponding author.*

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

signals and language semantics. Some previous studies have applied LMMs to related tasks such as region-level captioning and referring expression segmentation. However, these models are limited to performing either referring or segmentation tasks independently and fail to integrate these fine-grained perception capabilities into visual reasoning. To bridge this gap, we propose **UniPixel**, a large multi-modal model capable of flexibly comprehending visual prompt inputs and generating mask-grounded responses. Our model distinguishes itself by seamlessly integrating pixel-level perception with general visual understanding capabilities. Specifically, UniPixel processes visual prompts and generates relevant masks on demand, and performs subsequent reasoning conditioning on these intermediate pointers during inference, thereby enabling fine-grained **pixel-level reasoning**. The effectiveness of our approach has been verified on 10 benchmarks across a diverse set of tasks, including pixel-level referring/segmentation and object-centric understanding in images/videos. A novel **PixelQA** task that jointly requires referring, segmentation, and question answering is also designed to verify the flexibility of our method.

1 Introduction

Large Multi-modal Models (LMMs) have been the de facto standard for developing general-purpose assistants. By effectively aligning multi-modalities with language, their significance has been demonstrated across various applications, including multi-modal analysis [60, 20, 1, 44, 49], autonomous driving (AD) [17, 81, 103, 12], and Embodied AI [106, 23, 31, 96].

In the field of visual-language understanding, efforts have been dedicated to developing *holistic understanding models*, where simple projection layers between visual encoders and LLMs are utilized to bridge vision and language modalities. Supported by large-scale alignment pre-training and visual instruction tuning, such a straightforward paradigm achieves strong performance in holistic understanding tasks such as captioning [41, 7, 109] and general question answering [37, 25, 55, 50]. However, these models exhibit two fundamental limitations in fine-grained scenarios. **First**, their interactions with users are limited to text format, lacking support for more intuitive forms of communication such as drawing points/boxes as references or grounding model responses with key regions represented by masks. **Second**, the internal reasoning process of these models predominantly operates at a coarse level, directly perceiving the entire content rather than reasoning over specific objects/regions, making them hard to understand fine-grained details. Some previous studies have explored the application of LMMs to related tasks such as region-level captioning [13, 99, 100], referring expression segmentation [30, 56, 42, 22, 73, 64], and reasoning segmentation [33, 28, 93, 5, 107]. Nevertheless, their models are limited to performing either referring or segmentation tasks independently via rigidly defined input/output templates (*e.g.*, "It's `<SEG>`." in LISA [33]), lacking the flexibility to comprehend user-referred concepts and generate mask-grounded responses simultaneously. More importantly, these methods cannot integrate such fine-grained perception capabilities with their original multi-modal reasoning abilities, resulting in degraded performance on general visual understanding benchmarks [97, 90, 29].

In this work, we seek to bridge this gap by introducing **UniPixel**, a large multi-modal model that can flexibly comprehend visual prompt inputs (*i.e.*, points, boxes, and masks) and generate mask-grounded responses. Our model significantly differentiates itself from existing ones by unifying the internal representations of referred and segmented objects via a novel **object memory bank**, which is a hashmap storing the spatial-temporal information of object-of-interests. During inference, UniPixel initializes the object memory bank and updates it on demand by adding object-centric information according to the context. The model responses are then generated conditioning on the fine-grained object memory. Benefits from such unification, UniPixel is able to perform not only basic referring/segmentation tasks, but also flexible **pixel-level reasoning** tasks that require simultaneous visual prompt comprehension and mask prediction. As illustrated in Fig. 1 (the last row), given a video[2], a question, and optionally a visual prompt (*e.g.*, a point specified by a click on an object in any frame), UniPixel can (1) infer the mask for the referred object in the corresponding frame, (2) propagate it to all video frames containing the same instance, (3) extract the mask-grounded object features, and finally (4) answer the question conditioning on both the video-level and object-centric

[2] *Images are treated as single-frame videos, thus we do not explicitly differentiate them in this work.*

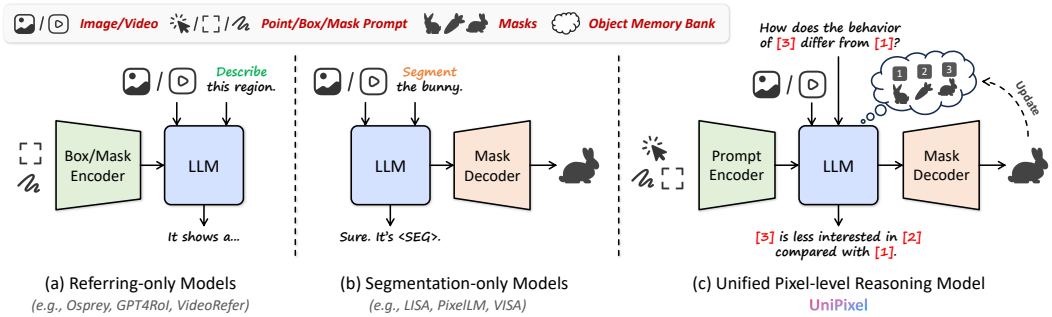

Figure 2: **Schematic comparison between UniPixel and its counterparts.** To the best of our knowledge, UniPixel is the first unified method supporting simultaneous object referring and segmentation.

information. All these operations are seamlessly conducted *within a single model*, eliminating the need for external frame samplers [93], mask generators [99, 100], or object trackers [5].

We evaluate the effectiveness of UniPixel from two aspects, *i.e.*, basic referring/segmentation capabilities and flexible pixel-level reasoning capabilities. For the first aspect, we conduct extensive experiments on 10 public benchmarks across 9 image/video referring/segmentation tasks. Our method achieves state-of-the-art performance in diverse scenarios. Notably, on the challenging video reasoning segmentation and referred video QA tasks, our 3B model obtains **62.1** $\mathcal{J}\&\mathcal{F}$ on ReVOS [93] and **72.8%** Acc on VideoRefer-Bench$^{\text{Q}}$ [100], surpassing strong counterparts with 7B $\sim$ 13B parameters. Further ablation studies also demonstrate the mutual reinforcement effect of referring and segmentation. For the second aspect, we introduce a novel **PixelQA** task that jointly requires object-centric referring, segmentation, and QA in videos, which cannot be handled by existing methods. UniPixel establishes a strong baseline for this novel setting. Our contributions are summarized below:

1. We propose **UniPixel**, a unified large multi-modal model that supports flexible object referring and segmentation in images and videos, via a novel **object memory bank** design.
2. Our model achieves state-of-the-art performance on 10 public benchmarks across 9 referring/segmentation tasks, verifying the **mutual reinforcement effect** of such unification.
3. We also introduce a novel **PixelQA** task that jointly requires object-centric referring, segmentation, and QA in videos, where UniPixel establishes a strong baseline for this setting.

2 Related Work

Large Multi-modal Models The remarkable success of large multi-modal models (LMMs) has shifted the paradigm of visual-language understanding from close-ended experts to open-ended task solvers. Early attempts [44, 43, 18, 110] involve an MLP projector or Q-Former [35] to align visual encoders to LLMs, enabling open-ended tasks such as visual question answering. With advanced designs such as dynamic resolution and data augmentation, open-source models, *e.g.*, Qwen-VL [3, 79, 4] and InternVL [15, 76, 14] series, have narrowed the gap with advanced proprietary models like the GPT [59, 60] and Gemini families [69, 19]. Recent studies [61, 26, 52, 38, 49] also explore test-time scaling on visual-language understanding. However, these methods are spatially coarse-grained. UniPixel can also be regarded as an object-centric test-time scaling approach, where key objects are first segmented then encoded to facilitate the subsequent reasoning process.

Visual Referring and Segmentation To meet the growing demand for fine-grained visual understanding [51, 47, 48, 46, 86], recent efforts have focused on enhancing LMMs with object referring and segmentation capabilities, as compared in Fig. 2. LISA [33] is a representative model that enables LMM-based segmentation by integrating SAM [32] as its decoder. They also introduced a novel reasoning segmentation task, requiring models to perform segmentation based on implicit queries. Other works in this direction [101, 71, 63, 105, 82, 67, 28] have explored more advanced mask decoders, more flexible tasks, and larger-scale datasets. Recent studies have also extended these capabilities to videos [5, 93, 98]. Additionally, some research has examined regional understanding through boxes [13] and masks [99, 100]. While recent approaches attempt to unify these two capabilities, they either support only images [67] or rely on sub-optimal, tool-based pipelines [24]. To the best of knowledge, UniPixel is the first end-to-end method unifying object referring and mask prediction.

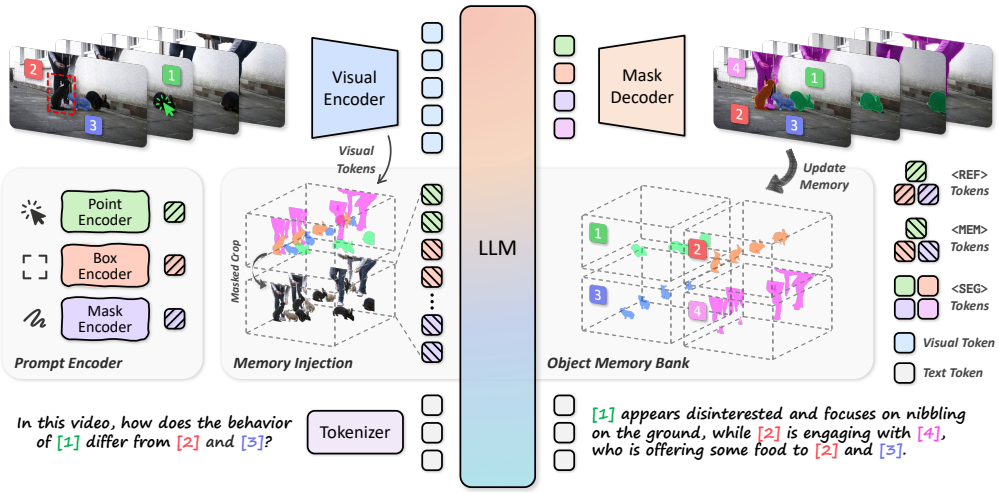

Figure 3: **The architecture of UniPixel.** Given a video, a question, and visual prompts, the model encodes them into tokens via the visual encoder, prompt encoder, and tokenizer, respectively, then predicts a spatial-temporal mask for each visual prompt via the mask decoder. The masks are updated into the object memory bank, and subsequently injected into the prompt for pixel-level reasoning.

3 Method

Problem Formulation We provide a unified definition for pixel-level reasoning tasks. Formally, the inputs are an image or a video $\mathcal{X}$, a text prompt $\mathcal{T}$, and N optional visual prompts $\{\mathcal{P}_i\}_{i=1}^{N}$ where each $\mathcal{P}_i$ could be a point, box, or mask on a specific frame. The outputs are textual responses to the prompt with K grounded spatial-temporal masks $\{\mathcal{M}_i\}_{i=1}^{K}$. Here, both N and K could be zero (degenerating to a normal visual understanding task) and K is not necessarily equal to N, as the model may segment extra objects/regions that are not specified by the visual prompts.

Overview Fig. 3 presents an overview of UniPixel. It is built upon the Qwen2.5-VL [4] framework, consisting of an LLM backbone and a ViT-based visual encoder that supports dynamic resolution inputs. Given a video and a text prompt, the model first tokenizes them via the visual encoder and text tokenizer, then sends them into the LLM for response generation. To boost this framework from holistic-level to pixel-level, we introduce (1) a *prompt encoder* (Sec.3.1) supporting three types of visual prompts, (2) an *object memory bank* (Sec.3.2) for storing object information and injecting it into the response generation process, and (3) a *mask decoder* (Sec.3.3) for generating spatial-temporal masks. We also extend the LLM's vocabulary by adding `<REF>`, `<MEM>`, and `<SEG>` tokens. The former two serve as placeholders in the input prompt that would be replaced by visual prompt and memory tokens, respectively, while the `<SEG>` token is utilized to trigger and guide the mask decoding process. Detailed designs and interactions among these components are illustrated as follows.

3.1 Prompt Encoder

This module aims to effectively encode each visual prompt into a single token that can be processed by the LLM. We denote a point prompt as a tuple (x, y, t) containing its spatial coordinates (x, y) and the corresponding frame index t. For box prompts, it is extended to (x_1, y_1, x_2, y_2, t) containing the positions of top-left and bottom-right corners. A mask prompt is densely represented by a 2D binary mask $\mathbf{m}_{ij} \in \{0, 1\}$ with the same shape as the encoded target frame.

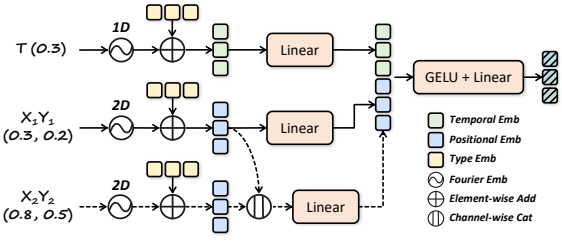

Figure 4: **Joint positional & temporal encoding** for point (X_1Y_1T) and box ($X_1Y_1X_2Y_2T$) prompts.

For sparse prompts (points and boxes), as shown in Fig. 4, we encode each position (x_i, y_i) as the sum of a 2D Fourier embedding [75] and a learnable type embedding (indicating whether it is a single point, top-left corner, or bottom-right corner). For box prompts, we merge the two positional embeddings

by concatenating them along the channel dimension and linearly projecting them back to the original size. Frame indices are also encoded similarly with 1D Fourier embeddings. The resulting positional and temporal embeddings are concatenated again, and then projected to the LLM's embedding space via a `GELU` → `Linear` block, such that the sparse coordinates in a point/box are encoded into a compact high-dimensional token. This design is inspired by [32, 68] with two key differences: (1) the spatial-only embeddings are extended to include temporal information, and (2) the negative points are discarded. For dense prompts (masks), we directly resize the binary masks and apply masked pooling on the outputs of the visual encoder. An M→L projector (`Linear` → `GELU` → `Linear`) is leveraged to project the pooled visual features to the LLM's embedding space.

3.2 Object Memory Bank

Although sparse prompts contain rich positional and temporal information indicating the objects that users are referring to, it is still hard for the model to focus on these important regions. Previous studies [13, 99, 100] also confirm that direct region cropping can generally provide better object awareness compared to positional pointers. To seamlessly integrate such a mechanism while preserving the flexibility of visual prompts (*e.g.*, allow pointing on a single frame instead of drawing complete masks on all frames), we propose an object memory bank to bridge sparse visual prompts and dense object masks. This is a hashmap where the keys are object IDs and the values are the corresponding spatial-temporal masks. It is initialized as an empty storage for every chat session, and is dynamically updated on demand. We define two operations for the object memory bank, namely *memory pre-filling* and *memory injection*. Below is an example of memory-enhanced multi-round conversation.

Prompt 1: How does the behavior of [1] <REF> differ from [2] <REF> and [3] <REF>?

<REF> detected, enhancing the prompt with object memory.

Memory Pre-filling Response:
The relevant regions for this question are [1] <SEG> [2] <SEG> [3] <SEG> [4] <SEG>. ← *4 objects saved into the memory*

Memory Injected Prompt:
Here is a video with 4 frames denoted as <1> to <4>. The highlighted regions are as follows:
[1]: <1> <MEM> <2> <MEM> <3> <MEM> ← *This object cannot be seen in the last frame*
[2]: <1> <MEM> <2> <MEM> <3> <MEM> <4> <MEM>
[3]: <1> <MEM> <2> <MEM> <3> <MEM> <4> <MEM>
[4]: <1> <MEM> <2> <MEM> <3> <MEM> <4> <MEM>
How does the behavior of [1] differ from [2] and [3]?

Response 1: [1] appears disinterested and focuses on nibbling on the ground, while [2] is engaging with [4], who is offering some food to [2] and [3].
Prompt 2: What food is [4] offering? ← *Users can directly refer to objects in the memory*
Response 2: [4] is offering carrots.

Memory Pre-filling This operation is triggered upon the detection of <REF> tokens in the input prompt, aiming to thoroughly analyze the referred objects and predict their corresponding masks. In this stage, the model responds with object IDs and <SEG> tokens for the relevant objects according to the context, and predicts their spatial-temporal masks accordingly. These object-mask pairs are then saved into the object memory bank.

Memory Injection We inject the features of the saved objects into the prompt to enhance object-awareness. Similar to the mask prompt encoder described in Sec. 3.1, each frame-level object mask is downsampled to match the resolution of visual tokens. We then apply masked pooling to aggregate object-centric features. Each frame-level mask is condensed into a single feature token, projected through the mask projector, and subsequently utilized to replace the corresponding <MEM> token in the memory-injected prompt. Through this *pre-filling and injection* mechanism, object-centric information is effectively integrated into the model inference process.

Why using object memory bank? An alternative is directly appending a <SEG> token to each <REF> token, followed by masked pooled features obtained during inference. However, we do not adopt this approach for two reasons: (1) During mask prediction, the <SEG> tokens, due to the unidirectional nature of causal self-attention, are unable to aggregate the full context of the prompt, thereby compromising the quality of predicted masks. (2) By utilizing the object memory bank, we can effectively decouple regional understanding and mask prediction, allowing each to benefit from referring and segmentation data during training, thus enhancing both capabilities.

Table 1: Comparison with state-of-the-art methods on ReVOS [93] `val` split. The best and second-best results are marked **bold** and underlined, respectively.

Method	Size	Referring			Reasoning			Overall			$\mathcal{R}$
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	
Non-LLM-based Specialists											
MTTR [6]	–	29.8	30.2	30.0	20.4	21.5	21.0	25.1	25.9	25.5	5.6
LMPM [22]	–	29.0	39.1	34.1	13.3	24.3	18.8	21.2	31.7	26.4	3.2
ReferFormer [87]	–	31.2	34.3	32.7	21.3	25.6	23.4	26.2	29.9	28.1	8.8
LLM-based Generalists											
LISA [33]	13B	45.2	47.9	46.6	34.3	39.1	36.7	39.8	43.5	41.6	8.6
TrackGPT [74]	13B	48.3	50.6	49.5	38.1	42.9	40.5	43.2	46.8	45.0	12.8
VISA [93]	13B	55.6	59.1	57.4	42.0	46.7	44.3	48.8	52.9	50.9	14.5
HyperSeg [83]	3B	56.0	60.9	58.5	50.2	55.8	53.0	53.1	58.4	55.7	–
InstructSeg [84]	3B	54.8	59.2	57.0	49.2	54.7	51.9	52.0	56.9	54.5	–
GLUS [40]	7B	56.0	60.7	58.3	48.8	53.9	51.4	52.4	57.3	54.9	17.9
ViLLa [107]	6B	–	–	–	–	–	–	54.9	59.1	57.0	–
Sa2VA [98]	4B	–	–	–	–	–	–	–	–	53.2	–
UniPixel (Ours)	3B	62.3	66.7	64.5	57.1	62.1	59.6	59.7	64.4	62.1	19.0
UniPixel (Ours)	7B	**64.2**	**68.5**	**66.4**	**59.6**	**63.9**	**61.8**	**61.9**	**66.1**	**64.0**	**19.1**

3.3 Mask Decoder

We adopt SAM 2.1 [68] as the mask decoder to disentangle the discrete language modeling and continuous mask prediction capabilities. For each `<SEG>` token, we extract its last-layer hidden states, downsample them via an L→M projector (architecturally identical to the M→L projector), and reshape them into two tokens. Using two tokens ensures better preservation of object information when downsampling from high- to low-dimensional channel space. These tokens prompt the mask decoder to predict the mask on the first frame, which is then propagated to the other frames.

3.4 Model Training

The training loss for UniPixel is a linear combination of language modeling loss and mask decoding losses [68], including a focal loss and dice loss for mask prediction, a mean-absolute-error (MAE) loss for IoU prediction, and a cross-entropy loss for objectness prediction. The loss weights are set to 1, 100, 5, 5, and 5, respectively. We train the model through a three-stage progressive alignment recipe. The datasets are listed in Tab. 12. In the first stage, we pre-train the sparse prompt encoder using 851K regional captioning data. Then, we align the LLM and mask decoder by training the L→M projector on 87K referring segmentation data. In the last stage, we further unfreeze the M→L projector and mask decoder, and apply LoRA [27] on the visual encoder and LLM. The model is jointly trained on a large-scale corpus with around 1M samples for diverse tasks.

4 Experiments

We evaluate the effectiveness of UniPixel by conducting extensive experiments across a diverse set of benchmarks. Specifically, we study the following research questions.

- **Q1.** Whether UniPixel is flexible and effective on basic image/video referring and segmentation tasks compared to the corresponding representative methods?
- **Q2.** How does it perform on the more challenging PixelQA task, which requires joint referring, segmentation, and question answering in videos?
- **Q3.** What effects does each architectural design contribute? More importantly, does the unified modeling of referring and segmentation lead to a mutual reinforcement effect?

Detailed information about the benchmarks, evaluation metrics, implementation details, and more experimental results can be found in the appendix.

4.1 Q1: Comparison with State-of-the-Arts on Referring and Segmentation Tasks

Reasoning Video Object Segmentation We begin with the most challenging ReVOS [93] dataset, which requires models to predict masks based on implicit text queries demanding complex reasoning abilities based on world knowledge. The results are shown in Tab. 1. Our 3B variant outperforms all

Table 2: Comparison with state-of-the-art methods on referring video object segmentation (RVOS) and motion-grounded video reasoning datasets, including MeViS [22] (`val`), Ref-YouTube-VOS [73] (`val`), Ref-DAVIS17 [64] (`val`), and GroundMoRe [21] (`test`). The best and second-best results are marked **bold** and underlined, respectively.

Method	Size	MeViS			Ref-YouTube-VOS			Ref-DAVIS17			GroundMoRe		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Non-LLM-based Specialists													
ReferFormer [87]	–	29.8	32.2	31.0	61.3	64.6	62.9	58.1	64.1	61.1	11.2	14.3	12.7
LMPM [22]	–	34.2	40.2	37.2	–	–	–	–	–	–	12.7	14.0	13.3
OnlineRefer [85]	–	–	–	–	61.6	65.5	63.5	61.6	67.7	64.8	–	–	–
LLM-based Generalists													
PixelLM [71]	7B	36.3	41.1	38.7	54.3	55.7	55.0	63.4	70.0	66.7	9.9	10.0	10.0
LISA [33]	13B	35.8	40.0	37.9	54.0	54.8	54.4	63.2	68.8	66.0	6.3	6.7	6.5
VISA [93]	13B	41.8	47.1	44.5	61.4	64.7	63.0	67.0	73.8	70.4	5.3	4.7	5.9
VideoLISA [5]	3.8B	41.3	47.6	44.4	61.7	65.7	63.7	64.9	72.7	68.8	–	–	–
VideoGLaMM [57]	3.8B	42.1	48.2	45.2	65.4	68.2	66.8	73.3	65.6	69.5	–	–	–
ViLLa [107]	6B	46.5	52.3	49.4	64.6	70.4	67.5	70.6	78.0	74.3	–	–	–
GLUS [40]	7B	48.5	54.2	51.3	65.5	69.0	67.3	–	–	–	–	–	–
Sa2VA [98]	4B	–	–	46.2	–	–	70.0	–	–	73.8	–	–	–
MoRA [21]	7B	–	–	–	–	–	–	–	–	–	27.4	26.9	27.2
UniPixel (Ours)	3B	50.4	55.7	53.1	68.6	72.3	70.5	70.7	77.8	74.2	36.0	38.7	37.4
UniPixel (Ours)	7B	**52.3**	**57.1**	**54.7**	**70.2**	**74.1**	**72.1**	**71.4**	**80.0**	**75.7**	**46.2**	**49.0**	**47.6**

Table 3: Comparison with state-of-the-art methods on image referring expression segmentation (RES) and reasoning segmentation datasets, including RefCOCO/+/g [30, 56] and ReasonSeg [33] (`val`). The best and second-best results are marked **bold** and underlined, respectively.

Method	Size	RefCOCO			RefCOCO+			RefCOCOg		ReasonSeg	
		val	testA	testB	val	testA	testB	val(U)	test(U)	gIoU	cIoU
Non-LLM-based Specialists											
ReLA [42]	–	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	–	–
X-Decoder [111]	–	–	–	–	–	–	–	64.6	–	22.6	17.9
SEEM [112]	–	–	–	–	–	–	–	65.7	–	25.5	21.2
LLM-based Image Generalists											
NExT-Chat [101]	7B	74.7	78.9	69.5	65.1	71.9	56.7	67.0	67.0	–	–
PixelLM [71]	7B	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	–	–
LISA [33]	7B	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	61.3	62.9
Groundhog [105]	7B	78.5	79.9	75.7	70.5	75.0	64.9	74.1	74.6	56.2	–
LaSagnA [82]	7B	76.8	78.7	73.8	66.4	70.6	60.1	70.6	71.9	48.8	47.2
M^2SA [28]	13B	74.6	77.6	71.0	64.0	68.1	57.6	69.0	69.3	–	–
LLM-based Video Generalists											
VideoLISA [5]	3.8B	73.8	76.6	68.8	63.4	68.8	56.2	68.3	68.8	61.4	**67.1**
VISA [93]	7B	72.4	75.5	68.1	59.8	64.8	53.1	65.5	66.4	52.7	57.8
Vitron [24]	7B	75.5	79.5	72.2	66.7	72.5	58.0	67.9	68.9	–	–
Sa2VA [98]	4B	78.9	–	–	71.7	–	–	74.1	–	–	–
UniPixel (Ours)	3B	80.5	82.6	76.9	74.3	78.9	68.4	76.3	77.0	64.0	56.2
UniPixel (Ours)	7B	**82.5**	**83.8**	**79.8**	**76.5**	**81.0**	**70.9**	**77.5**	**78.4**	**65.3**	58.0

existing methods with larger LLMs (including Sa2VA-4B [98] also with SAM 2 decoder), achieving 62.1 overall $\mathcal{J}\&\mathcal{F}$. The 7B model further boosts the performance to 64.0 $\mathcal{J}\&\mathcal{F}$ – an improvement of 12% over the previous state-of-the-art – demonstrating that UniPixel can effectively understand implicit queries based on its world knowledge, and accurately generate masks as responses.

Referring Video Object Segmentation The performance comparisons on MeViS [22], Ref-YouTube-VOS [73], and Ref-DAVIS17 [64] datasets are presented in Tab. 2. UniPixel consistently achieves the best performance among its counterparts. Its advantage is particularly evident on the more challenging MeViS dataset, where our 3B model outperforms GLUS-7B [40] by around 3.5%, as well as the similarly sized VideoGLaMM-3.8B [57] by 17%. More experimental results on MeViS [22] `val`$^{\text{u}}$ set and Ref-SAV [98] `val` set are provided in Tab. 4 and Tab. 5, respectively. Ref-SAV features long referring descriptions, large object motion, large camera motion, and heavy occlusion compared with existing datasets. Given these complex descriptions and video content, our method consistently performs better than counterparts, including those fine-tuned on the target dataset.

Motion-Grounded Video Reasoning We also evaluate our method on GroundMoRe [21] dataset (results shown in Tab. 2), which highlights visual answer generation that requires joint spatial and

Table 4: Experimental results on MeViS [22] `val`$^{\text{u}}$ set. Post means applying post optimization.

Method	Size	FT	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
LMPM [22]	–	✗	36.5	43.9	40.2
LISA [33]	7B	✗	39.9	46.5	43.2
LISA [33] + XMem [16]	7B	✗	41.9	49.3	45.6
VideoLISA [5]	7B	✗	48.4	54.9	51.7
VideoLISA [5] + Post	7B	✗	50.9	58.1	54.5
Sa2VA [98]	4B	✗	–	–	52.1
Sa2VA [98]	8B	✗	–	–	57.0
UniPixel (Ours)	3B	✗	56.1	**63.2**	59.7
UniPixel (Ours)	7B	✗	**56.9**	62.9	**59.9**

Table 5: Comparison on Ref-SAV [98] `val` set. FT means fine-tuning after pre-/co-training.

Method	Size	FT	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
UniRef++ [88]	–	✗	11.6	9.5	10.5
UNINEXT [92]	–	✗	8.8	6.4	7.6
LMPM [22]	–	✗	12.2	9.8	10.3
VISA [93]	7B	✗	13.2	11.3	11.8
Sa2VA [98]	8B	✗	39.6	43.0	41.3
UniRef++ [88]	–	✓	15.8	13.4	14.6
Sa2VA [98]	8B	✓	48.3	51.7	50.0
UniPixel (Ours)	3B	✗	66.9	67.6	67.2
UniPixel (Ours)	7B	✗	**72.0**	**73.6**	**72.8**

Table 6: Fine-tuned performance on referring expression segmentation (RES) datasets, including RefCOCO/+/g [30, 56]. The best and second-best results are marked **bold** and underlined, respectively.

Method	Size	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val(U)	test(U)
LISA [33]	7B	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6
GSVA [89]	7B	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3
OMG-LLaVA [104]	7B	78.0	80.3	74.1	69.1	73.1	63.0	72.9	72.9
GLaMM [67]	7B	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9
Sa2VA [98]	4B	80.4	–	–	74.3	–	–	75.7	–
UniPixel (Ours)	3B	81.9	83.5	78.6	75.3	80.3	70.6	77.2	78.5
UniPixel (Ours)	7B	**83.1**	**85.0**	**80.5**	**77.4**	**81.8**	**71.9**	**78.1**	**79.5**

Table 7: Experimental results on referring expression comprehension (REC) datasets, including RefCOCO/+/g [30, 56]. The best and second-best results are marked **bold** and underlined, respectively.

Method	Size	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val(U)	test(U)
OFA [80]	–	80.0	83.7	76.4	68.3	76.0	61.8	67.6	67.6
Shikra [10]	7B	87.0	90.6	80.2	81.6	87.4	72.1	82.3	82.2
MiniGPT-v2 [9]	7B	88.7	91.6	85.3	79.9	85.1	74.4	84.4	84.6
Vitron [24]	7B	90.9	93.2	89.3	83.7	89.1	76.9	86.4	87.0
UniPixel (Ours)	3B	91.8	93.8	87.5	86.3	90.8	80.3	88.0	88.2
UniPixel (Ours)	7B	**93.5**	**94.7**	**90.1**	**88.5**	**92.8**	**82.9**	**89.4**	**89.7**

temporal grounding. Note that we mainly compare the results with MoRA [21], which is fine-tuned on GroundMoRe while other methods are evaluated under the zero-shot setting. Benefit from the strong pixel-level reasoning capability, UniPixel significantly performs better than the baseline.

Referring Expression Segmentation and Reasoning Segmentation Tab. 3 compares the image segmentation capabilities using explicit and implicit queries. We evaluate our co-trained model on RefCOCO/+/g [30, 56] and ReasonSeg [33]. While state-of-the-art performance has been achieved on RES datasets, we observe that the reasoning segmentation data (239 samples) can be easily overwhelmed by the other samples during training due to its limited size. Tab. 6 presents the RES performance after fine-tuning. We follow the common practice that jointly fine-tunes the model on RefCOCO/+/g datasets [30, 56], and then evaluate on them separately. These results demonstrate the generalizability of UniPixel when facing both explicit and implicit queries.

Referring Expression Comprehension Our method also supports referring expression comprehension by inferring the bounding boxes from predicted masks. Its performance (accuracy with IoU $\geqslant 0.5$) is compared with representative methods in Tab. 7. Benefiting from the high-quality mask prediction, UniPixel can also achieve very competitive performance on this simpler task.

Referred Video Description and Question Answering We study UniPixel's regional understanding capabilities on VideoRefer-Bench [100], which contains two subsets for description and question answering tasks. The comparisons are in Tab. 8 and Tab. 9. BQ, SQ, RQ, CQ, and FP denote basic questions, sequential questions, relational questions, reasoning questions, and future predictions, respectively. Both tasks leverage mask prompts as inputs, where single-frame and multi-frame modes denote applying the masks only on a specific frame and on all frames, respectively. UniPixel can

Table 8: Comparison with state-of-the-art methods on VideoRefer-Bench^{D} [100]. The best and second-best results are marked **bold** and underlined, respectively.

Method	Size	Single-Frame					Multi-Frame				
		SC	AD	TD	HD	Avg.	SC	AD	TD	HD	Avg.
General LMMs											
LLaVA-OV [34]	7B	2.62	1.58	2.19	2.07	2.12	3.09	1.94	2.50	2.41	2.48
Qwen2-VL [79]	7B	2.97	2.24	2.03	2.31	2.39	3.30	2.54	2.22	2.12	2.55
InternVL2 [76]	26B	3.55	2.99	2.57	2.25	2.84	4.08	**3.35**	3.08	2.28	3.20
GPT-4o-mini [60]	–	3.56	2.85	2.87	2.38	2.92	3.89	3.18	2.62	2.50	3.05
GPT-4o [60]	–	3.34	2.96	3.01	2.50	2.95	4.15	3.31	3.11	2.43	3.25
Image Referring LMMs											
Ferret [95]	7B	3.08	2.01	1.54	2.14	2.19	3.20	2.38	1.97	1.38	2.23
Osprey [99]	7B	3.19	2.16	1.54	2.45	2.34	3.30	2.66	2.10	1.58	2.41
Video Referring LMMs											
Elysium [77]	7B	2.35	0.30	0.02	3.59	1.57	–	–	–	–	–
Artemis [65]	7B	–	–	–	–	–	3.42	1.34	1.39	2.90	2.26
VideoRefer [100]	7B	4.41	3.27	3.03	2.97	3.42	4.44	3.27	3.10	3.04	3.46
UniPixel (Ours)	3B	4.04	3.15	**3.10**	**3.37**	3.42	4.08	3.13	**3.13**	**3.42**	3.44
UniPixel (Ours)	7B	**4.45**	**3.32**	3.05	3.04	**3.47**	**4.48**	3.34	3.03	3.07	**3.48**

Table 9: Comparison with state-of-the-art methods on VideoRefer-Bench^{Q} [100] (*mask prompts*). MF denotes multi-frame mode. Full question types are in Sec. 4.1.

Method	Size	MF	BQ	SQ	RQ	CQ	FP	Avg.
General LMMs								
LLaVA-OV [34]	7B	✗	58.7	62.9	64.7	87.4	76.3	67.4
Qwen2-VL [79]	7B	✗	62.0	69.6	54.9	87.3	74.6	66.0
InternVL2 [76]	26B	✗	58.5	63.5	53.4	88.0	78.9	65.0
GPT-4o-mini [60]	–	✗	57.6	67.1	56.5	85.9	75.4	65.8
GPT-4o [60]	–	✗	62.3	**74.5**	**66.0**	88.0	73.7	71.3
Image Referring LMMs								
Ferret [95]	7B	✗	35.2	44.7	41.9	70.4	74.6	48.8
Osprey [99]	7B	✗	45.9	47.1	30.0	48.6	23.7	39.9
Video Referring LMMs								
VideoRefer [100]	7B	✗	**75.4**	68.6	59.3	89.4	78.1	71.9
UniPixel (Ours)	3B	✗	73.6	70.3	60.7	88.8	78.0	72.2
UniPixel (Ours)	7B	✗	71.7	73.2	64.6	**90.1**	**79.6**	**73.8**
VideoRefer [100]	7B	✓	–	70.6	60.5	–	–	72.1
UniPixel (Ours)	3B	✓	75.3	70.7	62.3	87.4	77.2	72.8
UniPixel (Ours)	7B	✓	**79.5**	**74.7**	**64.4**	**90.8**	**81.5**	**76.3**

Table 10: Evaluation results on our newly introduced PixelQA task. All the visual prompts are applied in a single frame. See Sec. 4.2 for detailed settings.

Method	Size	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Acc
Point Prompts					
InternVL2 [76]	26B	–	–	–	60.8
Qwen2-VL [79]	72B	–	–	–	69.3
UniPixel (Ours)	3B	**57.3**	64.4	**60.9**	71.1
UniPixel (Ours)	7B	56.9	**64.5**	60.7	**71.5**
Box Prompts					
InternVL2 [76]	26B	–	–	–	61.3
Qwen2-VL [79]	72B	–	–	–	69.0
UniPixel (Ours)	3B	57.8	64.7	61.3	70.3
UniPixel (Ours)	7B	**58.1**	**64.9**	**61.5**	**70.5**
Mixed (50% Points + 50% Boxes)					
InternVL2 [76]	26B	–	–	–	60.9
Qwen2-VL [79]	72B	–	–	–	69.1
UniPixel (Ours)	3B	57.2	64.1	60.6	70.8
UniPixel (Ours)	7B	**57.5**	**64.7**	**61.1**	**71.0**

effectively comprehend both types of prompts, and accurately respond with object-centric descriptions or answers, surpassing strong models including GPT-4o [60] and VideoRefer [100].

4.2 Q2: Pixel-Level Video Question Answering (PixelQA)

We design the new PixelQA task based on VideoRefer-Bench^{Q} [100], where the original mask prompts are replaced with more challenging point or box prompts. Given these ambiguous visual cues, models are expected to correctly identify the target object according to the question and the visual prompt, then respond with **both the textual answer and the corresponding object masks**. We report the mask prediction $\mathcal{J}\&\mathcal{F}$ and MCQ accuracy in Tab. 10. Note that none of the existing methods supports this scenario. Thus, we apply set-of-mark prompts [94] directly on video frames, and evaluate the QA accuracies of two strong LMMs [79, 76] as our baselines. Aside from point- or box-only prompts, we also explore a more flexible setting that randomly chooses different prompts for different objects. The results verify that our ***memory pre-filling & injection*** paradigm effectively enhances the model's reasoning capabilities. Visualizations of this task are shown in Fig. 5.

4.3 Q3: Key Ablation Studies

Effect of Task Unification We study the effect of task unification in Tab. 11 (a). Unifying referring and segmentation capabilities into a single model and training them jointly leads to better results

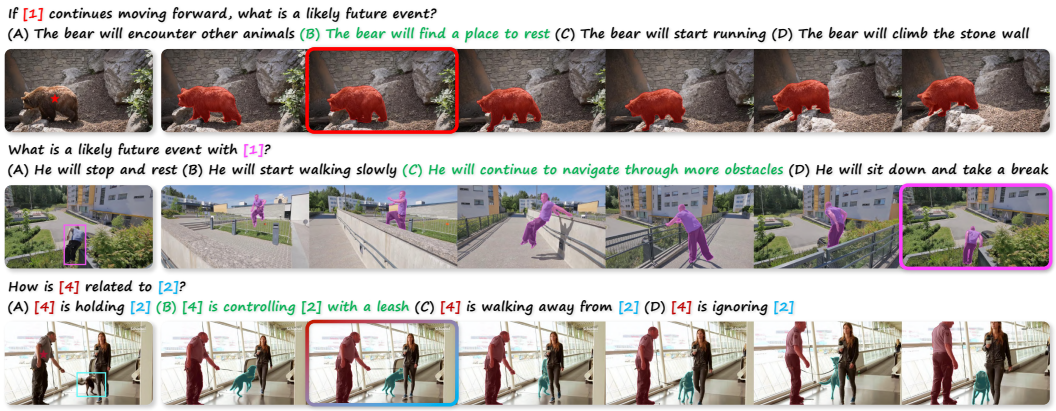

Figure 5: **Visualization of the outputs from UniPixel on PixelQA task**. Star marks and boxes refer to point and box prompts, respectively. The boxed frames denote where the visual prompts are applied. Given different types of visual prompts on a single frame, our method can flexibly infer the relevant object, track it across the entire video, and involve its features in reasoning.

Table 11: Key ablation studies with UniPixel-3B on PixelQA (*mixed*). See Sec. 4.3 for explanations.

(a) Task Unification

Refer	Segment	Memory	$\mathcal{J}\&\mathcal{F}$	Acc
✓			–	64.6
	✓		47.5	–
✓	✓		48.2	67.4
✓	✓	✓	**49.0**	**68.5**

(b) Object Memory Bank

Referring Method	$\mathcal{J}\&\mathcal{F}$	Acc
① <REF>	46.8	64.5
② <REF><SEG>	47.8	64.9
③ <REF><SEG> + Pooling	47.5	66.3
④ **Object Memory Bank**	**49.0**	**68.5**

(c) Prompt Encoder & Mask Decoder

Encoder	Decoder	$\mathcal{J}\&\mathcal{F}$	Acc
w/o Time	–	44.3	63.7
w/ Time	–	**49.0**	**68.5**
–	Independent	46.1	66.2
–	Propagation	**49.0**	**68.5**

on both tasks (first three rows), demonstrating the **mutual reinforcement effect** of such unification. Incorporating memory pre-filling as an auxiliary task (last row) brings extra improvements.

Effect of Object Memory Bank Tab. 11 (b) verifies the effectiveness of object memory bank. ① means using a single token for each referred object. ② means adding an extra segmentation token to segment it as an auxiliary task. ③ further appends masked-pooled visual tokens after it. The results show that (1) both adding auxiliary segmentation task and masked-pooled features help regional understanding, and (2) decoupling them via object memory bank can further boost the performance.

Design Space of Prompt Encoder & Mask Decoder We compare different prompt encoder and mask decoder designs in Tab. 11 (c). The performance significantly drops when the temporal encoding in the prompt encoder is removed (first two rows). For the mask decoder (last two rows), we explore an alternative strategy that treats video frames independently (as batched images), which could largely accelerate inference but lead to sub-optimal accuracies. We hypothesize that this is because the LLM-generated <SEG> token cannot well-capture the object information in all frames, thus disentangling the segmentation and tracking capabilities to an external module is reasonable.

5 Conclusion

In this work, we proposed **UniPixel**, a large multi-modal model that supports flexible pixel-level visual reasoning. It unifies the internal representations of referred and segmented objects through a novel **object memory bank**. We observe that by such unification, the performance of object referring and segmentation can be jointly enhanced. Extensive experiments on diverse pixel-level understanding tasks, including the **PixelQA** task, demonstrate the significance of the proposed method. We hope this work inspires future advancements in pixel-level visual understanding.

Acknowledgements

This study was supported by The Hong Kong RGC Grant (15229423) and a financial support from ARC Lab, Tencent PCG (ZGG9). We also acknowledge The University Research Facility in Big Data Analytics (UBDA) at The Hong Kong Polytechnic University for providing computing resources that have contributed to the research results reported within this paper.

Appendix

In this appendix, we provide more details about the training data, model implementation, and experimental settings to complement the main paper. Additional analysis, ablation studies, visualizations, and discussions are also incorporated. Below is the table of contents.

A Model

A.1 Implementation Details

We instantiate our base models with 3B and 7B versions of Qwen2.5-VL [4]. Both variants employ pre-trained SAM 2.1 [68] with Hiera Base+ [72] backbone as the mask decoder. The M→L projector is initialized with the weights from the V→L projector of Qwen2.5-VL. The hidden size inside the prompt encoder is 256. To reduce GPU memory and accelerate training, we randomly sample 8 frames per video, with each frame resized to $316^2 \sim 448^2$ pixels ($128 \sim 256$ tokens per frame). The frame sampling strategies follow the specifications of each benchmark during inference. The mask decoder has a fixed resolution of 768×768. For each segmentation sample, up to 5 objects are randomly selected to compute the mask prediction losses. During training, LoRA adapters [27] with `rank=128` and `alpha=256` are applied to all `QKVO` layers in the visual encoder and LLM. The input sequences are restricted to 4K tokens. We train the model with 8 RTX A6000 Ada (48G) GPUs, with a global batch size of 256 for stages 1 and 2, and 32 for stage 3. In the first two stages, the learning rates are set to 1e-3. In the last stage, it is set to 5e-6 for the mask decoder and 2e-5 for all the other parameters, respectively. A linear warmup in the first 3% steps followed by cosine decay is adopted in all stages. The configurations of datasets are introduced in the following section.

A.2 Training Recipe

The detailed distribution of training datasets for UniPixel is shown in Tab. 12. Within the three-stage training recipe, we first pre-train the sparse prompt encoder using short caption samples from Inst-IT [62] and VideoRefer [100]. For each sample, we randomly select a point inside the ground truth mask (50%) or generate an augmented box from it (50%). This stage aims to enable the model with simple visual prompt comprehension and regional captioning capabilities on images and videos. In the second stage, we align the LLM and mask decoder using referring object segmentation datasets [30, 56, 73]. We use short caption/query samples for the first two stages to focus on alignment rather than knowledge learning. For the last stage, we collect a large-scale, high-quality corpus called `UniPixel-SFT-1M`[3] to jointly train the model on diverse pixel-level tasks. The original annotations have been rewritten using task-specific templates to incorporate instructions. All the repurposed datasets and pre-processing pipelines will be publicly available to facilitate future research.

[3] `https://huggingface.co/datasets/PolyU-ChenLab/UniPixel-SFT-1M`

Table 12: The distribution of training datasets for UniPixel. We use different background colors to denote object referring, object segmentation, regional understanding, memory pre-filling, and general video understanding data, respectively.

Stage	Dataset	Inputs						Outputs		#Samples	#Repeat	Ratio
		Text	Image	Video	Point	Box	Mask	Text	Mask			
1	Inst-IT-Image-Short-Caption [62]	✓	✓		✓	✓		✓		351K	1	41.2%
	VideoRefer-Short-Caption [100]	✓		✓	✓	✓		✓		500K	1	58.8%
2	RefCOCO [30]	✓	✓					✓	✓	17K	5	20.8%
	RefCOCO+ [30]	✓	✓					✓	✓	17K	5	20.8%
	RefCOCOg [56]	✓	✓					✓	✓	22K	5	26.8%
	RefClef [30]	✓	✓					✓	✓	18K	5	22.0%
	Ref-YouTube-VOS [73]	✓		✓				✓	✓	13K	3	9.5%
3	Osprey-Conversation [99]	✓	✓				✓	✓		1.4K	5	0.1%
	Osprey-Detail-Description [99]	✓	✓				✓	✓		29K	5	2.5%
	Osprey-Pos-Neg [99]	✓	✓				✓	✓		20K	5	1.7%
	VideoRefer-Detailed-Caption [100]	✓		✓			✓	✓		120K	5	10.1%
	VideoRefer-QA [100]	✓		✓			✓	✓		69K	5	5.8%
	Inst-IT-Video-QA [62]	✓		✓			✓	✓		159K	5	13.4%
	VideoRefer-QA-Memory [100]	✓		✓	✓	✓		✓	✓	69K	3	3.5%
	Inst-IT-QA-Memory [62]	✓		✓	✓	✓		✓	✓	158K	3	8.0%
	RefCOCO [30]	✓	✓					✓	✓	17K	10	2.9%
	RefCOCO+ [30]	✓	✓					✓	✓	17K	10	2.9%
	RefCOCOg [56]	✓	✓					✓	✓	22K	10	3.7%
	RefClef [30]	✓	✓					✓	✓	18K	10	3.0%
	ReasonSeg [33]	✓	✓					✓	✓	1.6K	10	0.3%
	ADE20K [108]	✓	✓					✓	✓	20K	3	1.0%
	COCOStuff [8]	✓	✓					✓	✓	118K	3	6.0%
	Mapillary Vistas [58]	✓	✓					✓	✓	18K	3	0.9%
	PACO-LVIS [66]	✓	✓					✓	✓	46K	3	2.3%
	PASCAL-Part [11]	✓	✓					✓	✓	4.4K	3	0.2%
	Ref-YouTube-VOS [73]	✓		✓				✓	✓	13K	5	1.1%
	Ref-DAVIS17 [64]	✓		✓				✓	✓	0.6K	10	0.1%
	Ref-SAV [98]	✓		✓				✓	✓	56K	3	2.8%
	MeViS [22]	✓		✓				✓	✓	23K	5	1.9%
	LV-VIS [78]	✓		✓				✓	✓	11K	3	0.6%
	ViCaS [2]	✓		✓				✓	✓	41K	3	2.1%
	ReVOS [93]	✓		✓				✓	✓	29K	5	2.5%
	GroundMoRe [21]	✓		✓				✓	✓	5.6K	3	0.3%
	LLaVA-1.5-Mix-665K [43]	✓	✓					✓		647K	1	10.9%
	VideoGPT+ Instruct [53]	✓		✓				✓		573K	1	9.7%

B Experiments

B.1 Tasks and Benchmarks

Our method is extensively evaluated across 9 fine-grained image/video understanding tasks. The benchmark(s) used for each task are listed as follows:

1. **Reasoning Video Object Segmentation:** ReVOS [93]
2. **Referring Video Object Segmentation:** MeViS [22], Ref-YouTube-VOS [73], Ref-DAVIS17 [64], Ref-SAV [98]
3. **Motion-Grounded Video Reasoning:** GroundMoRe [21]
4. **Referring Expression Segmentation:** RefCOCO [30], RefCOCO+ [30], RefCOCOg [56]
5. **Reasoning Segmentation:** ReasonSeg [33]
6. **Referring Expression Comprehension:** RefCOCO [30], RefCOCO+ [30], RefCOCOg [56]
7. **Referred Video Description:** VideoRefer-Bench$^{\text{D}}$ [100]
8. **Referred Video Question Answering:** VideoRefer-Bench$^{\text{Q}}$ [100]
9. **Flexible Pixel-Level Understanding:** PixelQA (Ours)

B.2 Evaluation Metrics

For video segmentation tasks, we adopt $\mathcal{J}\&\mathcal{F}$ as the main metric to jointly consider region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$. Image segmentation is evaluated using cIoU (the cumulative intersection over the cumulative union) and gIoU (the average of all per-image IoUs) following existing work. For referred video description and question answering tasks, we follow the official evaluation protocols to

Table 13: Performance comparison on general video question answering (VideoQA) on MVBench [37]. Note that UniPixel is the only model supporting pixel-level referring & segmentation.

Model	Size	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI	Avg.
GPT-4V [59]	–	55.5	**63.5**	72.0	46.5	73.5	18.5	59.0	29.5	12.0	40.5	83.5	39.0	12.0	22.5	45.0	47.5	52.0	31.0	**59.0**	11.0	43.5
Video-ChatGPT [54]	7B	23.5	26.0	62.0	22.5	26.5	54.0	28.0	40.0	23.0	20.0	31.0	30.5	25.5	39.5	48.5	29.0	33.0	29.5	26.0	35.5	32.7
Video-LLaMA [102]	7B	27.5	25.5	51.0	29.0	39.0	48.0	40.5	38.0	22.5	22.5	43.0	34.0	22.5	32.5	45.5	32.5	40.0	30.0	21.0	37.0	34.1
VideoChat [36]	7B	33.5	26.5	56.0	33.5	40.5	53.0	40.5	30.0	25.5	27.0	48.5	35.0	20.5	42.5	46.0	26.5	41.0	23.5	23.5	36.0	35.5
Video-LLaVA [39]	7B	46.0	42.5	56.5	39.0	53.5	53.0	48.0	41.0	29.0	31.5	82.5	45.0	26.0	53.0	41.5	33.5	41.5	27.5	38.5	31.5	43.0
TimeChat [70]	7B	40.5	36.0	61.0	32.5	53.0	53.5	41.5	29.0	19.5	26.5	66.5	34.0	20.0	43.5	42.0	36.5	36.0	29.0	35.0	35.0	38.5
PLLaVA [91]	7B	58.0	49.0	55.5	41.0	61.0	56.0	61.0	36.0	23.5	26.0	82.0	39.5	42.0	52.0	45.0	42.0	53.5	30.5	48.0	31.0	46.6
ST-LLM [45]	7B	66.0	53.5	**84.0**	44.0	58.5	80.5	73.5	38.5	42.5	31.0	86.5	36.5	56.5	78.5	43.0	44.5	46.5	34.5	41.5	58.5	54.9
VideoGPT+ [53]	4B	69.0	60.0	83.0	48.5	66.5	85.5	**75.5**	36.0	44.0	34.0	**89.5**	39.5	71.0	90.5	45.0	53.0	50.0	29.5	44.0	60.0	58.7
VideoChat2 [37]	7B	75.5	58.0	83.5	**50.5**	60.5	**87.5**	74.5	**45.0**	47.5	**44.0**	82.5	37.0	64.5	87.5	**51.0**	**66.5**	47.0	**35.0**	37.0	72.5	60.4
UniPixel (Ours)	3B	**69.5**	62.5	83.0	48.5	**76.5**	86.5	66.5	38.0	**49.0**	40.5	87.0	**49.0**	**74.0**	**95.0**	49.0	45.0	**63.5**	34.5	58.0	**73.5**	**62.5**

Table 14: Effectiveness justification of multi-stage training. The best and second-best results are marked **bold** and underlined, respectively. The three-stage recipe leads to optimal performance.

Stage 1	Stage 2	Stage 3	ReVOS			MeViS (val^u)			VideoRefer-Bench^Q	
			$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Single-Frame	Multi-Frame
		✓	58.3	63.6	61.0	54.8	61.9	58.4	71.1	71.5
✓		✓	59.0	63.4	61.2	55.2	62.1	58.7	71.8	72.3
	✓	✓	59.6	63.5	61.6	55.7	62.5	59.1	71.2	71.6
✓	✓	✓	**59.7**	**64.4**	**62.1**	**56.1**	**63.2**	**59.7**	**72.2**	**72.8**

report GPT-4o [60] scores and MCQ accuracy, respectively. For referring expression comprehension, we leverage mean accuracies, where a predicted bounding box is considered correct when it has the intersection over union (IoU) with the ground truth no less than 0.5.

B.3 More Experimental Results

General Video Question Answering We also evaluate UniPixel on MVBench [37] to compare its general video understanding capabilities with existing methods. The results are illustrated in Tab. 13. Note that our method is the only one in the table that supports referring and segmentation. By jointly training on holistic-level and pixel-level data, UniPixel can effectively balance the capabilities under both scenarios, demonstrated by the strong performance compared with holistic-level models.

B.4 Ablation Studies

Effect of Multi-stage Training We investigate the effectiveness of multi-stage training in Tab. 14. As shown in the first line, directly training the model using large-scale data only leads to sub-optimal performance, due to the unaligned representations among prompt encoder, LLM, and mask decoder. We observe that pre-training either the sparse prompt encoder or the L→M projector (the second and third lines) brings performance gains on both tasks (referring and segmentation). We hypothesize that this is because pre-aligning either of them can alleviate the burden of joint-task learning in stage 3. The last row verifies that the performance can be further boosted by pre-aligning both of them.

Number of Hidden Tokens for Mask Decoder As mentioned in the main paper, there is a huge gap between the feature dimensions of the LLM and the mask decoder, thus splitting the `<SEG>` token into more hidden tokens can better preserve the object information from the LLM. We ablate this mechanism in Tab. 15. According to the results, using only 1 hidden token cannot fully preserve the object information, as the mask prediction performance is sub-optimal. However, we also observe that using more than 2 hidden tokens (*e.g.*, 4 or 8) only brings negligible performance gain. Therefore, we choose 2 hidden tokens per object in our final model.

Training Strategy for the M→L projector The M→L projector aims to project the masked-pooled object-centric features to the LLM's embedding space. Since the object features originate from the visual encoder, it is possible to re-use the pre-trained weights of the original V→L projector in Qwen2.5-VL. Its effects are studied in Tab. 16. We investigated two strategies: 1) re-using the weights and 2) adding an extra pre-training stage for better alignment. The comparison shows that directly re-using weights without extra pre-training can achieve the best results.

Table 15: Ablation study on the number of hidden tokens for each <SEG>. Performance gains are negligible with more than 2 tokens/object.

#Tokens	ReVOS			MeViS (val^u)		
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
1	59.6	63.5	61.6	55.8	62.5	59.2
2	**59.7**	**64.4**	**62.1**	56.1	**63.2**	59.7
4	59.8	63.9	61.9	**56.8**	63.1	**59.9**
8	59.5	64.0	61.8	56.4	62.8	59.6

Table 16: Ablation study on M→L projector. Init and PT denote weight initialization from V→L projector and extra pre-training, respectively.

Init	PT	VideoRefer-Bench^Q		PixelQA
		Single-Frame	Multi-Frame	Mixed Acc
	✓	71.4	71.9	67.7
✓		71.5	71.7	67.4
✓	✓	**72.4**	72.6	68.2
		72.2	**72.8**	**68.5**

Table 17: Ablation study on training data used in stage 3. The best and second-best results are marked **bold** and underlined, respectively. Gradually adding more pixel-level data brings performance gains.

Regional	Segmentation	Memory	General	ReVOS			MeViS (val^u)			VideoRefer-Bench^Q	
				$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Single-Frame	Multi-Frame
✓				–	–	–	–	–	–	72.1	72.0
	✓			58.9	63.8	61.4	56.0	63.2	59.6	–	–
✓	✓			59.2	63.7	61.5	55.8	63.1	59.5	72.3	72.6
✓	✓	✓		59.6	**64.5**	**62.1**	**56.3**	**63.5**	**59.9**	**72.4**	72.5
✓	✓	✓	✓	**59.7**	64.4	**62.1**	56.1	63.2	59.7	72.2	**72.8**

Combination of Training Data Tab. 17 studies the effect of the combination of multi-task co-training data in stage 3. Compared with training only on the regional or segmentation data, leveraging both of them leads to considerable performance on both tasks. Incorporating memory pre-filling data (requiring both referring and segmentation) can further boost the performance. We also mix some general holistic-level video understanding data to preserve the original capabilities of the pre-trained model, while it slightly affects the performance on pixel-level tasks.

B.5 Qualitative Results

Fig. 6 ∼ 11 present more visualizations of outputs from UniPixel on different pixel-level understanding tasks. Our method can effectively handle flexible visual prompts [100], implicit queries [33, 93], long queries [98], and motion-grounded questions [21].

C Discussion

C.1 Limitations & Future Work

Due to the limited computing resources, we did not further scale up the training data to incorporate more pixel-level tasks such as grounded caption generation (GCG) on images [67] or videos [57], which are interesting scenarios and their data may bring more performance gains. Besides, the mask decoder currently predicts the first mask on the first frame and propagates it to the following frames, while it potentially supports predicting on the best frame (defined as the frame with the best view of the target) and propagates it to both sides of the video. We will focus in our future work to explore more pixel-level understanding tasks and more flexible mechanisms for the mask decoder.

C.2 Potential Societal Impacts

This work introduces a new framework for pixel-level visual-language understanding, which could potentially be used in education, surveillance, and healthcare industries, where flexible interactions with the users and fine-grained understanding of images & videos are required. In other scenarios requiring multi-modal assistants, our method can also serve as a more advanced alternative. To the best of our knowledge, there are no potential negative societal impacts to declare.

D Licenses

Our model is built based on the pre-trained Qwen2.5-VL [4] and SAM 2.1 [68] models. They are both licensed under the Apache License 2.0 (https://www.apache.org/licenses/LICENSE-2.0).

Figure 6: Visualization of the predictions from UniPixel on PixelQA.

Figure 7: Visualization of the predictions from UniPixel on ReVOS [93].

Figure 8: Visualization of the predictions from UniPixel on Ref-DAVIS17 [64].

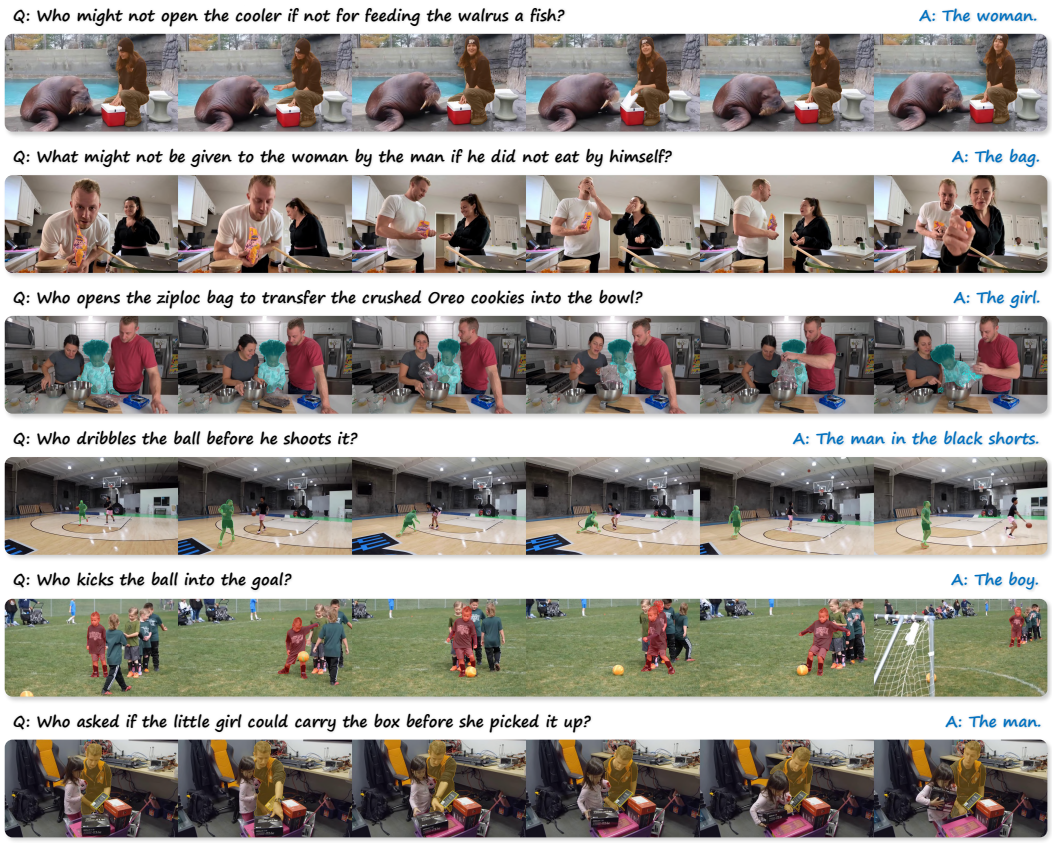

Figure 9: Visualization of the predictions from UniPixel on GroundMoRe [21].

Figure 10: Visualization of the predictions from UniPixel on Ref-SAV [98].

Figure 11: Visualization of the predictions from UniPixel on ReasonSeg [33].

References

[1] Anthropic. Claude 3.7 sonnet system card, 2025.

[2] Ali Athar, Xueqing Deng, and Liang-Chieh Chen. Vicas: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. *arXiv preprint arXiv:2412.09754*, 2024.

[3] 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. 2023.

[4] 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.

[5] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *NeurIPS*, 37:6833–6859, 2024.

[6] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, pages 4985–4995, 2022.

[7] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.

[8] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *CVPR*, pages 1209–1218, 2018.

[9] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[10] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[11] Xianjie Chen, Roozbeh Mottaghi, Xiaobai Liu, Sanja Fidler, Raquel Urtasun, and Alan Yuille. Detect what you can: Detecting and representing objects using holistic models and body parts. In *CVPR*, pages 1971–1978, 2014.

[12] Yuan Chen, Zi-han Ding, Ziqin Wang, Yan Wang, Lijun Zhang, and Si Liu. Asynchronous large language model enhanced planner for autonomous driving. In *ECCV*, pages 22–38. Springer, 2024.

[13] Zewen Chen, Juan Wang, Wen Wang, Sunhan Xu, Hang Xiong, Yun Zeng, Jian Guo, Shuxun Wang, Chunfeng Yuan, Bing Li, othersShilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*, 2023.

[14] 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.

[15] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

[16] Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, pages 640–658. Springer, 2022.

[17] Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 958–979, 2024.

[18] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *NeurIPS*, 36, 2023.

[19] Google DeepMind. Introducing gemini 2.0: our new ai model for the agentic era, 2024.

[20] Google DeepMind. Gemini 2.5: Our most intelligent ai model, 2025.

[21] Andong Deng, Tongjia Chen, Shoubin Yu, Taojiannan Yang, Lincoln Spencer, Yapeng Tian, Ajmal Saeed Mian, Mohit Bansal, and Chen Chen. Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. *arXiv preprint arXiv:2411.09921*, 2024.

[22] 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*, pages 2694–2703, 2023.

[23] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, et al. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*, 2023.

[24] Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. *arXiv preprint arXiv:2412.19806*, 2024.

[25] Chaoyou Fu, Yuhan Dai, Yongdong 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] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[28] Donggon Jang, Yucheol Cho, Suin Lee, Taehyeon Kim, and Dae-Shik Kim. Mmr: A large-scale benchmark dataset for multi-target and multi-granularity reasoning segmentation. *arXiv preprint arXiv:2503.13881*, 2025.

[29] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, pages 2758–2766, 2017.

[30] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[31] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023.

[33] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. In *CVPR*, pages 9579–9589, 2024.

[34] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[35] 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*, pages 19730–19742. PMLR, 2023.

[36] 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.

[37] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024.

[38] Zongzhao Li, Zongyang Ma, Mingze Li, Songyou Li, Yu Rong, Tingyang Xu, Ziqi Zhang, Deli Zhao, and Wenbing Huang. Star-r1: Spacial transformation reasoning by reinforcing multimodal llms. *arXiv preprint arXiv:2505.15804*, 2025.

[39] 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. *arXiv preprint arXiv:2311.10122*, 2023.

[40] Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. Glus: Global-local reasoning unified into a single large language model for video segmentation. *arXiv preprint arXiv:2504.07962*, 2025.

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755. Springer, 2014.

[42] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *CVPR*, pages 23592–23601, 2023.

[43] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.

[44] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023.

[45] Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. St-llm: Large language models are effective temporal learners. In *ECCV*, pages 1–18. Springer, 2024.

[46] Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. r^2-tuning: Efficient image-to-video transfer learning for video temporal grounding. In *ECCV*, 2024.

[47] Ye Liu, Huifang Li, Chao Hu, Shuang Luo, Yan Luo, and Chang Wen Chen. Learning to aggregate multi-scale context for instance segmentation in remote sensing images. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1):595–609, 2024.

[48] Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, pages 3042–3051, 2022.

[49] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. Videomind: A chain-of-lora agent for long video reasoning. *arXiv preprint arXiv:2503.13444*, 2025.

[50] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. E.t. bench: Towards open-ended event-level video-language understanding. *NeurIPS*, 37:32076–32110, 2024.

[51] Ye Liu, Junsong Yuan, and Chang Wen Chen. Consnet: Learning consistency graph for zero-shot human-object interaction detection. In *ACM MM*, pages 4235–4243, 2020.

[52] Zongyang Ma, Yuxin Chen, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Shaojie Zhu, Chengxiang Zhuo, Bing Li, Ye Liu, Zang Li, Ying Shan, and Weiming Hu. Visionmath: Vision-form mathematical problem-solving. In *ICCV*, 2025.

[53] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024.

[54] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[55] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *NeurIPS*, 36:46212–46244, 2023.

[56] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.

[57] 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.

[58] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulo, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, pages 4990–4999, 2017.

[59] OpenAI. Gpt-4v(ision) system card, 2023.

[60] OpenAI. Gpt-4o system card, 2024.

[61] OpenAI. Openai o1 system card, 2024.

[62] Wujian Peng, Lingchen Meng, Yitong Chen, Yiweng Xie, Yang Liu, Tao Gui, Hang Xu, Xipeng Qiu, Zuxuan Wu, and Yu-Gang Jiang. Inst-it: Boosting multimodal instance understanding via explicit visual prompt instruction tuning. *arXiv preprint arXiv:2412.03565*, 2024.

[63] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. In *CVPR*, pages 27124–27133, 2024.

[64] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[65] Jihao Qiu, Yuan Zhang, Xi Tang, Lingxi Xie, Tianren Ma, Pengyu Yan, David Doermann, Qixiang Ye, and Yunjie Tian. Artemis: Towards referential understanding in complex videos. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[66] Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, Yi Wen, Baixue Zheng, Baishan Guo, Rui Wang, Aaron Marquez, Rama Kovvuri, Abhishek Kadian, et al. Paco: Parts and attributes of common objects. In *CVPR*, pages 7141–7151, 2023.

[67] 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*, pages 13009–13018, 2024.

[68] 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.

[69] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

[70] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *CVPR*, pages 14313–14323, 2024.

[71] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. In *CVPR*, pages 26374–26383, 2024.

[72] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *ICML*, pages 29441–29454. PMLR, 2023.

[73] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *ECCV*, pages 208–223. Springer, 2020.

[74] Nicholas Stroh. Trackgpt–a generative pre-trained transformer for cross-domain entity trajectory forecasting. *arXiv preprint arXiv:2402.00066*, 2024.

[75] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 33:7537–7547, 2020.

[76] OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024.

[77] Han Wang, Yongjie Ye, Yanjie Wang, Yuxiang Nie, and Can Huang. Elysium: Exploring object-level perception in videos via mllm. In *ECCV*, pages 166–185. Springer, 2024.

[78] Haochen Wang, Cilin Yan, Shuai Wang, Xiaolong Jiang, Xu Tang, Yao Hu, Weidi Xie, and Efstratios Gavves. Towards open-vocabulary video instance segmentation. In *ICCV*, pages 4057–4066, 2023.

[79] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[80] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, pages 23318–23340. PMLR, 2022.

[81] Shihao Wang, Zhiding Yu, Xiaohui Jiang, Shiyi Lan, Min Shi, Nadine Chang, Jan Kautz, Ying Li, and Jose M Alvarez. Omnidrive: A holistic llm-agent framework for autonomous driving with 3d perception, reasoning and planning. *arXiv preprint arXiv:2405.01533*, 2024.

[82] 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.

[83] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model. *arXiv preprint arXiv:2411.17606*, 2024.

[84] 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 arXiv:2412.14006*, 2024.

[85] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *ICCV*, pages 2761–2770, 2023.

[86] Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. A survey on video temporal grounding with multimodal large language model. *arXiv preprint arXiv:2508.10922*, 2025.

[87] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, pages 4974–4984, 2022.

[88] Jiannan Wu, Yi Jiang, Bin Yan, Huchuan Lu, Zehuan Yuan, and Ping Luo. Segment every reference object in spatial and temporal spaces. In *ICCV*, pages 2538–2550, 2023.

[89] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. In *CVPR*, pages 3858–3869, 2024.

[90] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*, pages 1645–1653, 2017.

[91] 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.

[92] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. In *CVPR*, pages 15325–15336, 2023.

[93] 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*, pages 98–115. Springer, 2024.

[94] Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023.

[95] Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704*, 2023.

[96] Samson Yu, Kelvin Lin, Anxing Xiao, Jiafei Duan, and Harold Soh. Octopi: Object property reasoning with large tactile-language models. *arXiv preprint arXiv:2405.02794*, 2024.

[97] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *AAAI*, volume 33, pages 9127–9134, 2019.

[98] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

[99] 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*, pages 28202–28211, 2024.

[100] Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. *arXiv preprint arXiv:2501.00599*, 2024.

[101] Ao Zhang, Yuan Yao, Wei Ji, Zhiyuan Liu, and Tat-Seng Chua. Next-chat: An lmm for chat, detection and segmentation. *arXiv preprint arXiv:2311.04498*, 2023.

[102] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[103] Jiawei Zhang, Chejian Xu, and Bo Li. Chatscene: Knowledge-enabled safety-critical scenario generation for autonomous vehicles. In *CVPR*, pages 15459–15469, 2024.

[104] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *NeurIPS*, 37:71737–71767, 2024.

[105] Yichi Zhang, Ziqiao Ma, Xiaofeng Gao, Suhaila Shakiah, Qiaozi Gao, and Joyce Chai. Groundhog: Grounding large language models to holistic segmentation. In *CVPR*, pages 14227–14238, 2024.

[106] Xufeng Zhao, Mengdi Li, Cornelius Weber, Muhammad Burhan Hafez, and Stefan Wermter. Chat with the environment: Interactive multimodal perception using large language models. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3590–3596. IEEE, 2023.

[107] Rongkun Zheng, Lu Qi, Xi Chen, Yi Wang, Kun Wang, Yu Qiao, and Hengshuang Zhao. Villa: Video reasoning segmentation with large language model. *arXiv preprint arXiv:2407.14500*, 2024.

[108] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, pages 633–641, 2017.

[109] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[110] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[111] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. Generalized decoding for pixel, image, and language. In *CVPR*, pages 15116–15127, 2023.

[112] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *NeurIPS*, 36:19769–19782, 2023.