HyperSeg: Hybrid Segmentation Assistant with Fine-grained Visual Perceiver

Cong Wei[1,2], Yujie Zhong[2†], Haoxian Tan[2], Yong Liu[1], Jie Hu[2], Dengjie Li[2], Zheng Zhao[2], Yujiu Yang[1†]
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University [2]Meituan Inc.
weic22@mails.tsinghua.edu.cn, jaszhong@hotmail.com, yang.yujiu@sz.tsinghua.edu.cn

Figure 1. Illustration of our HyperSeg which can conduct image and video segmentation tasks with various language and visual instructions. Additionally, HyperSeg can handle complicated reasoning perception tasks compared with previous universal segmentation methods. To our knowledge, HyperSeg is the first VLLM-based universal segmentation model with perception and complex reasoning abilities in both image and video domains.

Abstract

This paper aims to address universal segmentation for image and video perception with the strong reasoning ability empowered by Visual Large Language Models (VLLMs). Despite significant progress in current unified segmentation methods, limitations in adaptation to both image and video scenarios, as well as the complex reasoning segmentation, make it difficult for them to handle various challenging instructions and achieve an accurate understanding of fine-grained vision-language correlations. We propose HyperSeg, the first VLLM-based universal segmentation model for pixel-level image and video perception, encompassing generic segmentation tasks and more complex reasoning perception tasks requiring powerful reasoning abilities and world knowledge. Besides, to fully leverage the recognition capabilities of VLLMs and the fine-grained visual information, HyperSeg incorporates hybrid entity recognition and fine-grained visual perceiver modules for various segmentation tasks. Combined with the temporal adapter, HyperSeg achieves a comprehensive understanding of temporal information. Experimental results validate the effectiveness of our insights in resolving universal image and video segmentation tasks, including the more complex reasoning perception tasks. Our code is available at https://github.com/congvvc/HyperSeg.

†Corresponding authors.

1. Introduction

Visual segmentation is one of the most significant tasks in computer vision research, which aims to perform accurate pixel-level semantic understanding. Many specialist models [7, 17, 19, 22] have made great progress in specific segmentation tasks while showing limitations in handling diverse and complicated scenarios since new training data, paradigms, and model architectures are required to adapt to new segmentation tasks. Recent works [24, 25, 57] propose a single framework to unify diverse segmentation tasks. Despite promising, they show the inability to tackle text instructions and complex reasoning segmentation tasks needing powerful reasoning capabilities and world knowledge.

Visual Large Language Models (VLLMs) have exhibited excellent reasoning and conversation abilities, which play a pivotal role in various vision-language co-understanding tasks [2, 8, 23, 29, 61]. **However, these methods are based on rudimentary vision-language alignment, which limits their ability to comprehend finer details in visual perception tasks, like pixel-level segmentation.** Recent studies [22, 45, 51, 58, 59] enables VLLMs to perform fine-grained visual understanding, like referring and reasoning segmentation. [22, 41, 45] use the special token [SEG] generated by VLLMs as the prompt for the mask decoder to generate segmentation masks while [58, 59] focus on incorporating instance-aware mask tokens into VLLMs. **Though impressive, they show limitations to the universal segmentation framework based on VLLMs for both image and video domains and the capabilities of handling more complex video reasoning segmentation tasks.**

To this end, we introduce HyperSeg, the first VLLM-based universal segmentation model for pixel-level image and video perception with complex reasoning and conversation capabilities. HyperSeg can conduct diverse image and video segmentation tasks with various elaborate prompts and temporal adapter module. Besides, HyperSeg shows excellent abilities in complicated vision-language reasoning perception tasks needing rich world knowledge, which is significant for real-world understanding and interactions. As shown in Fig. 1, the explored tasks contain both image and video domains. We organize the tasks into two unified prompt formats: (1) text prompts (class names, reasoning questions, and referring languages), (2) visual prompts (box, mask, etc.). Owing to such flexible and cohesive design, HyperSeg benefits from concurrent training on diverse segmentation tasks and vision domains, facilitating the intricate correlations between different instructions and visual concepts. To further enhance fine-grained object perception and video understanding, we introduce three distinct designs in the following.

Firstly, we incorporate a hybrid entity recognition strategy to enhance the exploitation of VLLM's recognition capacity. Generation-only works [22, 41, 49] solely rely on VLLM for object prediction leading to poor performance in complex multi-object segmentation scenarios. Decode-only methods [58, 59] use the prompt embedding and mask tokens decoded by VLLM to obtain class scores for each mask, which makes the mask tokens interact insufficiently with the semantic condition as they ignore the powerful generative capabilities of VLLM. The proposed hybrid entity recognition leverages the VLLM's powerful generative abilities to enhance the mask tokens' comprehension of category semantics while maintaining the final class scores decoding process.

Secondly, previous VLLMs usually use coarse-level visual features obtained from CLIP [38] series which primarily encode global visual information while overlooking visual details. To enhance VLLMs' ability of capturing visual details efficiently, we use the Fine-grained Visual Perceiver (FVP) to merge multi-scale visual features into fixed-length fine-grained tokens, allowing retrieval of rich visual details from various scales in the hierarchical vision encoder [7].

Thirdly, recent VLLM-based segmentation methods [22, 58, 59] demonstrate limitations in handling video perception tasks for video temporal understanding. To this end, we propose the temporal adapter for comprehensive video perception which incorporates global prompt aggregation and local space-time information injection for the coalescence of both long-term and short-term vision-language information.

Extensive experiments on various segmentation benchmarks demonstrate the preeminent segmentation ability of HyperSeg , providing strong evidence of the effectiveness of our insights. Our HyperSeg also exhibits promising performance on common Multi-modal benchmarks. Additionally, we explore the mutual influence among different tasks involving various visual and task types.

Our contributions are summarized as follows:

- We present HyperSeg, the first VLLM-based universal segmentation model for pixel-level image and video perception, covering a broad spectrum of common segmentation tasks, complex reasoning, and conversation-based vision-language understanding tasks.
- We incorporate hybrid entity recognition and fine-grained visual perceiver modules to VLLM, which allow full exploitation of VLLM's semantic recognition capacity and injection of fine-grained visual information to improve diverse detail-aware segmentation tasks. With the temporal adapter, HyperSeg can conduct more challenging video perception tasks, achieving universal segmentation.
- HyperSeg demonstrates superior capabilities on multiple segmentation tasks, achieving excellent performance on both generic and complex reasoning benchmarks with only one model.

2. Related Work

Visual Large Language Model. The emergence of Large Language Model (LLM) has significantly contributed to the development of VLMM. In this context, LLMs are enhanced with multimodal comprehension capabilities, allowing the vision-language co-understanding [1, 2, 23, 28, 29, 61]. Several notable examples of LLMs with multimodal comprehension include BLIP-2 [23], Flamingo [1], MiniGPT-4 [61], LLaVA [29], InstructBLIP [10], and Qwen-VL [2]. While these models have demonstrated impressive performance in vision-language tasks, they solely produce textual outputs that describe the entire image. This restricts their applicability in tasks that require the pixel-level detailed understanding.

Perception with VLLM. Several methods have been proposed to enhance VLLMs with a more detailed comprehension capability [5, 22, 36, 37, 39, 41, 43, 54]. Shikra [5], Ferret [54], Kosmos-2 [36], and VisionLLM [43] are examples that provide grounding capabilities through regression of box coordinates. Conversely, LISA [22], PixelLM [41], GLaMM [39], and PerceptionGPT [37] employ a mask decoder to predict object masks from special tokens. Most of the existing methods utilize a next-token-prediction approach, which restricts their applicability. PSALM [59] makes an important attempt to bring VLLM into visual perception tasks but fails to fully unleash the potential of VLLM. In contrast, our method propose to use a hybrid strategy to mitigate this problem and keep the capacity in high-level reasoning.

Unified segmentation model. Another line of studies focuses on the integration of various segmentation tasks into a single model. Mask2former [7] proposes a unified architecture that requires separate training on different segmentation tasks. OpenSeeD [57] introduces a text encoder and extends it to the Open-Set setting. Simultaneously, UNINEXT [25] supports referring segmentation with the assistance of text inputs and a text encoder. However, these works fall short of following complicated instructions and reasoning. In this work, we improve the understanding ability toward language by incorporating LLM, while also maintaining the original ability of vision-centric models.

3. Method

3.1. Overview

Overall architecture. The architecture of HyperSeg is illustrated in Fig. 2, which consists of a fine-grained pyramid visual encoder, a light-weight VLLM, and a segmentation predictor to generate segmentation masks, class scores, and instance embedding for video correspondence according to user's instruction. The proposed FVP module fuses multi-scale high-resolution visual features f_{img} into a set of fine-grained tokens to ensure the injection of fine-grained visual information (Sec 3.3). The VLLM takes three types of inputs: visual tokens encoded by the CLIP encoder, renewed fine-grained tokens, and prompt tokens for diverse instructions. The output embeddings of semantically enhanced mask tokens (Sec 3.2) and prompt tokens are further fed into the segmentation predictor for final segmentation results. Besides, we utilize the space-time information propagation and global prompt aggregation for comprehensive video understanding (Sec 3.4). We train the LLM with LoRA for efficient parameter tuning.

Visual Large Language Model. We take a light-weight VLLM as our powerful multi-modal feature encoder, which contains a low-resolution vision encoder like CLIP [38] and an efficient LLM.

Specifically, the model takes vision-prompt pairs $\{(\mathcal{V}, \mathcal{P})\}$ as inputs, where $\mathcal{V}$ is resized to low resolution and then encoded by CLIP encoder F_{CLIP} to get image features f_v. The f_v is further projected and concatenated with other task-specific tokens to ensure the comprehensive understanding of multi-modal inputs through the fusion process of LLM F_{LLM}, where G_c is the projection function and E_O denotes the output embeddings of LLM. Formally,

$$f_v = F_{CLIP}(\mathcal{V}), E_O = F_{LLM}(G_c(f_v), P, \mathcal{P}), \quad (1)$$

where P denotes fine-grained tokens. Furthermore, we manually extract semantic enhanced mask tokens $E_{\mathcal{Q}}$ and prompt embedding $E_{\mathcal{P}}$ from E_O, which are further fed into the pre-trained segmentation predictor [7] to generate masks, class scores, and instance embedding for final segmentation results.

Prompt design. In order to accommodate the different segmentation tasks, we propose a flexible design for prompt $\mathcal{P}$. As illustrated above, we divide $\mathcal{P}$ into two formats: text prompts and visual prompts. To be specific, $\mathcal{P}$ contains the instructions $\mathcal{P}_{\mathcal{I}}$ and task-specific conditions $\mathcal{P}_{\mathcal{C}}$, where $\mathcal{P}_{\mathcal{I}}$ instructs the model to perform different tasks while $\mathcal{P}_{\mathcal{C}}$ indicates diverse conditions which are further used as classifiers to calculate the class scores of predicted masks.

For class-based segmentation tasks like panoptic segmentation, open-vocabulary segmentation (OVS), and video instance segmentation (VIS), $\mathcal{P}$ can be demonstrated as $\mathcal{P}_{\mathcal{I}}$: *"Please segment all the positive objects according to the following potential categories."* $\mathcal{P}_{\mathcal{C}}$: *"[category 1, category 2, category 3, ...]"*

For referring and reasoning segmentation tasks like referring expression segmentation (RES), reasoning segmentation, referring video object segmentation (R-VOS), and ReasonVOS, $\mathcal{P}$ can be designed as $\mathcal{P}_{\mathcal{I}}$: *"Can you perform referring or reasoning segmentation according to the language expression?"* $\mathcal{P}_{\mathcal{C}}$: *"[referring / reasoning text]"*

For visual-guided segmentation tasks like interactive segmentation and video object segmentation (VOS), $\mathcal{P}$ can be designed as $\mathcal{P}_{\mathcal{I}}$: *"Please segment according to the given*

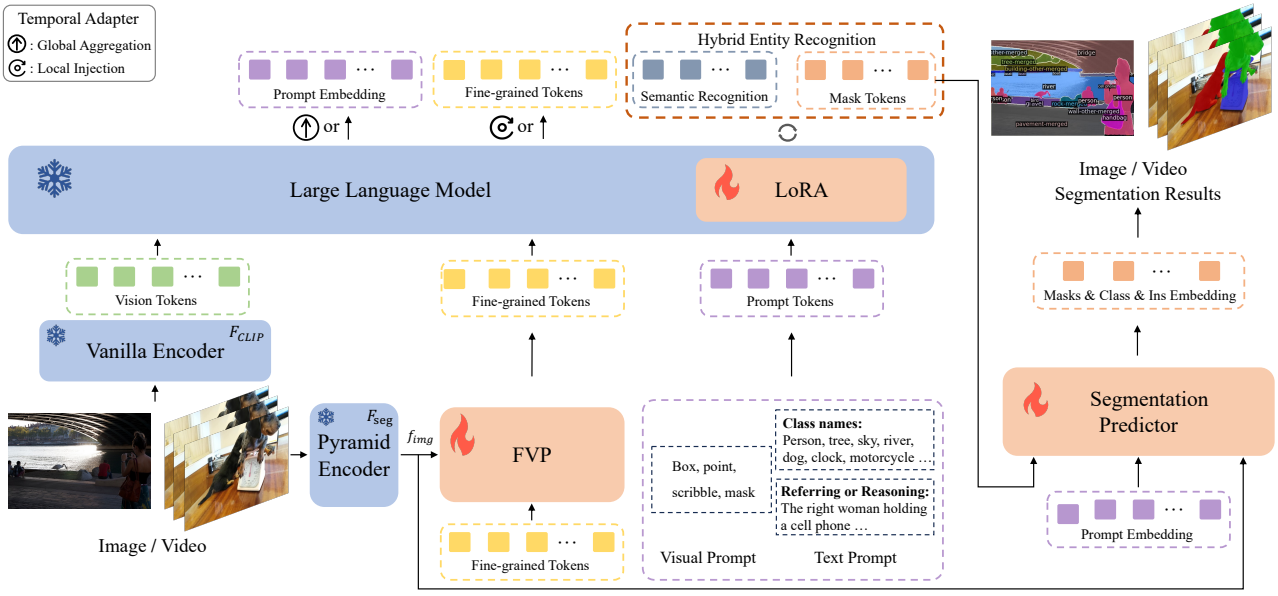

Figure 2. **Overview of HyperSeg**. HyperSeg encodes the visual input in a multi-grained manner and concatenates the prompt for different perception tasks. We feed learnable fine-grained tokens into a Fine-grained Visual Perceiver (FVP) to integrate multi-scale high-resolution image features into LLM for detailed visual learning and to facilitate space-time information propagation for video understanding. Additionally, we use the semantically enhanced mask tokens and prompt embedding to finally generate the segmentation masks and class scores for generic segmentation, and instance embedding for video instance association.

visual region reference" $\mathcal{P}_C$: "[vision 1, vision 2, vision 3, ...]". Instead of using an additional region encoder to extract visual reference features [25], we sample the CLIP visual features f_v in VLLM according to the region coordinates and perform adaptive average pooling on them to form the final reference features for each visual prompt.

Segmentation predictor. Segmentation predictor F_p generates the masks m, corresponding class scores z, and instance embedding e through the similar process [7, 15] of three inputs: task-specific prompt embedding $\{E^k_{\mathcal{P}}\}^K_{k=1}$, the semantically enhanced mask tokens $\{E^j_{\mathcal{Q}}\}^N_{j=1}$ and the multi-scale visual features f_{img}, where K and N denote K categories and N mask proposals. Formally,

$$\{m_j, z_j, e_j\}^N_{j=1} = F_p(\{E^k_{\mathcal{P}}\}^K_{k=1}, \{E^j_{\mathcal{Q}}\}^N_{j=1}, f_{img}), \quad (2)$$

where $m_j \in \mathbb{R}^{H\times W}$ is the j-th mask proposal, $z_j \in \mathbb{R}^K$ denotes the class scores of m_j, and $e_j \in \mathbb{R}^D$ denotes the j-th instance embedding obtained from an extra embedding head only for video domain. For video tasks, we adopt a frame-by-frame manner to get frame-level segmentation results for efficient training and inference processes.

Training objectives. The model can be trained jointly on multiple tasks using the unified loss $\mathcal{L}$. Specifically, we employ an autoregressive cross-entropy loss $\mathcal{L}_{text}$ for text prediction, a combination of per-pixel binary cross-entropy loss $\mathcal{L}_{bce}$ and DICE loss $\mathcal{L}_{dice}$ for mask supervision $\mathcal{L}_{mask}$, a cross-entropy loss $\mathcal{L}_{cls}$ for category classification, and a contrastive loss $\mathcal{L}_{ins}$ for instance association of video sequences following [48]. λ indicates their sum weight respectively. Formally,

$$\mathcal{L} = \mathcal{L}_{text} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{cls}\mathcal{L}_{cls} + \lambda_{ins}\mathcal{L}_{ins}, \quad (3)$$

$$\mathcal{L}_{mask} = \lambda_{bce}\mathcal{L}_{bce} + \lambda_{dice}\mathcal{L}_{dice}, \quad (4)$$

Differences between HyperSeg and previous methods. Previous universal segmentation methods [24, 25, 31] lacking of VLLMs show inability in reasoning perception tasks while our HyperSeg demonstrates brilliant reasoning segmentation capability in complex scenarios. Besides, we make a significant generalization of the current VLLM-based segmentation methods [22, 51, 58, 59] for more diverse segmentation tasks in both image and video domains using a single model framework. Moreover, HyperSeg differs from previous methods in the three designs elaborated in the following sections.

3.2. Hybrid Entity Recognition

As shown in Fig. 3 (a), predicting presented objects in the way of sequence generation (semantic prediction) tends to miss objects or produce repetitive predictions [45]. On the other hand, Fig. 3 (b), only using VLLM to embed class names (prompt tokens) as mask classifier at the decode stage disregards VLLM's powerful semantic recognition capability. Consequently, we propose a hybrid approach that leverages LLM in both generation and decoding processes.

Instead of integrating mask tokens in input sequences and extracting the corresponding embedding from the one-pass

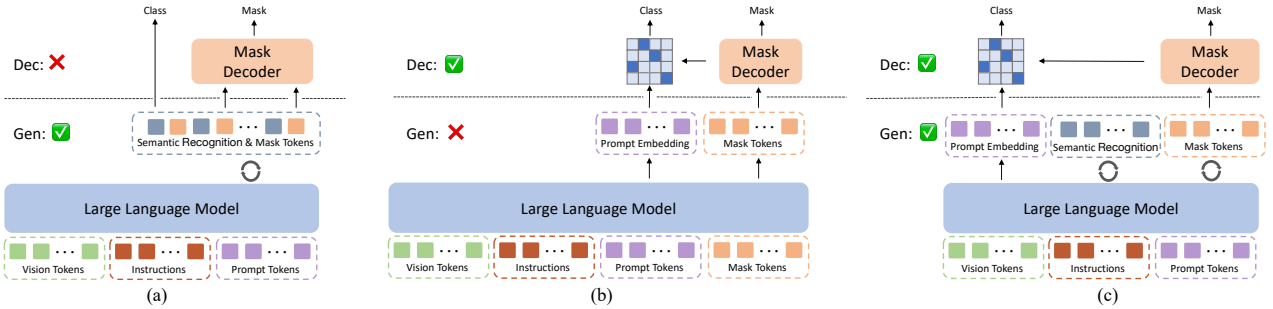

Figure 3. **The comparison of different recognition strategies**. (a) Generation-Only [22, 41]: both the semantic recognition (existing objects) and their mask tokens are generated by LLM. (b) Decode-Only [58, 59]: prompt embedding and mask tokens are decoded from LLM. The present objects are then determined by their similarity scores. (c) Hybrid (ours): prompt embedding is decoded from LLM while the semantically enhanced mask tokens are generated by LLM. Their similarity scores reflect the objects' presence.

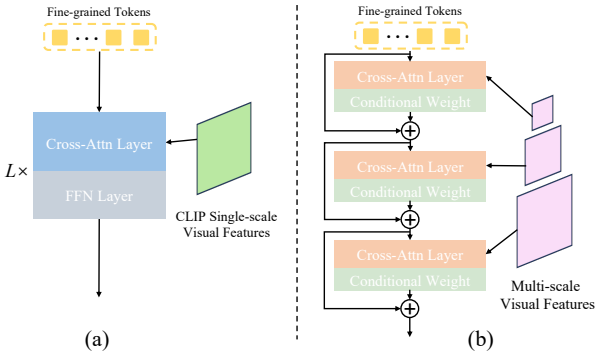

Figure 4. **Comparison between previous vision perceiver and our FVP**. (a): previous vision perceiver [2, 23] uses the coarse single-scale CLIP visual features which are inadequate for fine-grained perception tasks. (b): FVP encodes the multi-scale visual features into fine-grained tokens.

forward output of VLLM, we instruct VLLM to generate the mask tokens preceded by the estimated objects' names. As illustrated in Fig. 3 (c), VLLM is compelled to generate all the existing objects in the vision input and then the mask tokens. The semantically enhanced mask tokens contain valuable semantic integrated information about the image, which are subsequently used as input for the segmentation predictor to generate segmentation masks.

3.3. Fine-grained Visual Perceiver

Why twin-tower vision encoder? As shown in Fig. 4, previous VLLMs and VLLM-based segmentation methods usually utilize the pre-trained CLIP encoder to obtain single-scale and low-resolution vision features interacted with diverse languages, which is insufficient for fine-grained image and video segmentation tasks. Therefore, we adopt an extra pyramid vision encoder [7] to inject details-aware visual information.

Specifically, we fuse multi-scale visual features into fine-grained tokens (stated as P in Sec 3.1) which can inject rich fine-grained visual information into the pre-trained VLLMs without excessive computation cost. Formally, given the vision input $\mathcal{V}$, we leverage a pyramid vision encoder [7] F_{seg} to get details-aware image features f_{img}. For the j-th scale and the previous fine-grained tokens P_{j-1}, the FVP module enriches each token through conditional weighted cross-attention:

$$\hat{P}_j = \text{MHCA}(P_{j-1}, G_p(f_{img}^{(j)})), \tag{5}$$

$$P_j = P_{j-1} + \tanh(\text{MLP}(\hat{P}_j)) \cdot \hat{P}_j, \tag{6}$$

where MHCA denotes the Multi-Head Cross-Attention layer, G_p is the projection function, tanh is a normalization function and MLP is a Multilayer Perceptron. The component of $\tanh(\text{MLP}(\hat{P}_j))$ is the *conditional weight* used to multiply the enriched fine-grained tokens $\hat{P}_j$ before the residual connection to the previous tokens P_{j-1}. Additionally, we initialize the weight value to zero to ensure the adaptation to diverse multi-scale image features while retaining the training stability.

3.4. Temporal Adapter

Video segmentation entails distinct challenges, requiring reasoning across multiple frames and the maintenance of temporal coherence. Existing VLLM-based methods exhibit limitations in addressing video perception tasks and lack specialized designs for comprehending temporal dynamics in video analysis. To this end, we utilize global prompt aggregation and local space-time information injection in the time dimension to adapt to more complicated video perception tasks.

Global prompt aggregation. For the current prompt embedding $E_{\mathcal{P}}$ in the video object mask retrieval process, we leverage the adaptive average pooling strategy along the time dimension to aggregate global object and temporal information of previous T frames.

$$E_{\mathcal{P}} = AvgPool([E_{\mathcal{P}}^0, E_{\mathcal{P}}^1, ..., E_{\mathcal{P}}^T]), \tag{7}$$

Local space-time information injection. We propose a sequential renewal strategy for space-time information prop-

Table 1. Comparison with the state-of-the-art models on the RefCOCO series and more challenging generalized referring expression segmentation benchmark gRefCOCO. ‡ denotes models using pre-trained SAM [21] for mask generation. * means using gRefCOCO for training while other methods are evaluated in zero-shot manners.

Type	Method	RefCOCO			RefCOCO+			RefCOCOg		gRefCOCO		
		val	testA	testB	val	testA	testB	val(U)	test(U)	val	testA	testB
Segmentation Specialist	VLT [11]	67.5	70.5	65.2	56.3	61.0	50.1	55.0	57.7	52.5*	62.2*	50.5*
	CRIS [44]	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	55.3*	63.8*	51.0*
	LAVT [53]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	57.6*	65.3*	55.0*
	PolyFormer-B [30]	74.8	76.6	71.1	67.6	72.9	59.3	67.8	69.1	-	-	-
VLLM-based Segmentation Network	LISA-7B [22] ‡	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	38.7*	52.6*	44.8*
	PixelLM-7B [41]	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	-	-	-
	GSVA-7B [49] ‡	76.4	77.4	72.8	64.5	67.7	58.6	71.1	72.0	61.7*	69.2*	60.3*
	GroundHog-7B [33]	78.5	79.9	75.7	70.5	75.0	64.9	74.1	74.6	66.7*	-	-
	SAM4MLLM-7B [6] ‡	79.6	82.8	76.1	73.5	77.8	65.8	74.5	75.6	66.3*	70.1*	63.2*
	LaSagnA-7B [45] ‡	76.8	78.7	73.8	66.4	70.6	60.1	70.6	71.9	38.1	50.4	42.1
	OMG-LLaVA [58]	78.0	80.3	74.1	69.1	73.1	63.0	72.9	72.9	-	-	-
	GLaMM [40] ‡	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	-	-	-
	PSALM [59]	83.6	84.7	81.6	72.9	75.5	70.1	73.8	74.4	42.0	52.4	50.6
	HyperSeg	**84.8**	**85.7**	**83.4**	**79.0**	**83.5**	**75.2**	**79.4**	**78.9**	**47.5**	**57.3**	**52.5**

Table 2. Comparison with the state-of-the-art models on more complex and challenging reasoning segmentation benchmarks: ReVOS in video domain and ReasonSeg in image domain. ‡ denotes the same meaning as Tab. 1.

Method	Backbone	ReVOS-Reasoning			ReVOS-Referring			ReVOS-Overall			ReasonSeg	
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	gIoU	cIoU
LMPM [12]	Swin-T	13.3	24.3	18.8	29.0	39.1	34.1	21.2	31.7	26.4	-	-
ReferFormer [47]	Video-Swin-B	21.3	25.6	23.4	31.2	34.3	32.7	26.2	29.9	28.1	-	-
LISA-7B [22] ‡	ViT-H	33.8	38.4	36.1	44.3	47.1	45.7	39.1	42.7	40.9	52.9	54.0
LaSagnA-7B [45] ‡	ViT-H	-	-	-	-	-	-	-	-	-	48.8	47.2
SAM4MLLM-7B [6] ‡	EfficientViT-SAM-XL1	-	-	-	-	-	-	-	-	-	46.7	48.1
TrackGPT-13B [62] ‡	ViT-H	38.1	42.9	40.5	48.3	50.6	49.5	43.2	46.8	45.0	-	-
VISA-7B [51] ‡	ViT-H	36.7	41.7	39.2	51.1	54.7	52.9	43.9	48.2	46.1	52.7	**57.8**
VISA-13B [51] ‡	ViT-H	38.3	43.5	40.9	52.3	55.8	54.1	45.3	49.7	47.5	-	-
HyperSeg-3B	Swin-B	**50.2**	**55.8**	**53.0**	**56.0**	**60.9**	**58.5**	**53.1**	**58.4**	**55.7**	**59.2**	56.7

agation based on fine-grained tokens P to inject object information of adjacent frames. Formally,

$$P_t = G_l[F_{LLM}(P_{t-1})], \tag{8}$$

where P_t denotes the time-aware fine-grained tokens of the current t-th frame, G_l is the projection function to transfer the previous features to the current space and align the feature dimensions.

The proposed global prompt aggregation and local space-time information injection within our temporal adapter facilitate the coalescence of both long-term and short-term vision-language information, which is essential for comprehensive video perception. More details about the temporal adapter are provided in the Supplementary Material.

4. Experiments

Datasets. We use the one-stage training strategy to train HyperSeg with the multi-dataset and multi-task manners. For image segmentation, we use COCO Panoptic [26], RefCOCO series [35, 55], COCO-Interactive, and ReasonSeg [22]. For video segmentation, we utilize DAVIS-2017 datasets [4], Ref-Youtube-VOS [42], YouTube-VIS 2019 [52], and ReVOS [51]. Besides, we use LLAVA-150k [29] to maintain the vision-language conversation capability of VLLM (we show the results on Multi-modal benchmarks in the Supplementary Material).

Implementation details We load the pre-trained weights of Mipha [63] for our VLLM, and Maks2Former [7] for our segmentation predictor. We use three layers of FVP for fine-grained information fusion and utilize LoRA [18] to finetune the LLM efficiently. We train HyperSeg jointly for 160k iterations using a batch size of 32 on 8 NVIDIA A100 GPUs, which means each task takes approximately 16k iterations. We employ the AdamW optimizer with a learning rate of 4×10^{-5} and with a cosine schedule. All the hyper-parameters in the loss $\mathcal{L}$ are assigned values 1.0.

Table 3. Quantitative results on the closed-set COCO-Panoptic segmentation, open-vocabulary segmentation (-OV) benchmarks.

Type	Method	Backbone	COCO-Panoptic		ADE-OV		Citys-OV	PC59-OV	PAS20-OV
			PQ	mIoU	PQ	mIoU	PQ	mIoU	mIoU
Segmentation Specialist	Mask2former [7]	Swin-B	55.1	65.1	-	-	-	-	-
	OneFormer [19]	Swin-L	57.9	67.4	-	-	-	-	-
	SEEM [64]	DaViT-B	56.1	66.3	-	-	-	-	-
	MaskCLIP [13]	ViT-L	30.9	47.6	15.1	**23.7**	-	45.9	-
	DeOP [16]	ResNet-101c	-	-	-	22.9	-	48.8	91.7
	SimBaseline [50]	ViT-B	-	-	-	20.5	-	47.7	88.4
	DaTaSeg [15]	ViTDet-B	52.8	62.7	12.3	18.3	28.0	51.1	-
VLLM-based Segmentation Network	OMG-LLaVA [58]	ConvNeXt-L	53.8	-	-	-	-	-	-
	PSALM [22]	Swin-B	55.9	66.6	13.7	18.2	28.8	48.5	81.3
	HyperSeg	Swin-B	**61.2**	**77.2**	**16.1**	22.3	**31.1**	**64.6**	**92.1**

Table 4. Results of common video segmentation benchmarks, including DAVIS17, Ref-YouTube-VOS, Ref-DAVIS17, and YouTube-VIS 2019. ‡ denotes the same meaning as Tab. 1.

Method	Backbone	DAVIS17	Ref-YT	Ref-DAVIS	YT-VIS
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	mAP
SEEM [64]	DaViT-B	62.8	-	-	-
OMG-Seg [24]	ConvNeXt-L	74.3	-	-	56.4
ReferFormer [47]	Video-Swin-B	-	62.9	61.1	-
OnlineRefer [46]	Swin-L	-	63.5	64.8	-
UNINEXT [25]	ConvNeXt-L	77.2	66.2	66.7	**64.3**
LISA-7B [22] ‡	ViT-H	-	53.9	64.8	-
VISA-13B [51] ‡	ViT-H	-	63.0	70.4	-
VideoLISA-3.8B [3] ‡	ViT-H	-	63.7	68.8	-
HyperSeg-3B	Swin-B	**77.6**	**68.5**	**71.2**	53.8

4.1. Comparisons with State-of-the-Arts

Referring expression segmentation results. We compare HyperSeg with the state-of-the-art methods on the benchmarks RefCOCO/+/g [35, 55] and more challenging generalized referring expression segmentation benchmark gRefCOCO [27]. in Tab. 1. Based on the versatile and adaptable design of HyperSeg, our model achieves state-of-the-art performance on all the referring datasets. Specifically, HyperSeg surpasses the current SOTA by a large margin, reaching 79.7 cIoU on RefCOCO+ val (+6.8 over PSALM). Besides, Our model shows superiority in challenging G-RES tasks compared with previous **zero-shot** methods, demonstrating the robustness and generalization ability of HyperSeg.

Reasoning segmentation results. We compare HyperSeg with the state-of-the-art methods on image reasoning segmentation (ReasonSeg [22]) and reasoning video object segmentation (ReVOS [51]) in Tab. 2. Our HyperSeg achieves superior performance on reasoning tasks, significantly surpassing previous state-of-the-art methods (+12.1 on ReVOS-Reasoning), which shows HyperSeg powerful reasoning capability of tackling complex scenarios.

Generic image segmentation results. We show the performance of HyperSeg on COCO-Panoptic [26] and open-vocabulary segmentation [9, 14, 34, 60] tasks in Tab. 3. HyperSeg achieves excellent performance compared with both specialist models and VLLMs-based methods on both closed-set and open-vocabulary segmentation tasks. Specifically, HyperSeg surpasses the VLLM-based PSALM by a significant margin (+5.3 on COCO PQ, and +10.6 on mIoU), which demonstrates our powerful capabilities of handling complex semantic perception and segmentation tasks. Besides, we show the results of COCO-Interactive in the Supplementary Material.

Common video segmentation results. We compare HyperSeg with previous video segmentation methods in Tab. 4, including visual-prompted semi-supervised VOS (DAVIS17 val), text-prompted referring video object segmentation (Ref-YouTube-VOS, Ref-DAVIS17) and video instance segmentation (YouTube-VIS 2019). HyperSeg shows promising results over previous unified segmentation methods [24, 25]. Besides, HyperSeg performs more video perception tasks than previous VLLM-based models [3, 51].

4.2. Ablations

The mutual influence between different tasks. Our model can be trained and inferred across multiple tasks and datasets simultaneously. We evaluate the mutual impact of different tasks in Tab. 5. The results show that joint training can enhance the model performance compared with the task-specific model. Besides, the performance of video segmentation tasks can be improved significantly by adding the image training datasets. This demonstrates the generalization and self-consistency of our HyperSeg to perform universal segmentation.

Effect of different LLMs and vision backbone. In Tab. 6, we evaluate the effect of different sizes of LLMs and vision backbone. Our HyperSeg achieves excellent performance using smaller LLMs and vision encoder compared with the previous SOTA models like VISA[51] and PSALM[59]. Besides, the performance of HyperSeg can be further improved by using the more powerful LLM (Phi-2-2.7B [20]).

Ablation on the proposed components. We assess the effectiveness of our proposed FVP module and Hybrid Entity Recognition strategy. As shown in Tab. 7, with our fine-grained visual integration and hybrid entity semantic

Table 5. The mutual influence between different tasks. Task-specific means training task-specific models only on data from corresponding tasks, Refer+Reason denotes the model is trained on referring and reasoning segmentation data.

Task-specific	Refer+Reason	Video	Image	RefCOCO			COCO		ReVOS			YT-VIS
				val	testA	testB	PQ	mIoU	Reasoning	Referring	Overall	mAP
✓				83.8	85.9	82.2	60.8	75.1	51.2	56.6	53.9	50.7
	✓			83.3	84.9	80.9	-	-	53.1	57.3	55.2	-
			✓	85.6	86.1	82.4	60.9	76.5	-	-	-	-
		✓		-	-	-	-	-	51.1	57.0	54.1	50.4
	✓	✓	✓	84.8	85.7	83.4	61.2	77.2	53.0	58.5	55.7	53.8

Table 6. The comparison of different LLMs and backbone usages. w/o CLIP means without using CLIP vision encoder.

Method	LLM	COCO		ReVOS			ADE-OV	PC59-OV	PAS20-OV
		PQ	mIoU	Reasoning	Referring	Overall	mIoU	mIoU	mIoU
LISA [22]	Vicuna-7B	-	-	36.1	45.7	40.9	-	-	-
VISA [51]	Vicuna-13B	-	-	40.9	54.1	47.5	-	-	-
PSALM(w/o CLIP) [22]	Phi-1.5-1.3B	55.9	66.6	-	-	-	18.2	48.5	81.3
HyperSeg (w/o CLIP)	Phi-1.5-1.3B	61.1	76.0	44.0	49.7	46.9	18.9	60.0	90.6
HyperSeg	Phi-1.5-1.3B	60.9	76.7	50.8	57.0	53.9	20.3	61.5	90.8
HyperSeg	Phi-2-2.7B	**61.2**	**77.2**	**53.0**	**58.5**	**55.7**	**22.3**	**64.6**	**92.1**

Table 7. Ablation on the core components of HyperSeg. FVP and HER denote the proposed Fine-grained Visual Perceiver and Hybrid Entity Recognition modules.

FVP	HER	YT-VIS	COCO		RefCOCO
		mAP	PQ	mIoU	cIoU
		48.4	54.8	66.2	82.8
✓		50.8	55.8	66.6	84.6
	✓	52.0	59.7	74.6	84.3
✓	✓	**53.8**	**61.2**	**77.2**	**84.8**

Table 8. Ablation on the Fine-grained Visual Perceiver design. CW denotes the Conditional Weight illustrated in Sec. 3.3, and Scale denotes the total scale in the proposed FVP module.

CW	Scale	YT-VIS	COCO		RefCOCO
		mAP	PQ	mIoU	cIoU
	single-layer	49.7	55.8	68.0	83.7
	multi-layers	50.4	58.9	73.4	84.5
✓	multi-layers	**53.8**	**61.2**	**77.2**	**84.8**

enhancement, the segmentation accuracy can be enhanced significantly (+5.4 on YT-VIS, +6.4 on COCO panoptic PQ).

Design of the Fine-grained Visual Perceiver. In the FVP module, we combine multi-scale visual features into fixed perception queries using the condition-wise cross-attention layers to extract rich visual details from different scales of the pyramid encoder. As shown in Tab. 8, together with the conditional weight and the multi-scale design, our model makes a significant improvement on both image and video segmentation tasks.

Effect of temporal adapter. We evaluate the effectiveness of the proposed temporal adapter including global prompt aggregation (global) and local space-time information injection (local) in Tab. 9. Incorporating both global and local components, the temporal adapter significantly enhances model performance across multiple video segmentation tasks.

Table 9. Ablation on the temporal adapter for video tasks.

Global	Local	Ref-DAVIS17	ReVOS	YT-VIS
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	mAP
		67.3	54.1	47.9
✓		68.8	54.5	48.5
	✓	69.3	54.8	50.2
✓	✓	**71.2**	**55.7**	**53.8**

5. Conclusion

In this study, we aim to present HyperSeg, the first VLLM-based universal segmentation model designed for pixel-level image and video perception, encompassing a wide range of generic segmentation and complex reasoning tasks. We propose the Hybrid Entity Recognition and Fine-grained Visual Perceiver to leverage the recognition capacity of VLLMs more effectively and enhances the VLLM's ability by capturing diverse levels of visual information without incurring excessive computational costs. With additional Temporal Adapter, HyperSeg can tackle challenging video tasks by incorporating global and local information. HyperSeg surpasses existing methods on complex reasoning segmentation and traditional perception tasks. The insights presented in this work expand the possibilities of VLLMs in visual perception and lay a foundation for future research on the integration of vision-language models.

6. Acknowledgments

This work was partly supported by the National Key Research and Development Program of China (No. 2024YFB2808903) and the Shenzhen Science and Technology Program (JCYJ20220818101001004).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022. 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 3, 5, 1

[3] Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Zheng Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *arXiv preprint arXiv:2409.19603*, 2024. 7

[4] Sergi Caelles, Alberto Montes, Kevis-Kokitsi Maninis, Yuhua Chen, Luc Van Gool, Federico Perazzi, and Jordi Pont-Tuset. The 2018 davis challenge on video object segmentation. *arXiv preprint arXiv:1803.00557*, 2018. 6

[5] 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. 3, 1

[6] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2025. 6

[7] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 3, 4, 5, 6, 7

[8] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3, 1

[11] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021. 6, 3

[12] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2694–2703, 2023. 6

[13] Zheng Ding, Jieke Wang, and Zhuowen Tu. Open-vocabulary universal image segmentation with maskclip. *arXiv preprint arXiv:2208.08984*, 2022. 7

[14] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88:303–338, 2010. 7

[15] Xiuye Gu, Yin Cui, Jonathan Huang, Abdullah Rashwan, Xuan Yang, Xingyi Zhou, Golnaz Ghiasi, Weicheng Kuo, Huizhong Chen, Liang-Chieh Chen, et al. Dataseg: Taming a universal multi-dataset multi-task segmentation model. *Advances in Neural Information Processing Systems*, 36, 2024. 4, 7

[16] Cong Han, Yujie Zhong, Dengjie Li, Kai Han, and Lin Ma. Open-vocabulary semantic segmentation with decoupled one-pass network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1086–1096, 2023. 7

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[18] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 6

[19] Jitesh Jain, Jiachen Li, Mang Tik Chiu, Ali Hassani, Nikita Orlov, and Humphrey Shi. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2989–2998, 2023. 2, 7, 3

[20] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1:3, 2023. 7, 1

[21] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 6, 1, 2

[22] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. Lisa: Reasoning segmentation via large language model. *ArXiv*, abs/2308.00692, 2023. 2, 3, 4, 5, 6, 7, 8

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 2, 3, 5, 1

[24] Xiangtai Li, Haobo Yuan, Wei Li, Henghui Ding, Size Wu, Wenwei Zhang, Yining Li, Kai Chen, and Chen Change Loy. Omg-seg: Is one model good enough for all segmentation? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27948–27959, 2024. 2, 4, 7, 3

[25] Fangjian Lin, Jianlong Yuan, Sitong Wu, Fan Wang, and Zhibin Wang. Uninext: Exploring a unified architecture for vision recognition. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3200–3208, 2023. 2, 3, 4, 7

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014. 6, 7

[27] Chang Liu, Henghui Ding, and Xudong Jiang. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23592–23601, 2023. 7

[28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 3

[29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2, 3, 6, 1

[30] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023. 6, 3

[31] Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3459–3469, 2024. 4

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 1

[33] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 920–930, 2023. 6

[34] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 891–898, 2014. 7

[35] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 792–807. Springer, 2016. 6, 7

[36] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 3

[37] Renjie Pi, Lewei Yao, Jiahui Gao, Jipeng Zhang, and Tong Zhang. Perceptiongpt: Effectively fusing visual perception into llm. *arXiv preprint arXiv:2311.06612*, 2023. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[39] Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Erix Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. *arXiv preprint arXiv:2311.03356*, 2023. 3

[40] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13009–13018, 2024. 6

[41] Zhongwei Ren, Zhicheng Huang, Yunchao Wei, Yao Zhao, Dongmei Fu, Jiashi Feng, and Xiaojie Jin. Pixellm: Pixel reasoning with large multimodal model. *ArXiv*, abs/2312.02228, 2023. 2, 3, 5, 6

[42] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*, pages 208–223. Springer, 2020. 6

[43] Wenhai Wang, Zhe Chen, Xiaokang Chen, Jiannan Wu, Xizhou Zhu, Gang Zeng, Ping Luo, Tong Lu, Jie Zhou, Yu Qiao, et al. Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[44] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022. 6

[45] 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. 2, 4, 6, 3

[46] Dongming Wu, Tiancai Wang, Yuang Zhang, Xiangyu Zhang, and Jianbing Shen. Onlinerefer: A simple online baseline for referring video object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2761–2770, 2023. 7, 3

[47] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 6, 7, 3

[48] Junfeng Wu, Qihao Liu, Yi Jiang, Song Bai, Alan Yuille, and Xiang Bai. In defense of online models for video instance segmentation. In *ECCV*, 2022. 4

[49] Zhuofan Xia, Dongchen Han, Yizeng Han, Xuran Pan, Shiji Song, and Gao Huang. Gsva: Generalized segmentation via multimodal large language models. *arXiv preprint arXiv:2312.10103*, 2023. 2, 6, 3

[50] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022. 7

[51] 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. *arXiv preprint arXiv:2407.11325*, 2024. 2, 4, 6, 7, 8, 3

[52] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5188–5197, 2019. 6

[53] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022. 6, 3

[54] 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. 3

[55] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 6, 7

[56] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986, 2023. 1

[57] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1020–1031, 2023. 2, 3

[58] 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. *arXiv preprint arXiv:2406.19389*, 2024. 2, 4, 5, 6, 7, 3

[59] Zheng Zhang, Yeyao Ma, Enming Zhang, and Xiang Bai. Psalm: Pixelwise segmentation with large multi-modal model. *arXiv preprint arXiv:2403.14598*, 2024. 2, 3, 4, 5, 6, 7, 1

[60] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ade20k dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019. 7

[61] 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. 2, 3

[62] Jiawen Zhu, Zhi-Qi Cheng, Jun-Yan He, Chenyang Li, Bin Luo, Huchuan Lu, Yifeng Geng, and Xuansong Xie. Tracking with human-intent reasoning. *arXiv preprint arXiv:2312.17448*, 2023. 6

[63] Minjie Zhu, Yichen Zhu, Xin Liu, Ning Liu, Zhiyuan Xu, Chaomin Shen, Yaxin Peng, Zhicai Ou, Feifei Feng, and Jian Tang. A comprehensive overhaul of multimodal assistant with small language models. *arXiv preprint arXiv:2403.06199*, 2024. 6

[64] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 7, 1, 2, 3