VISA: Reasoning Video Object Segmentation via Large Language Models

Cilin Yan[§1], Haochen Wang[§2], Shilin Yan[3], Xiaolong Jiang[3], Yao Hu[3], Guoliang Kang[*1], Weidi Xie[4], and Efstratios Gavves[2]

[1] Beihang University
[2] University of Amsterdam
[3] Xiaohongshu Inc.
[4] Shanghai Jiao Tong University

Abstract. Existing Video Object Segmentation (VOS) relies on explicit user instructions, such as categories, masks, or short phrases, restricting their ability to perform complex video segmentation requiring reasoning with world knowledge. In this paper, we introduce a new task, Reasoning Video Object Segmentation (ReasonVOS). This task aims to generate a sequence of segmentation masks in response to implicit text queries that require complex reasoning abilities based on world knowledge and video contexts, which is crucial for structured environment understanding and object-centric interactions, pivotal in the development of embodied AI. To tackle ReasonVOS, we introduce VISA (Video-based large language Instructed Segmentation Assistant), to leverage the world knowledge reasoning capabilities of multi-modal LLMs while possessing the ability to segment and track objects in videos with a mask decoder. Moreover, we establish a comprehensive benchmark consisting of 35,074 instruction-mask sequence pairs from 1,042 diverse videos, which incorporates complex world knowledge reasoning into segmentation tasks for instruction-tuning and evaluation purposes of ReasonVOS models. Experiments conducted on 8 datasets demonstrate the effectiveness of VISA in tackling complex reasoning segmentation and vanilla referring segmentation in both video and image domains. The code and dataset are available at `https://github.com/cilinyan/VISA`.

Keywords: reasoning video object segmentation · video-based instructed segmentation assistant · dataset collection

1 Introduction

Existing video object segmentation relies on explicit queries, such as pre-defined categories [3, 41, 45], masks of certain frames [5, 47], or explicit short phrases describing intuitive features [2, 34, 44]. Such systems lack the capacity to reason and infer users' intentions based on implicit instructions. For instance, it is more intuitive for the users to give instructions like "Find my favorite cup" instead

*Corresponding author. § Equal contribution.

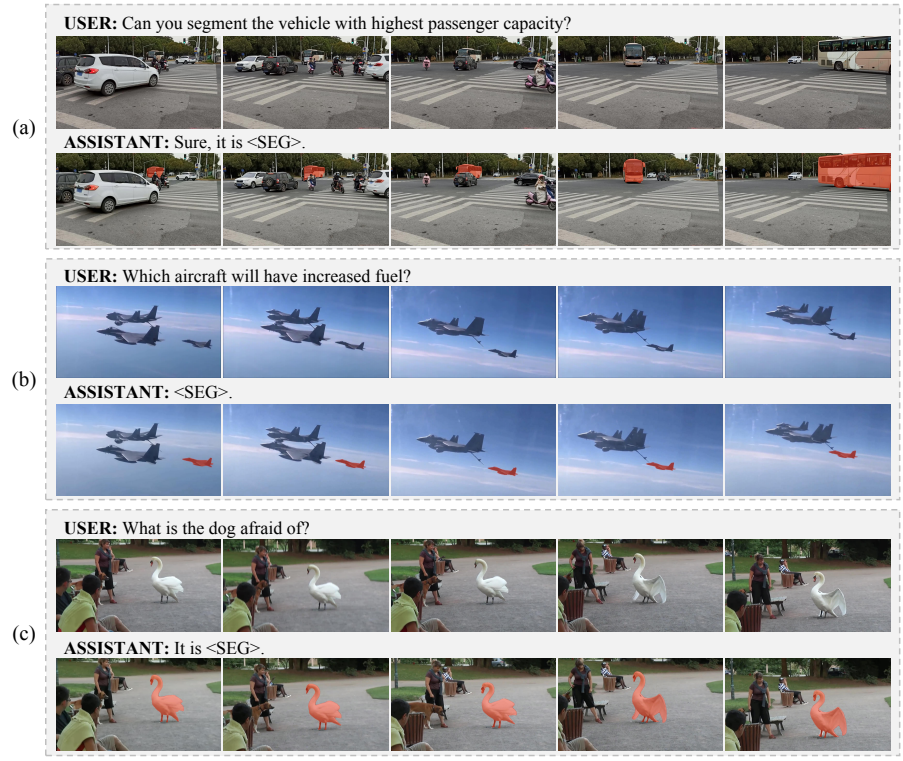

Fig. 1: We enable the reasoning video object segmentation capabilities for current multi-modal LLMs. The proposed VISA is capable of segmenting and tracking objects given text descriptions involving: (a) complex reasoning of world knowledge; (b) inference of upcoming events; and (c) comprehensive understanding of video content.

of "Find the red cup located second to the left on the table". To accomplish the first instruction, the model needs to understand that "favorite" means "most frequently used" to some degree, and callback the history temporal information to localize the cup.

In this work, we propose Reasoning Video Object Segmentation (ReasonVOS), which aims to generate a binary mask sequence given complex and implicit text instruction in videos. Notably, the text instruction is not limited to a straightforward reference (e.g., the running car), but a more complex description including reasoning of world knowledge (e.g., the car powered by electricity). This task requires the integration of reasoning ability with long-term video understanding to accurately localize target objects in videos, which is a crucial ability for Embodied AI systems that enable robots to fulfill effective interaction with objects in dynamic environments given user instructions.

To tackle Reasoning Segmentation in images, recent work LISA [16] leverages the language generation prowess of multi-modal LLMs, complemented by a mask decoder for generating segmentation results. However, the Reasoning Segmentation in videos demands temporal information for comprehensive video understanding and spatial details for producing high-quality segmentation mask sequences. Therefore, the multi-modal LLMs need to simultaneously process multiple frames with a substantial number of tokens for each frame. Considering the numerous visual tokens to be processed simultaneously, it is computationally intractable to directly broadcast an image reasoning segmentation model to the video domain.

To this end, we introduce VISA (Video-based large language Instructed Segmentation Assistant), designed to efficiently encode long-term video features while preserving spatial details to enable reasoning in video object segmentation. Specifically, we start by designing a Text-guided Frame Sampler (TFS) to select frames that are most relevant to the task based on textual instructions, focusing the model on the most significant moments for object identification. TFS reduces the requirement of visual token numbers to be handled, enabling VISA to process long-term videos. These selected frames, along with the text queries, are tokenized and processed concurrently by a multi-modal Large Language Model (LLM), enabling sophisticated reasoning over video content and facilitating the generation of precise textual outputs. To equip VISA with robust segmentation capabilities, we incorporate a special token <SEG> in the output text, inspired by the approach in LISA [16]. The hidden embedding of <SEG> is leveraged to produce segmentation masks of selected frames using a SAM [15] decoder. The segmentation process is completed by deriving the masks for the remaining frames with an object tracker [4]. As illustrated in Fig. 1, VISA demonstrates remarkable proficiency in handling complex segmentation tasks that require: (a) reasoning based on world knowledge; (b) inference of future events; and (c) a comprehensive understanding of video content.

To evaluate the effectiveness of the proposed VISA, we create a benchmark dataset named ReVOS. This dataset comprises 35,074 pairs of instruction-mask sequences derived from 1,042 diverse videos. In contrast to traditional referring video segmentation datasets, such as Ref-YouTube-VOS [34] and MeViS [8], which primarily contain explicit short phrases, ReVOS includes text instructions that necessitate a sophisticated understanding of both video content and general world knowledge. We carry out comprehensive experiments on the ReVOS dataset as well as on seven existing segmentation datasets. The results in Fig. 2 demonstrate that VISA not only facilitates advanced reasoning segmentation in both video and image domains but also achieves competitive performance on referring segmentation tasks.

Our main contributions could be summarized as follows: (i) We introduce a new task ReasonVOS (Reasoning Video Object Segmentation), which aims to segment and track objects in videos given implicit texts. ReasonVOS emphasizes the requirements of reasoning, summary, and inference ability based on video content and world knowledge, crucial for an intelligent perception sys-

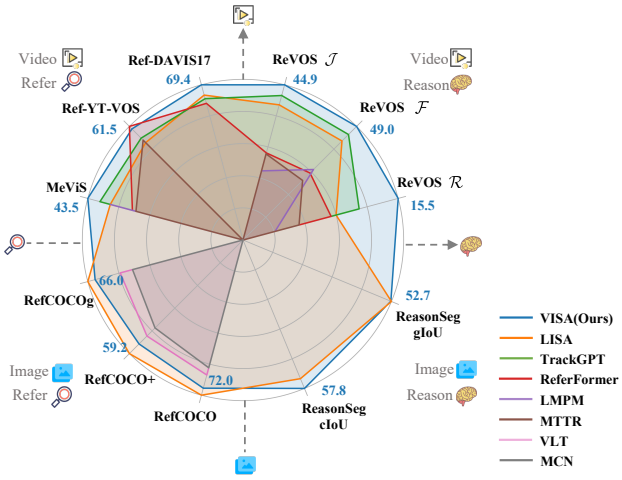

Fig. 2: Our proposed VISA consistently achieves state-of-the-art performances on video and image datasets over reasoning and referring segmentation tasks. $\mathcal{J}$ is region similarity [34], $\mathcal{F}$ is contour accuracy [34], and $\mathcal{R}$ is robustness score [19].

tem to interact with dynamic environments. (ii) We propose VISA (Video-based large language Instructed Segmentation Assistant), which efficiently integrates long-term video features and complex text queries to enable the reasoning video object segmentation ability. (iii) We collect a large-scale dataset ReVOS, comprising 1,042 videos and 35,074 object descriptions for instruction tuning and evaluation purposes of ReasonVOS models. The experiments on ReVOS and existing datasets show that our proposed VISA performs robustly in reasoning segmentation tasks of both image and video domains.

2 Related Work

Video Object Segmentation. Video Object Segmentation (VOS) is designed to segment and track objects in videos based on specific references, including categories [3, 41, 45, 52–55], segmentation masks [4, 6, 28, 30], or explicit text descriptions [2, 8, 21, 34, 44]. VOS plays a critical role in structured video representation learning and Embodied AI. Category-based VOS methods (or Video Instance Segmentation), such as Mask2Former [3], SeqFormer [45], and VisTR [41], segment and associate objects in videos given a pre-defined category list. Mask-based VOS methods (or semi-supervised VOS), such as STM [28] and XMem [4], segment and track objects in videos based on the segmentation mask given in certain frames. The utility of the aforementioned approaches is constrained by their reliance on structured and straightforward input, resulting in limited generalizability in real-world scenarios that necessitate complex reasoning and flexible input formulation.

In contrast, the text-based VOS (Referring VOS) [2, 8, 34], aims to segment objects in videos given text description. However, the text descriptions in Referring VOS fall into short phrases indicating the explicit object information, such as action, localization, and appearance. This system lacks the ability to handle complex sentences that involve common sense reasoning or inference based on video content. In this work, we introduce ReasonVOS, extending the short phrases to complex sentences requiring reasoning and the inference of world knowledge alongside video content. This advancement significantly enhances the practical utility of VOS across various tasks.

Multi-Modal Large Language Model. Inspired by the impressive reasoning capabilities of Large Language Models (LLMs), researchers are investigating methods to transpose these abilities into the vision domain, leading to the development of multi-modal LLMs [1, 38, 51]. Flamingo [1] utilizes a cross-attention structure to attend to visual contexts, facilitating visual in-context learning. Meanwhile, models like BLIP-2 [18] and mPLUG-OWL [48] propose the encoding of image features using a visual encoder, which are subsequently integrated into the LLM along with text embeddings. Otter [17] further incorporates robust few-shot capabilities through in-context instruction tuning on the proposed MIMIC-IT dataset. LLaVA [23] and MiniGPT-4 [56] first conduct image-text feature alignment followed by instruction tuning and also investigate image retrieval for LLMs.

Recent studies have delved into the confluence of multi-modal Large Language Models (LLMs) and vision tasks. VisionLLM [40] provides a versatile interface for engaging with various vision-centric tasks through instruction tuning but fails to fully exploit LLMs for complex reasoning. Kosmos-2 [29] builds a large-scale dataset of grounded image-text pairs, thereby injecting grounding capabilities into LLMs. DetGPT [31] connects the multi-modal LLMs and open-vocabulary detectors, facilitating detection tasks based on user instructions. GPT4RoI [50] innovates by incorporating spatial boxes as input and training the model on region-text pairings. LISA [16] efficiently enables segmentation capabilities of multi-modal LLMs in the image domain by introducing a special <SEG> token. All the above-mentioned methods focus on downstream tasks in the image domain. The concurrent work TrackGPT [35], made the first attempt to tackle reasoning segmentation in videos. However, TrackGPT processes single frames at one time and segments the objects frame-by-frame without any temporal correspondence, which falls in complex scenarios requiring long-term video understanding. On the contrary, our proposed VISA handles multiple frames at one time to obtain long-term awareness.

Video Multi-Modal Large Language Model. To support video understanding in LLMs, Video-LLaMA [49] attempts to utilize BLIP-2 for video embedding extraction, while Video-ChatGPT [25] proposes spatial and temporal pooling for video features. However, given the substantial number of tokens required for each frame, LLMs encounter significant challenges when processing extensive video sequences. It prevents previous work [25, 49] from representing long video sequences that exceed a duration of one hour in LLMs. To solve the issue,

LLaMA-VID [20] proposes to efficiently encode each frame with only 2 tokens, which supports long video understanding in existing LLMs. Those works either use pooling or projection to abstract each frame into a few visual tokens, which is inadequate to provide detailed spatial information for segmentation. In this work, we first select significant frames for identifying the target objects and simultaneously process the selected frames with a large number of visual tokens, avoiding the spatial pooling or projection and thus benefiting the segmentation tasks.

3 Method

3.1 Task Setting

In this section, we start by defining the task of interest, termed ReasonVOS. Specifically, given a high-level query text instruction $\mathbf{x}_t$ for which reasoning with world knowledge is required, and an input video $\mathbf{x}_v$, we aim to build a model $\varphi_\theta(\cdot)$ that outputs a binary mask sequence $\mathcal{M}$ representing the described object in the input video:

$$\mathcal{M} = \varphi_\theta(\mathbf{x}_t, \mathbf{x}_v), \tag{1}$$

where the input video $\mathbf{x}_v = \{f_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W\times 3}$ contains T frames, and each frame f_t has a size of $H \times W$. The output binary mask sequence $\mathcal{M} = \{m_t\}_{t=1}^{T} \in \mathbb{R}^{T\times H\times W}$ has the same frame number and size.

ReasonVOS shares a similar formulation of input and output with the Referring VOS [34] but is far more challenging. The principal difference stems from the complexity of the query text in ReasonVOS. Unlike simple phrases in Referring VOS that directly describe the appearance, action, or localization characteristics (e.g., *"the running car"*), the query texts in ReasonVOS involve more complex expressions that require world knowledge and common sense (e.g., *"the car powered by electricity"*) or expressions that require complex understanding and inference about video content and upcoming events (e.g., *"Which car is most likely to win the race?"*).

3.2 Architecture of VISA

Overview. As shown in Fig. 3, VISA consists of three main components, namely Text-guided Frame Sampler, multi-modal Large Language Model (LLM), and Object Tracker. Specifically, (a) the input video $\mathbf{x}_v$ is fed to the Text-guided Frame Sampler (TFS), which outputs a target frame f_{tgt} for segmentation and T_r corresponding reference frames $\mathbf{x}_r$ to gain long-term information, guided by the text instructions $\mathbf{x}_t$. (b) Then the selected frames are fed into the multi-modal LLM, generating text output including a special token <SEG> for segmentation of the target frame f_{tgt}. (c) Finally, an Object Tracker is utilized to generate the segmentation masks of all frames $\mathcal{M}$ via bi-directional mask propagation.

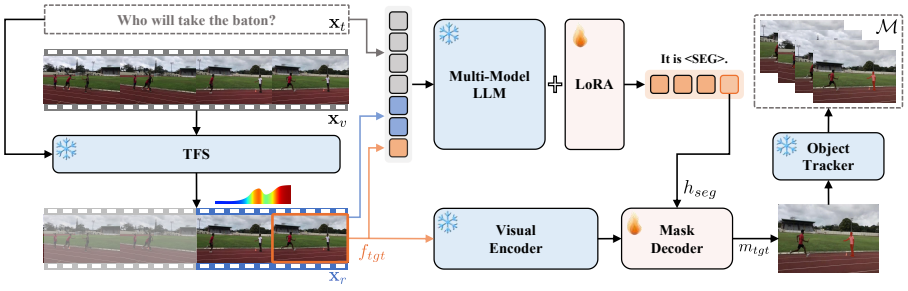

Fig. 3: Overview of VISA. (a) Given a video $\mathbf{x}_v$ and a text description $\mathbf{x}_t$, a Text-guided Frame Sampler (TFS) is proposed to sample the most distinguishing frame f_{tgt} as the target to be segmented and corresponding reference frames $\mathbf{x}_r$. (b) Then f_{tgt}, $\mathbf{x}_r$, and $\mathbf{x}_t$ are tokenized and fed to a Multi-Modal LLM to generate text output, including a special token <SEG>. The last-layer embedding of <SEG> token h_{seg} is then decoded into the segmentation mask m_{tgt} of frame f_{tgt} via the mask decoder. (c) Finally, the segmentation masks of all frames $\mathcal{M}$ are generated by propagation with an Object Tracker. The modules in blue are frozen during the training, while the modules in pink are trainable.

Text-guided Frame Sampler. Given input video $\mathbf{x}_v$ comprising T frames where each frame is represented by L visual tokens, the total number of tokens to be processed by the multi-modal LLM is $T \times L$. For segmentation purposes, L should be large enough to maintain spatial details, instead of pooling into a few tokens such as in Video-ChatGPT [25]. Consequently, it is computationally intractable to directly feed such numerous visual tokens to the multi-modal LLM.

As shown in Fig. 3, the text query *"Which person will take the baton?"* could be answered within the last few frames of the video, while the rest frames are irrelevant to the question. Inspired by this, we adopt LLaMA-VID [20], a multi-modal LLM that abstracts each input frame into two visual tokens and enables long video processes, to serve as a Text-guided Frame Sampler (TFS). TFS generates the most distinguishing frame f_{tgt} and corresponding reference frames $\mathbf{x}_r$ for identifying the described object. Specifically, a task-specific template is designed: *"<VIDEO> To find {description}, which percentage mark of the video should I check? Please respond with a number between 0% and 100%."* We extract the percentage values p_i in the top K responses and use the average value to obtain the target frame $f_{tgt} = f_{T/K \sum p_i}$. K is set to 10 in this work. Based on f_{tgt}, T_r frames are sampled as reference frames $\mathbf{x}_r$ to obtain long-term temporal correspondence and help with the segmentation of the described object in frame f_{tgt}. We adopt multiple reference sampling strategies in this work, such as Local sampling and Global sampling. The details and ablation studies of different reference sampling strategies are shown in Ablation Study Sec. 4.3.

Multi-Modal Large Language Model. Each frame in $\mathbf{x}_r$ and f_{tgt} are encoded via ViT [12] and tokenized into L visual embeddings by Spatial Merging [13], yielding visual tokens $<\mathbf{x}_r>$ and $<f_{tgt}>$. Then, the concatenated visual and text

tokens are fed to a Multi-Modal LLM to generate the text output containing a special token <SEG>. The task-specific template is designed as: *"USER: <f_{tgt}> <$\mathbf{x}_r$> Can you segment the {description}? ASSISTANT: Yes, it is <SEG>."*, where {description} will be replaced by the text description, and the text will be tokenized before being fed to multi-modal LLMs. We extract the last-layer embedding corresponding to the <SEG> token and apply an MLP projection layer to generate h_{seg}, which serves as the prompt embedding in SAM decoder [15].

Simultaneously, the vision backbone $\mathcal{E}_{\text{v}}$ extracts the visual features of target frame f_{tgt}, which is utilized along with the prompt embedding h_{seg} to produce the segmentation mask m_{tgt}:

$$m_{tgt} = \text{SAM}(\mathcal{E}_{\text{v}}(f_{tgt}), h_{seg}). \tag{2}$$

Finally, an Object Tracking method [4] is adopted to propagate m_{tgt} bidirectionally to all rest frames and obtain the mask sequence $\mathcal{M}$:

$$\mathcal{M} = \{m_t\}_{t=1}^{T} = \text{OT}(m_{tgt}, \mathbf{x}_v). \tag{3}$$

Training. Following LISA [16], our model is trained end-to-end using the standard text generation loss $\mathcal{L}_{\text{txt}}$ and the segmentation mask loss $\mathcal{L}_{\text{mask}}$. The overall objective $\mathcal{L}$ is the weighted sum of $\mathcal{L}_{\text{txt}}$ and $\mathcal{L}_{\text{mask}}$:

$$\mathcal{L} = \lambda_{\text{txt}}\mathcal{L}_{\text{txt}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}. \tag{4}$$

Specifically, $\mathcal{L}_{\text{txt}}$ is the auto-regressive cross-entropy loss for text generation, and $\mathcal{L}_{\text{mask}}$ is the combination of per-pixel binary cross-entropy (BCE) loss and DICE loss [27], with corresponding loss weights λ_{bce} and λ_{dice}. Given the ground-truth targets ($\hat{\mathbf{y}}_{\text{txt}}$, $\hat{m}_{tgt}$) and the predictions ($\mathbf{y}_{\text{txt}}$, m_{tgt}), $\mathcal{L}_{\text{txt}}$ and $\mathcal{L}_{\text{mask}}$ can be formulated as:

$$\mathcal{L}_{\text{txt}} = \text{CE}(\hat{\mathbf{y}}_{\text{txt}}, \mathbf{y}_{\text{txt}}), \mathcal{L}_{\text{mask}} = \lambda_{\text{bce}}\text{BCE}(\hat{m}_{tgt}, m_{tgt}) + \lambda_{\text{dice}}\text{DICE}(\hat{m}_{tgt}, m_{tgt}). \tag{5}$$

3.3 ReVOS Dataset

For the quantitative evaluation of ReasonVOS, it is essential to establish a benchmark characterized by implicit object descriptions and high-quality mask sequences. To this end, we collect ReVOS, a dataset containing complex text instructions and corresponding high-quality masks in videos for both instruction tuning and evaluation of ReasonVOS. To guarantee reliable assessment, we collect a diverse set of videos from LV-VIS [36, 37], MOSE [9], OVIS [33], TAO [7] and UVO [39]. Subsequently, we annotate the objects in the videos with complex text instructions and match these instructions with the corresponding target mask sequences.

Dataset Statistics Overall, our dataset comprises a total of 35,074 object-instruction pairs from 1,042 videos. All the videos are divided into a training (instruction tuning) set and a validation set containing 626 videos and 416 videos respectively. The text instructions consist of (1) 14,678 implicit descriptions requiring world knowledge and video content reasoning and inference to evaluate the

ReasonVOS; (2) 20,071 explicit descriptions to evaluate the generalization ability in traditional Referring VOS task; (3) 325 descriptions of nonexistent objects for the hallucination evaluation. Check `https://github.com/cilinyan/ReVOS-api` for more detailed dataset information.

Evaluation Metrics We follow most previous works on Referring VOS [8, 34] to adopt $\mathcal{J}\&\mathcal{F}$ as the main evaluation metric, which is the average of region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$. As for the evaluation of hallucination, we adopt the robustness score $\mathcal{R}$ introduced in R2VOS [19].

4 Experiments

4.1 Dataset

Training Dataset. Our training data consists of vanilla Referring VOS datasets, Video Question-Answering datasets, Image datasets, and the ReVOS dataset. The details are as follows: (1) Referring VOS datasets. We use Ref-YouTube-VOS [34], MeViS [8], and Ref-DAVIS17 [32] during training to learn the projections between objects in videos and text expressions. Those Referring VOS datasets provide input videos, explicit short descriptions, and corresponding object masks. (2) Video Question-Answering datasets. To achieve better reasoning and question-answering ability in videos of the multi-modal LLM, we include the video instruction data from Video-ChatGPT [25]. The answer template "It's <SEG>" is replaced by the original annotated answers in those datasets, and the corresponding segmentation loss is ignored during training. (3) Image datasets. Images could be regarded as one-frame videos. Thereby, we adopt all the vanilla datasets used by LISA [16] in our work to achieve more stable training. (4) ReVOS dataset. The above-mentioned training datasets contain no ReasonVOS samples. Therefore, we include the ReVOS dataset during training to achieve more comprehensive reasoning and object segmentation ability in videos. The implementation details are shown in Supplementary Material.

Evaluation Dataset We evaluate VISA on both Video datasets and Image datasets. (1) Video datasets. We use the ReVOS dataset to evaluate the performance of ReasonVOS; we use Ref-YouTube-VOS [34], MeViS [8], and Ref-DAVIS17 [32] to evaluate the performance of vanilla Referring VOS. (2) Image datasets. We use ReasonSeg [16], refCOCO [14], refCOCO+ [14], and refCOCOg [26] to evaluate the generalization ability of VISA on image-level segmentation tasks.

4.2 Comparison

ReVOS. The results comparison on ReVOS are shown in Tab. 1. Compared with traditional methods (even with extremely large visual backbones), our proposed VISA(IT)-7B generally achieves over 20 $\mathcal{J}\&\mathcal{F}$ improvements in terms of reasoning. Those traditional works are limited to short explicit references and have no capability of reasoning and understanding the implicit text queries.

Fig. 4: Visualizations of VISA on ReVOS dataset.

VISA(IT)-7B outperforms the single frame method LISA-7B [16] by 6.0 $\mathcal{J}\&\mathcal{F}$ in terms of overall performance, which indicates the ability of VISA to conduct video-level segmentation. Recent work TrackGPT [35] incorporates tracking with LLMs, yet the multi-modal LLMs in TrackGPT only process a single frame at one time, leading to poor temporal information gathering. As a consequence, VISA(IT)-7B outperforms TrackGPT(IT)-7B by 3.3 $\mathcal{J}\&\mathcal{F}$ overall. Moreover, as we include plenty of negative samples (text queries of nonexistent objects) in the ReVOS training set, the hallucination of VITA is much lower than in existing methods. As shown, the robustness scores $\mathcal{R}$ of VISA are much higher than existing methods.

VQA+RerferringVOS could serve as a baseline model, but can not solve ReasonVOS well. As suggested in Tab. 1, we use LLaMA-VID, a Video-VQA method, to transfer the complex questions (e.g., scared dog) into low-level descriptions (e.g., dog on left), and then employ LMPM, a RerferringVOS method, to segment the described objects. As shown below, LLaMA-VID + LMPM performs worse than LMPM on ReVOS reasoning set. That is because the existing video-VQA methods take only a few visual tokens per frame, which is too vague to localize the described objects and brings mistakes when converting complex questions into low-level expressions.

Note that VISA with LLaVA-7B [23] and Chat-UniVi-7B [13] achieve similar performance. Chat-UniVi could process a flexible number of visual tokens via

Table 1: Performance comparison on ReVOS dataset. $*$ means the method is reproduced in this work. (IT) means instruction tuning with the ReVOS training set. $\mathcal{R}$ is the robustness score.

Method	Backbone	referring			reasoning			overall			$\mathcal{R}$
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	
ReferFormer [44]	Resnet50	16.6	17.1	16.9	11.9	13.8	12.8	14.3	15.4	14.9	4.9
MTTR [2]	Video-Swin-T	29.8	30.2	30.0	20.4	21.5	21.0	25.1	25.9	25.5	5.6
LMPM [8]	Swin-T	29.0	39.1	34.1	13.3	24.3	18.8	21.2	31.7	26.4	3.2
ReferFormer [44]	Video-Swin-B	31.2	34.3	32.7	21.3	25.6	23.4	26.2	29.9	28.1	8.8
LLaMA-VID [20]+LMPM	Swin-T	29.0	39.1	34.1	12.8	23.7	18.2	20.9	31.4	26.1	3.4
LISA [16]	LLaVA-7B	44.3	47.1	45.7	33.8	38.4	36.1	39.1	42.7	40.9	9.3
LISA* [16]	LLaVA-13B	45.2	47.9	46.6	34.3	39.1	36.7	39.8	43.5	41.6	8.6
TrackGPT(IT)* [35]	LLaVA-7B	46.7	49.7	48.2	36.8	41.2	39.0	41.8	45.5	43.6	11.6
TrackGPT(IT)* [35]	LLaVA-13B	48.3	50.6	49.5	38.1	42.9	40.5	43.2	46.8	45.0	12.8
VISA	Chat-UniVi-7B	51.1	54.7	52.9	36.7	41.7	39.2	43.9	48.2	46.1	7.9
VISA	Chat-UniVi-13B	52.3	55.8	54.1	38.3	43.5	40.9	45.3	49.7	47.5	8.3
VISA(IT)	LLaVA-7B	49.4	52.6	51.0	40.5	45.8	43.2	44.9	49.2	47.1	15.3
VISA(IT)	LLaVA-13B	**55.7**	59.0	**57.4**	41.9	46.5	44.2	**48.8**	52.8	50.8	15.1
VISA(IT)	Chat-UniVi-7B	49.2	52.6	50.9	40.6	45.4	43.0	44.9	49.0	46.9	**15.5**
VISA(IT)	Chat-UniVi-13B	55.6	**59.1**	**57.4**	**42.0**	**46.7**	**44.3**	**48.8**	**52.9**	**50.9**	14.5

Table 2: Performance comparison on Referring VOS datasets. The results on MeViS above the horizontal line are provided in LMPM [8], which are all obtained with the Swin-T backbone. The results of TrackGPT on MeVIS are generated by our reproduced model.

Methods	Backbone	MeViS			Ref-YT-VOS			Ref-DAVIS17		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
URVOS [34]	ResNet50	25.7	29.9	27.8	45.3	49.2	47.2	47.3	56.0	51.6
LBDT [11]	ResNet50	27.8	30.8	29.3	48.2	50.6	49.4	-	-	54.1
MTTR [2]	Video-Swin-T	28.8	31.2	30.0	54.0	56.6	55.3	-	-	-
ReferFormer [44]	Video-Swin-B	29.8	32.2	31.0	61.3	64.6	62.9	58.1	64.1	61.1
LMPM [8]	Swin-T	34.2	40.2	37.2	-	-	-	-	-	-
OnlineRefer [43]	Swin-L	-	-	-	**61.6**	**65.5**	**63.5**	61.6	67.7	64.8
LISA [16]	LLaVA-7B	35.1	39.4	37.2	53.4	54.3	53.9	62.2	67.3	64.8
LISA [16]	LLaVA-13B	35.8	40.0	37.9	54.0	54.8	54.4	63.2	68.8	66.0
TrackGPT [35]	LLaVA-7B	37.6	42.6	40.1	55.3	57.4	56.4	59.4	67.0	63.2
TrackGPT [35]	LLaVA-13B	39.2	43.1	41.2	58.1	60.8	59.5	62.7	70.4	66.5
VISA (Ours)	Chat-UniVi-7B	40.7	46.3	43.5	59.8	63.2	61.5	66.3	72.5	69.4
VISA (Ours)	Chat-UniVi-13B	**41.8**	**47.1**	**44.5**	61.4	64.7	63.0	**67.0**	**73.8**	**70.4**

spatial merging, thus we chose it in this work. We visualize the results of VISA on the ReVOS dataset in Fig. 4.

Referring VOS. To demonstrate that VISA generalizes well in the vanilla Referring VOS task, we compare VISA with the existing methods in Tab. 2. As

Table 3: Performance comparison on Image Segmentation datasets. † denotes the results obtained from the model we trained via LISA's official GitHub repository.

Methods	Backbone	refCOCO			refCOCO+			refCOCOg		ReasonSeg	
		val	testA	testB	val	testA	testB	val(U)	test(U)	gIoU	cIoU
MCN [24]	Darknet53	62.4	64.2	59.7	50.6	55.0	44.7	49.2	49.4	-	-
VLT [10]	Darknet53	65.7	68.3	62.7	55.5	59.2	49.4	53.0	56.7	-	-
CRIS [42]	ResNet101	70.5	73.2	66.1	62.3	68.1	53.7	59.9	60.4	-	-
LAVT [46]	Swin-B	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	-	-
ReLA [22]	Swin-B	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	-	-
X-Decoder [57]	DaViT-L	-	-	-	-	-	-	64.6	-	22.6	17.9
SEEM [58]	DaViT-L	-	-	-	-	-	-	65.7	-	25.5	21.2
LISA [16]	LLaVA-7B	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	52.9	54.0
LISA†	LLaVA-7B	70.8	73.7	66.3	58.1	63.2	51.2	63.8	64.8	48.1	53.7
VISA (Ours)	Chat-UniVi-7B	72.4	75.5	68.1	59.8	64.8	53.1	65.5	66.4	52.7	57.8

Table 4: Performance on ReVOS validation set with different training datasets. The columns with ✓ mean the corresponding datasets are adopted during training.

ReferringVOS	VQA	Image	ReVOS	referring			reasoning		
				$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
	✓	✓	✓	45.9	49.3	47.6	37.4	42.3	39.9
✓		✓	✓	48.6	52.1	50.3	38.9	43.9	41.4
✓	✓		✓	32.3	36.2	34.2	30.9	36.1	33.5
✓	✓	✓		51.1	54.7	52.9	36.7	41.7	39.2
✓	✓	✓	✓	49.2	52.6	50.9	40.6	45.4	43.0

shown, VISA achieves the SOTA results over three widely used Referring VOS datasets.

Image Datasets. Images could be regarded as single-frame videos. Therefore, VISA could be directly applied to image datasets without any modification. As shown in Tab. 3, VISA achieves comparable performance with LISA on three referring image segmentation datasets, while significantly outperforming traditional methods by over 20% on the ReasonSeg [16] dataset. The results indicate that VISA has a strong generalization ability on vanilla referring image segmentation and reasoning image segmentation.

4.3 Ablation Studies

Training Datasets In Tab. 4, we show the contribution of each type of dataset during training. As shown, without Referring VOS datasets, the performance drops by 3.3% $\mathcal{J}\&\mathcal{F}$ and 3.1% $\mathcal{J}\&\mathcal{F}$ in terms of referring and reasoning segmentation, respectively. That is because Referring VOS datasets provide text expressions and mask sequence pairs in videos, aligning the video domain and linguistic domain, thus generally helping with the vanilla referring segmenta-

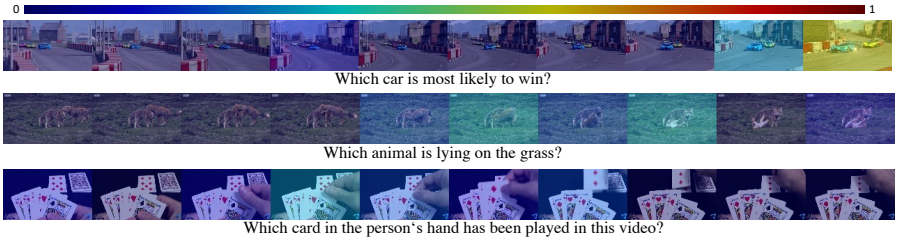

Fig. 5: Heatmaps of the target frame f_{tgt}. To draw the heatmap, we generate 10 responses with the Text-guided Frame Sampler (TFS) and obtain the normalized distribution. As shown, the highlighted frames are related to the text queries.

tion and reasoning segmentation tasks in videos. Without Image datasets, the performance of VISA significantly drops by 16.7% and 9.5%. Generally, image datasets have much larger scales than video datasets, leading to more robust feature alignment and stronger generalization ability. The models tend to be overfitting during training without image datasets. By instruction tuning on ReVOS, VISA further gains 3.8% $\mathcal{J}\&\mathcal{F}$ performance improvements of reasoning segmentation, while the performance of referring segmentation barely changes, which shows the effectiveness of our collected ReVOS dataset to improve the complex video reasoning ability.

Table 5: Overall $\mathcal{J}\&\mathcal{F}$ on ReVOS with different number T_r of reference frames $\mathbf{x}_r$ and different sampling strategies.

	$\lvert T_r \rvert$	w/o Sample	Global	Local	Global-Local
f_0	0	42.6	-	-	-
	6	-	43.9	44.5	44.6
	12	-	44.5	44.9	45.0
f_{tgt}	0	44.3	-	-	-
	6	-	46.0	46.1	46.3
	12	-	46.7	46.3	46.9

Table 6: The performance comparison on ReVOS with different number L of visual tokens per frame.

L	backbone	referring			reasoning		
		$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
256	LLaVA-7B	49.4	52.6	51.0	40.5	45.8	43.2
112	Chat-UniVi-7B	49.2	52.6	50.9	40.6	45.4	43.0
56	Chat-UniVi-7B	44.9	48.5	46.7	36.5	41.5	39.0

Target Frame f_{tgt}. We visualize the heatmaps of the target frame f_{tgt} in Fig. 5. As shown in the figure, the frames related to given text expressions are highlighted. We show the ablation of f_{tgt} in Tab. 5. f_0 means we directly segment the object in the first frame and propagate to the rest frames, while f_{tgt} means we use TFS to obtain the target frame f_{tgt} and segment f_{tgt} in consequence. As shown, the performance with f_{tgt} generally outperforms f_0 by around 2% under different settings, which indicates the effectiveness of TFS in obtaining the significant moments related to described objects.

Reference Sampling Strategies. In Tab. 5, Global means we uniformly sample frames through the whole video as reference frames $\mathbf{x}_r$, while Local means we sample contiguous frames centered by f_{tgt} as reference frames. Global-Local means the combination of $T_r/2$ frames from Global and $T_r/2$ frames from Local. As shown, Global-Local sampling slightly outperforms the separate ones, thus

we adopt it in VISA. As the number T_r of reference frames $\mathbf{x}_r$ increases, the performance gradually improves. To keep feasible training and inference, we adopt T_r=12 in VISA. Overall, with f_{tgt} and Global-Local sampling, VISA achieves 4.3% $\mathcal{J}\&\mathcal{F}$ improvements on the ReVOS dataset.

Number L of visual tokens. The performance comparison under different numbers L of visual tokens per frame is shown in Tab 6. For L=256, we adopt LLaVA-7B [23] as the backbone, which takes 256 visual tokens for the input image. For L=112 and L=52, we use Chat-UniVi-7B [13] as the backbone, and utilize the Spatial Merging [13] to project visual tokens to corresponding numbers. As shown, VISA with 256 tokens and 112 tokens per frame achieves comparable performance on ReVOS. When L is set to 52, the performance of VISA significantly drops. Therefore, we adopt L=112 in this work.

4.4 Limitations

Small Objects Limited by the number of visual tokens per frame (for instance, 256 in LISA [23], and 112 in VISA), the current methods have a poor ability to capture very small objects. As shown in Fig. 4 (d), the small paddles are not segmented. A multi-modal LLM with more input visual tokens could relieve this issue, but will lead to more computational burden and complex training process.

Temporal Information Gathering In this work, we intuitively adopt a Text-guided Frame Sampler to select a feasible number of important frames for the multi-modal LLM. The performance highly relies on the accuracy of located frames. Some objects may only appear in a few frames, which is hard to locate. As shown in Fig. 4 (e), the person with a fire tank only appears in one frame, while VISA falls to locate this frame and segment another person in consequence. Moreover, the text description could require extremely long temporal correspondence, but VISA could only handle a few selected frames at the same time. To this end, a more effective way to gather long-term temporal information while maintaining spatial details is required. We leave those issues to our future work.

5 Conclusion

In this work, we propose a new task, ReasonVOS, which aims to generate object mask sequences in response to text queries that require complex reasoning and inference abilities within video contexts. To tackle ReasonVOS, we design VISA (Video-based large language Instructed Segmentation Assistant), to leverage the world knowledge and reasoning capabilities of multi-modal LLMs while possessing the ability to segment and track objects in videos. Moreover, we collect a large-scale dataset ReVOS, containing 35,074 expression-mask pairs from 1,042 videos for the instruction tuning and evaluation of ReasonVOS methods. Experiments on eight various datasets show that our proposed VISA not only enables the reasoning segmentation ability in videos but also generally provides SOTA performance on traditional video and image segmentation tasks.

Acknowledgements

This project is supported by National Natural Science Foundation of China under Grant 92370114 and European Union (ERC, EVA, 950086).

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in Neural Information Processing Systems **35**, 23716–23736 (2022)
2. Botach, A., Zheltonozhskii, E., Baskin, C.: End-to-end referring video object segmentation with multimodal transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4985–4995 (2022)
3. Cheng, B., Choudhuri, A., Misra, I., Kirillov, A., Girdhar, R., Schwing, A.G.: Mask2former for video instance segmentation. arXiv preprint arXiv:2112.10764 (2021)
4. Cheng, H.K., Schwing, A.G.: Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In: European Conference on Computer Vision. pp. 640–658. Springer (2022)
5. Cheng, H.K., Tai, Y.W., Tang, C.K.: Rethinking space-time networks with improved memory coverage for efficient video object segmentation. Advances in Neural Information Processing Systems **34**, 11781–11794 (2021)
6. Cheng, J., Tsai, Y.H., Hung, W.C., Wang, S., Yang, M.H.: Fast and accurate online video object segmentation via tracking parts. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7415–7424 (2018)
7. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 436–454. Springer (2020)
8. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: Mevis: A large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2694–2703 (2023)
9. Ding, H., Liu, C., He, S., Jiang, X., Torr, P.H., Bai, S.: Mose: A new dataset for video object segmentation in complex scenes. arXiv preprint arXiv:2302.01872 (2023)
10. Ding, H., Liu, C., Wang, S., Jiang, X.: Vision-language transformer and query generation for referring segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16321–16330 (2021)
11. Ding, Z., Hui, T., Huang, J., Wei, X., Han, J., Liu, S.: Language-bridged spatial-temporal interaction for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4964–4973 (2022)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Jin, P., Takanobu, R., Zhang, C., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. arXiv preprint arXiv:2311.08046 (2023)

14. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 787–798 (2014)
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
16. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model. arXiv preprint arXiv:2308.00692 (2023)
17. Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z.: Otter: A multi-modal model with in-context instruction tuning. arXiv preprint arXiv:2305.03726 (2023)
18. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. arXiv preprint arXiv:2301.12597 (2023)
19. Li, X., Wang, J., Xu, X., Li, X., Lu, Y., Raj, B.: Rˆ 2vos: Robust referring video object segmentation via relational multimodal cycle consistency. arXiv preprint arXiv:2207.01203 (2022)
20. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. arXiv preprint arXiv:2311.17043 (2023)
21. Li, Z., Tao, R., Gavves, E., Snoek, C.G.M., Smeulders, A.W.M.: Tracking by natural language specification. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7350–7358 (2017). https://doi.org/10.1109/CVPR.2017.777
22. Liu, C., Ding, H., Jiang, X.: Gres: Generalized referring expression segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23592–23601 (2023)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
24. Luo, G., Zhou, Y., Sun, X., Cao, L., Wu, C., Deng, C., Ji, R.: Multi-task collaborative network for joint referring expression comprehension and segmentation. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 10034–10043 (2020)
25. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)
26. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 11–20 (2016)
27. Milletari, F., Navab, N., Ahmadi, S.A.: V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: 2016 fourth international conference on 3D vision. pp. 565–571. IEEE (2016)
28. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9226–9235 (2019)
29. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
30. Perazzi, F., Khoreva, A., Benenson, R., Schiele, B., Sorkine-Hornung, A.: Learning video object segmentation from static images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2663–2672 (2017)

31. Pi, R., Gao, J., Diao, S., Pan, R., Dong, H., Zhang, J., Yao, L., Han, J., Xu, H., Zhang, L.K.T.: Detgpt: Detect what you need via reasoning. arXiv preprint arXiv:2305.14167 (2023)
32. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
33. Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X., Belongie, S., Yuille, A., Torr, P.H., Bai, S.: Occluded video instance segmentation: Dataset and iccv 2021 challenge. arXiv preprint arXiv:2111.07950 (2021)
34. Seo, S., Lee, J.Y., Han, B.: 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. pp. 208–223. Springer (2020)
35. Stroh, N.: Trackgpt–a generative pre-trained transformer for cross-domain entity trajectory forecasting. arXiv preprint arXiv:2402.00066 (2024)
36. Wang, H., Yan, C., Chen, K., Jiang, X., Tang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: Ov-vis: Open-vocabulary video instance segmentation. International Journal of Computer Vision pp. 1–18 (2024)
37. Wang, H., Yan, C., Wang, S., Jiang, X., Tang, X., Hu, Y., Xie, W., Gavves, E.: Towards open-vocabulary video instance segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4057–4066 (2023)
38. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. arXiv preprint arXiv:2311.03079 (2023)
39. Wang, W., Feiszli, M., Wang, H., Tran, D.: Unidentified video objects: A benchmark for dense, open-world segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10776–10785 (2021)
40. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., et al.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks. Advances in Neural Information Processing Systems **36** (2024)
41. Wang, Y., Xu, Z., Wang, X., Shen, C., Cheng, B., Shen, H., Xia, H.: End-to-end video instance segmentation with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8741–8750 (2021)
42. Wang, Z., Lu, Y., Li, Q., Tao, X., Guo, Y., Gong, M., Liu, T.: Cris: Clip-driven referring image segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11686–11695 (2022)
43. Wu, D., Wang, T., Zhang, Y., Zhang, X., Shen, J.: Onlinerefer: A simple online baseline for referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2761–2770 (2023)
44. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4974–4984 (2022)
45. Wu, J., Jiang, Y., Zhang, W., Bai, X., Bai, S.: Seqformer: a frustratingly simple model for video instance segmentation. arXiv preprint arXiv:2112.08275 **1**(2), 6 (2021)
46. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: Lavt: Language-aware vision transformer for referring image segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18155–18165 (2022)

47. Yang, Z., Wei, Y., Yang, Y.: Associating objects with transformers for video object segmentation. Advances in Neural Information Processing Systems **34**, 2491–2502 (2021)
48. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
49. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023)
50. Zhang, S., Sun, P., Chen, S., Xiao, M., Shao, W., Zhang, W., Chen, K., Luo, P.: Gpt4roi: Instruction tuning large language model on region-of-interest. arXiv preprint arXiv:2307.03601 (2023)
51. Zhang, T., Li, X., Fei, H., Yuan, H., Wu, S., Ji, S., Loy, C.C., Yan, S.: Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. arXiv preprint arXiv:2406.19389 (2024)
52. Zhang, T., Tian, X., Wu, Y., Ji, S., Wang, X., Zhang, Y., Wan, P.: Dvis: Decoupled video instance segmentation framework. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1282–1291 (2023)
53. Zhang, T., Tian, X., Zhou, Y., Ji, S., Wang, X., Tao, X., Zhang, Y., Wan, P., Wang, Z., Wu, Y.: Dvis++: Improved decoupled framework for universal video segmentation. arXiv preprint arXiv:2312.13305 (2023)
54. Zhang, T., Tian, X., Zhou, Y., Wu, Y., Ji, S., Yan, C., Wang, X., Tao, X., Zhang, Y., Wan, P.: 1st place solution for the 5th lsvos challenge: Video instance segmentation. arXiv preprint arXiv:2308.14392 (2023)
55. Zhou, Y., Zhang, T., Ji, S., Yan, S., Li, X.: Dvis-daq: Improving video segmentation via dynamic anchor queries. arXiv preprint arXiv:2404.00086 (2024)
56. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)
57. Zou, X., Dou, Z.Y., Yang, J., Gan, Z., Li, L., Li, C., Dai, X., Behl, H., Wang, J., Yuan, L., et al.: Generalized decoding for pixel, image, and language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15116–15127 (2023)
58. Zou, X., Yang, J., Zhang, H., Li, F., Li, L., Wang, J., Wang, L., Gao, J., Lee, Y.J.: Segment everything everywhere all at once. Advances in Neural Information Processing Systems **36** (2024)