FeVOS: Foresight Expression Video Object Segmentation

Kehan Lan, Kaining Ying, and Henghui Ding ✉

Institute of Big Data, College of Computer Science and Artificial Intelligence, Fudan University, China
{khlan24, knying24}@m.fudan.edu.cn, hhding@fudan.edu.cn
https://henghuiding.com/FeVOS/

Abstract. Existing Referring Video Object Segmentation tasks focus on referring expressions describing events, actions or appearances of relevant objects within the observed frames, lacking evaluation in scenarios that require pre-decisive spatio-temporal reasoning, thereby limiting their applicability. To address this, we propose **Foresight Expression Video Object Segmentation**, a task that queries future events in upcoming video segments and requires masks of the objects in the observed frames as visual answers. For example, in ego-centric scenes, the question *"What tool will be used?"* demands reasoning over spatio-temporal cues to predict the masks of the next tool to be used, which helps with the understanding of future actions and decisions. To support this task, we introduce **FeVOS**, a dataset with 968 video clips, 14,525 foresight expressions, and 2,904 chain-of-thought annotations to provide explicit and interpretable reasoning steps. We further develop **FeVOS-R1**, an MLLM-based model trained on our dataset via a two-stage pipeline of supervised fine-tuning and reinforcement learning. FeVOS-R1 not only achieves state-of-the-art performance on FeVOS, but also demonstrates strong generalization to existing RVOS benchmarks. We hope this work can inspire more research on predictive reasoning in video perception.

Keywords: Referring Video Object Segmentation · Foresight Expression · Multimodal Large Language Model

1 Introduction

Referring Video Object Segmentation (RVOS) [3, 7, 8, 20, 31, 42, 43] is a challenging task that requires vision-language understanding with pixel-level grounding. Given a video clip and a referring expression, the model needs to segment the object corresponding to the referring expression throughout the entire video. It has shown great potential in fields like video editing [37], autonomous driving [23], and language-guided robotic planning [17, 34, 48]. However, existing RVOS tasks and datasets focus solely on grounding expressions within observed frames, lacking the capability to anticipate future events, which is a critical requirement for proactive decision-making in real-world applications.

✉ Corresponding author

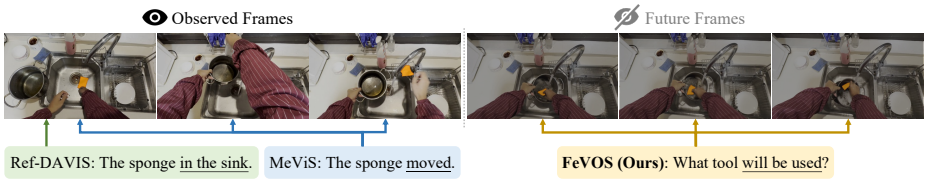

Fig. 1: Comparison of related datasets with **F**oresight **e**xpression **V**ideo **O**bject **S**egmentation (**FeVOS**). Unlike existing datasets (Ref-DAVIS [20], MeViS [7]) that ground expressions describing *observable* events (*e.g.*, in the sink, moved), our FeVOS requires predicting which object *will be involved* in future events based on observed visual cues. In this case, given the foresight expression "*What tool will be used?*", the model must analyze temporal context (dirty pot) and spatial cues (hand states) to anticipate the correct target. Zoom in for a better view.

This limitation is reflected in design paradigms of existing RVOS datasets. Earlier RVOS datasets, *e.g.*, Ref-DAVIS [20] and Ref-YouTube-VOS [31], provide videos in various scenarios with diverse object categories, but mainly focus on static attributes of objects (*e.g.*, appearances, categories, positions) that can be inferred from a single frame. Later, MeViS [7, 8] was proposed to involve spatio-temporal understanding by introducing motion expressions that describe motion-related attributes of objects across different frames. However, the referring scope of expressions in these datasets is limited to the *observed* frames provided. As shown in Figure 1, existing benchmarks like Ref-DAVIS ground expressions such as *"the sponge in the sink"* based on observable attributes, while MeViS handles motion-related expressions like *"the sponge moved"*, both describing events already present in the provided frames. In contrast, robotic applications often require models to anticipate which object(s) might be involved in the next scene or action before it occurs. This motivates our predictive reformulation of the task, where only visual cues prior to the referred action are provided. In the kitchen scene shown in Figure 1, the expression *"What tool will be used?"* requires the model to analyze both temporal context and spatial cues in the observed frames to predict which object will be involved in the future cleaning action. In this case, visual evidence of the dirty pot with dish soap applied signals a temporal transition toward cleaning action, allowing the model to infer that a sponge will be used next. Beyond tool category prediction, the model must also perform spatial disambiguation to locate the sponge on the right side, which is more convenient for manipulation since the left hand remains occupied. While existing video understanding [22, 27, 38, 46] tasks have explored temporal reasoning and future prediction, they typically follow a simple QA paradigm, lacking the fine-grained pixel-level grounding necessary for actionable decision-making.

To address these limitations, we propose Foresight Expression Video Object Segmentation (FeVOS), a novel task that queries future events in upcoming video segments and requires segmentation masks of the relevant objects in the observed frames. To support this task, we introduce FeVOS, a carefully curated dataset with 968 video clips spanning diverse scenes, 14,525 foresight expressions, and corresponding segmentation masks. Unlike traditional RVOS or existing

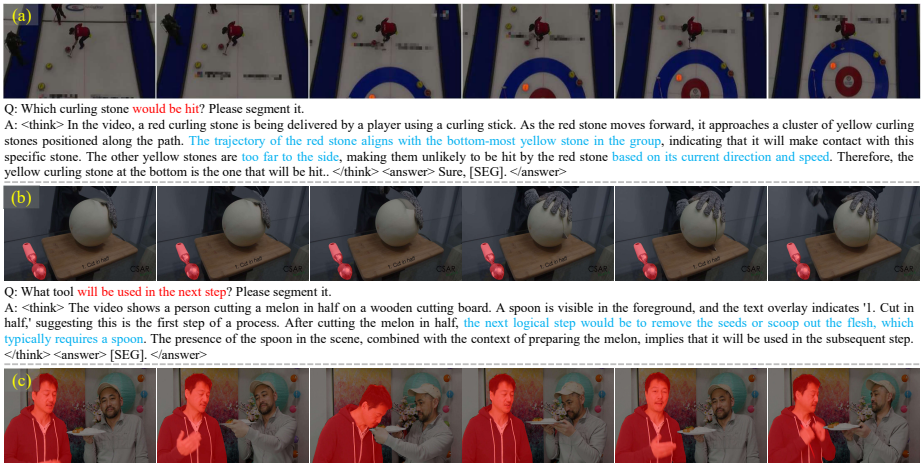

Fig. 2: Samples from FeVOS with chain-of-thought annotations. These examples present representative reasoning challenges included in our task: (a) Physically-Aligned Prediction, (b) Procedure-Grounded Prediction, (c) Intention-Guided Prediction. Red text highlights foresight expressions describing future, while blue text indicates key reasoning steps that lead to identifying the target objects. Zoom in for a better view.

video understanding tasks, our task introduces additional challenges: (I) The information provided in the question is limited to the future, making it difficult to directly align text expressions with the currently observed visual content. (II) It requires analyzing both temporal context and spatial cues to predict which object(s) will be involved in future actions, demanding a combination of visual reasoning and world knowledge. (III) The model must accurately provide pixel-level segmentation masks for the relevant object(s) in the observed frames, enriching future prediction with fine-grained spatio-temporal understanding.

While existing MLLM-based methods [3, 13, 24, 42, 47] achieve strong performance on traditional benchmarks that involve observable events, they struggle with our predictive reasoning task that requires anticipating future events from observed visual cues. To address this challenge, we draw inspiration from recent advances in reasoning-oriented reinforcement learning [14] and develop FeVOS-R1, a model trained via a two-stage pipeline. In Stage 1, we perform chain-of-thought supervised fine-tuning as a cold start, leveraging 2,904 synthetic CoT annotations we generated for FeVOS that provide step-by-step reasoning paths (as shown in Figure 2). In Stage 2, we employ Group Relative Policy Optimization (GRPO) [32] with proper reward, enabling the model to directly optimize for task performance while maintaining interpretable reasoning. Experiments demonstrate that FeVOS-R1 achieves state-of-the-art performance on FeVOS while generalizing well to existing RVOS benchmarks [7, 42] without additional fine-tuning. In a nutshell, our main contributions are as follows:

- We propose **Foresight Expression Video Object Segmentation**, a task requiring prediction of which object(s) will be involved in future events from observed visual cues, advancing RVOS toward predictive reasoning.
- We introduce **FeVOS dataset**, containing 968 videos, 14,525 foresight expressions, and corresponding pixel-level masks, along with 2,904 synthetic chain-of-thought annotations that provide explicit reasoning supervision.
- We develop **FeVOS-R1**, a two-stage training framework combining supervised fine-tuning with CoT data and reinforcement learning with end-to-end rewards, achieving state-of-the-art performance on predictive segmentation while generalizing well to traditional RVOS benchmarks.

2 Related Work

2.1 Referring Video Object Segmentation

Referring Video Object Segmentation (RVOS) [3, 7, 20, 31, 42] aims to segment objects corresponding to given language expressions in videos. Early RVOS datasets [11, 20, 31] predominantly focus on static attributes that can be inferred from single frames. MeViS [7] advances the field by introducing motion expressions that require spatio-temporal understanding across frames. Recent works like ReVOS [42] and ReasonVOS [3] further incorporate reasoning capabilities and world knowledge. However, all of these datasets perform reasoning on *observed* frames and focus on grounding expressions describing events already present in the video. In contrast, our FeVOS requires models to predict which object(s) *will be involved* in future events based solely on observed visual cues, introducing a fundamentally different predictive reasoning challenge.

2.2 MLLMs for Segmentation

As MLLMs [1, 2, 4, 5, 18, 25, 26, 39, 45] advance in vision-language understanding and reasoning, several works [13, 21, 24, 42, 49] uncover their significant potential for visual grounding tasks like referring segmentation. LISA [21] pioneers the introduction of a special token `[SEG]` into MLLMs and appends a segmentation head, thereby unleashing their segmentation capabilities for referring image segmentation tasks. VideoLISA [3] and TrackGPT [49] further extend this MLLM-based approach to the video domain. VISA [42] proposes a two-stage pipeline that first localizes the frames where the target object appears and then performs segmentation. Subsequent works [13, 24] have significantly boosted the performance by refining the frame-selection stage. Sa2VA [47] proposes a unified framework that seamlessly integrates an MLLM with SAM2 [30], delivering superior performance on both video understanding and segmentation tasks.

2.3 Reinforcement Learning for Vision-Language

Deepseek-R1 [14] demonstrates that Reinforcement Learning (RL) [35] effectively enhances LLM reasoning via GRPO [32]. VLM-R1 [33] extends the training

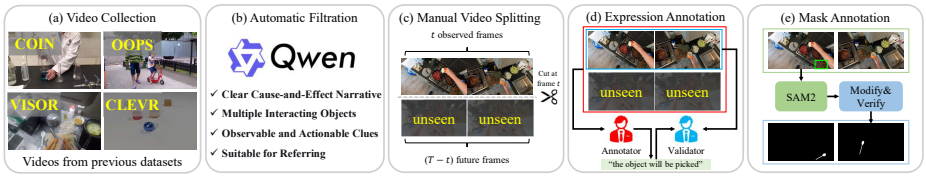

Fig. 3: Data Construction Pipeline. (a) Video Collection. Diverse videos were gathered from multiple sources [6, 10, 19, 36, 40]. (b) Automatic Filtration. We used Qwen2.5-VL [2] to filter videos based on carefully designed rules. (c) Manual Video Splitting. We manually split the videos that meet our criteria to observed frames and future frames. (d) Expression Annotation. We designed foresight expressions through annotation and validation to minimize the ambiguity while keeping the task challenging. (e). Mask Annotation. We used an interactive tool to annotate masks in the observed frames.

paradigm to more general vision-language tasks (referring expression comprehension and open-vocabulary object detection). SegAgent [50] trains MLLMs to mimic the annotation trajectories of human annotators. Seg-Zero [28] implements RL training from scratch by asking MLLMs to directly generate the coordinates of bounding boxes to prompt SAM2. Concurrently, Veason-R1 [12] explores RL for video segmentation. While concurrent work focuses on segmenting objects with expressions in traditional RVOS settings, we focus on predictive reasoning that requires anticipating future events from observed visual cues.

3 Benchmark: FeVOS

3.1 Task Formulation

Given a video clip $V \in \mathbb{R}^{T\times 3\times H\times W}$, where T, H, and W denote the length, height, and width of frames, and a predictive expression Q describing *future events*, the model outputs pixel-level segmentation masks $M \in \mathbb{R}^{T\times H\times W}$ for objects in the observed frames. Unlike traditional RVOS, our task requires reasoning about future actions or events based solely on cues from observed frames.

3.2 Construction of FeVOS

We demonstrate the five-stage FeVOS construction pipeline in Figure 3 and describe each stage in detail below.

(a) Video Collection. We collected candidate videos from various established benchmarks, including COIN [36], STAR [40], CLEVR [19], OOPS [10], and EPIC-KITCHENS-VISOR [6]. These datasets cover diverse scenarios such as egocentric activities, instructional videos, and dynamic outdoor environments. They are well-suited for our predictive reasoning and grounding task as they exhibit rich spatio-temporal dynamics and causal relationships, where early visual cues could naturally foreshadow subsequent events.

(b) Automatic Filtration. We employed Qwen2.5-VL [2] to automatically filter videos for predictive suitability. The model evaluated whether each video

could be temporally divided into two clips where: (1) the first clip contains a clear cause-and-effect narrative that enables deterministic inference of events in the second clip; (2) multiple interacting objects or agents are present; (3) the scene contains observable and actionable clues; and (4) the video is suitable for designing referring expressions. Around 30.5% videos are retained in this stage.

(c) Manual Video Splitting. We designed an online tool which can be used to load videos and decide to label time points at 0.2-second intervals or discard a video. Expert annotators manually reviewed the filtered videos and determined the optimal temporal split points for each video. The split point was chosen to ensure that: (1) the observation clip contains visible target objects with implicit visual cues indicating their future involvement; (2) the future clip contains clear, predictable events or actions; and (3) the split maximizes the predictive challenge while maintaining reasonable inference. Videos that could not be meaningfully split according to these criteria were excluded (around 44.0% retained).

(d) Expression Annotation. To ensure high-quality predictive expressions, we adopted a two-stage annotation process. In the first stage, annotators with access to both observed and unseen clips selected target objects in the observed frames and designed expressions describing future events, *e.g.*, *"the object that will be picked"*. In the second stage, independent validators, provided with the expression and observed frames without access to unseen clips, attempted to identify the target objects through predictive reasoning. Crucially, expressions must describe future events rather than directly describing observable attributes in the current frames. Around 85.9% of expressions are retained to ensure that they both required genuine predictive reasoning and allowed validators to correctly identify targets. This two-stage process ensures that (1) our dataset truly demands anticipatory reasoning rather than simple object recognition and (2) enough information is provided to predict the answers. Notably, although the future has its ambiguity and there might be multiple plausible outcomes in real-world scenarios, we only retain video-expression pairs with target object(s) that can be reasonably inferred with observable spatio-temporal cues. This can be further ensured in this expression annotation stage by adding restrictions (*e.g.* "first" when the referred events might happen multiple times).

(e) Mask Annotation. We generated pixel-level segmentation masks using an interactive annotation tool built upon SAM2 [30]. Annotators carefully modified and verified the masks of the target objects across all frames in the observed part, ensuring spatial and temporal consistency.

3.3 Automatic CoT Annotation Generation

After obtaining the video-object-expression triplets through the aforementioned pipeline, we generated chain-of-thought annotations to enable explicit reasoning processes. Specifically, we leveraged Qwen2.5-VL [2] to automatically produce CoT annotations using visual prompting. For each triplet, we overlaid the ground-truth segmentation masks onto the video frames as visual prompts, highlighting the target objects. The model was then provided with both the masked

Fig. 4: Expressions cloud.

Fig. 5: Reasoning cloud.

Stat.	Train	Val	All
Videos	779	189	968
Expr.	11,708	2,817	14,525
Frames	24,689	5,436	30,125
Obj.	1,137	282	1,419
Masks	30,694	6,718	37,412
CoT	2,337	567	2,904

Table 1: FeVOS statistics.

video and the foresight expression, and prompted to generate step-by-step reasoning that explains why the highlighted object is the answer to the given predictive question. This reasoning chain typically includes analysis of visual cues, temporal context, and causal relationships observed in the frames. We generated 3 CoT annotations per video to provide diverse reasoning perspectives. As shown in Figure 2, the CoT annotations demonstrate how to analyze temporal context and causal relationships to identify target objects that will participate in future events. This automatic annotation process enables our model to learn interpretable reasoning patterns that bridge the gap between visual observations and predictive conclusions and helps with the reinforcement learning stage.

3.4 Dataset Statistics & Analysis

Through the aforementioned annotation pipeline, we curated FeVOS, which contains 968 video clips with 14,525 foresight expressions and corresponding pixel-level segmentation masks. Each expression is paired with precise annotations identifying target objects across all frames in the observation segment. Additionally, we generated 2,904 synthetic chain-of-thought annotations to provide explicit reasoning supervision. We split FeVOS into training and validation subsets. The detailed dataset statistics can be found in Table 1.

Word Cloud. We provide word clouds illustrating the distribution of grounding expressions and reasoning processes in Figures 4 and 5. Notably, our foresight expressions frequently contain phrases like *"the next part of the video"* and words like *"will"*, which are intentionally used to encourage models to anticipate forthcoming events and actions before segmenting. Additionally, in reasoning processes, terms like *"Given"* appear frequently, guiding the model to align its inferences with the visual evidence provided by observed frames.

Predictive Reasoning Challenges. Compared with former RVOS datasets [7, 20,31], our task needs the model to perform predictive spatio-temporal reasoning. We further analyze the three representative examples in Figure 2 to illustrate the diverse reasoning challenges posed by our task including: (a) Physically-Aligned Prediction. The model must use the cross-frame information to predict the trajectory of the red curling stone to identify the right yellow curling stone that will be hit. (b) Procedure-Grounded Prediction. The model must understand the current cutting operation and reason about the whole logical process in the

procedure to decide that the spoon would be involved in the next step in removing the flesh. (c) Intention-Guided Prediction. The model should distinguish between the presenter and the potential taster, aligning the tasting action with the correct individual by inferring their intentions through their interactions. These challenges require a combination of spatio-temporally grounded reasoning and world knowledge to make the right mask prediction.

3.5 Evaluation Metrics

Following standard practice in video object segmentation [7,9,16,42,44], we employ two complementary metrics: $\mathcal{J}$ (region similarity via IoU) and $\mathcal{F}$ (boundary accuracy), and use their mean $\mathcal{J}\&\mathcal{F}$ as the overall performance metric.

4 Baseline: FeVOS-R1

4.1 Preliminary

Sa2VA. Sa2VA [47] is a unified framework that integrates MLLM [4] with SAM2 [30] for referring video object segmentation. The architecture consists of three key components: (1) a vision encoder that extracts visual features from video frames, (2) a large language model that processes both visual embeddings and text prompts to generate responses with special segmentation tokens, and (3) SAM2's mask decoder that produces pixel-wise segmentation masks conditioned on the hidden states of segmentation tokens. By leveraging the reasoning capabilities of LLMs and the strong segmentation performance of SAM2, Sa2VA achieves state-of-the-art results on traditional RVOS benchmarks. However, its supervised fine-tuning paradigm struggles with predictive reasoning tasks that require anticipating future events from observed visual cues, as it lacks explicit reasoning processes that lead to better optimization for segmentation quality. To address this limitation, we propose a two-stage training paradigm that equips the model with interpretable chain-of-thought reasoning processes and aligns it with task-specific objectives through reinforcement learning.

Group Relative Policy Optimization. GRPO [32] is a value-free reinforcement learning algorithm for efficient policy optimization. For each query q, GRPO samples a group of outputs $G = \{o_i\}_{i=1}^{|G|}$ and computes their rewards $\{r_i\}$. The advantage function is normalized relative to group statistics: $A_i = (r_i - \bar{r}_G)/\sigma_G$, where $\bar{r}_G$ and σ_G denote the mean and standard deviation of rewards. The policy π_θ is updated by maximizing the following objective:

$$\mathcal{J}(\theta) = \mathbb{E}_G\left[\frac{1}{|G|}\sum_{i\in G}\Big(\min\left(s_i A_i, \text{clip}\big(s_i, 1-\epsilon, 1+\epsilon\big)A_i\right) - \beta\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{ref})\Big)\right], \quad (1)$$

where $s_i = \frac{\pi_\theta(o_i|q)}{\pi_{\text{old}}(o_i|q)}$ denotes the probability ratio between the new and old policies, and ϵ is the clipping parameter controlling the update range. $\mathbb{D}_{\text{KL}}(\pi_\theta||\pi_{ref})$ is a KL divergence regularization term with a coefficient β that penalizes the model from deviating excessively from a reference model π_{ref}.

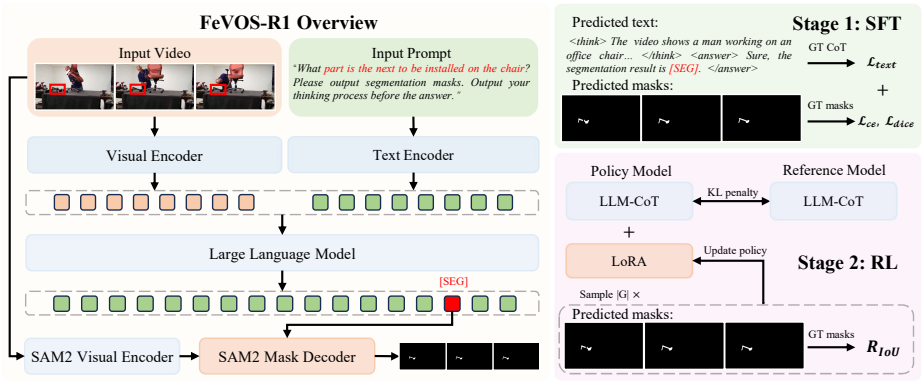

Fig. 6: Overview of FeVOS-R1. Our two-stage training framework: Stage 1 performs SFT with chain-of-thought annotations, and Stage 2 employs GRPO with IoU-based rewards to optimize segmentation quality through reasoning optimization.

4.2 Two-Stage Training Pipeline

We implement our method FeVOS-R1 based on Sa2VA via a two-stage training paradigm. As shown in Figure 6, given an input video V, we first sample a sequence of frames $\{I_t\}_{t=1}^{T}$ and encode them using the vision encoder to obtain visual embeddings. These embeddings, along with a text prompt P, are processed by the LLM to generate a response containing a special token `[SEG]`. The hidden states of `[SEG]` are then projected and fed into SAM2's mask decoder to predict a segmentation mask sequence $\{\hat{M}_t\}_{t=1}^{T}$ for the input frames.

Stage 1: Supervised Fine-Tuning with Chain-of-Thought. To equip the model with basic reasoning capabilities, we perform supervised fine-tuning using our synthetic CoT dataset introduced in Section 3.3. During training, the model learns to generate step-by-step reasoning processes that analyze temporal context and causal relationships before producing the `[SEG]` token. We employ three loss functions: pixel-wise cross-entropy loss $\mathcal{L}_{ce}$ and Dice loss $\mathcal{L}_{dice}$ for evaluating mask quality, and a text generation loss $\mathcal{L}_{text}$ for supervising the reasoning processes. The total loss is formulated as:

$$\mathcal{L}_{total} = \alpha_{ce}\mathcal{L}_{ce} + \alpha_{dice}\mathcal{L}_{dice} + \alpha_{text}\mathcal{L}_{text}, \tag{2}$$

where α_{ce}, α_{dice}, α_{text} represent the weights of the three losses, respectively. This stage enables the model to generate interpretable reasoning in the correct format while producing segmentation masks, laying the foundation for subsequent training with reinforcement learning.

Stage 2: Reinforcement Learning with End-to-End Rewards. While supervised fine-tuning provides preliminary knowledge of reasoning, the reasoning process remains suboptimal in quality and weakly aligned with segmentation objectives. We employ GRPO to further refine the reasoning process that leads to high segmentation quality with task-specific objectives. Unlike existing visual

grounding methods [12,28] that require models to output intermediate representations (*e.g.*, bounding box coordinates in JSON format), we leverage the `[SEG]` token to support end-to-end optimization that directly maximizes segmentation accuracy. Specifically, we define an accuracy reward R_{IoU} from the IoU between predicted and ground-truth masks:

$$R_{IoU} = \frac{1}{T}\sum_{t=1}^{T} \text{IoU}(\hat{M}_t, M_t), \tag{3}$$

where $\hat{M}_t$ and M_t denote the predicted and ground-truth masks at frame t, respectively. While previous works [12,14,28] typically incorporate a format reward to enforce structured output with `<think></think>` and `<answer></answer>` tags and correct JSON format, we find that using the accuracy reward alone achieves better performance. Our ablation studies in Section 5.3 reveal that format constraints provide no additional benefit and can even impede optimization. This is largely because our reasoning format is lightweight and already well-learned in the SFT stage. Consequently, our GRPO stage focuses exclusively on segmentation-aware reasoning refinement, guided solely by R_{IoU}, enabling tighter alignment between reasoning quality and segmentation performance.

5 Experiments

5.1 Implementation Details

Supervised Fine-Tuning. This stage serves as the CoT cold start before RL. We adopt pretrained Sa2VA-4B [47] as our base model, which integrates InternVL2.5 [4] as the MLLM and SAM2-L [30] as the segmentation module. During training, we only update the LLM and the SAM2 mask decoder while keeping other parts frozen, and apply LoRA [15] (rank = 128) for efficient parameter tuning. The optimization is performed using a learning rate of 2×10^{-5} with the cosine annealing schedule. We set the batch size to 4 with gradient accumulation over 4 steps and train on FeVOS dataset with CoT annotations for 4 epochs.

Reinforcement Learning. We freeze SAM2 mask decoder and exclusively fine-tune the LLM component using the same LoRA configuration as in the SFT stage. During training, the model generates $|G| = 4$ responses per input for GRPO. We set the learning rate to 1×10^{-5} with a batch size of 4 and gradient accumulation over 2 steps, training on the same expressions as SFT stage for 2 epochs. All experiments are conducted on 4 NVIDIA RTX 4090 GPUs.

Inference. We get a CoT response followed by a segmentation answer from the model and prompt SAM2 with the output `[SEG]` token to get final masks.

5.2 Quantitative Results

Results on FeVOS. As shown in Table 2, we comprehensively benchmark a range of recent video segmentation models on the proposed FeVOS dataset. Zero-shot models exhibit substantial difficulty with our predictive reasoning task, with

Table 2: Main results on the FeVOS dataset. * indicates the model is fine-tuned on FeVOS. **Bold** indicates the best performance.

Method	Backbone	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
Zero-shot baselines				
ReferFormer [41]	ResNet-50	16.4	20.0	18.2
LMPM [7]	Swin-T	17.1	20.7	18.9
VISA [42]	Chat-UniVi-7B	22.9	28.2	25.6
VideoLISA [3]	LLaVA-Phi-3-V-3.8B	22.9	29.4	26.1
VideoGLaMM [29]	Phi3-Mini-3.8B	21.7	26.7	24.2
VRS-HQ [13]	Chat-UniVi-7B	28.8	33.3	31.0
GLUS [24]	Chat-UniVi-7B	27.4	31.7	29.6
Sa2VA [47]	InternVL2.5-4B	23.7	27.2	25.4
Fine-tuned on FeVOS				
GLUS* [24]	Chat-UniVi-7B	31.0	35.9	33.5
Sa2VA* [47]	InternVL2.5-4B	33.1	38.4	35.8
FeVOS-R1 (ours)	InternVL2.5-4B	**39.5**	**45.1**	**42.3**

$\mathcal{J}\&\mathcal{F}$ scores below 31.0. Traditional RVOS methods such as ReferFormer (18.2) and LMPM (18.9) struggle significantly, as they are designed for grounding explicit descriptions rather than anticipating future events. Recent video-based MLLMs including VideoLISA (26.1) and VideoGLaMM (24.2) show moderate improvements through stronger video-language understanding. Models with advanced reasoning capabilities, particularly VRS-HQ (31.0) and GLUS (29.6), achieve the best zero-shot performance, yet remain substantially below fine-tuned models. When directly fine-tuning Sa2VA on FeVOS using standard SFT without CoT enhancement, performance improves markedly from 25.4 to 35.8, representing a substantial gain of +10.4. While fine-tuning GLUS yields a +3.9 performance gain, it continues to underperform relative to fine-tuned Sa2VA. This confirms that domain-specific adaptation enables models to better capture the temporal dynamics and causal relationships in predictive scenarios, and Sa2VA better learns these patterns and thus serves as our primary baseline. Our complete training pipeline, incorporating both CoT-guided reasoning and RL-based optimization, further pushes the performance to 42.3, achieving an additional +6.5 gain over the SFT baseline. This demonstrates that explicit reasoning chains and reward-guided optimization are essential for capturing subtle visual cues required for accurate future event prediction. Notably, the performance on FeVOS (42.3 $\mathcal{J}\&\mathcal{F}$) is substantially lower than on ReVOS (60.3) and MeViS (49.5), highlighting the increased complexity and challenge posed by our predictive segmentation task, which requires models to anticipate future events from visual cues rather than grounding expressions about observable events.

Generalization to Related Benchmarks. To assess generalization capability, we evaluate on ReVOS and MeViS benchmarks in Table 3. On ReVOS, the directly fine-tuned baseline Sa2VA* suffers a performance drop from 59.1 to 58.1 compared to its zero-shot counterpart, suggesting overfitting to FeVOS, while our FeVOS-R1 achieves 60.3, outperforming both baselines with particularly

Table 3: Quantitative results on ReVOS (Referring, Reasoning and Overall $\mathcal{J}\&\mathcal{F}$) and MeViS datasets. "-" denotes results not reported in the original papers.

Method	Backbone	ReVOS			MeViS		
		Ref.	Reas.	All	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
ReferFormer [41]	ResNet50	16.9	12.8	14.9	-	-	-
ReferFormer [41]	Video-Swin-B	32.7	23.4	28.1	29.8	32.2	31.0
LMPM [7]	Swin-T	34.1	18.8	26.4	34.2	40.2	37.2
LISA [21]	LLaVA-7B	45.7	36.1	40.9	35.1	39.4	37.2
TrackGPT [49]	LLaVA-7B	-	-	-	37.6	42.6	40.1
VISA [42]	Chat-UniVi-7B	50.9	43.0	46.9	40.7	46.3	43.5
VISA [42]	LLaVA-7B	51.0	43.2	47.1	-	-	-
VideoLISA [3]	LLaVA-Phi-3-V-3.8B	-	-	-	41.3	47.6	44.4
VideoGLaMM [29]	Phi3-Mini-3.8B	-	-	-	42.1	48.2	45.2
VRS-HQ [13]	Chat-UniVi-7B	62.1	56.1	59.1	47.6	53.7	50.6
GLUS [24]	Chat-UniVi-7B	58.3	51.4	54.9	**48.5**	**54.2**	**51.3**
Sa2VA [47]	InternVL2.5-4B	62.5	55.6	59.1	-	-	46.4
Sa2VA* [47]	InternVL2.5-4B	61.0	55.2	58.1	43.4	49.7	46.5
FeVOS-R1 (ours)	InternVL2.5-4B	**62.8**	**57.8**	**60.3**	46.2	52.7	49.5

Table 4: Ablation on training stages.

ID	Training Stages		FeVOS		
	CoT-SFT	RL	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
I	✓	✗	34.7	39.8	37.2
II	✗	✓	33.5	38.6	36.0
III	✓	✓	**39.5**	**45.1**	**42.3**

Table 5: Ablation on rewards.

ID	Reward		FeVOS		
	IoU	Format	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
I	✗	✓	35.1	40.3	37.7
II	✓	✓	38.3	43.5	40.9
III	✓	✗	**39.5**	**45.1**	**42.3**

strong gains on the reasoning subset (57.8 vs. 55.2 for baseline). On MeViS, our method reaches 49.5, representing a +3.0 improvement over the baseline (46.5), and surpassing all comparable-sized models including VideoLISA (44.4) and VideoGLaMM (45.2), though slightly below larger 7B models like GLUS (51.3) due to model scale differences. These results demonstrate that our CoT-augmented and RL-enhanced training strategy not only improves in-domain performance but also substantially enhances cross-domain generalization, particularly in reasoning-intensive scenes (*e.g.* Reasoning subset of ReVOS).

5.3 Ablation Studies

Two-stage Training Pipeline. To endow the model with reasoning capabilities and ensure correct output formatting, we adopt a two-stage training pipeline. In the first stage, the model is fine-tuned on FeVOS with CoT annotation using supervised learning to acquire basic reasoning skills and learn the expected output structure. In the second stage, we employ GRPO to further refine the model's reasoning chain to progressively enhance its segmentation performance. To evaluate the effectiveness of our pipeline, we compare three training strategies

as shown in Table 4: (I) SFT-only training with CoT data, (II) pure RL training from scratch, and (III) our proposed two-stage pipeline. SFT-only (37.2 $\mathcal{J}\&\mathcal{F}$) outperforms pure RL (36.0 $\mathcal{J}\&\mathcal{F}$), validating that explicit reasoning supervision provides a strong foundation. Pure RL struggles as the model produces trivial responses without meaningful reasoning trajectories. Our two-stage pipeline achieves 42.3 $\mathcal{J}\&\mathcal{F}$, representing substantial gains of +5.1 over SFT-only and +6.3 over RL-only, demonstrating that the two stages are highly complementary: SFT establishes reasoning patterns and output formatting, while RL further optimizes the reasoning chain through reward-guided exploration and exploitation.

Reward Design. Previous works [12, 28, 33] using GRPO for vision-language reasoning tasks typically rely on explicit format rewards to encourage structured reasoning outputs and ensure models adhere to specific JSON formats. This design is considered crucial for guiding models in generating proper reasoning processes before structured answers. However, we observe that since our method does not need JSON formats and the reasoning format can be well-learned in SFT, format rewards may become redundant in our experiment setting. To validate this hypothesis, we design a format reward that enforces structured outputs properly separated with reasoning and answer tags. We compare three different reward configurations: IoU reward only, format reward only, and their combination. As shown in Table 5, training with the IoU reward alone achieves the best overall performance (42.3 $\mathcal{J}\&\mathcal{F}$), outperforming the joint reward setting by 1.4 points and the format reward alone by 4.6 points. This reveals that when our model is properly initialized with basic knowledge of the output format through SFT with CoT data, explicit format rewards become unnecessary and may even divert valuable model capacity from learning task-relevant objectives.

5.4 Qualitative Results

Figure 7 shows qualitative comparisons demonstrating the benefit of explicit reasoning. In both cases, the finetuned baseline Sa2VA produces incorrect predictions by failing to analyze temporal context and causal relationships. In contrast, our FeVOS-R1 generates detailed reasoning chains that identify key visual cues and predict future events. For example, when asked *"What will fly out?"*, the baseline incorrectly segments the entire bottle, while our method reasons about internal pressure and cork behavior to correctly identify the cork. Similarly, for *"What will be full of clothes soon?"*, our model analyzes the ongoing action of removing clothes to accurately predict the laundry bag. These results demonstrate that explicit reasoning enables more accurate predictions aligned with human intuition and enhances the interpretability.

6 Conclusion and Discussion

We introduce Foresight Expression Video Object Segmentation, a novel task that advances pixel-level video understanding from observation to anticipation through predictive reasoning. FeVOS requires models to predict future events

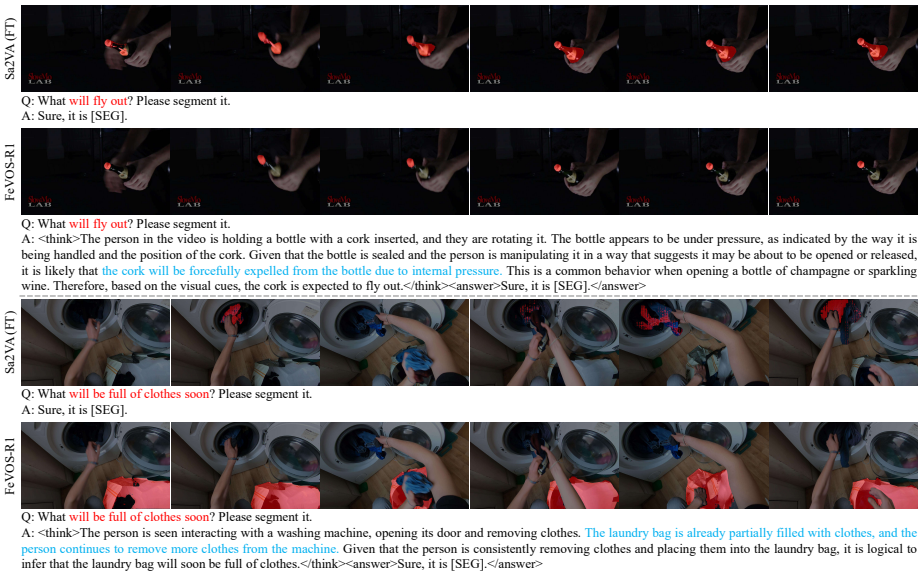

Fig. 7: Qualitative comparison. Our FeVOS-R1 generates explicit reasoning with key components (blue text) to analyze visual cues and predict future events, leading to more accurate segmentation results compared to the finetuned baseline Sa2VA.

and segment relevant objects based on implicit visual cues, posing fundamentally new challenges in spatio-temporal reasoning and visual grounding. To support this task, we construct FeVOS, containing 968 video clips with 14,525 predictive expressions and corresponding pixel-level segmentation masks, along with 2,904 synthetic chain-of-thought annotations to enable explicit reasoning. We further develop FeVOS-R1, a reasoning-enhanced model trained through a two-stage pipeline combining supervised fine-tuning with CoT data and reinforcement learning via GRPO. Experiments demonstrate that FeVOS-R1 achieves strong performance on FeVOS and exhibits robust generalization to existing RVOS benchmarks. Despite significant progress, the absolute performance on FeVOS remains substantially lower than on traditional RVOS benchmarks, underscoring the inherent difficulty of predictive reasoning and highlighting promising directions for future research in anticipatory visual understanding and grounding.

Future Directions. Though FeVOS-R1 achieves promising results on FeVOS, several directions remain open for future exploration: (I) Scaling beyond the current frame limit of MLLMs to leverage richer global context. (II) Modeling transient visual cues or local information that are critical for predictive reasoning via optimizing sampling strategy. (III) Performing CoT-SFT and RL training on more RVOS datasets to improve generalization. (IV) Mitigating hallucinations during the reasoning process to improve its reliability and alignment. (V) Modeling uncertainty in future predictions to handle multiple plausible outcomes. (VI) Enabling real-time inference to support broader downstream applications.

Acknowledgements This work was supported by the National Natural Science Foundation of China (NSFC) under Grant No. 62472104 and the Science and Technology Commission of Shanghai Municipality under Grant No. 25511103600.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. arXiv preprint arXiv:2308.12966 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. Advances in Neural Information Processing Systems **37**, 6833–6859 (2024)
4. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
5. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24185–24198 (2024)
6. Darkhalil, A., Shan, D., Zhu, B., Ma, J., Kar, A., Higgins, R., Fidler, S., Fouhey, D., Damen, D.: Epic-kitchens visor benchmark: Video segmentations and object relations. Advances in Neural Information Processing Systems **35**, 13745–13758 (2022)
7. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: MeViS: A large-scale benchmark for video segmentation with motion expressions. In: ICCV (2023)
8. Ding, H., Liu, C., He, S., Ying, K., Jiang, X., Loy, C.C., Jiang, Y.G.: Mevis: A multi-modal dataset for referring motion expression video segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
9. Ding, H., Ying, K., Liu, C., He, S., Jiang, X., Jiang, Y.G., Torr, P.H., Bai, S.: Mosev2: A more challenging dataset for video object segmentation in complex scenes. arXiv preprint arXiv:2508.05630 (2025)
10. Epstein, D., Chen, B., Vondrick, C.: Oops! predicting unintentional action in video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 919–929 (2020)
11. Gavrilyuk, K., Ghodrati, A., Li, Z., Snoek, C.G.: Actor and Action Video Segmentation from a Sentence. In: CVPR (2018)
12. Gong, S., Zhang, L., Zhuge, Y., Jia, X., Zhang, P., Lu, H.: Reinforcing video reasoning segmentation to think before it segments. arXiv preprint arXiv:2508.11538 (2025)
13. Gong, S., Zhuge, Y., Zhang, L., Yang, Z., Zhang, P., Lu, H.: The devil is in temporal token: High quality video reasoning segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29183–29192 (2025)
14. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)

15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR (2022)
16. Hu, H., Ying, K., Ding, H.: Segment anything across shots: a method and benchmark. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 4825–4833 (2026)
17. Huang, H., Chen, X., Chen, Y., Li, H., Han, X., Wang, Z., Wang, T., Pang, J., Zhao, Z.: Roboground: Robotic manipulation with grounded vision-language priors. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22540–22550 (2025)
18. Jin, P., Takanobu, R., Zhang, W., Cao, X., Yuan, L.: Chat-univi: Unified visual representation empowers large language models with image and video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13700–13710 (2024)
19. Johnson, J., Hariharan, B., Van Der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., Girshick, R.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2901–2910 (2017)
20. Khoreva, A., Rohrbach, A., Schiele, B.: Video Object Segmentation with Language Referring Expressions. In: ACCV (2019)
21. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: LISA: Reasoning Segmentation via Large Language Model. In: CVPR (2024)
22. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
23. Lin, J., Chen, J., Peng, K., He, X., Li, Z., Stiefelhagen, R., Yang, K.: Echotrack: Auditory referring multi-object tracking for autonomous driving. IEEE Transactions on Intelligent Transportation Systems (2024)
24. Lin, L., Yu, X., Pang, Z., Wang, Y.X.: Glus: Global-local reasoning unified into a single large language model for video segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual Instruction Tuning. In: NeurIPS (2023)
26. Liu, S., Ying, K., Zhang, H., Yang, Y., Lin, Y., Zhang, T., Li, C., Qiao, Y., Luo, P., Shao, W., et al.: ConvBench: A Multi-Turn Conversation Evaluation Benchmark with Hierarchical Ablation Capability for Large Vision-Language Models. In: Adv. Neural Inform. Process. Syst. Datasets Benchmarks Track (2024)
27. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
28. Liu, Y., Peng, B., Zhong, Z., Yue, Z., Lu, F., Yu, B., Jia, J.: Seg-zero: Reasoning-chain guided segmentation via cognitive reinforcement. arXiv preprint arXiv:2503.06520 (2025)
29. Munasinghe, S., Gani, H., Zhu, W., Cao, J., Xing, E., Khan, F.S., Khan, S.: Videoglamm: A large multimodal model for pixel-level visual grounding in videos. In: CVPR (2025)
30. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment Anything in Images and Videos. arXiv preprint arXiv:2408.00714 (2024)
31. Seo, S., Lee, J.Y., Han, B.: URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark. In: ECCV (2020)

32. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
33. Shen, H., Liu, P., Li, J., Fang, C., Ma, Y., Liao, J., Shen, Q., Zhang, Z., Zhao, K., Zhang, Q., et al.: Vlm-r1: A stable and generalizable r1-style large vision-language model. arXiv preprint arXiv:2504.07615 (2025)
34. Stone, A., Xiao, T., Lu, Y., Gopalakrishnan, K., Lee, K.H., Vuong, Q., Wohlhart, P., Kirmani, S., Zitkovich, B., Xia, F., et al.: Open-world object manipulation using pre-trained vision-language models. arXiv preprint arXiv:2303.00905 (2023)
35. Sutton, R.S., Barto, A.G., et al.: Reinforcement learning: An introduction. MIT press Cambridge (1998)
36. Tang, Y., Ding, D., Rao, Y., Zheng, Y., Zhang, D., Zhao, L., Lu, J., Zhou, J.: Coin: A large-scale dataset for comprehensive instructional video analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1207–1216 (2019)
37. Tilekbay, B., Yang, S., Lewkowicz, M.A., Suryapranata, A., Kim, J.: Expressedit: Video editing with natural language and sketching. In: Proceedings of the 29th International Conference on Intelligent User Interfaces. pp. 515–536 (2024)
38. Wang, H., Liu, H., Liu, X., Du, C., Kawaguchi, K., Wang, Y., Pang, T.: Fostering video reasoning via next-event prediction. arXiv preprint arXiv:2505.22457 (2025)
39. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. arXiv preprint arXiv:2409.12191 (2024)
40. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: Star: A benchmark for situated reasoning in real-world videos. arXiv preprint arXiv:2405.09711 (2024)
41. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as Queries for Referring Video Object Segmentation. In: CVPR (2022)
42. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: VISA: Reasoning Video Object Segmentation via Large Language Models. In: ECCV (2024)
43. Ying, K., Ding, H., Jie, G., Jiang, Y.G.: Towards omnimodal expressions and reasoning in referring audio-visual segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22575–22585 (2025)
44. Ying, K., Hu, H., Ding, H.: MOVE: Motion-guided few-shot video object segmentation. In: ICCV (2025)
45. Ying, K., Meng, F., Wang, J., Li, Z., Lin, H., Yang, Y., Zhang, H., Zhang, W., Lin, Y., Liu, S., Lei, J., Lu, Q., Chen, R., Xu, P., Zhang, R., Zhang, H., Gao, P., Wang, Y., Qiao, Y., Luo, P., Zhang, K., Shao, W.: MMT-Bench: A Comprehensive Multimodal Benchmark for Evaluating Large Vision-Language Models Towards Multitask AGI. In: Int. Conf. Mach. Learn. (2024)
46. Yu, E., Zhao, L., Wei, Y., Yang, J., Wu, D., Kong, L., Wei, H., Wang, T., Ge, Z., Zhang, X., et al.: Merlin: Empowering multimodal llms with foresight minds. In: European Conference on Computer Vision. pp. 425–443. Springer (2024)
47. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025)
48. Zhou, E., An, J., Chi, C., Han, Y., Rong, S., Zhang, C., Wang, P., Wang, Z., Huang, T., Sheng, L., et al.: Roborefer: Towards spatial referring with reasoning in vision-language models for robotics. arXiv preprint arXiv:2506.04308 (2025)
49. Zhu, J., Cheng, Z.Q., He, J.Y., Li, C., Luo, B., Lu, H., Geng, Y., Xie, X.: Tracking with human-intent reasoning. arXiv preprint arXiv:2312.17448 (2023)

50. Zhu, M., Tian, Y., Chen, H., Zhou, C., Guo, Q., Liu, Y., Yang, M., Shen, C.: Segagent: Exploring pixel understanding capabilities in mllms by imitating human annotator trajectories. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3686–3696 (2025)