THINK VISUALLY ACROSS SPACE AND TIME: GROUNDED REASONING FOR VIDEO SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Video Reasoning Segmentation (VRS) requires a model to identify an implicitly described target and predict its pixel-level mask trajectory, while resolving the spatial relations and temporal events that determine the answer. Text-only Chain-of-Thought (CoT) can explain a target choice without making the underlying visual references explicit, leaving a *video reference gap* between linguistic inference and the evidence distributed across frames. We introduce **Temporal Chain of Video Primitives (TCoVP)**, a spatiotemporally grounded form of Video CoT that organizes query-relevant observations into a chronological reasoning chain. Each primitive associates an entity reference with temporal support and timestamped spatial evidence, and the chain interleaves these references with language to guide reasoning about which entity is relevant, when the queried event holds, and where its evidence appears. The selected target primitives then provide box prompts and temporal support for video mask prediction. We construct TCoVP-SFT-15k and TCoVP-RL-13k to learn this representation through supervised fine-tuning and chain-aligned reinforcement learning. Extensive experiments across five VRS benchmarks and three model backbones demonstrate the strong performance of TCoVP, highlighting the effectiveness of organizing spatial and temporal evidence within the reasoning process. All models and code will be publicly released.

1 INTRODUCTION

Video Reasoning Segmentation (VRS) (Yan et al., 2024; Bai et al., 2024) asks a model to resolve an implicit natural-language query into the mask trajectory of its intended target. Because the target may be specified through an interaction, an event sequence, or a transient action, solving the task requires connecting semantic interpretation with visual evidence across space and time. Chain-of-Thought (CoT) reasoning offers a way to make these intermediate interpretations explicit, and recent VRS methods use it to reason about a query before producing segmentation prompts (Gong et al., 2026; Wen et al., 2026). The central challenge is to ground the reasoning itself in the evolving scene, so that the entities and events used to identify a target remain tied to the observations that support them.

Existing approaches connect language to masks through latent segmentation tokens (Zheng et al., 2024; Wei et al., 2024) or through keyframe selection and geometric prompts (Gong et al., 2026; Xu et al., 2026). For VRS, establishing this connection requires distinguishing the entities that provide contextual evidence from the target to segment, while separating the queried action from the longer period during which an object is visible. In Figure 1, the query asks who clasps hands with the athlete with dyed hair before hugging that athlete. The athlete provides the identifying context, whereas the coach involved in the hand-clasp is the target. As the coach remains visible beyond the hand-clasp, attending to the later hug or propagating masks over the full video can produce the wrong answer. This example illustrates the *video reference gap*: linguistic descriptions can leave unresolved which visual observations support a target choice and when that choice satisfies the query.

Our goal is to give Video CoT **explicit spatial and temporal guidance** throughout this reasoning process. The key idea is to organize visual references as a temporal chain, so that observations of relevant entities and events form the evidence from which the target is inferred. Spatial references specify the participants in an interaction, while timestamps and temporal support distinguish successive events and delimit the action of interest. Interleaving these references with language gives subsequent reasoning steps access to explicitly localized evidence and makes intermediate grounding

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

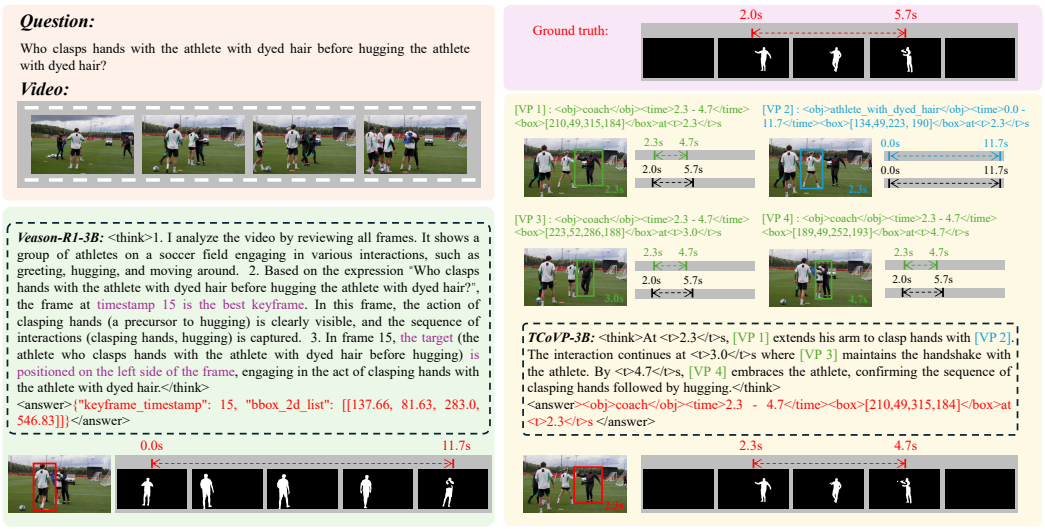

Figure 1: **Reasoning with Temporal Chain of Video Primitives.** A text-only keyframe-based prediction focuses on the later hugging event and localizes the wrong athlete. TCoVP progressively records the relevant entities as timestamped video primitives, distinguishes the hand-clasp interval from the longer presence interval, and converts the selected target primitive into a mask trajectory.

available for supervision. The same representation also connects the resulting target decision to the spatial prompts and temporal support needed for mask prediction.

We realize this idea through **Temporal Chain of Video Primitives (TCoVP)**, which represents query-conditioned visual evidence as a chronological chain of timestamped video primitives. Each primitive combines an entity reference, a temporal span, and a bounding box at an observed timestamp, together with a semantic description. Entity identifiers link references within the generated chain, while chronological ordering records how the supporting observations unfold over the video. A canonical serialization interleaves these primitives with natural-language reasoning, allowing contextual observations and target evidence to be expressed in a common form. The selected target primitives supply timestamped box prompts for mask propagation, and their temporal support restricts the output when the query specifies an action interval.

We train TCoVP through supervised fine-tuning on **TCoVP-SFT-15k** and chain-aligned reinforcement learning on **TCoVP-RL-13k**, with rewards for structural validity, temporal alignment, spatiotemporal grounding, and target primitive accuracy. Evaluation spans five VRS benchmarks and Qwen2.5-VL, Qwen3-VL, and Molmo2 backbones. With Qwen2.5-VL-3B and the data volume fixed, TCoVP and its intermediate grounding supervision improve the five-benchmark macro-average $\mathcal{J}\&\mathcal{F}$ from 57.8 to 62.7 over text-only CoT with the same answer-primitive format. The largest gain of 7.4 points occurs on GroundMoRe, where queries require explicit temporal event localization.

Our contributions are summarized as follows:

- We formulate the **video reference gap** and introduce **Temporal Chain of Video Primitives (TCoVP)**, which constructs entity identity and timestamped observations progressively over the video in a single reasoning trace, and makes the selected target primitives executable for VRS.
- We construct **TCoVP-SFT-15k** and **TCoVP-RL-13k** and train TCoVP with chain-aligned rewards for structural validity, temporal alignment, spatiotemporal grounding, and target primitive accuracy.
- We evaluate TCoVP across five VRS benchmarks and multiple MLLM backbones. Controlled comparisons show consistent improvements over text-only CoT, especially on queries that require explicit temporal event localization.

2 Related Work

Video Reasoning Segmentation. VRS extends referring video segmentation from explicit object descriptions to queries that encode intent, commonsense, and temporal logic (Yan et al., 2024; Bai

et al., 2024). Supervised approaches map MLLM representations to segmentation tokens or mask decoders (Zheng et al., 2024; Wei et al., 2024; Lin et al., 2025; Gong et al., 2025; Yuan et al., 2025; Liu et al., 2025; Dai et al., 2025). RL-based methods instead optimize intermediate reasoning, keyframe selection, or spatial grounding (Gong et al., 2026; Xu et al., 2026; Li et al., 2025; Wen et al., 2026). VideoSEG-O3 uses multi-turn tool interaction to refine temporal and spatial predictions (Dai et al., 2026), while VIRST fuses video-language and segmentation representations in a unified model (Hong & Do, 2026). TCoVP targets a complementary interface problem: it makes references to relevant entities and events explicit through a single-turn temporally ordered chain of video primitives before dense mask propagation.

Visual primitives and grounded video reasoning. Thinking-with-image methods manipulate crops, regions, or visual tokens during reasoning (OpenAI, 2025; Zheng et al., 2025; Wang et al., 2025a;b; Fan et al., 2025). Thinking with Visual Primitives instead treats points and boxes as units of image reasoning that close the reference gap between linguistic thoughts and visual entities (Lu et al., 2026). Video introduces temporal structure to this gap. Open-o3 Video generates timestamped spatial evidence for video question answering (Meng et al., 2025), STVG-o1 produces bounding-box CoT for spatio-temporal grounding (Gu et al., 2025), and EventCoT grounds answers in selected temporal events (Song et al., 2026). TCoVP focuses on implicit-query segmentation and organizes visual evidence as a chronological chain of timestamped video primitives. This structure distinguishes entity identity from individual observations, represents their temporal ordering, and exposes the selected target primitives to a video mask propagator.

3 TCoVP: Temporal Chain of Video Primitives

In this section, we introduce TCoVP to address the video reference gap by grounding reasoning in explicit spatial and temporal evidence. We present the temporal primitive chain representation (Section 3.1), data construction (Section 3.2), training strategy (Section 3.3), and primitive-to-mask execution (Section 3.4).

3.1 Temporal primitive chain representation

Let a video be $\mathcal{V} = \{I_f\}_{f=1}^{F}$, from which we uniformly sample T observations $\{I_{t_k}\}_{k=1}^{T}$, and let q denote an implicit reasoning query. VRS predicts a full mask trajectory $\mathcal{M} = \{M_f\}_{f=1}^{F}$ for the object or object set that satisfies q. The central challenge is to maintain a consistent reference to an entity while its appearance, visibility, and relations change over time.

We represent the query-conditioned visual evidence as an ordered temporal chain

$$\mathcal{C}_q = (p_1, p_2, \ldots, p_L), \qquad \tau_1 \le \tau_2 \le \cdots \le \tau_L, \tag{1}$$

where each video primitive is defined as

$$p_\ell = (\mathrm{id}_\ell, c_\ell, \tau_\ell, b_\ell, [s_\ell, e_\ell], r_\ell)\,. \tag{2}$$

Here, id_ℓ denotes the identity of the referenced entity, c_ℓ describes its semantic state, τ_ℓ is the timestamp of the supporting observation, b_ℓ is the corresponding bounding box, $[s_\ell, e_\ell]$ denotes the temporal support of the entity or event, and r_ℓ optionally describes an action or relation. Reusing the same identifier at different timestamps preserves entity continuity, while the ordering of primitives records the temporal evolution of the observed evidence.

Chronological chain rollout. TCoVP externalizes video understanding as a chronological rollout of video primitives. Let $\mathcal{C}_q^{(0)}$ denote an initially empty chain. At each sampled timestamp t_k, the model either emits no primitive or appends one or more primitives:

$$\mathcal{C}_q^{(k)} = \mathrm{Append}\left(\mathcal{C}_q^{(k-1)}, \mathrm{Emit}(I_{t_k}, q)\right), \qquad 0 = t_1 < t_2 < \cdots < t_T, \tag{3}$$

where $\mathrm{Emit}(I_{t_k}, q)$ may return an empty set when the observation does not provide query-relevant evidence. Each emitted primitive may introduce a new entity, add a timestamped observation for an existing identifier, or update the temporal support of an entity or event. The rollout specifies the order of the externalized reasoning trace rather than restricting model attention to past frames.

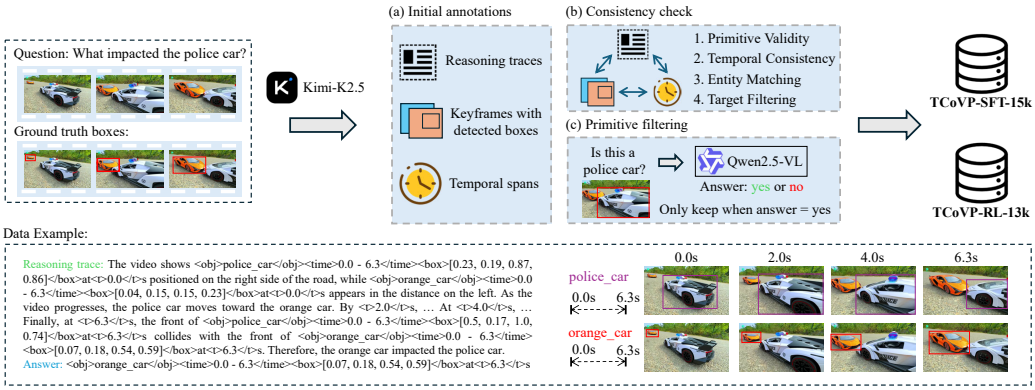

Figure 2: **Constructing TCoVP supervision.** Kimi-K2.5 proposes language traces interleaved with temporally ordered video primitives. Chain-consistency checking and Qwen2.5-VL primitive filtering retain primitives that can be grounded in the sampled video. The filtered examples form TCoVP-SFT-15k and TCoVP-RL-13k.

Autoregressive chain serialization. TCoVP is a structured sequence-generation target rather than an auxiliary neural network. Each primitive is serialized using `<obj>`, `<time>`, `<box>`, and `<t>` tags, interleaved with the language that explains the query-specific event logic. Reusing the same identifier across primitives preserves entity continuity, while the order of the serialized records follows the temporal order of the corresponding observations. This format remains compatible with standard autoregressive decoding and allows each primitive to be parsed and verified independently.

Target primitive selection. The reasoning trace may contain contextual primitives that help resolve the query but should not be segmented. After the temporal rollout is complete, the final `<answer>` block selects one or more query-relevant primitives from the chain. These selected primitives form the target primitive set and provide timestamped boxes for mask propagation:

$$\widehat{\mathcal{M}} = \bigvee_{p_\ell \in \mathcal{P}_{\text{ans}}} \Phi_{\text{seg}}\left(\mathcal{V}; \tau_\ell, b_\ell\right), \tag{4}$$

where $\mathcal{P}_{\text{ans}} \subseteq \mathcal{C}_q$ denotes the selected target primitives and Φ_{seg} denotes a video mask propagation operator. The resulting masks are merged by pixel-wise union. For motion-grounded queries, propagation is restricted to the predicted temporal support of the selected primitives.

3.2 Data Construction

Existing video object segmentation datasets provide masks for target objects, but they do not directly supervise a model to use visual references during reasoning. In particular, they rarely associate target and contextual entities with temporal supports and timestamped spatial evidence, even when these entities and their interactions determine the answer to an implicit query. A language-only trace therefore cannot be checked for entity identity, event timing, or spatial grounding. To address this limitation, we construct supervision in the form of chronologically ordered chains of timestamped video primitives.

We derive two training corpora from the same pool of temporal-chain-annotated examples: **TCoVP-SFT-15k** provides complete chain-serialized reasoning trajectories for supervised fine-tuning, and **TCoVP-RL-13k** retains examples with reward-compatible video primitive annotations for reinforcement learning. Each instance contains a question and answer, the timestamps of query-relevant detected keyframes, boxes for target and contextual entities, temporal supports, a text-only trace for controlled comparison, and a serialized TCoVP chain.

Our datasets are constructed on top of the training splits of existing video object segmentation datasets, including ReVOS (Yan et al., 2024), MeViS (Ding et al., 2023), Long-RVOS (Liang et al., 2026), and GroundMoRe (Deng et al., 2025), and cover single-target, multi-target, and zero-target samples. Specifically, we design the following automated data annotation pipeline:

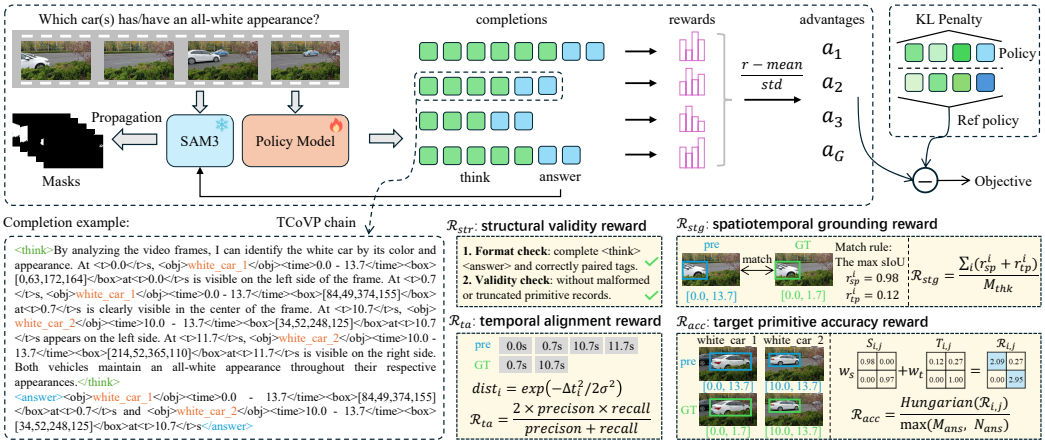

Figure 3: **Chain-aligned training and mask execution.** The policy generates multiple serialized TCoVP completions. Structural validity, temporal alignment, spatiotemporal grounding, and target primitive accuracy rewards produce group-normalized advantages, while the reference policy provides KL regularization. The selected target primitives are passed to SAM3 for video mask propagation.

Data preprocessing. We uniformly sample 16 frames from each video while always retaining the first and last frames. For samples with annotated targets, we remove instances whose target is absent from every sampled frame. Target masks are converted into boxes at the sampled timestamps. The frame with the largest visible target region supplies the spatial evidence used in the expected answer, while the remaining sampled frames provide evidence for contextual primitives.

Temporal-chain annotation. Kimi-K2.5 (Team et al., 2026) generates an initial reasoning trace from the sampled video and task annotations. Whenever the trace refers to a visually relevant entity, it emits a video primitive containing an entity identifier, temporal support, a detected keyframe timestamp, and a bounding box. Subsequent observations of the same entity reuse its identifier, thereby preserving entity continuity across time. The emitted primitives follow the chronological order of their keyframe timestamps. The trace typically starts from the earliest available evidence and progresses along the video, while allowing revisits to earlier primitives when a later event changes their interpretation. The resulting serialization interleaves natural-language reasoning with temporally ordered primitive records. We retain a text-only counterpart to isolate the effect of the temporal-primitive representation.

Chain consistency and primitive filtering. We parse every serialized primitive and require it to contain a valid entity identifier, temporal support, detected keyframe timestamp, and bounding box. We additionally check timestamp order and consistent reuse of entity identifiers across observations. References that cannot be matched to an annotated entity are removed, and samples with invalid or unresolved target primitives are discarded. Qwen2.5-VL-7B (Bai et al., 2025b) performs a second-stage semantic check on bounding boxes for contextual entities. This pipeline converts mask-oriented VRS data into supervision for spatiotemporally grounded referential reasoning, as summarized in Figure 2.

3.3 Training Strategy

3.3.1 Cold Start Initialization via SFT

We initialize TCoVP with Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a), and Molmo2 (Clark et al., 2026) backbones on TCoVP-SFT-15k. The supervised objective teaches the canonical serialization of a temporal primitive chain together with natural-language reasoning. The model learns to emit primitives only when sampled observations provide query-relevant evidence, preserve chronological order, and reuse entity identifiers across timestamps. This stage also provides a text-only counterpart with the same questions and answers for the representation ablation in Table 3.

3.3.2 CHAIN-ALIGNED REINFORCEMENT LEARNING

Figure 3 summarizes the chain-aligned training objective and the primitive-to-mask execution path. Given a video-query pair x, the current policy samples a group of G candidate chain serializations $\{y_i\}_{i=1}^{G}$. Each response receives a chain-aligned reward $\mathcal{R}(x, y_i)$, which is normalized within the group following group-relative policy optimization (Shao et al., 2024):

$$\hat{A}_i = \frac{\mathcal{R}(x, y_i) - \text{mean}(\{\mathcal{R}(x, y_j)\}_{j=1}^{G})}{\text{std}(\{\mathcal{R}(x, y_j)\}_{j=1}^{G})}. \tag{5}$$

We optimize an on-policy group-relative objective with length-normalized log likelihood and a KL penalty to a fixed reference policy:

$$\mathcal{L}_{\text{RL}}(\theta) = -\mathbb{E}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{\hat{A}_i}{|y_i|}\sum_{t=1}^{|y_i|}\log \pi_\theta(y_{i,t} \mid x, y_{i,<t})\right] + \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}). \tag{6}$$

The group normalization compares candidate TCoVP chain serializations for the same video-query pair, while the KL term limits drift from the supervised model.

For each query-completion pair (x, y), the chain-aligned reward is defined as follows:

$$\mathcal{R}(x, y) = \mathcal{R}_{str}(x, y) + \mathcal{R}_{ta}(x, y) + \mathcal{R}_{stg}(x, y) + \mathcal{R}_{acc}(x, y). \tag{7}$$

Here, $\mathcal{R}_{str}$ combines format and validity checks to ensure that the generated TCoVP chain is well-formed and parseable. The remaining terms evaluate temporal alignment, spatiotemporal grounding, and target primitive accuracy, respectively.

Structural validity reward $\mathcal{R}_{str}$. It verifies that the response contains complete `<think>` and `<answer>` blocks, correctly paired primitive tags, valid entity identifiers, temporal supports, keyframe timestamps, and in-range bounding boxes. Malformed or truncated primitive records receive a lower reward.

Temporal alignment reward $\mathcal{R}_{ta}$. Let M denote the number of timestamps $\{t_m\}_{m=1}^{M}$ associated with the video primitives emitted in `<think>`, and N denote the number of ground-truth evidence timestamps $\{t_n^{\text{gt}}\}_{n=1}^{N}$. We measure the temporal alignment between the generated primitives and the annotated evidence using soft precision and recall:

$$\text{Prec}(x, y) = \frac{1}{M}\sum_{m=1}^{M}\exp\left(-\frac{\Delta t_m^2}{2\sigma^2}\right), \quad \Delta t_m = \min_{1\le n\le N}|t_m - t_n^{\text{gt}}|, \tag{8}$$

$$\text{Rec}(x, y) = \frac{1}{N}\sum_{n=1}^{N}\exp\left(-\frac{\Delta t_n^2}{2\sigma^2}\right), \quad \Delta t_n = \min_{1\le m\le M}|t_n^{\text{gt}} - t_m|, \tag{9}$$

$$\mathcal{R}_{ta}(x, y) = 2 \cdot \frac{\text{Prec}(x, y) \cdot \text{Rec}(x, y)}{\text{Prec}(x, y) + \text{Rec}(x, y)}. \tag{10}$$

We set $\mathcal{R}_{ta} = 0$ when no valid predicted or reference timestamp is available. Combining temporal precision and recall penalizes both irrelevant and missing temporal evidence, preventing the generated chain from collapsing to a single late timestamp and encouraging it to recover all query-relevant evidence across the video.

Spatiotemporal grounding reward $\mathcal{R}_{stg}$. Let $\hat{\mathcal{P}}(y) = \{\hat{p}_m\}_{m=1}^{M_{\text{thk}}}$ denote the set of video primitives generated in the reasoning trace, and $\mathcal{P}^{\text{gt}} = \{p_n^{\text{gt}}\}_{n=1}^{N_{\text{thk}}}$ denote the annotated primitives. For each $\hat{p}_m$, we compute its best spatial match among annotated primitives with nearby timestamps:

$$r_{\text{sp}}^{m} = \max_{n:\,|\hat{\tau}_m - \tau_n^{\text{gt}}|\le\tau} \text{IoU}\left(\hat{b}_m, b_n^{\text{gt}}\right), \tag{11}$$

where $\hat{\tau}_m$ and $\hat{b}_m$ are the timestamp and bounding box of $\hat{p}_m$. If no ground-truth primitive falls within the temporal gate, we set $r_{\text{sp}}^{m} = 0$ and omit the temporal-IoU term for that primitive. Let $n^*(m)$ denote the index of the matched ground-truth primitive. The spatiotemporal grounding reward is:

$$\mathcal{R}_{stg}(x, y) = \frac{1}{M_{\text{thk}}}\sum_{m=1}^{M_{\text{thk}}}\left[r_{\text{sp}}^{m} + \text{tIoU}\left(\hat{\mathcal{T}}_m, \mathcal{T}_{n^*(m)}^{\text{gt}}\right)\right]. \tag{12}$$

Table 1: Performance comparison on the motion-grounded video reasoning segmentation benchmark **GroundMoRe**. Best results are **bolded**, and the second best results are underlined.

Methods	Overall			Causal			Sequential			Counterfactual			Descriptive		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Two-Stage Baseline															
ViLA (Lin et al., 2024)+ReferFormer	16.37	14.84	17.89	15.92	14.33	17.52	16.30	14.94	17.67	13.63	12.35	14.91	19.61	17.78	21.45
ViLA (Lin et al., 2024)+SgMg	17.07	15.25	18.89	17.34	15.43	19.26	16.30	14.57	18.04	15.31	13.96	16.67	19.18	16.96	21.40
ViLA (Lin et al., 2024)+HTR	15.72	13.90	17.53	16.46	14.60	18.31	15.11	13.39	16.83	12.61	11.22	13.99	18.37	16.10	20.63
VideoChat2 (Li et al., 2024)+ReferFormer	15.38	13.84	16.93	14.74	13.10	16.37	13.95	12.69	15.21	13.12	11.80	14.43	19.82	17.89	21.74
VideoChat2 (Li et al., 2024)+SgMg	16.65	14.83	18.48	16.64	14.75	18.54	14.70	13.02	16.38	15.22	13.87	16.58	20.01	17.68	22.34
VideoChat2 (Li et al., 2024)+HTR	15.08	13.25	16.90	15.91	14.00	17.81	13.91	12.33	15.50	11.97	10.56	13.37	18.17	15.82	20.52
SeViLA (Yu et al., 2023)+ReferFormer	20.37	18.99	21.75	21.29	19.86	22.72	18.27	17.06	19.48	17.56	16.37	18.75	24.01	22.35	25.66
SeViLA (Yu et al., 2023)+SgMg	22.34	20.92	23.75	21.60	20.17	23.02	21.68	20.47	22.89	20.46	19.32	21.61	24.53	22.83	26.22
SeViLA (Yu et al., 2023)+HTR	21.75	19.85	23.64	22.28	20.42	24.15	20.37	18.60	22.14	19.70	18.03	21.36	24.41	22.11	26.71
Video Reasoning Segmentation Baseline															
PG-Video-LLaVA (Munasinghe et al., 2023)	11.17	11.47	10.89	10.57	10.82	10.33	10.93	11.15	10.71	10.36	10.37	10.34	12.51	13.04	11.99
PG-Video-LLaVA+SAM2 (Ravi et al., 2024)	11.85	11.15	12.55	13.35	12.84	13.85	3.85	2.49	5.22	15.52	15.16	15.89	12.11	11.37	12.85
VISA (Yan et al., 2024) (Chat-UniVi-13B)	5.31	4.72	5.90	6.07	5.50	6.64	4.48	3.80	5.15	5.10	4.34	5.86	5.39	5.02	5.77
MORA (Deng et al., 2025) (LLaVA-13B)	23.13	22.79	23.46	21.62	21.46	21.79	22.43	22.03	22.83	26.37	25.87	26.88	22.63	22.30	22.97
RL-based VRS Methods															
Veason-R1-3B (Gong et al., 2026) (Qwen2.5-VL-3B)	23.53	22.03	25.04	23.73	22.04	25.43	21.06	19.80	22.31	22.78	21.50	24.06	26.51	24.77	28.26
VideoSEG-O3-2B (Dai et al., 2026) (Qwen3-VL-2B)	29.06	27.39	30.72	30.93	29.21	32.65	26.97	25.42	28.52	22.04	20.62	23.46	35.49	33.54	37.44
VideoSEG-O3-4B (Dai et al., 2026) (Qwen3-VL-4B)	31.96	30.19	33.73	34.68	32.88	36.47	29.71	28.14	31.28	25.06	23.51	26.60	37.39	35.21	39.56
TCoVP-3B (Ours) (Qwen2.5-VL-3B)	53.67	52.57	54.76	52.95	51.75	**54.15**	57.04	56.06	58.02	50.82	49.86	51.78	53.86	52.64	55.08
TCoVP-4B (Ours) (Qwen3-VL-4B)	55.21	54.40	56.01	52.46	51.67	53.26	59.65	58.87	60.42	**55.21**	**54.47**	**55.95**	54.18	53.27	55.09
TCoVP-M-4B (Ours) (Molmo2-4B)	**55.37**	**54.50**	**56.25**	**53.00**	**52.09**	53.92	**60.17**	**59.28**	**61.05**	54.51	53.78	55.24	**54.33**	**53.37**	**55.28**

Target primitive accuracy reward $\mathcal{R}_{acc}$. The final <answer> block selects a set of target primitives $\hat{\mathcal{P}}_{\text{ans}} = \{\hat{p}_i\}_{i=1}^{M_{\text{ans}}}$ from the generated temporal chain. We align these predictions with the ground-truth target primitives $\mathcal{P}_{\text{ans}}^{\text{gt}} = \{p_j^{\text{gt}}\}_{j=1}^{N_{\text{ans}}}$. For the i-th predicted primitive and the j-th ground-truth primitive, we define their temporal and spatial matching scores as:

$$T_{i,j} = \text{tIoU}\left(\hat{\mathcal{T}}_i, \mathcal{T}_j^{\text{gt}}\right), \qquad S_{i,j} = \text{IoU}\left(\hat{b}_i, b_j^{\text{gt}}(\hat{\tau}_i)\right), \tag{13}$$

where $\hat{\mathcal{T}}_i$ and $\mathcal{T}_j^{\text{gt}}$ denote the predicted and ground-truth temporal spans, $\hat{b}_i$ is the predicted bounding box, and $b_j^{\text{gt}}(\hat{\tau}_i)$ is the ground-truth box evaluated at the keyframe timestamp closest to the predicted timestamp $\hat{\tau}_i$. These scores are aggregated into a total pairwise reward matrix $R_{i,j}$ using source-specific weights w_t and w_s to balance temporal localization and spatial precision:

$$R_{i,j} = w_t T_{i,j} + w_s S_{i,j}, \qquad \mathcal{R}_{acc}(x, y) = \frac{1}{\max(M_{\text{ans}}, N_{\text{ans}})} \sum_{(i,j)\in O^*} R_{i,j}. \tag{14}$$

The optimal one-to-one assignment O^* is obtained by maximizing the total pairwise reward matrix R using the Hungarian algorithm. Normalizing by $\max(M_{\text{ans}}, N_{\text{ans}})$ penalizes both hallucinated target primitives and missed ground-truth targets. For zero-target cases, the reward receives the maximum score when both the predicted and ground-truth target sets are empty, and receives zero when only one of them is empty.

3.4 Primitive-to-Mask Execution

At inference, the target primitives in the final <answer> block define the mask prompts. The box is mapped to the corresponding video frame and passed to SAM3 (Carion et al., 2025), which propagates the initialized mask in both temporal directions. Masks initialized from multiple primitives are merged by pixel-wise union. General VRS benchmarks retain masks over the full video, whereas GroundMoRe restricts the output masks to the predicted query-relevant action interval. If the final answer contains no valid target primitive, the model produces an empty mask trajectory.

4 Experiments

4.1 Implementation Details

We instantiate TCoVP with Qwen2.5-VL-3B (Bai et al., 2025b), Qwen3-VL-4B (Bai et al., 2025a), and Molmo2-4B (Clark et al., 2026). All models use 16 uniformly sampled frames, one SFT epoch, two reinforcement-learning epochs, and eight sampled responses per query. The Qwen models use LoRA (Hu et al., 2022) for SFT with a learning rate of 4×10^{-5} and a reinforcement-learning rate of

Table 2: Performance comparison on Video Reasoning Segmentation benchmarks. Best results are **bolded**, and the second best results are underlined.

Methods	MLLM	Long-RVOS			ReasonVOS			ReVOS			MeViS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SFT-based Methods													
VISA-7B (Yan et al., 2024)	Chat-UniVi-7B	-	-	-	-	-	-	46.9	44.9	49.0	43.5	40.7	46.3
VISA-13B (Yan et al., 2024)	Chat-UniVi-13B	-	-	-	-	-	-	50.9	48.8	52.9	44.5	41.8	47.1
VideoLISA-3.8B (Bai et al., 2024)	LLaVA-Phi-3-V	14.0	-	-	47.5	45.1	49.9	-	-	-	44.4	41.3	47.6
InstructSeg-3B (Wei et al., 2025)	Mipha-3B	-	-	-	-	-	-	54.5	52.0	56.9	-	-	-
HyperSeg-3B (Wei et al., 2024)	Mipha-3B	-	-	-	-	-	-	55.7	53.1	58.4	-	-	-
GLUS-7B (Lin et al., 2025)	LLaVA-7B	25.7	-	-	49.9	47.5	52.4	54.8	52.4	57.3	51.3	48.5	54.2
VRS-HQ-7B (Gong et al., 2025)	Chat-UniVi-7B	-	-	-	-	-	-	59.1	56.6	61.6	50.6	47.6	53.7
VRS-HQ-13B (Gong et al., 2025)	Chat-UniVi-13B	-	-	-	-	-	-	60.0	57.6	62.5	50.9	48.0	53.7
RL-based Methods													
Omni-R1-8B (Zhong et al., 2025)	Qwen2.5-Omni-7B	-	-	-	-	-	-	56.2	-	-	-	-	-
Videoseg-R1-3B (Xu et al., 2026)	Qwen2.5-VL-3B	-	-	-	-	-	-	58.6	56.4	60.8	53.0	50.7	55.3
ReVSeg-7B (Li et al., 2025)	Qwen2.5-VL-7B	-	-	-	64.8	61.8	67.7	62.1	59.3	65.0	59.8	56.1	63.4
RCoT-Seg-3B (Wen et al., 2026)	Qwen2.5-VL-3B	-	-	-	58.2	54.6	61.8	61.2	58.2	64.3	53.7	50.7	56.7
Veason-R1-3B (Gong et al., 2026)	Qwen2.5-VL-3B	51.8	50.8	52.7	55.2	51.8	58.5	59.9	56.9	62.8	51.2	48.2	54.2
VideoSEG-O3-2B (Dai et al., 2026)	Qwen3-VL-2B	54.8	-	-	60.2	-	-	64.8	62.0	67.5	55.6	52.5	58.6
VideoSEG-O3-4B (Dai et al., 2026)	Qwen3-VL-4B	57.4	-	-	62.9	-	-	67.7	65.0	70.4	60.0	57.0	63.0
TCoVP-3B (Ours)	Qwen2.5-VL-3B	65.6	64.7	66.5	71.6	68.8	74.4	65.3	62.8	67.9	57.3	54.2	60.5
TCoVP-4B (Ours)	Qwen3-VL-4B	65.1	64.5	65.7	**72.1**	**69.1**	**75.0**	68.0	65.3	70.7	59.5	56.3	62.6
TCoVP-M-4B (Ours)	Molmo2-4B	**67.4**	**66.8**	**67.9**	71.5	68.5	74.5	**68.3**	**65.7**	**70.9**	**65.2**	**62.4**	**67.9**

5×10^{-7}. The Molmo2 models use full-parameter SFT at 2×10^{-6} and reinforcement learning at 2×10^{-7}. Target primitives in the final answer block provide timestamped bounding-box prompts to SAM3 for video mask propagation.

4.2 Evaluation Datasets and Metrics

We evaluate on Long-RVOS (Liang et al., 2026), ReasonVOS (Bai et al., 2024), ReVOS (Yan et al., 2024), MeViS (Ding et al., 2023), and the motion-grounded GroundMoRe benchmark (Deng et al., 2025). Following prior VRS protocols (Yan et al., 2024), we report region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their average $\mathcal{J}\&\mathcal{F}$. The first four benchmarks test target resolution from implicit video queries, while GroundMoRe additionally requires the model to distinguish the action interval that satisfies the query from the longer interval in which an entity is merely visible.

4.3 Main Results

Motion-grounded VRS. GroundMoRe evaluates the temporal component of the video reference gap, as the target is defined by an action or relation that holds over a bounded interval. As shown in Table 1, TCoVP-M-4B achieves the highest overall $\mathcal{J}\&\mathcal{F}$ of 55.37, followed by TCoVP-4B at 55.21 and TCoVP-3B at 53.67. All three variants outperform the strongest listed prior VRS baseline, which scores 31.96. These results demonstrate the effectiveness of TCoVP on queries that require identifying the target together with its relevant action interval.

General VRS performance. Table 2 shows that TCoVP performs well across multiple MLLM backbones. The Molmo2-4B variant achieves the highest listed $\mathcal{J}\&\mathcal{F}$ scores on Long-RVOS, ReVOS, and MeViS, reaching 67.4, 68.3, and 65.2, respectively. On ReasonVOS, the Qwen3-VL-4B variant leads with 72.1, followed by Qwen2.5-VL-3B at 71.6 and Molmo2-4B at 71.5. All three variants consistently outperform the strongest listed prior baseline on Long-RVOS and ReasonVOS, with gains of 7.7 to 10.0 and 6.7 to 7.3 points, respectively. The consistent gains on Long-RVOS and ReasonVOS across all three backbones support the broad applicability of TCoVP.

4.4 Ablation Studies

All ablation studies below use TCoVP-3B instantiated with Qwen2.5-VL-3B (Bai et al., 2025b) to examine the reasoning representation, the reinforcement-learning stage, and the temporal alignment and spatiotemporal grounding rewards. Additional experiments are provided in Appendix A.2.

TCoVP versus text-only CoT. Table 3 keeps the data volume, training configuration, and backbone fixed while varying the reasoning representation. TCoVP raises the five-benchmark macro average from 57.8 to 62.7, with a gain of 7.4 points on GroundMoRe and gains ranging from 2.5 to 6.4 points

Table 3: Ablation of the reasoning representation. All entries are $\mathcal{J}\&\mathcal{F}$. The macro average uses five benchmarks once each.

Model	ReVOS	ReasonVOS	Long-RVOS	GroundMoRe	MeViS	Macro avg.
Veason-R1-3B	59.9	55.2	51.8	23.5	51.2	48.3
Text-CoT with answer-VP	62.0	65.2	60.5	46.3	54.8	57.8
TCoVP-3B	**65.3**	**71.6**	**65.6**	**53.7**	**57.3**	**62.7**

Table 4: Ablation of chain-aligned reinforcement learning and the temporal alignment and spatiotemporal grounding rewards.

Model	Long-RVOS			ReasonVOS			ReVOS			GroundMoRe		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Different Training Strategies												
SFT only	56.8	56.2	57.4	63.1	60.3	65.9	59.9	57.1	62.7	43.8	42.5	45.0
SFT + chain-aligned RL	**65.6**	**64.7**	**66.5**	**71.6**	**68.8**	**74.4**	**65.3**	**62.8**	**67.9**	**53.7**	**52.6**	**54.8**
Different Reward Functions												
w/o $\mathcal{R}_{ta}$ and $\mathcal{R}_{stg}$	62.5	62.1	62.9	67.4	64.5	70.3	63.1	60.4	65.8	47.9	46.8	49.1
w/o $\mathcal{R}_{stg}$	63.2	62.6	63.7	65.6	62.7	68.4	63.5	60.8	66.2	49.2	48.0	50.4
w/o $\mathcal{R}_{ta}$	64.0	63.6	64.4	68.5	65.7	71.4	63.9	61.3	66.6	48.4	47.4	49.5
TCoVP-3B	**65.6**	**64.7**	**66.5**	**71.6**	**68.8**	**74.4**	**65.3**	**62.8**	**67.9**	**53.7**	**52.6**	**54.8**

on the other benchmarks. Relative to Veason-R1, the macro average increases from 48.3 to 62.7. Incorporating explicit temporal and spatial references into intermediate reasoning, together with their grounding supervision, improves segmentation under the same backbone and data budget.

Effect of chain-aligned reinforcement learning. Adding chain-aligned reinforcement learning to SFT consistently improves performance across all four benchmarks, yielding gains of 8.8, 8.5, 5.4, and 9.9 points on Long-RVOS, ReasonVOS, ReVOS, and GroundMoRe, respectively. Chain-aligned reinforcement learning therefore provides substantial gains beyond supervised fine-tuning, with the largest improvement on GroundMoRe.

Effect of temporal alignment and spatiotemporal grounding rewards. Removing either $\mathcal{R}_{ta}$ or $\mathcal{R}_{stg}$ decreases performance across all four benchmarks. Using both rewards yields the best performance across all four benchmarks, indicating that each contributes to the final segmentation quality.

5 Limitations

TCoVP reasons over 16 uniformly sampled frames, so brief events may occur between observations and remain unobserved. Although the temporal primitive chain provides explicit entity references and spatiotemporal evidence, cross-entity actions and relations are expressed in natural language without a dedicated structured representation. The final masks also inherit errors from SAM3 propagation. Extending TCoVP to more complex video scenarios may require adaptive temporal sampling, denser spatial evidence, and structured representations of interactions.

6 Conclusion

We introduce TCoVP, a reasoning framework that extends primitive-based visual reasoning to videos through a temporal chain of video primitives. The chain accumulates query-relevant evidence in chronological order, while repeated entity identifiers link observations across time. Each primitive provides explicit temporal support and spatial grounding, making intermediate reasoning structured and verifiable. Interleaving these primitives with language enables autoregressive MLLMs to generate grounded reasoning traces, whose final target primitives provide timestamped bounding-box prompts to SAM3 for mask propagation. Improvements across multiple backbones and benchmarks, particularly on motion-grounded queries, support the value of maintaining entity continuity and explicit temporal and spatial references throughout video reasoning.

AI USE STATEMENT

Generative AI systems were used to produce the initial temporal-chain annotations described in Figure 2 and to assist with manuscript language editing; structural checks and visual verification filter the generated records. The authors reviewed the resulting data, text, claims, and citations and take responsibility for the final content.

ETHICS STATEMENT

This study uses existing video segmentation datasets and introduces no new human-subject data collection, but the derived annotations may inherit errors and social biases from the source datasets and annotation models. Any release of derived data will retain source attribution and follow the licenses and redistribution terms of the underlying datasets.

REPRODUCIBILITY STATEMENT

Section 3 describes the temporal primitive representation, data construction pipeline, training objectives, and primitive-to-mask inference procedure. Sections 4.1 and 4.2 specify the model backbones, main training settings, evaluation datasets, and metrics. Appendix A.1 provides additional details on dataset composition, annotation prompts, task-specific system prompts, optimization settings, reward hyperparameters, and inference configurations. Appendix A.2 reports ablations and oracle analyses that examine the effects of mask propagation backends and intermediate spatial and temporal grounding.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Zechen Bai, Tong He, Haiyang Mei, Pichao Wang, Ziteng Gao, Joya Chen, Lei Liu, Zheng Zhang, and Mike Z Shou. One token to seg them all: Language instructed reasoning segmentation in videos. *Advances in Neural Information Processing Systems*, 37:6833–6859, 2024.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Ming Dai, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. Momentseg: Moment-centric sampling for enhanced video pixel understanding. *arXiv preprint arXiv:2510.09274*, 2025.

Ming Dai, Sen Yang, Boqiang Duan, Boyuan Tong, Jiedong Zhuang, Wankou Yang, and Jingdong Wang. Videoseg-o3: A multi-turn reinforcement learning framework for reasoning video object segmentation. In *ICML*, 2026.

Andong Deng, Tongjia Chen, Shoubin Yu, Taojiannan Yang, Lincoln Spencer, Yapeng Tian, Ajmal Saeed Mian, Mohit Bansal, and Chen Chen. Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8625–8636, 2025.

Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. Mevis: A large-scale benchmark for video segmentation with motion expressions. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2694–2703. IEEE, 2023.

Yue Fan, Xuehai He, Diji Yang, Kaizhi Zheng, Ching-Chen Kuo, Yuting Zheng, Sravana Jyothi Narayanaraju, Xinze Guan, and Xin Eric Wang. Grit: Teaching mllms to think with images. *arXiv preprint arXiv:2505.15879*, 2025.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Zongxin Yang, Pingping Zhang, and Huchuan Lu. 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.

Sitong Gong, Yunzhi Zhuge, Lu Zhang, Jiazuo Yu, Xu Jia, Pingping Zhang, and Huchuan Lu. Reinforcing video reasoning segmentation to think before it segments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xin Gu, Haoji Zhang, Qihang Fan, Jingxuan Niu, Zhipeng Zhang, Libo Zhang, Guang Chen, Fan Chen, Longyin Wen, and Sijie Zhu. Thinking with bounding boxes: Enhancing spatio-temporal video grounding via reinforcement fine-tuning. *arXiv preprint arXiv:2511.21375*, 2025.

Jihwan Hong and Jaeyoung Do. Virst: Video-instructed reasoning assistant for spatio-temporal segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Yifan Li, Yingda Yin, Lingting Zhu, Weikai Chen, Shengju Qian, Xin Wang, and Yanwei Fu. Revseg: Incentivizing the reasoning chain for video segmentation with reinforcement learning. *arXiv preprint arXiv:2512.02835*, 2025.

Tianming Liang, Haichao Jiang, Yuting Yang, Chaolei Tan, Shuai Li, Wei-Shi Zheng, and Jian-Fang Hu. Long-rvos: A comprehensive benchmark for long-term referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39497–39507, 2026.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Lang Lin, Xueyang Yu, Ziqi Pang, and Yu-Xiong Wang. Glus: Global-local reasoning unified into a single large language model for video segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 8658–8667, 2025.

Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Chen. Unipixel: Unified object referring and segmentation for pixel-level visual reasoning. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 126078–126108, 2025.

Ruijie Lu, Yiyang Ma, Xiaokang Chen, Lingxiao Luo, Zhiyu Wu, Zizheng Pan, Xingchao Liu, Yutong Lin, Hao Li, Wen Liu, et al. Thinking with visual primitives. Technical report, DeepSeek-AI, 2026.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, et al. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025.

Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023.

OpenAI. Openai-o3. https://openai.com/index/introducing-o3-and-o4-mini/, 2025.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Urvos: Unified referring video object segmentation network with a large-scale benchmark. In *European conference on computer vision*, pp. 208–223. Springer, 2020.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Youngkil Song, Yoonjae Baek, Dongwon Kim, Inho Kim, Dongkeun Kim, and Suha Kwak. Event-cot: Event-centric video chain-of-thought for reasoning temporal localization. *arXiv preprint arXiv:2607.04872*, 2026.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Ziwei Chai, Y Charles, HS Che, Cheng Chen, et al. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, et al. Traceable evidence enhanced visual grounded reasoning: Evaluation and methodology. *arXiv preprint arXiv:2507.07999*, 2025a.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, et al. Vgr: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025b.

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

Cong Wei, Yujie Zhong, Haoxian Tan, Yingsen Zeng, Yong Liu, Hongfa Wang, and Yujiu Yang. Instructseg: Unifying instructed visual segmentation with multi-modal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20193–20203, 2025.

Junwei Wen, Deshui Miao, Guangming Lu, Xin Li, and Wenjie Pei. Rcot-seg: Reinforced chain-of-thought for video reasoning and segmentation. *arXiv preprint arXiv:2605.07334*, 2026.

Zishan Xu, Yifu Guo, Yuquan Lu, Fengyu Yang, Junxin Li, and Lihua Cai. Videoseg-r1: reasoning video object segmentation via reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11496–11504, 2026.

Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pp. 98–115. Springer, 2024.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. *Advances in Neural Information Processing Systems*, 36:76749–76771, 2023.

Haobo Yuan, Xiangtai Li, Tao Zhang, Yueyi Sun, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, et al. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025.

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

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing" thinking with images" via reinforcement learning. *arXiv preprint arXiv:2505.14362*, 2025.

Hao Zhong, Muzhi Zhu, Zongze Du, Zheng Huang, Canyu Zhao, Mingyu Liu, Wen Wang, Hao Chen, and Chunhua Shen. Omni-r1: Reinforcement learning for omnimodal reasoning via two-system collaboration. *arXiv preprint arXiv:2505.20256*, 2025.

A APPENDIX

The appendix is organized as follows:

1. We provide additional implementation details on dataset construction, training, and inference.

2. We present additional experiments on mask propagation backends, oracle analysis, intermediate spatial and temporal grounding, and conventional referring VOS performance.

3. More visualizations and failure cases are shown.

A.1 MORE IMPLEMENTATION DETAILS

A.1.1 DATA CONSTRUCTION

As shown in Figure 5, we use a structured prompt to guide Kimi-K2.5 (Team et al., 2026) in generating spatiotemporally grounded annotations. Along with the sampled video frames, we provide the video duration, the question, the number of distinct target objects, the annotated bounding boxes for each target in each sampled frame, and the expected answer with object descriptions omitted. The prompt then directs the model to select keyframes, identify target and contextual objects, estimate their temporal intervals, generate grounded and text-only reasoning traces, and add object references to the final answer, while enforcing the specified formatting and consistency constraints. The model returns these annotations in a prescribed JSON schema containing the completed answer, keyframes and associated object boxes, temporal intervals, and both reasoning traces. For zero-target samples, the prompt explicitly states that no object in the video satisfies the query. For GroundMoRe (Deng et al., 2025) samples, it further specifies that the temporal intervals of target objects should be restricted to the duration of the query-relevant action, whereas those of non-target objects retain their original definition as the periods during which the objects appear in the video.

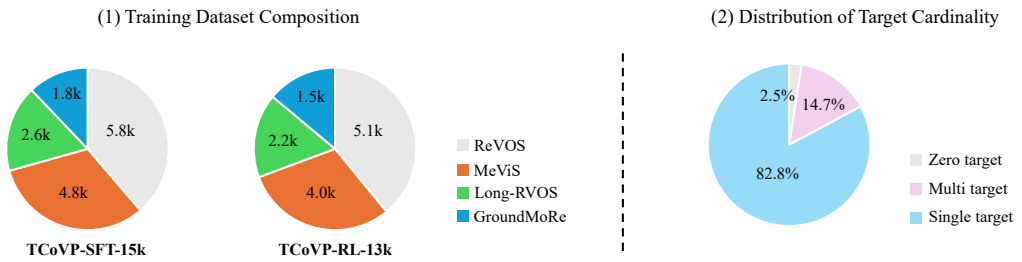

Figure 4: Composition of the TCoVP training datasets

Figure 4 summarizes the composition of the resulting datasets. TCoVP-SFT-15k contains approximately 5.8k samples from ReVOS (Yan et al., 2024), 4.8k from MeViS (Ding et al., 2023), 2.6k from Long-RVOS (Liang et al., 2026), and 1.8k from GroundMoRe (Deng et al., 2025), while TCoVP-RL-13k contains approximately 5.1k, 4.0k, 2.2k, and 1.5k samples from the same four datasets, respectively. In terms of target cardinality, single-target instances account for 82.8% of the data, followed by multi-target instances at 14.7% and zero-target instances at 2.5%. This composition reflects the predominance of conventional single-target video segmentation while retaining multi-target and zero-target cases to provide supervision for relational reasoning and the recognition of queries without a valid target.

A.1.2 TRAINING SETTINGS

Training proceeds in two stages. The first stage performs supervised fine-tuning to initialize the model with the canonical TCoVP serialization and grounded reasoning format. The second stage applies chain-aligned reinforcement learning to optimize structural validity, temporal alignment, spatiotemporal grounding, and target primitive accuracy. The visual input is constrained by a maximum image-pixel budget of $128 \times 28 \times 28$ for Qwen2.5-VL-3B (Bai et al., 2025b) and $128 \times 32 \times 32$ for Qwen3-VL-4B (Bai et al., 2025a) and Molmo2-4B (Clark et al., 2026) in both stages.

Prompt template for Kimi K2.5:

```
The video has a duration of {duration} seconds. The spatial grounding annotation for the video is as follows:
Question: {question}
Distinct target object(s) number: {obj_num}. Annotated bounding boxes: {bboxes}
Answer: {answer}
Based on this annotation, please complete the following tasks:

1. Select at least ONE and at most FIVE keyframes.
- Each timestamp MUST be the same as the timestamp of one of the input video frames and be written as a float rounded to exactly one decimal place.

2. For each keyframe, include at least ONE and at most FIVE detected objects with corresponding bounding boxes.
- The detected objects are not only the target objects, but also objects associated with the target that will appear during the reasoning process(not provided, you need to detect).
- Each bounding box coordinates are normalized floats (rounded to exactly two decimal place) in [x_min, y_min, x_max, y_max] format.
- If a detected object is one of the target objects, its corresponding bounding box must match the provided annotated bounding box.

3. For all detected objects in the aforementioned keyframes, accurately identify the start_time and the end_time of their appearance.
- Object names must strictly match those provided in the keyframes.
- All object names must be identical. If multiple objects share the same name, add a numerical suffix to differentiate between them.

4. Generate a grounded reasoning process for solving this spatial grounding problem naturally using visual evidence (Assuming the question and the video are the only you know):
- Reasoning must use visual evidence(detected objects above should all be included and their name should be consistent) grounded in the video.
- Whenever referencing any object, person, or visual element, you MUST follow the following strict format: "<obj>object_name</obj><time>start_time - end_time</time><box>[x_min, y_min, x_max, y_max]</box>at<t>time_in_seconds</t>s"
- "<time>start_time - end_time</time>" represents the time period when the object appears in the video.
- Both the `time_in_seconds` and the box coordinates MUST be consistent with the info in the key frames.
- The reasoning must not exceed 300 words.

5. Generate a pure-text reasoning process for solving this spatial grounding problem naturally (Assuming the question and the video are the only you know) .

6. Add object names to the final answer in the same strict format: <obj>object_name</obj><time>start_time - end_time</time><box>[x_min, y_min, x_max, y_max]</box>at<t>time_in_seconds</t>s
- Just add the name of the object; the rest of the answer should remain unchanged.

Your output content must strictly follow the following JSON format below (with no additional text outside the JSON):
{
  "answer": "...",
  "key_frames": [
    {
      "timestamp": time in second,
      "items": {
        "object1_name": box1,
        "object2_name": box2,
      }
    },
    ...
  ],
  "appear_time": [
    {
      "object_name": "...",
      "start_time": time in second,
      "end_time": time in second
    },
    ...
  ],
  "reasoning_process": "...",
  "grounded_reasoning_process": "..."
}
```

Figure 5: Prompt template for spatiotemporally grounded annotation generation

System prompt for VRS tasks:

A conversation between user and assistant. The user provides a video and asks for specific target object(s), and the Assistant locates them. The assistant MUST first think about the reasoning process in the mind and then provide the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively. All reasoning must be grounded in visual evidence from the video. When you mention any related object, person, or specific visual element, you must strictly follow the following format: `<obj>object_name</obj><time>start_time - end_time</time><box>bounding_box</box>at<t>time_in_seconds</t>s`. Between the <answer> and </answer> tags, the assistant MUST provide all detected instance(s) of the target object(s). Each instance must strictly follow the format: `<obj>object_name</obj><time>start_time - end_time </time><box>bounding_box</box>at<t>time_in_seconds</t>s`. If multiple instances are found, they MUST be connected using the word 'and'.

System prompt for motion-grounded VRS task:

A conversation between user and assistant. The user provides a video and asks for a specific target object based on a specific action or interaction, and the Assistant locates it. The assistant MUST first think about the reasoning process in the mind and then provide the user with the answer. The reasoning process and answer are enclosed within <think> </think> and <answer> </answer> tags, respectively. All reasoning must be grounded in visual evidence from the video. CRITICAL TEMPORAL RULE: The <time> definition depends on the object's role: 1. For the TARGET object queried by the user, `<time>start_time - end_time</time>` MUST strictly represent the duration of the specific ACTION being queried. Do not use the full appearance time even if it is visible longer. 2. For NON-TARGET objects used as context in reasoning, `<time>start_time - end_time</time>` MUST represent their full appearance duration in the video. When you mention any related object, person, or specific visual element in the reasoning process, you must strictly follow the following format: `<obj>object_name </obj><time>start_time - end_time</time><box>bounding_box</box>at<t>time_in_seconds</t>s`, ensuring the `<time>` strictly follows the rule above based on whether it is a target or non-target object. Between the <answer> and </answer> tags, the assistant MUST provide the single detected target object. This object must strictly follow the format: `<obj>object_name</obj><time>start_time - end_time</time><box>bounding_box</box>at<t>time_in_seconds</t>s`, where the `<time>` MUST strictly be the action duration. Do not output multiple target objects or use 'and', as there is strictly only one target object for each query.

Figure 6: **Task-specific system prompts for TCoVP.** The conventional VRS prompt permits multiple target objects, whereas the GroundMoRe prompt restricts target temporal spans to the duration of the query-relevant action and requires a single target object.

During supervised fine-tuning, the LoRA rank and scaling factor are set to $r = 64$ and $\alpha = 128$, respectively. During reinforcement learning, the policy samples $G = 8$ completions per query. The KL coefficient is $\beta = 0.04$ for Qwen2.5-VL-3B and $\beta = 0.1$ for Qwen3-VL-4B and Molmo2-4B. The temporal-alignment reward uses a Gaussian tolerance of $\sigma = 1.0$, and the spatiotemporal-grounding reward uses a temporal matching gate of $\tau = 1s$. In the target primitive accuracy reward, we use source-specific weights to balance temporal and spatial matching. For datasets other than GroundMoRe, we set $w_t = 1.0$ and $w_s = 2.0$, emphasizing spatial precision. For GroundMoRe, we reverse these weights and set $w_t = 2.0$ and $w_s = 1.0$, emphasizing temporal localization. For response generation during reinforcement learning, we set the temperature to 1.0 and top-p to 0.95.

The task-specific system prompts are shown in Figure 6. For conventional VRS, the prompt requires the model to ground visual references with consistent object identities, appearance intervals, timestamps, and bounding boxes, and allows multiple target objects in the final answer. For GroundMoRe, the prompt requires the temporal span of the target object to represent the duration of the query-relevant action, while the span of each contextual non-target object covers its full appearance in the video. It also specifies that each query has a single target object and therefore prohibits multiple target outputs in the final answer.

A.1.3 Inference details

At inference time, we uniformly sample 16 frames from each video and apply the same model-specific image-pixel constraints as during training: $128 \times 28 \times 28$ for Qwen2.5-VL-3B and $128 \times 32 \times 32$ for Qwen3-VL-4B and Molmo2-4B. We use the same system prompt as in training and perform MLLM inference with vLLM to generate the TCoVP reasoning trace and final answer. We then parse the target primitives in the `<answer>` block and extract their timestamps and bounding boxes as prompts for SAM3 (Carion et al., 2025). Each timestamped box initializes a mask track on the corresponding video frame, and the resulting tracks are propagated through the video and merged by pixel-wise union. For GroundMoRe, mask propagation is restricted to the predicted query-relevant action interval; if the answer contains no valid target primitive, the model returns an empty mask trajectory.

A.2 Additional experiments

A.2.1 Mask propagation backends and oracle analysis

We first examine the effect of the mask propagation backend and oracle spatial prompts, as shown in Table 5. With SAM2 (Ravi et al., 2024), TCoVP-3B substantially outperforms Veason-R1-3B (Gong et al., 2026) across all four benchmarks, improving $\mathcal{J}\&\mathcal{F}$ by 12.9, 14.2, 4.3, and 30.6 points on Long-RVOS, ReasonVOS, ReVOS, and GroundMoRe, respectively. Under the SAM3 setting, TCoVP-3B consistently outperforms the SFT-only variant, yielding gains of 8.8, 8.5, 5.4, and 9.9 points on the same benchmarks. Replacing SAM2 with SAM3 provides modest improvements on the first three benchmarks, while producing comparable performance on GroundMoRe. The oracle rows replace the predicted boxes with ground-truth boxes while propagating masks over the full video. This substantially improves performance on Long-RVOS, ReasonVOS, and ReVOS, reaching $\mathcal{J}\&\mathcal{F}$ scores of 76.4, 87.3, and 85.6 with SAM3, indicating remaining room for improving spatial grounding. On GroundMoRe, however, the full-video oracle configuration performs below TCoVP-3B. This result arises because GroundMoRe requires restricting mask propagation to the query-relevant action interval; using the full video includes temporally irrelevant frames even when the spatial box is oracle-correct. Thus, the comparison highlights the complementary roles of accurate spatial prompts and temporally appropriate propagation.

Table 5: Comparison of SAM2 and SAM3 mask propagation with oracle spatial prompts.

Model	Long-RVOS			ReasonVOS			ReVOS			GroundMoRe		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SAM2-based Mask Propagation												
Veason-R1-3B	51.8	50.8	52.7	55.2	51.8	58.5	59.9	56.9	62.8	23.5	22.0	25.0
TCoVP-3B	**64.7**	**63.9**	**65.5**	**69.4**	**66.7**	**72.1**	**64.2**	**61.7**	**66.8**	**54.1**	**52.9**	**55.2**
SAM3-based Mask Propagation												
SFT only	56.8	56.2	57.4	63.1	60.3	65.9	59.9	57.1	62.7	43.8	42.5	45.0
TCoVP-3B	**65.6**	**64.7**	**66.5**	**71.6**	**68.8**	**74.4**	**65.3**	**62.8**	**67.9**	**53.7**	**52.6**	**54.8**
Oracle Test (Full video + oracle box)												
SAM2-oracle	73.8	73.4	74.2	83.6	81.0	86.3	83.9	81.7	86.2	47.1	45.8	48.5
SAM3-oracle	76.4	76.0	76.7	87.3	84.2	90.4	85.6	83.4	87.8	46.3	45.0	47.7

As illustrated in Table 6, we further conduct a fine-grained oracle analysis on GroundMoRe by independently replacing the predicted temporal intervals and bounding boxes with ground-truth

annotations. Under SAM2, restricting mask propagation from the full video to the predicted timespan improves the overall $\mathcal{J}\&\mathcal{F}$ score from 39.22 to 54.06 with predicted boxes, and from 47.14 to 60.41 with oracle boxes. Similar improvements are observed with SAM3, where the corresponding scores increase from 38.24 to 53.67 and from 46.32 to 60.10, respectively. These results show that restricting propagation to the query-relevant interval is essential for GroundMoRe. Replacing the predicted timespan with the oracle timespan produces a much larger improvement, increasing $\mathcal{J}\&\mathcal{F}$ from 54.06 to 85.60 with SAM2 and from 53.67 to 85.54 with SAM3, whereas replacing only the predicted box with an oracle box yields gains of 6.35 and 6.43 points, respectively. Thus, temporal localization represents the primary source of error, although spatial grounding also leaves measurable room for improvement. When both the timespan and the box are oracle, the performance reaches 92.97 with SAM2 and 93.10 with SAM3, indicating substantial headroom for improving the predicted grounding cues. The same trends hold across the causal, sequential, counterfactual, and descriptive subsets, and the close results obtained with SAM2 and SAM3 suggest that these conclusions are largely independent of the choice of propagation backbone.

Table 6: Fine-grained oracle analysis of temporal and spatial grounding on GroundMoRe.

Methods	Overall			Causal			Sequential			Counterfactual			Descriptive		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
SAM2-based Mask Propagation															
Full video + predicted box	39.22	37.80	40.65	40.74	39.21	42.27	36.22	34.90	37.54	35.40	34.09	36.70	43.99	42.49	45.50
Predicted timespan + predicted box	54.06	52.93	55.20	53.44	52.22	54.67	57.46	56.46	58.45	51.39	50.36	52.41	53.96	52.70	55.22
Oracle timespan + predicted box	85.60	84.18	87.03	86.83	85.30	88.37	85.85	84.53	87.17	84.66	83.35	85.96	84.70	83.19	86.20
Full video + **oracle** box	47.14	45.83	48.46	47.17	45.80	48.54	44.01	42.82	45.21	44.46	43.23	45.69	52.84	51.38	54.31
Predicted timespan + **oracle** box	60.41	59.39	61.43	58.56	57.43	59.68	63.46	62.63	64.29	58.18	57.26	59.11	61.78	60.61	62.95
Oracle timespan + **oracle** box	92.97	91.65	94.29	92.76	91.39	94.13	93.04	91.85	94.23	92.85	91.62	94.08	93.28	91.82	94.75
SAM3-based Mask Propagation															
Full video + predicted box	38.24	36.84	39.64	39.48	37.96	41.01	35.43	34.12	36.74	34.29	33.06	35.53	43.28	41.78	44.77
Predicted timespan + predicted box	53.67	52.57	54.76	52.95	51.75	54.15	57.04	56.06	58.02	50.82	49.86	51.78	53.86	52.64	55.08
Oracle timespan + predicted box	85.54	84.14	86.95	86.53	85.01	88.05	85.91	84.60	87.22	84.69	83.45	85.92	84.75	83.25	86.24
Full video + **oracle** box	46.32	44.98	47.66	46.28	44.92	47.64	43.02	41.79	44.26	43.49	42.22	44.76	52.42	50.92	53.91
Predicted timespan + **oracle** box	60.10	59.07	61.12	58.25	57.15	59.35	63.03	62.16	63.89	57.68	56.72	58.63	61.77	60.60	62.94
Oracle timespan + **oracle** box	93.10	91.76	94.44	92.87	91.51	94.23	93.09	91.86	94.33	92.83	91.57	94.10	93.66	92.16	95.15

A.2.2 Intermediate spatial grounding quality

We directly evaluate the quality of intermediate spatial grounding using bounding-box IoU. The think and answer columns measure the IoU of the boxes generated during the reasoning process and those included in the final answer, respectively. As shown in Table 7, compared with SFT-only training, chain-aligned reinforcement learning consistently improves both types of grounding across all four benchmarks. The gains are particularly clear on Long-RVOS, where the answer-box IoU increases from 54.9 to 64.3, and on GroundMoRe, where it increases from 52.6 to 59.4. The reward ablations further show that removing either the temporal-alignment reward $\mathcal{R}_{ta}$ or the spatiotemporal-grounding reward $\mathcal{R}_{stg}$ degrades the intermediate grounding quality, while removing both rewards generally leads to larger degradation. On GroundMoRe, removing $\mathcal{R}_{ta}$ reduces the answer-box IoU from 59.4 to 55.4, whereas removing $\mathcal{R}_{stg}$ reduces it to 57.4, indicating that temporal alignment is particularly important for the final grounding of motion-related queries. Overall, the complete TCoVP objective achieves the best results across all reported columns, demonstrating that chain-aligned reinforcement learning improves spatial references before mask propagation.

Table 7: Intermediate bounding-box IoU across four VRS benchmarks.

Model	Long-RVOS		ReasonVOS		ReVOS		GroundMoRe	
	think	answer	think	answer	think	answer	think	answer
Different Training Strategies								
SFT only	59.8	54.9	59.7	59.4	63.4	60.3	57.5	52.6
SFT + chain-aligned RL	**66.9**	**64.3**	**64.5**	**65.1**	**66.7**	**65.4**	**62.5**	**59.4**
Different Reward Functions								
w/o $\mathcal{R}_{ta}$ and $\mathcal{R}_{stg}$	63.6	60.8	59.9	61.6	65.1	62.9	59.0	53.8
w/o $\mathcal{R}_{stg}$	64.1	61.6	59.8	61.0	64.9	63.6	60.4	57.4
w/o $\mathcal{R}_{ta}$	65.0	62.1	61.9	62.2	66.4	63.8	60.6	55.4
TCoVP-3B	**66.9**	**64.3**	**64.5**	**65.1**	**66.7**	**65.4**	**62.5**	**59.4**

A.2.3 INTERMEDIATE TEMPORAL GROUNDING QUALITY ON GROUNDMORE

We evaluate temporal grounding on GroundMoRe by measuring the temporal intersection-over-union (tIoU) between the predicted and annotated query-relevant intervals, together with the downstream $\mathcal{J}\&\mathcal{F}$ score. As illustrated in Table 8, compared with SFT-only training, chain-aligned reinforcement learning improves the overall tIoU from 44.29 to 50.20 and the overall $\mathcal{J}\&\mathcal{F}$ score from 43.76 to 53.67. Improvements are consistently observed across all five query categories, with particularly large gains in segmentation quality for sequential and counterfactual queries. The reward ablations further demonstrate the importance of both temporal alignment and spatiotemporal grounding. Removing both rewards reduces the overall tIoU and $\mathcal{J}\&\mathcal{F}$ to 47.72 and 47.90, respectively. Removing only $\mathcal{R}ta$ causes a larger decline in overall tIoU than removing only $\mathcal{R}stg$, indicating that the temporal-alignment reward is especially important for accurately localizing the relevant action interval. However, both rewards contribute to the final segmentation quality, and their relative effects vary across query categories. The complete TCoVP objective achieves the best results across all reported metrics, confirming that accurate temporal grounding is closely related to improved mask prediction on motion-grounded queries.

Table 8: Temporal IoU and segmentation performance on GroundMoRe.

Model	Overall		Causal		Sequential		Counterfactual		Descriptive	
	tIoU	$\mathcal{J}\&\mathcal{F}$	tIoU	$\mathcal{J}\&\mathcal{F}$	tIoU	$\mathcal{J}\&\mathcal{F}$	tIoU	$\mathcal{J}\&\mathcal{F}$	tIoU	$\mathcal{J}\&\mathcal{F}$
Different Training Strategies										
SFT only	44.29	43.76	44.89	43.73	42.03	44.50	41.00	39.71	48.94	46.89
SFT + chain-aligned RL	**50.20**	**53.67**	**50.57**	**52.95**	**48.08**	**57.04**	**46.99**	**50.82**	**54.94**	**53.86**
Different Reward Functions										
w/o $\mathcal{R}_{ta}$ and $\mathcal{R}_{stg}$	47.72	47.90	48.22	46.50	44.74	51.13	44.23	43.24	53.42	50.85
w/o $\mathcal{R}_{stg}$	49.25	49.23	50.12	48.83	46.11	52.12	45.54	44.66	54.87	51.17
w/o $\mathcal{R}_{ta}$	47.92	48.42	48.90	47.35	44.89	54.09	43.70	44.02	53.77	48.23
TCoVP-3B	**50.20**	**53.67**	**50.57**	**52.95**	**48.08**	**57.04**	**46.99**	**50.82**	**54.94**	**53.86**

A.2.4 PERFORMANCE ON CONVENTIONAL REFERRING VOS BENCHMARKS

We further evaluate TCoVP on conventional referring video object segmentation benchmarks under a zero-shot transfer setting. Neither the training split of Ref-DAVIS17 (Pont-Tuset et al., 2017) nor that of Ref-YouTube-VOS (Seo et al., 2020) is used to train TCoVP. As shown in Table 9, under the zero-shot setting, TCoVP-4B achieves 81.7 $\mathcal{J}\&\mathcal{F}$ on Ref-DAVIS17 and 73.6 on Ref-YouTube-VOS, while TCoVP-M-4B obtains 80.0 and 73.2, respectively. TCoVP-4B achieves the best Ref-DAVIS17 result among all reported methods and remains competitive with VideoSEG-O3 on Ref-YouTube-VOS despite not using the target benchmark's training data. These results demonstrate the transferability of TCoVP to conventional referring VOS.

Table 9: Performance on conventional referring video object segmentation benchmarks. † denotes zero-shot transfer, where the training split of the corresponding evaluation benchmark is not used.

Model	MLLM	Ref-DAVIS17			Ref-YouTube-VOS		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Veason-R1-3B† (Gong et al., 2026)	Qwen2.5-VL-3B	71.2	68.3	74.2	66.6	64.9	68.2
VideoSEG-O3-2B (Dai et al., 2026)	Qwen3-VL-2B	<u>80.0</u>	76.0	<u>84.0</u>	70.5	68.3	72.7
VideoSEG-O3-4B (Dai et al., 2026)	Qwen3-VL-4B	79.4	75.5	83.2	**74.1**	<u>71.5</u>	**76.6**
TCoVP-3B† (Ours)	Qwen2.5-VL-3B	79.3	75.6	83.1	72.6	71.1	74.1
TCoVP-4B† (Ours)	Qwen3-VL-4B	**81.7**	**78.1**	**85.4**	<u>73.6</u>	**71.7**	<u>75.4</u>
TCoVP-M-4B† (Ours)	Molmo2-4B	<u>80.0</u>	<u>76.4</u>	83.7	73.2	71.4	75.0

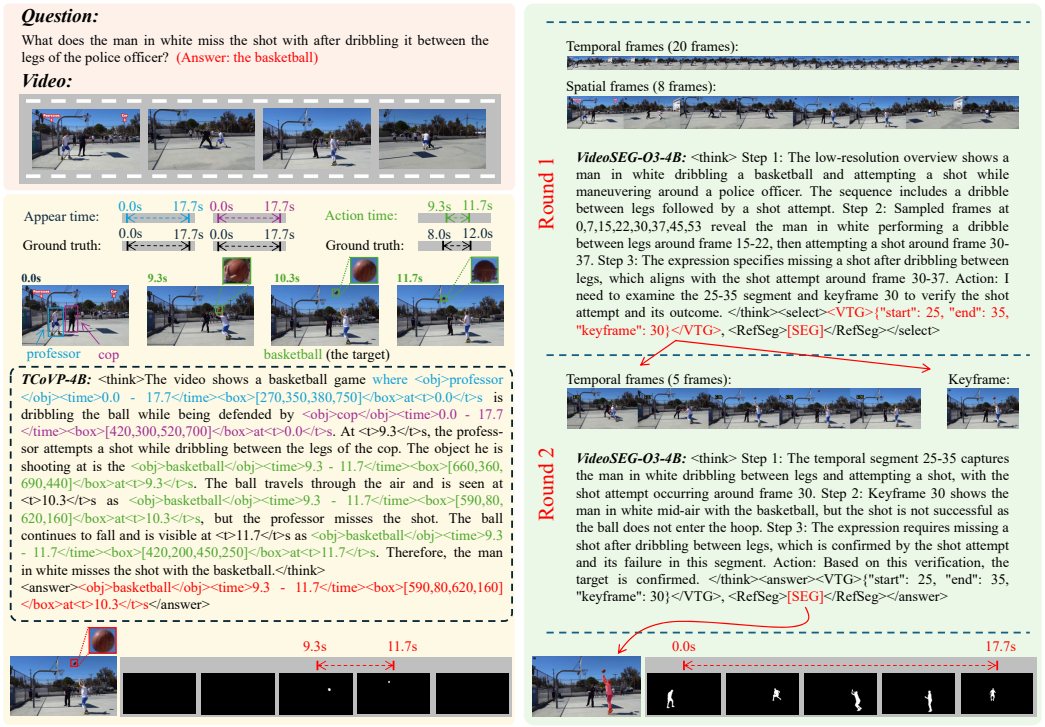

Figure 7: **Qualitative comparison with VideoSEG-O3-4B.** VideoSEG-O3-4B identifies the relevant shot segment but produces a mask that follows the player rather than the queried basketball. TCoVP-4B explicitly distinguishes the contextual participants from the basketball, grounds the target over the relevant interval, and provides timestamped bounding-box prompts for mask propagation.

A.3 More qualitative comparisons and failure case analysis

A.3.1 Comparison with VideoSEG-O3-4B

Figure 7 compares TCoVP-4B with VideoSEG-O3-4B (Dai et al., 2026) on a query that asks what the man in white misses after dribbling between the legs of the police officer. VideoSEG-O3-4B uses a two-round procedure: it first analyzes a coarse set of temporal and spatial frames to identify the relevant action segment, and then refines the segment using fewer frames and a selected keyframe. Although this procedure localizes the shot attempt, it does not explicitly represent the identity, temporal support, and spatial trajectory of the queried object. Its final mask is therefore not tightly aligned with the basketball. In contrast, TCoVP-4B grounds the contextual participants and the basketball as separate entities. It identifies the relevant interval as 9.3–11.7 seconds and provides timestamped bounding boxes for the basketball at multiple moments during the shot attempt. The resulting answer primitive directly refers to the basketball and supplies a temporally and spatially localized prompt for mask propagation. This example illustrates how explicit temporal chains and object-level grounding provide more precise evidence for queries whose target is an object involved in a short action.

A.3.2 Comparison with Veason-R1-3B

We next compare TCoVP-3B with Veason-R1-3B (Gong et al., 2026) on two representative cases. The first example in Figure 8 uses the query “the turtle in the water,” although no turtle appears in the video. Veason-R1-3B hallucinates a turtle, selects an unsupported keyframe, and produces an erroneous mask trajectory. In contrast, TCoVP-3B explicitly grounds the visible entities, including the water and multiple fish, while maintaining their appearance intervals. Based on this evidence, it correctly concludes that the queried target is absent and returns an empty answer. This example demonstrates that explicit object grounding helps the model distinguish between a plausible linguistic description and an object that is actually present in the video.

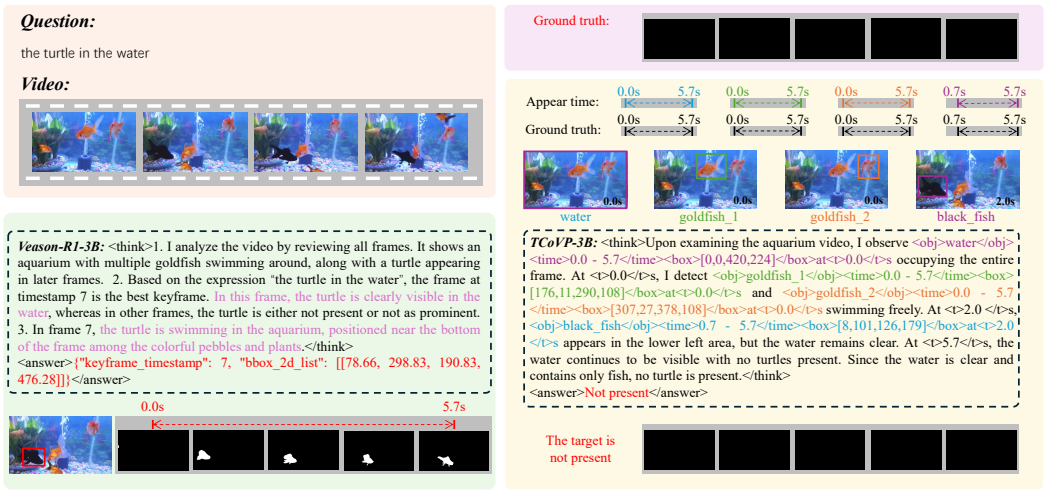

Figure 8: **Zero-target recognition.** TCoVP-3B identifies that the queried turtle is absent from the video, whereas Veason-R1-3B hallucinates the target and produces an erroneous mask trajectory.

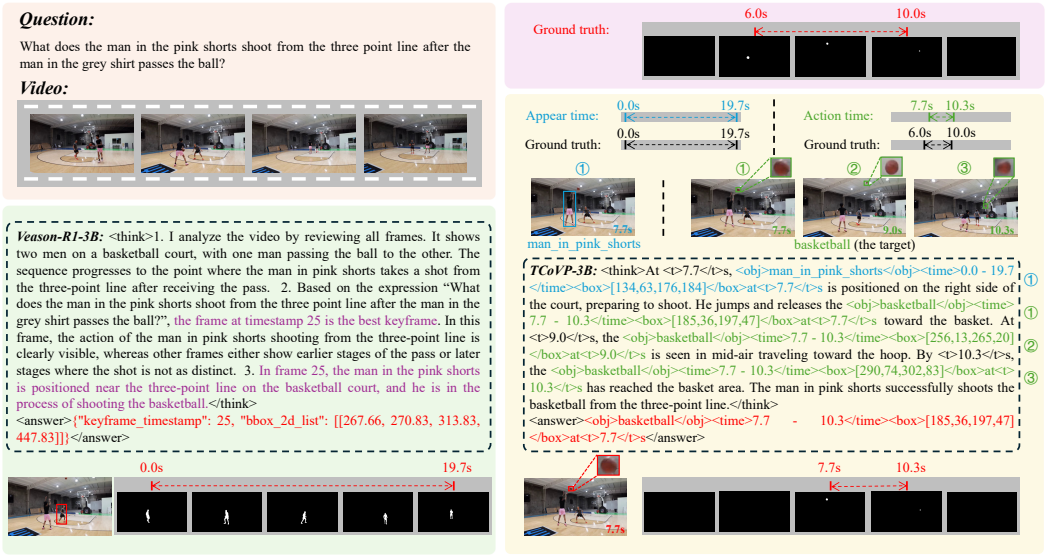

Figure 9: **Relational and action-based grounding.** TCoVP-3B tracks the basketball during the relevant action interval and answers the query correctly, while Veason-R1-3B incorrectly grounds the acting person.

The second example in Figure 9 asks what the man in pink shorts shoots after receiving a pass. Veason-R1-3B focuses on the acting person and outputs a box and mask for the man, rather than identifying the queried object. TCoVP-3B correctly distinguishes the person performing the action from the queried object. It identifies the man in pink shorts as contextual evidence and grounds the basketball during the relevant action interval from 7.7 to 10.3 seconds. The predicted temporal chain follows the ball from the player to the basket, leading to the correct answer, "basketball." This example shows how timestamped object references allow TCoVP to resolve relational queries whose target cannot be determined from the subject alone.

A.3.3 Failure Case Analysis

Figure 10 presents a representative target-selection failure. The query asks which player cannot be stopped by the man in white from scoring, and the ground-truth target is the man in red. TCoVP-3B

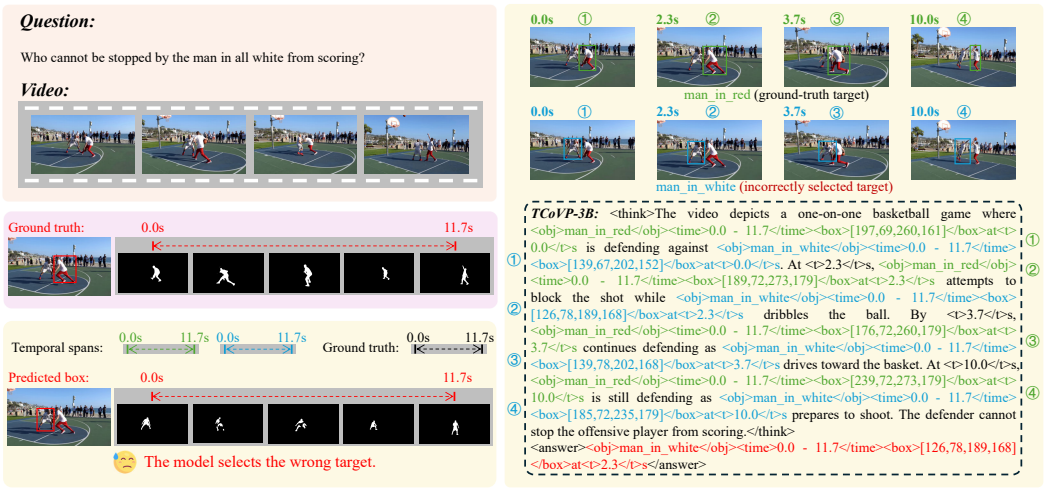

Figure 10: **Failure case caused by relational-role confusion.** TCoVP-3B correctly identifies and tracks both players but reverses their roles in the query, selecting the man in white instead of the ground-truth target, the man in red.

correctly identifies both players and tracks them over the full video, preserving their object identities, temporal spans, and spatial references. However, it assigns the action roles incorrectly: it interprets the man in red as the defender and the man in white as the offensive player. This role inversion causes the model to select "man_in_white" in the final answer, although the correct target is "man_in_red". The predicted mask therefore follows the wrong player throughout the video. This example shows that accurate object localization and temporal grounding do not always guarantee correct target selection when the query depends on fine-grained relational roles.