RESEARCH PREPRINT

BiSeek

Repurposing Generative Video–Language Models for Fine-Grained Perception

Deshui Miao[1] **Wenzhao Gao**[2] **Xingsen Huang**[3] **Xin Li**[4,5*] **Haijun Zhang**[1*] **Ming-Hsuan Yang**[6]

1 Harbin Institute of Technology, Shenzhen 2 Alibaba 3 Hainan University
4 Peng Cheng Laboratory 5 Pazhou Lab 6 University of California, Merced

* Corresponding authors

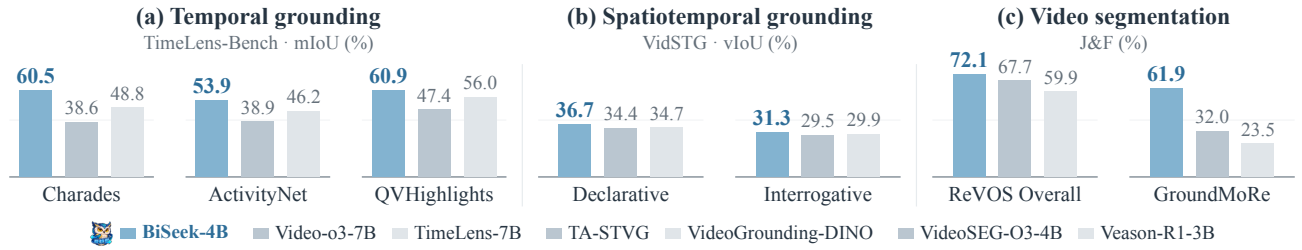

Figure 1. Fine-grained perception across three tasks. BiSeek-4B (blue) is compared with representative methods (gray) on (a) temporal grounding with TimeLens-Bench mIoU, (b) spatiotemporal grounding with video IoU on VidSTG declarative and interrogative queries, and (c) video segmentation with $\mathcal{J}\&\mathcal{F}$ on ReVOS Overall and GroundMoRe. Bar heights and labels show the original percentages.

ABSTRACT

Fine-grained video perception requires language understanding to be associated with specific times and regions. Generative video–language models provide semantic context through autoregressive response states, while their native video positions retain temporal and spatial indices. Under a video-first input order, however, causal attention prevents these visual positions from incorporating the subsequent instruction. We introduce BiSeek, which brings language conditioning into native video states through the joint design of bidirectional contextualization and structure-preserving readouts. The encoder allows video and query positions to integrate the complete input before prediction. The readouts use unit boundaries for temporal extent, time-addressed spatial patches for box geometry, and query–boundary conditions for mask-set prediction with independently encoded image features. All output times reuse one backbone encoding of the video–query pair, without an intermediate generated response. BiSeek achieves strong results on Charades-TimeLens and leads the listed comparisons on ReVOS and GroundMoRe. All models and code will be publicly released.

1 Introduction

Language-guided video perception links referring instructions to temporal intervals, box trajectories, and pixel masks. For "an adult in white clothes hugs a brown dog," these tasks locate the interaction, follow the person, and delineate the target. The query determines which visual evidence matters.

Generative video–language models connect semantic understanding to structured predictions through causal response generation. For temporal grounding, Time-R1 generates reasoning followed by event timestamps [36], and TimeLens outputs start and end timestamps as text [48]. For spatiotemporal grounding, STVG-R1 generates a target ID and time range to select a precomputed spatial track [49], while STVG-o1 autoregressively generates

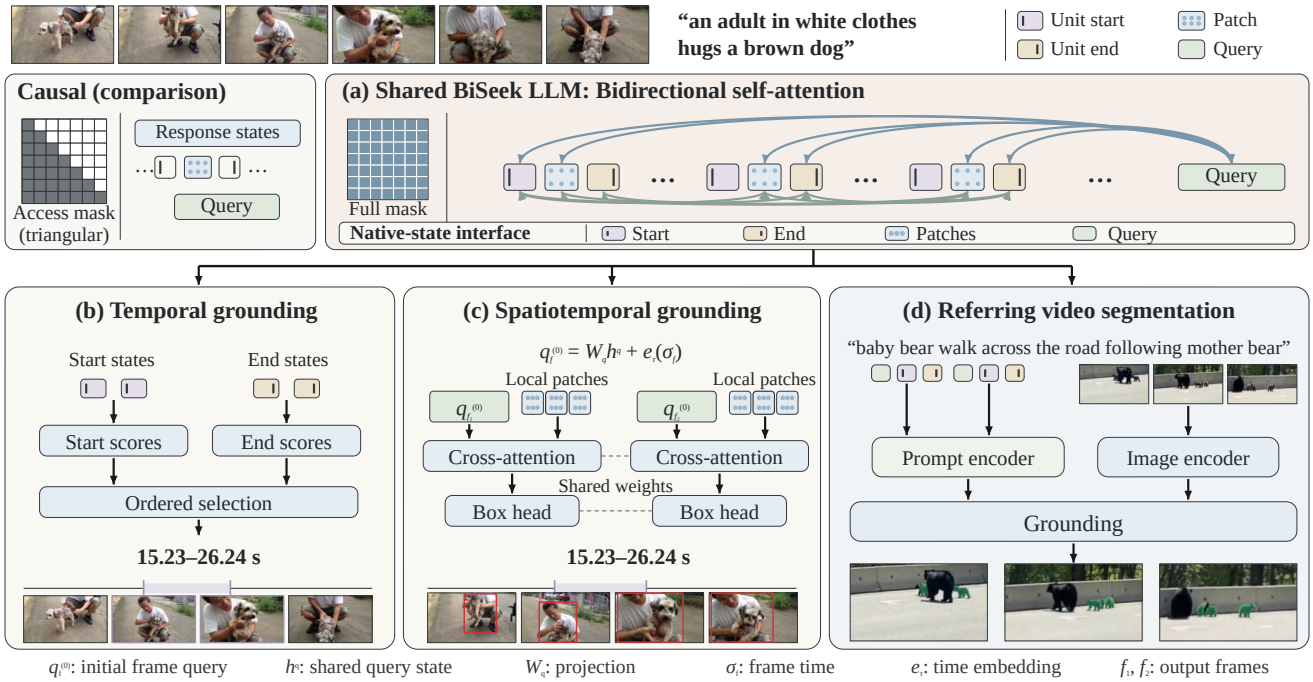

Figure 2. Overview of BiSeek. (a) Bidirectional video–query encoding supports (b) interval localization, (c) box-tube prediction, and (d) SAM 3 grounding and set decoding. BiSeek-2B examples: VidSTG (b,c) and MeViS (d).

intermediate box reasoning before a final temporally aligned box sequence [16]. For segmentation, VideoLISA uses response-token states [3], and SetCon conditions mask sets on generated concept states [52]. These methods route task predictions through generated responses.

Under video-first causal attention, native video positions cannot incorporate the subsequent instruction. Temporal queries such as "picking up a cup" and "handing it over" specify different intervals. Referring to "the giver" or "the receiver" selects different box trajectories, while "the receiver" and "both people" require different mask sets. Yet for a fixed sampled visual input, these query changes leave the native boundary and patch states unchanged. Response states integrate video and instruction, but index generated content rather than native times and patches. We therefore seek to contextualize native positions with the query before predicting each output.

We propose BiSeek, a bidirectional video–query encoder for fine-grained perception (Figure 2). Every valid video and query position can incorporate the complete input, bringing referring semantics and full-clip context to native temporal boundaries and spatial patches. These contextualized states retain their indices and serve as the shared interface for the three tasks.

Structure-preserving readouts recover the queried event from unit boundaries, the referred object's trajectory from time-addressed patches, and its masks from query–boundary conditions with independently encoded image features. All output times reuse one video–query encoding, without an intermediate generated response.

Experiments cover temporal grounding, spatiotemporal grounding, and referring video segmentation at 2B and 4B scales. BiSeek-4B reaches 60.5 mean temporal IoU on Charades-TimeLens, 36.7/31.3 mean video IoU on VidSTG declarative/interrogative queries, and 72.1/61.9 $\mathcal{J}\&\mathcal{F}$ on ReVOS/GroundMoRe. Controlled ablations examine encoding, readouts, and joint training. Figure 1 summarizes the comparisons.

Our main contributions are as follows:

- We propose BiSeek, a bidirectional video–query encoder that contextualizes native temporal and spatial states with the complete referring instruction.
- We introduce structure-preserving readouts that recover intervals, box tubes, and mask sets from unit boundaries, time-addressed patches, and query–boundary conditions.
- Extensive experiments and ablations across three tasks validate BiSeek, with strong results on Charades-TimeLens and leading scores among the listed methods on ReVOS and GroundMoRe.

2 Related Work

2.1 Generative models for perception

A common interface between MLLMs and dense prediction is to use language-model states as semantic conditions for visual decoders. For images, LISA introduces a generated segmentation token [20], while PixelLM extends this interface to multiple targets through a segmentation codebook and lightweight pixel decoder [31]. Video methods further organize these conditions across frames: VideoLISA learns a shared tracking-token state with multi-frame supervision [3], and Sa2VA connects language-model states to SAM 2 for image and video segmentation [45]. Object-level representations provide another way to maintain semantic conditions over time. UniPixel combines dedicated tokens with an object memory bank [27]; SetCon generates clip-level concept states through a causal LVLM and uses them to condition SAM 3 set prediction across frames [52].

Alongside these semantic interfaces, explicit reasoning guides evidence collection and alignment. VideoSEG-O3 gathers keyframe evidence through multi-turn temporal-spatial reasoning [8], while Veason-R1 uses reasoning supervision and rewards for temporal and spatial alignment [13]. These approaches motivate the use of generative models' semantic capabilities for fine-grained perception. BiSeek instead develops query-conditioned representations at native video positions, from which prediction heads directly recover temporal and spatial structure.

2.2 Bidirectional representation learning

Adapting a generative model for representation learning also changes the context available to its hidden states. LLM2Vec combines bidirectional attention with representation-oriented training [5], and NV-Embed removes the causal mask during contrastive training and introduces latent-attention pooling for text embeddings [21]. Both adapt decoder models to encode context for representation tasks.

For visual representations, Wu et al. [39] use linear probes and attention interventions to trace segmentation quality through the vision encoder, adapter, and LLM. Their bidirectional attention among image tokens alleviates limited context at early patch positions, while other token pairs retain causal masking. For a video followed by a referring query, causal attention also prevents native visual states from incorporating the query. BiSeek jointly contextualizes video and query positions, bringing the referring instruction into temporal boundaries and spatial patches before prediction.

2.3 Structured decoding for video grounding

Video grounding connects language to several output structures through different decoding mechanisms. For temporal intervals, Time-R1 and TimeLens develop training recipes for language-based timestamp prediction [36, 48]. For spatio-temporal tubes, TubeDETR couples temporal queries with time-aligned cross-attention [44], while CG-STVG refines localization by extracting and filtering visual context [14].

For video masks, MTTR decodes learned object queries against multimodal features [6], and ReferFormer initializes its queries from language features [40]. Promptable segmentation models offer complementary decoding capabilities: SAM 2 supports segmentation and propagation [30], while SAM 3 supports concept-conditioned object-set prediction [7]. BiSeek connects interval, tube, and mask-set prediction to a common query-conditioned encoding.

3 Method

3.1 Perception at native video positions

Given a video $\mathcal{V}$ and a referring instruction q, BiSeek predicts a task-specific interval, single-target box tube, or mask set from their joint encoding. We study offline perception, with the sampled video and complete instruction available at prediction time.

As shown in Figure 2, the Qwen3-VL encoder [1] jointly contextualizes native video and query positions. The selected readout uses unit boundaries for temporal extent, spatial patches for box geometry, or query–boundary conditions for mask-set prediction.

3.2 Bidirectional video–query contextualization

The processor groups sampled frames into temporal units, each containing a spatial visual-token block between vision-start and vision-end positions. Omitting chat delimiters, the input sequence is

$$X = [\tau_1, s_1, V_1, e_1, \ldots, \tau_T, s_T, V_T, e_T, Q], \tag{1}$$

where τ_t denotes the native timestamp representation of unit t, V_t is its spatial token block, and s_t, e_t are its existing vision delimiters. In readout formulas, τ_t denotes the corresponding clip-relative time in seconds. Q contains the complete query-bearing instruction. Each native unit groups two sampled context frames.

Causal attention makes the information available at these positions depend on their sequence order. With Q after the video, boundary and patch states cannot incorporate the instruction. We instead allow every valid input position to attend to every other valid position. For attention scores S, the attention matrix is

$$\begin{gathered} A = \operatorname{softmax}(S + M), \\ M_{ij} = 0 \text{ for valid } j, \qquad M_{ij} = -\infty \text{ for padding } j. \end{gathered} \tag{2}$$

Native positional encoding, visual processing, and padding masks are retained. The encoder E_θ, with parameters θ, maps the L-token input to normalized states of hidden width d:

$$H = E_\theta(X) \in \mathbb{R}^{L \times d}. \tag{3}$$

We retain each unit's N_p post-merge patch states $P_t \in \mathbb{R}^{N_p \times d}$ and gather its native boundary states. A shared query state is read at the end of the query-bearing user turn. For sequence indices p_t^s, p_t^e, p^q,

$$h_t^s = H_{p_t^s}, \qquad h_t^e = H_{p_t^e}, \qquad h^q = H_{p^q}. \tag{4}$$

3.3 Structure-preserving readouts

Temporal extent. Two trainable output-only rows of the language-model head, $w_s, w_e \in \mathbb{R}^d$, map the start and end boundary states h_t^s and h_t^e to endpoint logits. They are initialized from the language-head rows for start and end. The instruction influences these linear scores through the contextualized boundary states. With temperature ϑ,

$$\begin{aligned} \ell_t^s &= w_s^\top h_t^s, \quad p_s(t) = \operatorname{softmax}_t(\ell^s/\vartheta), \\ \ell_t^e &= w_e^\top h_t^e, \quad p_e(t) = \operatorname{softmax}_t(\ell^e/\vartheta). \end{aligned} \tag{5}$$

The two distributions are normalized independently over valid units. Order-constrained decoding selects a valid endpoint pair,

$$(\hat{i}, \hat{j}) = \arg\max_{i \le j} (\ell_i^s + \ell_j^e), \tag{6}$$

which is mapped to the selected temporal coordinates. If both endpoints select one unit, its temporal-cell bounds define a positive span. The same readout predicts event extent for temporal and spatiotemporal grounding.

Time-addressed box geometry. The box readout aligns each requested output frame f at time σ_f with the encoded temporal units and uses the shared query state to specify the referent. Let $u(f) = \arg\min_t |\sigma_f - \tau_t|$. The spatial memory includes every patch from the nearest temporal unit and each valid adjacent unit,

$$\mathcal{N}_f = \{u(f) - 1, u(f), u(f) + 1\} \cap \{1, \ldots, T\}, \tag{7}$$

The initial frame query $q_f^{(0)}$ uses a learned projection W_q and time encoding e_τ:

$$q_f^{(0)} = W_q h^q + e_\tau(\sigma_f). \tag{8}$$

For patch m in unit $t \in \mathcal{N}_f$, the learned projection W_p is combined with spatial, absolute-time, and relative-time encodings e_{xy}, e_τ, e_Δ:

$$\begin{aligned} z_{t,m}^{(f)} = W_p P_{t,m} + e_{xy}(x_m, y_m) \\ + e_\tau(\tau_t) + e_\Delta(\sigma_f - \tau_t), \end{aligned} \tag{9}$$

Patch centers $(x_m, y_m) \in [0,1]^2$ lie on the post-merge grid; absolute timestamps and relative offsets are normalized by clip duration. The spatial memory $\mathcal{Z}_f$ supplies cross-attention keys and values, followed by residual feed-forward updates,

$$q_f^{(r+1)} = \mathrm{Block}_r(q_f^{(r)}, \mathcal{Z}_f), \qquad r = 0, 1. \tag{10}$$

A box head maps the final query to a normalized box. All requested-time queries read the same backbone encoding and can be decoded in parallel. Their local memories retain access to full-video semantic context through the already contextualized patch states. At inference, the temporal readout proposes intervals. We rerank interval–tube pairs by temporal confidence, box continuity, and coverage.

Temporally indexed mask conditioning. Dense mask prediction combines semantic context from the video with spatial evidence from the current image. We form a condition for each native temporal unit by combining the shared query state with its two boundary states,

$$\begin{aligned} R_t &= [h^q; h_t^s; h_t^e] \in \mathbb{R}^{3\times d}, \\ C_t &= \phi(R_t) \in \mathbb{R}^{3\times 256}, \end{aligned} \tag{11}$$

The shared two-layer MLP ϕ, with ReLU activation, projects each row independently. It preserves three conditions per unit: shared referring context and two unit-specific states. SAM 3's prediction slots represent individual objects, allowing the condition length to stay fixed as the referred target count changes [7].

Grounding frames sampled in the context retain their native unit membership. For other frames I_f at time σ_f, nearest-unit assignment gives

$$u(f) = \arg\min_t |\sigma_f - \tau_t|, \qquad \widetilde{C}_f = C_{u(f)}, \tag{12}$$

The frame is independently encoded to supply its spatial evidence. The matched condition enters the SAM 3 grounding encoder alongside these image features, followed by set decoding:

$$\begin{aligned} Z_f &= E_{\mathrm{img}}(I_f), \\ \widehat{\mathcal{S}}_f &= D_{\mathrm{set}}\big(G(Z_f, \widetilde{C}_f)\big). \end{aligned} \tag{13}$$

The projected conditions supply the grounding encoder's language features. The set decoder predicts candidate masks, boxes, and confidence scores, which are filtered into the output target set. Video association and propagation connect predictions across frames while reusing the encoded semantic conditions.

3.4 Learning for structured perception

Training alternates task-homogeneous updates. Temporal examples supervise endpoint classification, interval overlap, and ordering. The continuous interval losses use soft endpoints $\bar{b}_s = \sum_t p_s(t)\tau_t$ and $\bar{b}_e = \sum_t p_e(t)\tau_t$, allowing gradients to reach both endpoint distributions. Tube examples additionally supervise box coordinates with L1 and generalized-IoU losses. Video and image segmentation examples use the SAM 3 set objective, with visible referred instances as the target set.

LoRA adapts the language backbone [18] together with the temporal output rows, tube readout, and condition projection. Mask supervision also trains the SAM 3 grounding transformer and prediction heads. The base language weights and the Qwen and SAM 3 visual backbones remain frozen. Each task updates the shared language adapters and its active output branch.

4 Experiments

4.1 Experimental setup

Implementation details. We jointly train each model variant on the same task mixture, initializing its language backbone from Qwen3-VL-Instruct [1] and segmentation branch from SAM 3 [7]. The 2B and 4B variants use identical optimization settings. We train for 30,210 updates on eight GPUs with global batch size 32 and bfloat16 precision. AdamW uses learning rate 4×10^{-5}, weight decay 0.05, and 906 warmup updates followed by cosine decay. LoRA [18] uses rank 64, scale 128, and dropout 0.05. Temporal and tube parameters use a 2.5× learning rate; SAM 3 grounding parameters use 0.25×. Updates alternate among RVOS, image segmentation, VTG, and STVG in a 30/10/30/30 ratio. Localization samples complete clips at 2 FPS. RVOS training uses 32 context frames and up to eight grounding frames; inference uses 32 uniformly sampled context frames. The tube decoder has width 256 and two cross-attention layers with eight heads. Temporal inference uses order-constrained decoding with softmax temperature 0.05 and a one-second output grid; STVG reranks five candidate intervals using temporal confidence, tube continuity, and coverage.

Datasets and metrics. RVOS is evaluated on MeViS `valid` [10], ReVOS [42], GroundMoRe test [9], and ReasonVOS, using region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their mean $\mathcal{J}\&\mathcal{F}$. ReVOS All equally averages the Referring and Reasoning group means. VTG uses Charades-TimeLens, ActivityNet-TimeLens, and QVHighlights-TimeLens [48], reporting mean temporal IoU and Recall@1 at IoU thresholds 0.3/0.5/0.7. STVG uses VidSTG test [51], HC-STVG-v1 test, and HC-STVG-v2 validation, reporting mean temporal and video IoU. VidSTG results separate declarative and interrogative queries.

4.2 Video segmentation

BiSeek improves segmentation across motion-based reference and queries that require reasoning (Table 1). On MeViS `valid` [10], the 4B model reaches 65.3 $\mathcal{J}\&\mathcal{F}$, surpassing VIRST-7B [17] by 2.4 points. Its advantage extends to ReVOS All [42], where it exceeds SetCon-8B [52] by 2.6 points and outperforms VIRST-7B in both the Referring and Reasoning groups. The improvement also holds at the smaller model size: BiSeek-2B surpasses VideoSEG-O3-2B [8] by 6.5/4.8 points on MeViS/ReVOS.

GroundMoRe [9] provides a closer view of the reasoning gains. Across its 2,005 test queries, BiSeek-4B scores 61.9 $\mathcal{J}\&\mathcal{F}$, improving over VideoSEG-O3-4B [8] by 29.9 points. The advantage reaches 27.6 points on causal questions and 19.1 on descriptive questions, showing stronger gains when target selection requires reasoning about causal relations. On ReasonVOS, BiSeek-4B reaches 72.2 $\mathcal{J}\&\mathcal{F}$, exceeding SetCon-8B [52] by 4.2 points.

Table 1. Video segmentation results (%). Blue/orange mark best/second-best reported scores. Green values below BiSeek-4B scores give gains over the strongest listed baseline per column (percentage points). Our ReasonVOS evaluation uses 458 queries. BiSeek-4B achieves the highest score in every metric column.

Method	Venue/Year	MeViS `valid`			GroundMoRe			ReVOS									ReasonVOS
								Referring			Reasoning			Overall			
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
LMPM (Swin-T) [10]	ICCV 2023	37.2	34.3	40.2	–	–	–	–	–	–	–	–	–	–	–	–	–
VideoLISA-3.8B [3]	NeurIPS 2024	44.4	41.3	47.6	–	–	–	–	–	–	–	–	–	–	–	–	47.5
GLUS-7B [25]	CVPR 2025	51.3	48.5	54.2	–	–	–	58.3	56.0	60.7	51.4	48.8	53.9	54.9	–	–	49.9
VRS-HQ-7B [12]	CVPR 2025	50.6	47.6	53.7	–	–	–	62.1	59.8	64.5	56.1	53.5	58.7	59.1	56.6	61.6	–
UniPixel-7B [27]	NeurIPS 2025	55.8	53.2	58.3	–	–	–	65.8	63.9	67.8	61.5	59.4	63.7	63.7	61.7	65.7	–
Veason-R1-3B [13, 38]	CVPR 2026	–	–	–	23.5	22.0	25.0	63.0	60.3	65.6	56.8	53.6	60.0	59.9	56.9	62.8	55.2
Veason-R1-7B [13]	CVPR 2026	–	–	–	–	–	–	63.6	60.7	66.5	59.0	55.8	62.2	61.3	58.2	64.4	59.9
VIRST-7B [17]	CVPR 2026	62.9	60.4	65.4	–	–	–	70.8	68.8	72.8	66.1	63.9	68.3	68.4	66.3	70.6	–
VideoSEG-O3-2B [8]	ICML 2026	55.6	52.5	58.6	29.1	27.4	30.7	67.5	65.0	70.0	62.0	59.0	65.0	64.8	62.0	67.5	60.2
VideoSEG-O3-4B [8]	ICML 2026	60.0	57.0	63.0	32.0	30.2	33.7	70.3	67.7	72.9	65.1	62.4	67.9	67.7	65.0	70.4	62.9
SetCon-8B [52]	NeurIPS 2026	62.4	–	–	–	–	–	–	–	–	–	–	–	69.5	–	–	68.0
BiSeek-2B	–	62.1	58.0	66.3	58.0	57.9	58.1	71.2	67.8	74.6	68.0	64.6	71.4	69.6	66.2	73.0	68.5
		65.3	**61.7**	**68.8**	**61.9**	**61.1**	**62.7**	**74.0**	**70.9**	**77.1**	**70.2**	**66.9**	**73.6**	**72.1**	**68.9**	**75.3**	**72.2**
BiSeek-4B	–	+2.4	+1.3	+3.4	+29.9	+30.9	+29.0	+3.2	+2.1	+4.2	+4.1	+3.0	+5.3	+2.6	+2.6	+4.7	+4.2

4.3 Temporal grounding

We compare direct interval prediction with general video LLMs and specialized temporal grounders (Table 2). Compared with Qwen3-VL-2B [53], BiSeek-2B gains 15.1, 11.9, and 7.2 mean-IoU points on Charades, ActivityNet, and QVHighlights. On Charades, BiSeek-4B reaches 60.5 mean IoU, exceeding TimeLens-7B [48] by 11.7 points. The improvement holds at the stricter $R_{.7}$ threshold: its score of 40.7 exceeds TimeLens-7B by 12.3 points, indicating better overlap with the event boundaries. On ActivityNet, BiSeek-4B reaches 53.9 mean IoU, 6.7 points above Qwen3-VL-4B. On QVHighlights, it reaches 60.9 mean IoU, 0.8 points below Qwen3-VL-4B.

Table 2. Temporal grounding on TimeLens-Bench (%). R_τ denotes Recall@1 at temporal IoU threshold τ. BiSeek-4B improves mean IoU over Qwen3-VL-4B on Charades-TimeLens and ActivityNet-TimeLens.

Method	Venue/Year	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
		$R_{.3}$	$R_{.5}$	$R_{.7}$	mIoU	$R_{.3}$	$R_{.5}$	$R_{.7}$	mIoU	$R_{.3}$	$R_{.5}$	$R_{.7}$	mIoU
Qwen3-VL-2B [1, 53]	arXiv 2025	62.7	44.4	23.4	43.4	53.7	41.0	26.5	39.9	67.8	53.4	39.3	52.2
Qwen3-VL-4B [1, 50]	arXiv 2025	67.1	47.5	24.1	45.6	62.9	51.0	34.3	47.2	**77.6**	**66.0**	**50.9**	**61.7**
Video-o3-7B [47, 53]	ICML 2026	58.7	34.2	16.2	38.6	53.2	41.0	24.4	38.9	62.7	49.5	34.4	47.4
VideoChat-Flash-7B [23, 48]	ICLR 2026	60.2	37.9	17.8	39.7	35.5	21.8	10.5	24.8	45.2	30.6	16.7	32.7
MiMo-VL-7B [41, 48]	arXiv 2025	57.9	42.6	20.5	39.6	49.3	38.7	22.4	35.5	57.1	42.6	28.4	41.5
TimeSuite-7B [46, 53]	ICLR 2025	56.3	35.5	18.0	38.1	27.1	17.5	8.6	19.8	27.1	16.9	9.9	21.7
VideoChat-R1-7B [22, 48]	arXiv 2025	51.9	30.8	11.7	33.7	35.0	23.9	11.3	25.0	29.3	19.1	9.4	21.5
Time-R1-7B [36, 48]	NeurIPS 2025	57.9	32.0	16.9	36.6	44.8	31.0	19.0	33.1	65.8	51.5	36.1	49.2
Qwen2.5-VL-7B [2, 48]	arXiv 2025	59.7	37.8	16.6	39.3	44.1	31.0	16.1	31.4	41.5	27.8	15.2	31.6
TimeLens-7B [48]	CVPR 2026	70.5	55.6	28.4	48.8	62.8	51.0	32.6	46.2	74.1	62.7	43.1	56.0
Qwen3-VL-8B [1, 48]	arXiv 2025	69.2	53.4	27.5	48.3	62.1	51.2	34.4	46.8	74.2	**64.6**	**49.3**	59.4
InternVideo3-8B [43, 53]	arXiv 2026	75.8	61.7	32.8	53.2	60.0	51.4	33.4	46.3	72.2	62.6	48.9	59.2
BiSeek-2B	–	**79.8**	**68.0**	**38.7**	**58.5**	**66.8**	**56.4**	**37.8**	**51.8**	74.2	62.4	45.8	59.4
BiSeek-4B	–	**82.1**	**70.4**	**40.7**	**60.5**	**69.2**	**59.4**	**40.0**	**53.9**	**75.4**	63.1	48.0	**60.9**

4.4 Spatiotemporal grounding

Both variants use a 2-FPS evaluation grid capped at 64 frames. On HC-STVG v1 (Table 3), BiSeek-4B reaches 72.2/53.4 mean temporal/video IoU, exceeding STVG-o1-7B [16] by 11.9/9.3 points. On v2, it improves temporal IoU over STVG-o1-7B by 4.1 points and video IoU over DEViL-7B [11] by 7.6 points, showing gains in both event extent and tube overlap.

VidSTG shows consistent gains for declarative and interrogative queries. Relative to VideoGrounding-DINO [37], BiSeek-4B improves temporal IoU by 2.4/1.7 points and video IoU by 2.0/1.4 points. This advantage extends to the 7B language-model grounder STVG-o1, with video-IoU gains of 3.2/3.4 points. Scaling from 2B to 4B adds 0.9/1.1 video-IoU points and 0.6/0.2 temporal-IoU points.

Table 3. Spatiotemporal grounding (%). BiSeek-4B achieves the best temporal and video IoU on VidSTG and both HC-STVG benchmarks among the methods listed in this table.

Method	Venue/Year	VidSTG (Declarative)		VidSTG (Interrogative)		HC-STVG v1 test		HC-STVG v2 val	
		m_{tIoU}	m_{vIoU}	m_{tIoU}	m_{vIoU}	m_{tIoU}	m_{vIoU}	m_{tIoU}	m_{vIoU}
STGRN [51]	CVPR 2020	48.5	19.8	47.0	18.3	–	–	–	–
STGVT [34]	TCSVT 2022	–	21.6	–	–	–	18.2	–	–
STVGBert [33]	ICCV 2021	–	24.0	–	22.5	–	20.4	–	–
TubeDETR [44]	CVPR 2022	48.1	30.4	46.9	25.7	43.7	32.4	53.9	36.4
STCAT [19]	NeurIPS 2022	50.8	33.1	49.7	28.2	49.4	35.1	–	–
CSDVL / STVGFormer [26]	CVPR 2023	–	33.7	–	28.5	–	36.9	58.1	38.7
CG-STVG [14]	CVPR 2024	51.4	34.0	49.9	29.0	52.8	38.4	60.0	39.5
TA-STVG [15]	ICLR 2025	51.7	34.4	50.2	29.5	53.0	39.1	60.4	40.2
VideoGrounding-DINO [37]	CVPR 2024	52.0	34.7	50.8	29.9	–	38.3	–	39.9
E3M [4]	ECCV 2024	–	16.2	–	10.6	–	19.1	–	–
SpaceVLLM-7B [35]	AAAI 2026	47.7	27.4	48.5	25.4	56.9	39.3	58.0	34.0
STVG-o1-7B [16]	arXiv 2025	52.1	33.5	50.5	27.9	60.3	44.1	**63.8**	41.2
STVG-R1-7B [49]	ICLR 2026	–	–	–	–	56.9	39.1	61.3	40.8
DEViL-7B [11]	ACM MM 2026	50.2	33.6	48.5	28.8	59.0	43.1	61.7	42.5
BiSeek-2B	–	**53.8**	**35.8**	**52.3**	**30.2**	**68.0**	**50.1**	63.3	**48.5**
BiSeek-4B	–	**54.4**	**36.7**	**52.5**	**31.3**	**72.2**	**53.4**	**67.9**	**50.1**

4.5 Ablation studies

All ablations use BiSeek-2B. Table 4 compares jointly retrained causal and bidirectional interfaces with the same video-first input order. VTG and STVG retain their readout architectures; the causal RVOS baseline uses a SEG token. Table 7b isolates the RVOS encoder effect under fixed conditions.

Effect of bidirectional encoding. The bidirectional interface improves all five benchmarks by 7.9–18.7 points over the causal interfaces in Table 4. The RVOS comparison also changes the conditioning interface. With query–boundary conditions fixed, Table 7b gives encoder gains of 1.5/1.6 points on MeViS/ReVOS.

Table 4. Comparison of causal and bidirectional perception interfaces. Both variants are jointly retrained; the causal RVOS baseline uses a SEG token. Char. denotes Charades-TimeLens.

	RVOS		VTG	STVG	
	MeViS	ReVOS	Char.	VidSTG	HC-v2
Encoder	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	m_t	m_v	m_t
Causal	54.2	58.2	45.4	19.2	44.6
Bidirectional	**62.1**	**69.6**	**58.5**	**33.0**	**63.3**

Effect of temporal readout. We separate the choice of boundary states from the scoring head (Table 5a,b). Under matched training budgets, separate start/end states improve temporal IoU over a shared end state by 6.5/5.2/5.9 points on Charades/VidSTG/HC-STVG v2, and over mean-pooled patches by 1.4/1.0/1.9 points. The default model uses the linear head in Eq. 5. Table 5b additionally retrains query-conditioned scores, $\ell_t^s = (w_s + W_s h^q)^\top h_t^s$, and analogous end scores. Under causal attention, query conditioning adds 9.1/13.7 temporal-IoU points on Charades/HC-STVG v2 and 10.3 video-IoU points on VidSTG. With query-conditioned scoring fixed, bidirectional encoding adds another 3.5–5.0 points, matching the reported linear-readout scores on these three metrics.

Initialization, training losses, and inference decoding are tested separately (Table 5c,d). Random output initialization yields much lower scores than lexical initialization within the matched budget. Relative to the budget-matched start/end control, removing the IoU or ordering loss lowers temporal IoU by 1.0–2.3 or 1.4–2.1 points. On the same checkpoint, order-constrained decoding improves temporal IoU over independent argmax by 1.3–1.6 points and over expected endpoints by 1.4–2.3 points.

Effect of spatial readout. Table 6 groups spatial memory, query conditioning, and tube reranking. Under joint training, radius-one memory improves spatial IoU over nearest-unit memory by 9.6/4.5 points on VidSTG/HC-STVG v2. The matched STVG controls favor local memory over larger or all-unit memories and retain both relative-time and spatial encodings. The four encoder–query combinations show complementary contributions: removing the query term costs 8.8/9.5 spatial-IoU points with bidirectional encoding and 6.0/7.7 with causal encoding. Tube reranking is an inference-only change; its effect is mainly temporal, with spatial IoU changing by at most 0.1 point.

Effect of mask conditions. We first compare conditioning signals with three slots per unit (Table 7a). Combining the query with unit-specific boundaries improves $\mathcal{J}\&\mathcal{F}$ over the query alone by 2.8–3.2 points and over temporal averaging by 1.7–1.9 points; boundaries alone are substantially weaker. Temporal averaging follows projection. Encoder controls keep either query–boundary or boundary-only conditions fixed (Table 7b). Bidirectional encoding adds 1.5–1.6 points with the query–boundary condition. In joint training (Table 7c), unit-indexed conditions also improve MeViS and ReVOS over temporal averaging.

Effect of joint task training. We progressively add VTG and STVG supervision while matching the number of RVOS updates (Table 8). Adding VTG to RVOS-only training raises MeViS/ReVOS $\mathcal{J}\&\mathcal{F}$ by 1.3/0.3 points; adding STVG gives a further 0.8/0.5 points. The cumulative gains of 2.1/0.8 points show that localization supervision also benefits mask prediction.

Table 5. Temporal readout (%). (a,c) use the budget-matched start/end control in (a). (b) jointly retrains each head–encoder pair. (d) uses one checkpoint; expected endpoints retain order-constrained fallback.

(a) Boundary states

	Charades		VidSTG		HC-v2	
Variant	m_t	$R_{.7}$	m_t	m_v	m_t	m_v
Start/end (control)	**58.3**	**38.4**	**52.8**	**32.7**	**63.1**	**48.1**
Patch mean	56.9	32.9	51.8	32.2	61.2	46.9
Shared end state	51.8	22.0	47.6	29.7	57.2	43.8

(b) Query-conditioned scoring

Encoder	Head	Char. m_t	VidSTG m_v	HC-v2 m_t
Causal	Linear	45.4	19.2	44.6
Causal	Query-cond.	54.5	29.5	58.3
Bidirectional	Query-cond.	**58.5**	**33.0**	**63.3**

	Charades		VidSTG		HC-v2	
Variant	m_t	$R_{.7}$	m_t	m_v	m_t	m_v
(c) Initialization and losses (matched training)						
Random output init.	9.2	0.6	8.0	1.5	9.5	3.0
No IoU loss	57.2	36.1	51.8	31.8	60.8	46.6
No ordering loss	56.6	33.0	51.4	32.0	61.0	46.7
(d) Endpoint decoding (inference only)						
Order-constrained decoding	**58.5**	**38.7**	**53.0**	**33.0**	**63.3**	**48.5**
Independent argmax	57.0	34.2	51.7	32.2	61.7	47.3
Expected endpoints	56.8	35.0	51.6	32.1	61.0	46.7

Table 6. Spatial readout (%). r is the native-unit radius; s is spatial IoU within the ground-truth interval. (b,c) use matched STVG-only training, with the bidirectional query-conditioned row as their common control. (d) changes only inference on the default checkpoint.

	VidSTG			HC-v2		
Variant	m_t	m_v	s	m_t	m_v	s
(a) Spatial memory (joint training)						
Nearest unit ($r = 0$)	52.6	27.7	52.7	62.4	45.0	72.1
Local memory ($r = 1$)	**53.0**	**33.0**	**62.3**	**63.3**	**48.5**	**76.6**
(b) Memory and position (matched STVG)						
Bidir., query ($r = 1$)	**53.0**	**33.0**	**62.3**	**63.3**	**48.5**	**76.6**
Radius $r = 2$	52.9	32.4	61.2	63.2	47.9	75.8
All-unit memory	52.7	31.2	59.2	63.0	46.5	73.8
No relative time	52.6	30.6	58.2	62.9	46.0	73.1
No spatial encoding	52.3	29.3	56.0	62.5	43.7	69.9
(c) Encoder and query (same STVG control)						
Bidir., no query	52.0	27.8	53.5	62.1	41.7	67.1
Causal, query	52.4	30.3	57.8	62.6	45.6	72.8
Causal, no query	51.7	26.8	51.8	61.6	40.1	65.1
(d) Tube reranking (inference only)						
No tube reranking	52.4	32.6	62.2	62.8	48.1	76.6

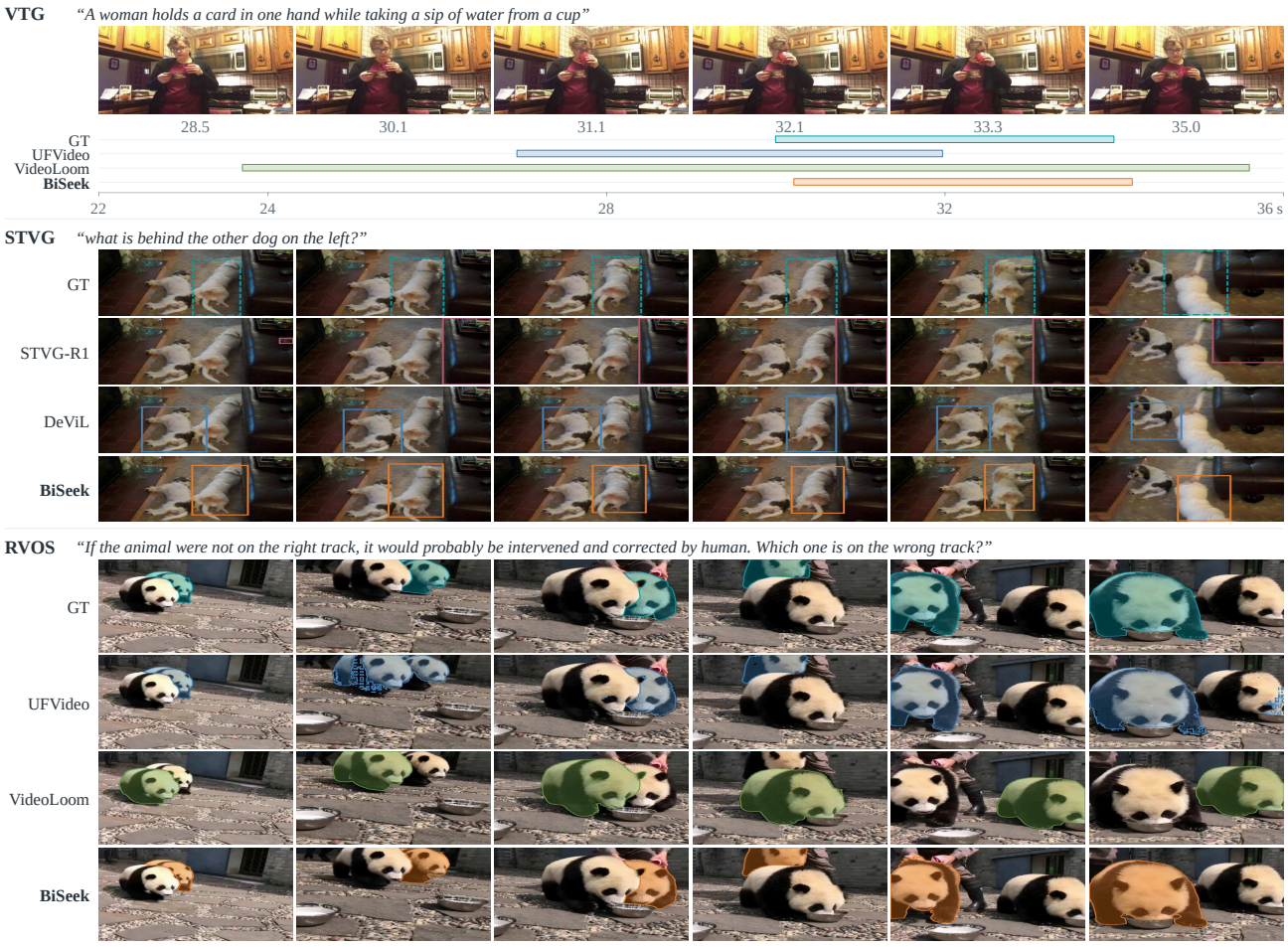

Figure 3. Qualitative comparison with SOTA methods. BiSeek-4B on Charades-TimeLens (VTG), VidSTG (STVG), and ReasonVOS (RVOS). Each example shows six frames from left to right. STVG and RVOS use the same frames for every method, with one method per row. VTG intervals use continuous clip-relative time in seconds. Cyan denotes ground truth and orange denotes BiSeek predictions.

Table 7. Mask conditions (%). (a) compares conditioning signals; (b) changes the encoder at fixed conditions. All variants retain three slots. (c) tests temporal granularity under joint training. GroundMoRe reports question-type $\mathcal{J}\&\mathcal{F}$.

	MeViS	ReVOS	GroundMoRe	
Variant	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	Seq.	Causal
(a) Conditioning signals				
Query + boundaries	**62.1**	**69.6**	**62.8**	**56.4**
Query only	58.9	66.5	59.7	53.6
Boundary states only	47.2	55.4	49.8	44.1
Query + pooled patches	60.2	67.6	60.7	54.6
Average over time	60.4	67.8	60.9	54.6
(b) Encoder controls at fixed conditions				
Causal, query + boundaries	60.6	68.0	61.2	54.9
Causal, boundaries only	45.8	53.9	48.3	42.8
(c) Temporal granularity (joint training)				
Average over time	60.4	67.8	–	–
Unit-indexed	**62.1**	**69.6**	**62.8**	**56.4**



Figure 4. Query attention. Causal/query-only and bidirectional/query–boundary BiSeek-2B variants on the same frames. Red boxes mark the target in the original frames. Maps average all 16 heads at layer 14 and are divided by each unit's spatial mean (uniform = 1). Two temporal units are shown per query, with one shared color scale across models and cases.

Table 8. Effect of adding task data on RVOS (BiSeek-2B; %). Runs match the number of RVOS updates.

Training data	MeViS $\mathcal{J}\&\mathcal{F}$	ReVOS $\mathcal{J}\&\mathcal{F}$
RVOS	60.0	68.8
RVOS + VTG	61.3	69.1
RVOS + VTG + STVG	62.1	69.6

4.6 Qualitative results

In the VTG example of Figure 3, BiSeek predicts 30.2–34.2 s for holding a card while drinking, close to the annotated 30–34 s interval. UFVideo shifts the prediction earlier to 26.9–32.0 s, missing the event ending, while VideoLoom includes surrounding activity with a 23.7–35.6 s interval.

For STVG, BiSeek follows the dog specified by its relationship to the other dog as their poses and relative positions change. STVG-R1 places boxes on background regions, and DeViL often selects the other dog or only part of the target. For RVOS, “corrected by human” identifies the panda involved in the interaction. BiSeek follows this panda with more complete masks. VideoLoom selects the foreground animal, while UFVideo's target masks exhibit missing regions and fragmented boundaries.

Figure 4 compares attention maps from the causal/query-only and bidirectional/query–boundary variants. In the black-helmet cyclist case, the causal map peaks on the yellow motorcyclist, whereas the bidirectional map concentrates on the marked cyclist across both temporal units. The two riders share yellow clothing, making the helmet an important distinguishing cue. In the yellow-jacket case, both variants respond to the target rider, with

stronger target-aligned spatial peaks in the bidirectional maps.

5 Conclusion

We presented BiSeek, a bidirectional video–query encoder for fine-grained perception. Joint video–query contextualization associates instruction semantics with native temporal and spatial positions. Structure-preserving readouts use these positions to predict intervals, box tubes, and mask sets from one backbone encoding. Both 2B and 4B variants support all three tasks. The 4B model achieves strong results on Charades-TimeLens and leads the listed comparisons on ReVOS and GroundMoRe. Controlled comparisons show complementary benefits from bidirectional encoding and task-specific readouts. Native unit boundaries improve event localization, while local patch memories preserve the spatial evidence used for box prediction. Query–boundary conditions combine referring semantics with temporally indexed evidence for mask prediction. Localization supervision also improves segmentation on MeViS and ReVOS under joint training. Native video states thus provide a common interface for language-guided video perception.

LLM Usage Statement

During the preparation of this manuscript, the authors used ChatGPT to assist with organizing research materials, conducting literature reviews, supporting code development, and improving the clarity, grammar, and overall presentation of the manuscript. Additionally, GPT-Image was used to provide suggestions for preliminary method illustrations and conceptual figure drafts. All AI-generated content was carefully reviewed, verified, and, where necessary, revised by the authors. The authors take full responsibility for the accuracy, integrity, and originality of the manuscript.

References

[1] 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, 2025.

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

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

[4] Peijun Bao, Zihao Shao, Wenhan Yang, Boon Poh Ng, and Alex C. Kot. E3M: Zero-shot spatio-temporal video grounding with expectation-maximization multimodal modulation. In *ECCV 2024*, pages 227–243. Springer, 2025.

[5] Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. LLM2Vec: Large language models are secretly powerful text encoders. In *COLM*, 2024.

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

[7] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *ICLR*, 2026.

[8] 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*, pages 22590–22616. PMLR, 2026.

[9] 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 *CVPR*, pages 8625–8636, 2025.

[10] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, pages 2694–2703, 2023.

[11] Shida Gao, Feng Xue, Xiangfeng Wang, Anlong Ming, Teng Long, Yihua Shao, Haozhe Wang, Zhaowen Lin, Haiyang Zhang, Wei Wang, et al. Detector-empowered video large language model for efficient spatio-temporal grounding. In *ACM MM*, pages 2285–2294. ACM, 2026.

[12] 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 *CVPR*, pages 29183–29192, 2025.

[13] Sitong Gong, Yunzhi Zhuge, Lu Zhang, Jiazuo Yu, Pingping Zhang, Xu Jia, and Huchuan Lu. Reinforcing video object segmentation to think before it segments. In *CVPR*, pages 3835–3844, 2026.

[14] Xin Gu, Heng Fan, Yan Huang, Tiejian Luo, and Libo Zhang. Context-guided spatio-temporal video grounding. In *CVPR*, pages 18330–18339, 2024.

[15] Xin Gu, Yaojie Shen, Chenxi Luo, Tiejian Luo, Yan Huang, Yuewei Lin, Heng Fan, and Libo Zhang. Knowing your target: Target-aware transformer makes better spatio-temporal video grounding. In *ICLR*, 2025.

[16] 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:2511.21375v1, 2025.

[17] Jihwan Hong and Jaeyoung Do. VIRST: Video-instructed reasoning assistant for SpatioTemporal segmentation. In *CVPR*, pages 3410–3420, 2026.

[18] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022.

[19] Yang Jin, Yongzhi Li, Zehuan Yuan, and Yadong Mu. Embracing consistency: A one-stage approach for spatio-temporal video grounding. In *NeurIPS*, pages 29192–29204, 2022.

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

[21] Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. In *ICLR*, 2025.

[22] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958, 2025.

[23] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. VideoChat-Flash: Hierarchical compression for long-context video modeling. In *ICLR*, pages 109089–109117, 2026.

[24] 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 *CVPR*, pages 39497–39507, 2026.

[25] 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 *CVPR*, pages 8658–8667, 2025.

[26] Zihang Lin, Chaolei Tan, Jian-Fang Hu, Zhi Jin, Tiancai Ye, and Wei-Shi Zheng. Collaborative static and dynamic vision-language streams for spatio-temporal video grounding. In *CVPR*, pages 23100–23109, 2023.

[27] Ye Liu, Zongyang Ma, Junfu Pu, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. UniPixel: Unified object referring and segmentation for pixel-level visual reasoning. In *NeurIPS*, pages 139744–139774, 2025.

[28] Hewen Pan, Cong Wei, Dashuang Liang, Zepeng Huang, Pengfei Gao, Ziqi Zhou, Lulu Xue, Pengfei Yan, Xiaoming Wei, Minghui Li, et al. Ufvideo: Towards unified fine-grained video cooperative understanding with large language models. *arXiv preprint arXiv:2512.11336*, 2025.

[29] Hanoona Rasheed, Haania Siddiqui, Ming-Hsuan Yang, Fahad Shahbaz Khan, and Salman Khan. Locate anything in videos: Rethinking efficient generative spatio-temporal video grounding. arXiv preprint arXiv:2608.28192v1, 2026.

[30] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.

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

[32] Jiapeng Shi, Junke Wang, Zuyao You, Bo He, and Zuxuan Wu. Videoloom: A video large language model for joint spatial-temporal understanding. In *Forty-third International Conference on Machine Learning*, 2026.

[33] Rui Su, Qian Yu, and Dong Xu. STVGBert: A visual-linguistic transformer based framework for spatio-temporal video grounding. In *ICCV*, pages 1533–1542, 2021.

[34] Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *TCSVT*, 32(12):8238–8249, 2022.

[35] Jiankang Wang, Zhihan Zhang, Zhihang Liu, Yang Li, Jiannan Ge, Hongtao Xie, and Yongdong Zhang. SpaceVLLM: Endowing multimodal large language model with spatio-temporal video grounding capability. *AAAI*, 40(12):9912–9920, 2026.

[36] Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, et al. Time-R1: Post-training large vision language model for temporal video grounding. In *NeurIPS*, pages 83330–83364, 2025.

[37] Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. VideoGrounding-DINO: Towards open-vocabulary spatio-temporal video grounding. In *CVPR*, pages 18909–18918, 2024.

[38] Junwei Wen, Deshui Miao, Guangming Lu, Ming-Hsuan Yang, Xin Li, and Wenjie Pei. Think visually across space and time: Grounded reasoning for video segmentation. arXiv preprint arXiv:2610.25090, 2026.

[39] Boyong Wu, Sanghwan Kim, and Zeynep Akata. From drop-off to recovery: A mechanistic analysis of segmentation in MLLMs. In *ECCV 2026*, pages 323–341. Springer, 2026.

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

[41] Xiaomi LLM-Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, et al. MiMo-VL technical report. arXiv preprint arXiv:2506.03569, 2025.

[42] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. VISA: Reasoning video object segmentation via large language models. In *ECCV 2024*, pages 98–115. Springer, 2025.

[43] Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, et al. InternVideo3: Agentify foundation models with multimodal contextual reasoning. arXiv preprint arXiv:2606.12195, 2026.

[44] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. TubeDETR: Spatio-temporal video grounding with transformers. In *CVPR*, pages 16442–16453, 2022.

[45] 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 MLLM for dense grounded understanding of images and videos. *TPAMI*, 48(11):14922–14936, 2026.

[46] Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *ICLR*, 2025.

[47] Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. In *ICML*, pages 153787–153813. PMLR, 2026.

[48] Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs. In *CVPR*, pages 10419–10429, 2026.

[49] Xiaowen Zhang, Zhi Gao, Licheng Jiao, Lingling Li, and Qing Li. STVG-R1: Incentivizing instance-level reasoning and grounding in videos via reinforcement learning. In *ICLR*, 2026.

[50] Xin Zhang, Haochen Wang, Beibei Lin, Zhaoxiang Zhang, Robby T. Tan, Jason Li, and Dacheng Tao. Generalized CycleGRPO for fine-grained multimodal understanding and grounding. arXiv preprint arXiv:2610.29552, 2026.

[51] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *CVPR*, pages 10668–10677, 2020.

[52] Zhixiong Zhang, Yizhuo Li, Shuangrui Ding, Yuhang Zang, Shengyuan Ding, Long Xing, Yibin Wang, Qiaosheng Zhang, and Jiaqi Wang. SetCon: Towards open-ended referring segmentation via set-level concept prediction. In *NeurIPS*, 2026.

[53] Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, et al. TimeLens2: Generalist video temporal grounding with multimodal LLMs. arXiv preprint arXiv:2607.17423, 2026.

Appendix

This appendix provides the implementation and evaluation details for BiSeek, complete quantitative results, and additional visual comparisons. The 2B and 4B variants share the same three perception interfaces. All times in localization results refer to the evaluated clip, and IoU, recall, and segmentation scores are percentages unless stated otherwise.

Reading guide. Sections A–B describe how native video states are read and trained. Section C specifies sampling, coordinates, and benchmark scopes. Section D expands the main-paper result tables. Sections E and F examine predictions and attention maps, followed by limitations in Section G.

A Readout implementation details

A.1 Input sequence and native state indexing

BiSeek takes a sampled video followed by a query-bearing user instruction. The native Qwen3-VL processor [1] groups two sampled context frames into each temporal unit. These frames can be far apart in the original video when the context is sampled sparsely. Each unit retains its timestamp, vision-start position, post-merge spatial grid, and vision-end position. Bidirectional language self-attention lets every valid input position incorporate the complete video and instruction. Native positional encoding and the padding mask are retained.

Let $H \in \mathbb{R}^{L\times d}$ denote the final normalized language-backbone states. For unit t, we gather its start and end states $h_t^s, h_t^e \in \mathbb{R}^d$ and patch grid $P_t \in \mathbb{R}^{N_p\times d}$. The shared query state h^q is read immediately before the closing delimiter of the query-bearing user turn. This location includes the instruction wrapper surrounding the referring expression. The state width d follows the selected backbone. The readouts keep each unit's temporal address and each patch's spatial address throughout prediction.

A.2 Temporal endpoint scoring and decoding

Two output-only rows w_s, w_e score the native start and end states independently. They are initialized from the language-head rows for *start* and *end*. With temperature $\vartheta = 0.05$ and valid temporal coordinates τ_t, the endpoint probabilities are

$$p_s(t) = \frac{\exp(w_s^\top h_t^s/\vartheta)}{\sum_{k=1}^{T}\exp(w_s^\top h_k^s/\vartheta)}, \qquad p_e(t) = \frac{\exp(w_e^\top h_t^e/\vartheta)}{\sum_{k=1}^{T}\exp(w_e^\top h_k^e/\vartheta)}. \tag{S1}$$

Temporal training combines endpoint classification with continuous interval-overlap and ordering objectives. The latter use differentiable soft coordinates

$$\bar{b}_s = \sum_t p_s(t)\tau_t, \qquad \bar{b}_e = \sum_t p_e(t)\tau_t. \tag{S2}$$

At inference, order-constrained decoding selects

$$(\hat{i},\hat{j}) = \arg\max_{1\le i\le j\le T}\left(w_s^\top h_i^s + w_e^\top h_j^e\right). \tag{S3}$$

This also maximizes $p_s(i)p_e(j)$ over valid pairs, since both softmax normalizers are constant with respect to the candidate indices. A prefix-maximum scan computes the solution in $O(T)$ time. For each end index j, the scan retains the highest start logit among indices $i \le j$, adds the end logit, and updates the best pair.

The temporal cells partition the clip at the midpoints of adjacent unique unit timestamps, with outer edges at zero and clip duration. When both endpoints select the same unit, its cell bounds provide a positive span. For different

units, their selected timestamps define the endpoints. The benchmark-specific output grid is applied after decoding as described in Section C.

For a ground-truth interval b^*, the temporal training objective uses

$$\begin{aligned}\mathcal{L}_{\text{time}} &= \mathcal{L}_{\text{CE},s} + \mathcal{L}_{\text{CE},e} \\ &\quad + 5\big[1 - \text{tIoU}(\bar{b}, b^*) \\ &\qquad + \text{ReLU}(\bar{b}_s - \bar{b}_e)\big], \\ \bar{b} &= (\bar{b}_s, \bar{b}_e).\end{aligned} \tag{S4}$$

The ordering term uses endpoint coordinates in seconds. Tube supervision uses $5\mathcal{L}_1 + 2\mathcal{L}_{\text{GIoU}}$ on valid annotated frames.

A.3 Time-addressed tube decoding

For an output frame at time σ_f, the tube head finds the nearest native unit $u(f) = \arg\min_t |\sigma_f - \tau_t|$. Its memory contains every patch from $u(f)$ and the valid radius-one neighbors. At a clip boundary, only existing units are included. A frame query combines a projection of h^q with the requested-time encoding. Each memory patch combines its projected state with spatial, absolute-time, and relative-time encodings. Patch centers use the post-merge grid in normalized image coordinates, and time encodings use clip-duration normalization.

The decoder has width 256, eight attention heads, and two cross-attention layers. A box head maps each resulting frame query to a normalized box. Multiple requested-time queries can read the same backbone states in parallel. The radius-one memory supplies local spatial detail, while the patch states already contain full-input semantic context from bidirectional encoding. STVG inference considers five temporal candidates and reranks their interval–tube pairs using temporal confidence, tube continuity, and coverage, with tube and coverage weights 0.25 and 0.1.

A.4 Mask conditions and frame assignment

For unit t, a shared two-layer MLP with ReLU projects three rows independently,

$$R_t = [h^q; h_t^s; h_t^e] \in \mathbb{R}^{3\times d}, \qquad C_t = \phi(R_t) \in \mathbb{R}^{3\times 256}. \tag{S5}$$

The three rows carry shared referring context and two unit-specific boundary states. Condition length stays fixed across expressions referring to different numbers of objects. SAM 3's prediction slots represent the individual targets [7].

During training, a grounding frame already sampled into the context retains its exact native-unit membership. Other grounding frames use nearest-unit assignment. At dense inference times, the corresponding condition is paired with image features independently encoded from the current frame. The projected rows enter the SAM 3 grounding encoder as language features, and the set decoder predicts masks, boxes, and scores. Association and propagation connect frame predictions. The shared video–query encoding is reused across output times, while image encoding and dense prediction still operate on the requested frames.

Table S1. Interfaces between the shared encoder and perception branches. T is the number of native units, N_p the patch count of unit t, and d the backbone hidden width.

Branch	Encoder states	Output and alignment
Temporal extent	$\{h_t^s, h_t^e\}_{t=1}^T$	One ordered interval, shared by VTG and STVG
Box tube	h^q and $\{P_t\}_{t\in\mathcal{N}_f}$	One box per requested time, nearest-unit radius-one memory
Mask set	$\{[h^q; h_t^s; h_t^e]\}_{t=1}^T$	Three 256-dimensional conditions per unit, assigned to image frames

B Training data and optimization

B.1 Task mixture and supervision

Both model scales are initialized from the corresponding Qwen3-VL-Instruct backbone and SAM 3 foundation weights. They use the same task mixture and optimization schedule. Training alternates task-homogeneous updates with proportions 30/10/30/30 for RVOS, image segmentation, VTG, and STVG. Image examples use the mask-set objective through a single visual unit. The datasets and their supervision are summarized in Table S2.

Table S2. Training sources and supervision. Proportions refer to task-homogeneous updates.

Task	Sources	Supervision	Proportion
RVOS	MeViS, Ref-DAVIS17, Ref-YouTube-VOS, ReVOS	Visible referred-instance masks	30%
Image segmentation	MUSE, RefCOCO/+/g, gRefCOCO, ReasonSeg	Image target sets	10%
VTG	TimeLens-100K	Event start and end times	30%
STVG	Processed VidSTG and HC-STVG training queries	Event extent and box trajectories	30%

Temporal examples supervise endpoint classification, interval overlap, and ordering. STVG adds L1 and generalized-IoU box supervision. Image and video segmentation use the SAM 3 set objective, including matched classification and presence, box L1 and generalized IoU, focal and Dice masks, and one-to-many matching. The matching targets are the visible referred instances in each grounding frame.

B.2 Optimization and trainable parameters

The training budget is 30,210 optimizer updates. Eight GPUs process two samples per GPU with two accumulation steps, giving global batch size 32. Task-routed AdamW uses the settings in Table S3. The 906-update warmup is followed by cosine decay. Language adaptation uses LoRA [18].

Table S3. Shared optimization and architecture settings for BiSeek-2B and BiSeek-4B.

Setting	Value
Optimizer and base learning rate	AdamW, 4×10^{-5}
Weight decay, Adam betas	0.05, $(0.9, 0.999)$
Global batch size	$8 \times 2 \times 2 = 32$
Optimizer updates, warmup updates	30,210, 906
Gradient norm clipping	1
LoRA rank, scale, dropout	64, 128, 0.05
Temporal and tube learning-rate multiplier	2.5×
SAM 3 grounding learning-rate multiplier	0.25×
Precision and execution	bfloat16, FlashAttention 2, DeepSpeed ZeRO-2
Tube width, heads, layers	256, 8, 2
Tube temporal neighborhood	Nearest unit and radius-one neighbors
Temporal softmax temperature	0.05
STVG candidate count	5
Tube continuity and coverage weights	0.25, 0.1

The base language weights, Qwen's visual encoder, and SAM 3's image backbone are frozen. The language LoRA adapters, condition projection, temporal output rows, tube decoder, SAM 3 grounding transformer including its decoder, and mask/score heads are trainable. Each task updates the shared language adapters together with its active output branch. Consequently, a task-homogeneous update routes its supervision through the readout used by that task.

B.3 Context and grounding sampling

Localization samples the complete evaluated clip at 2 FPS. Its nominal visual-token budget is 14,336, with at least 64 spatial tokens per native unit and sequence limit 65,536. RVOS training samples 32 context frames and up to eight grounding frames. RVOS inference uses 32 uniformly sampled context frames, visual budget 6,144, and sequence limit 8,192. Native units group pairs of sampled context frames. Grounding images supply their own spatial features, so the mask decoder can operate on frames beyond those selected as context.

C Evaluation protocols

C.1 Benchmarks, splits, and expression counts

Table S4 summarizes the evaluation scopes. VTG reports mean temporal IoU and Recall@1 at thresholds 0.3, 0.5, and 0.7. STVG reports mean temporal and video IoU, with video-IoU threshold metrics where available. Segmentation uses region similarity $\mathcal{J}$, contour accuracy $\mathcal{F}$, and their mean $\mathcal{J}\&\mathcal{F}$. Each benchmark retains its specified aggregation protocol.

Table S4. Evaluation splits and units. Additional HC-STVG scopes appear in Table S9. MeViS `valid_u` is a diagnostic split.

Task	Benchmark / split	Queries or expressions	Principal metrics
VTG	Charades-TimeLens	3,363	Temporal IoU, Recall@1
VTG	ActivityNet-TimeLens	4,500	Temporal IoU, Recall@1
VTG	QVHighlights-TimeLens	1,541	Temporal IoU, Recall@1
STVG	VidSTG test, declarative	4,610	Temporal and video IoU
STVG	VidSTG test, interrogative	5,693	Temporal and video IoU
STVG	HC-STVG v1 test	1,160	Temporal and video IoU
STVG	HC-STVG v2 validation	2,000	Temporal and video IoU
RVOS	MeViS `valid`	2,236	$\mathcal{J}$, $\mathcal{F}$, $\mathcal{J}\&\mathcal{F}$
RVOS	MeViS `valid_u`	793	$\mathcal{J}$, $\mathcal{F}$, $\mathcal{J}\&\mathcal{F}$
RVOS	GroundMoRe test	2,005	$\mathcal{J}$, $\mathcal{F}$, $\mathcal{J}\&\mathcal{F}$
RVOS	ReVOS evaluation pool	5,822	Group-averaged $\mathcal{J}$, $\mathcal{F}$, $\mathcal{J}\&\mathcal{F}$
RVOS	ReasonVOS	458	$\mathcal{J}$, $\mathcal{F}$, $\mathcal{J}\&\mathcal{F}$
RVOS	Ref-YouTube-VOS validation	834	$\mathcal{J}$, $\mathcal{F}$, $\mathcal{J}\&\mathcal{F}$
RVOS	Ref-DAVIS17 validation	244	$\mathcal{J}$, $\mathcal{F}$, $\mathcal{J}\&\mathcal{F}$
RVOS	Long-RVOS validation	1,326	$\mathcal{J}\&\mathcal{F}$, mask-derived tIoU/vIoU

C.2 Temporal coordinates and endpoint quantization

TimeLens-Bench uses the revised Charades, ActivityNet, and QVHighlights queries [48]. Order-constrained endpoint predictions are mapped to clip-relative seconds and quantized to a one-second grid using ties-to-even rounding. The continuous-time metrics then compare these outputs with the original ground-truth intervals. For example, an endpoint at 2.5 s maps to 2 s and one at 3.5 s maps to 4 s. Ground-truth endpoints retain their original coordinates. Endpoint mean absolute error is measured in seconds and supplements overlap and recall.

For the visual comparisons in Section E, VTG timelines show the saved continuous-time predictions. Their per-case overlap can therefore differ from the one-second-grid benchmark protocol. Clip-relative time also applies to the VidSTG examples. Long-RVOS retains original frame indices for temporal coordinates.

C.3 Spatial sampling and evaluation scopes

STVG outputs and tube-consistency reranking use a 2-FPS evaluation grid capped at 64 frames. The backbone receives complete-clip context. Boxes are mapped to original-image coordinates for comparison. The VidSTG figure in the main paper uses boxes normalized by original image dimensions.

HC-STVG v1 full test and v2 standard validation are reported separately from the additional evaluation pool and the subsets whose videos were unseen during training. Table S9 gives each scope's query count and overlap note.

Scores are interpreted within the scope named in their row. VidSTG declarative and interrogative results are also retained separately.

C.4 Segmentation aggregation and comparison settings

MeViS and ReVOS evaluate expression-level target unions. ReVOS All equally averages the Referring and Reasoning group means. The evaluation pool contains 5,822 expressions. ReasonVOS uses 458 queries, whereas the Veason-R1 reference uses 253 long queries [13]. GLUS includes additional supervised fine-tuning [25], and VideoLISA includes propagation post-optimization [3]. These settings are marked with the corresponding comparison tables.

Long-RVOS validation contains 112 videos and 1,326 expressions. Its temporal and video IoU are derived from predicted mask sequences. The static, dynamic, and hybrid categories contain 460, 427, and 439 expressions, respectively.

Main-paper comparison profiles. The task overview compares BiSeek-4B with Video-o3-7B and TimeLens-7B on TimeLens-Bench, TA-STVG and VideoGrounding-DINO on VidSTG, and VideoSEG-O3-4B and Veason-R1-3B on ReVOS and GroundMoRe. Bar heights use the original percentages on a shared zero-based scale. The MeViS architecture example applies a common crop to frames and predicted masks.

D Complete quantitative results

D.1 Temporal grounding and endpoint errors

Table S5 gives the full TimeLens-Bench comparison, including recall at three overlap thresholds. Table S6 adds endpoint errors for BiSeek-4B on the three VTG benchmarks and both VidSTG query types. Start/end errors are 1.69/2.13 s on Charades-TimeLens, 11.54/18.12 s on ActivityNet-TimeLens, and 7.60/15.56 s on QVHighlights-TimeLens. The absolute errors complement IoU, which depends on both endpoint displacement and event duration. Across these three datasets, end MAE is higher than start MAE.

Table S5. Temporal grounding on TimeLens-Bench (%). R_τ denotes Recall@1 at temporal IoU threshold τ. BiSeek-4B improves mean IoU over Qwen3-VL-4B on Charades-TimeLens and ActivityNet-TimeLens.

Method	Venue/Year	Charades-TimeLens				ActivityNet-TimeLens				QVHighlights-TimeLens			
		$R_{.3}$	$R_{.5}$	$R_{.7}$	mIoU	$R_{.3}$	$R_{.5}$	$R_{.7}$	mIoU	$R_{.3}$	$R_{.5}$	$R_{.7}$	mIoU
Qwen3-VL-2B [1, 53]	arXiv 2025	62.7	44.4	23.4	43.4	53.7	41.0	26.5	39.9	67.8	53.4	39.3	52.2
Qwen3-VL-4B [1, 50]	arXiv 2025	67.1	47.5	24.1	45.6	62.9	51.0	34.3	47.2	**77.6**	**66.0**	**50.9**	**61.7**
Video-o3-7B [47, 53]	ICML 2026	58.7	34.2	16.2	38.6	53.2	41.0	24.4	38.9	62.7	49.5	34.4	47.4
VideoChat-Flash-7B [23, 48]	ICLR 2026	60.2	37.9	17.8	39.7	35.5	21.8	10.5	24.8	45.2	30.6	16.7	32.7
MiMo-VL-7B [41, 48]	arXiv 2025	57.9	42.6	20.5	39.6	49.3	38.7	22.4	35.5	57.1	42.6	28.4	41.5
TimeSuite-7B [46, 53]	ICLR 2025	56.3	35.5	18.0	38.1	27.1	17.5	8.6	19.8	27.1	16.9	9.9	21.7
VideoChat-R1-7B [22, 48]	arXiv 2025	51.9	30.8	11.7	33.7	35.0	23.9	11.3	25.0	29.3	19.1	9.4	21.5
Time-R1-7B [36, 48]	NeurIPS 2025	57.9	32.0	16.9	36.6	44.8	31.0	19.0	33.1	65.8	51.5	36.1	49.2
Qwen2.5-VL-7B [2, 48]	arXiv 2025	59.7	37.8	16.6	39.3	44.1	31.0	16.1	31.4	41.5	27.8	15.2	31.6
TimeLens-7B [48]	CVPR 2026	70.5	55.6	28.4	48.8	62.8	51.0	32.6	46.2	74.1	62.7	43.1	56.0
Qwen3-VL-8B [1, 48]	arXiv 2025	69.2	53.4	27.5	48.3	62.1	51.2	34.4	46.8	74.2	**64.6**	**49.3**	59.4
InternVideo3-8B [43, 53]	arXiv 2026	75.8	61.7	32.8	53.2	60.0	51.4	33.4	46.3	72.2	62.6	48.9	59.2
BiSeek-2B	–	**79.8**	**68.0**	**38.7**	**58.5**	**66.8**	**56.4**	**37.8**	**51.8**	74.2	62.4	45.8	59.4
BiSeek-4B	–	**82.1**	**70.4**	**40.7**	**60.5**	**69.2**	**59.4**	**40.0**	**53.9**	**75.4**	63.1	48.0	**60.9**

Table S6. Temporal localization on TimeLens-Bench and VidSTG test. IoU and recall are percentages. Mean absolute endpoint errors (MAE) are in seconds.

Model	Benchmark	Queries	m_{tIoU}	$R_{.3}$	$R_{.5}$	$R_{.7}$	Start MAE	End MAE
BiSeek-4B	Charades-TimeLens	3,363	60.5	82.1	70.4	40.7	1.69	2.13
BiSeek-4B	ActivityNet-TimeLens	4,500	53.9	69.2	59.4	40.0	11.54	18.12
BiSeek-4B	QVHighlights-TimeLens	1,541	60.9	75.4	63.1	48.0	7.60	15.56
BiSeek-4B	VidSTG (declarative)	4,610	54.4	70.4	57.8	42.0	3.67	3.94
BiSeek-4B	VidSTG (interrogative)	5,693	52.5	67.5	55.6	40.5	4.10	4.29

D.2 Spatiotemporal grounding and evaluation scopes

Tables S7 and S8 expand the STVG comparison with threshold metrics and query-type breakdowns. On VidSTG, BiSeek-4B obtains 36.7/31.3 mean video IoU for declarative/interrogative queries, compared with 35.8/30.2 for the 2B variant. The 4B model reaches 53.4 video IoU on HC-STVG v1 test and 50.1 on v2 standard validation.

Table S9 names each HC-STVG evaluation scope explicitly. The v1 full test, v2 standard validation, additional pool, and training-unseen subsets have different query compositions. The overlap notes accompany the corresponding rows. Missing entries retain a dash when the current result tables provide no measurement.

Table S7. Spatiotemporal grounding on VidSTG test (%). The declarative and interrogative subsets contain 4,610 and 5,693 queries, respectively.

Method	Declarative				Interrogative			
	m_{tIoU}	m_{vIoU}	$v_{.3}$	$v_{.5}$	m_{tIoU}	m_{vIoU}	$v_{.3}$	$v_{.5}$
STGRN [51]	48.5	19.8	25.8	14.6	47.0	18.3	21.1	12.8
STGVT [34]	–	21.6	29.8	18.9	–	–	–	–
STVGBert [33]	–	24.0	30.9	18.4	–	22.5	26.0	16.0
TubeDETR [44]	48.1	30.4	42.5	28.2	46.9	25.7	35.7	23.2
STCAT [19]	50.8	33.1	46.2	32.6	49.7	28.2	39.2	26.6
CSDVL / STVGFormer [26]	–	33.7	47.2	32.8	–	28.5	39.9	26.2
CG-STVG [14]	51.4	34.0	47.7	33.1	49.9	29.0	40.5	27.5
TA-STVG [15]	51.7	34.4	48.2	33.5	50.2	29.5	**41.5**	28.0
VideoGrounding-DINO [37]	52.0	34.7	48.1	**34.0**	50.8	29.9	41.0	27.6
E3M [4]	–	16.2	20.5	11.9	–	10.6	12.2	5.4
PTD-4B (GRPO) [29]	53.7	**38.3**	**53.0**	**37.5**	52.2	**32.4**	**44.9**	**30.3**
SpaceVLLM-7B [35]	47.7	27.4	39.1	26.2	48.5	25.4	35.9	22.2
STVG-o1-7B [16]	52.1	33.5	**48.4**	32.0	50.5	27.9	39.8	26.0
DEViL-7B [11]	50.2	33.6	46.5	**34.0**	48.5	28.8	38.7	**28.2**
BiSeek-2B	**53.8**	35.8	–	–	**52.3**	30.2	–	–
BiSeek-4B	**54.4**	**36.7**	–	–	**52.5**	**31.3**	–	–

v_τ denotes video-IoU recall at threshold τ. This expanded comparison additionally includes PTD-4B (GRPO), which is not in the condensed main-paper table. Best/second-best highlighting is computed over the methods listed in each table.

D.3 Complete segmentation metrics

Table S10 reports $\mathcal{J}$, $\mathcal{F}$, and $\mathcal{J}\&\mathcal{F}$ for both model scales. The table retains MeViS `valid` and `valid_u` as separate rows and decomposes ReVOS into Referring and Reasoning groups. Scaling from 2B to 4B improves ReVOS All from 69.6 to 72.1 $\mathcal{J}\&\mathcal{F}$, ReasonVOS from 68.5 to 72.2, and Long-RVOS from 45.0 to 54.9. Values in each metric column are rounded independently from the reported results.

Table S11 compares Ref-YouTube-VOS and Ref-DAVIS17 results. The 4B model reaches 77.9/79.9 $\mathcal{J}\&\mathcal{F}$ on these datasets. The best comparison result depends on the dataset and metric. SetCon-8B reports 78.8/80.8 $\mathcal{J}\&\mathcal{F}$, and VideoSEG-O3-2B reports 80.0 on Ref-DAVIS17. Fine-tuning and post-optimization notes remain attached to the corresponding methods.

D.4 Long-video segmentation

Table S12 gives the Long-RVOS category breakdown available for BiSeek-2B. Its $\mathcal{J}\&\mathcal{F}$ scores are 43.1, 45.1, and 46.7 for static, dynamic, and hybrid expressions. Table S13 adds comparison results from Long-RVOS [24]. The 4B model has the highest reported $\mathcal{J}\&\mathcal{F}$ among these rows at 54.9. For the 2B model, the mask-derived temporal/video IoU scores of 39.6/25.1 remain below the comparison methods. These support metrics expose temporal-support errors alongside region and boundary quality.

Table S8. Spatiotemporal grounding on HC-STVG (%). The standard v1 test and v2 validation sets contain 1,160 and 2,000 queries.

Method	HC-STVG v1 test				HC-STVG v2 val			
	m_{tIoU}	m_{vIoU}	$v_{.3}$	$v_{.5}$	m_{tIoU}	m_{vIoU}	$v_{.3}$	$v_{.5}$
STGVT [34]	–	18.2	26.8	9.5	–	–	–	–
STVGBert [33]	–	20.4	29.4	11.3	–	–	–	–
TubeDETR [44]	43.7	32.4	49.8	23.5	53.9	36.4	58.8	30.6
STCAT [19]	49.4	35.1	57.7	30.1	–	–	–	–
CSDVL / STVGFormer [26]	–	36.9	62.2	34.8	58.1	38.7	65.5	33.8
CG-STVG [14]	52.8	38.4	61.5	36.3	60.0	39.5	64.5	36.3
TA-STVG [15]	53.0	39.1	63.1	36.8	60.4	40.2	65.8	36.7
VideoGrounding-DINO [37]	–	38.3	62.5	36.1	–	39.9	67.1	34.5
E3M [4]	–	19.1	29.4	10.6	–	–	–	–
PTD-4B (GRPO) [29]	59.4	45.6	73.8	46.7	63.1	43.6	**71.0**	42.2
SpaceVLLM-7B [35]	56.9	39.3	66.6	36.9	58.0	34.0	56.9	24.7
STVG-o1-7B [16]	60.3	44.1	73.3	43.5	**63.8**	41.2	68.5	39.6
STVG-R1-7B [49]	56.9	39.1	66.7	38.6	61.3	40.8	67.9	38.8
DEViL-7B [11]	59.0	43.1	70.5	44.3	61.7	42.5	67.3	42.2
BiSeek-2B	**68.0**	**50.1**	**83.4**	**56.1**	63.3	**48.5**	61.4	**49.2**
BiSeek-4B	**72.2**	**53.4**	**87.0**	**62.6**	**67.9**	**50.1**	**72.2**	**50.6**

Table S9. HC-STVG evaluation scopes for BiSeek (%).

Dataset	Scope	Queries	Model	m_{tIoU}	m_{vIoU}	$v_{.3}$	$v_{.5}$
HC-STVG v1	Full test	1,160	BiSeek-2B	68.0	50.1	83.4	56.1
HC-STVG v1	Full test	1,160	BiSeek-4B	72.2	53.4	87.0	62.6
HC-STVG v1	Unseen training videos	447	BiSeek-2B	61.3	41.7	74.3	36.7
HC-STVG v2	Standard validation	2,000	BiSeek-2B	63.3	48.5	61.4	49.2
HC-STVG v2	Standard validation	2,000	BiSeek-4B	67.9	50.1	72.2	50.6
HC-STVG v2	Additional evaluation pool	3,482	BiSeek-2B	66.1	40.3	67.4	35.7
HC-STVG v2	Additional evaluation pool	3,482	BiSeek-4B	67.9	41.2	69.7	37.2
HC-STVG v2	Unseen training videos	1,768	BiSeek-2B	63.9	41.3	70.2	36.9

For BiSeek-2B, the v1 full test includes 713 queries from training-seen videos, and the v2 additional pool includes 1,714.

D.5 GroundMoRe question types

Table S14 reports overall, causal, sequential, counterfactual, and descriptive scores. The 4B model improves $\mathcal{J}\&\mathcal{F}$ over 2B by 5.9 points on causal questions, 3.9 on sequential questions, and 5.4 on descriptive questions. Counterfactual $\mathcal{J}\&\mathcal{F}$ is 62.3/62.2 for 2B/4B, showing that the scaling trend varies across question types. The overall 4B score is 61.9.

Table S10. Complete segmentation metrics (%). MeViS `valid_u` is reported separately from the official `valid` split.

Dataset / protocol	Model	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
MeViS `valid`	BiSeek-2B	58.0	66.3	62.1
MeViS `valid`	BiSeek-4B	61.7	68.8	65.3
MeViS `valid_u` (diag.)	BiSeek-2B	61.7	70.7	66.2
MeViS `valid_u` (diag.)	BiSeek-4B	67.0	75.2	71.1
ReasonVOS	BiSeek-2B	64.7	72.3	68.5
ReasonVOS	BiSeek-4B	68.6	75.7	72.2
ReVOS Referring	BiSeek-2B	67.8	74.6	71.2
ReVOS Referring	BiSeek-4B	70.9	77.1	74.0
ReVOS Reasoning	BiSeek-2B	64.6	71.4	68.0
ReVOS Reasoning	BiSeek-4B	66.9	73.6	70.2
ReVOS All	BiSeek-2B	66.2	73.0	69.6
ReVOS All	BiSeek-4B	68.9	75.3	72.1
Ref-YT-VOS valid	BiSeek-2B	73.8	79.1	76.5
Ref-YT-VOS valid	BiSeek-4B	75.4	80.5	77.9
Ref-DAVIS17 valid	BiSeek-2B	72.6	82.4	77.5
Ref-DAVIS17 valid	BiSeek-4B	75.7	84.1	79.9
Long-RVOS valid	BiSeek-2B	44.0	45.9	45.0
Long-RVOS valid	BiSeek-4B	54.0	55.8	54.9

Table S11. Additional referring video segmentation comparisons (%). Blue and orange values mark the best and second-best scores in each column.

Method	Ref-YT-VOS			Ref-DAVIS17		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
VideoSEG-O3-2B [8]	70.5	68.3	72.7	**80.0**	**76.0**	**84.0**
VideoLISA-3.8B[B] [3]	63.7	61.7	65.7	68.8	64.9	72.7
VideoSEG-O3-4B [8]	74.1	71.5	76.6	79.4	75.5	83.2
GLUS[A]-7B [25]	67.3	65.5	69.0	–	–	–
VRS-HQ-7B [12]	70.4	68.3	72.5	76.0	72.6	79.4
UniPixel-7B [27]	71.0	69.5	72.4	76.4	72.7	80.1
VIRST-7B [17]	74.2	72.2	76.1	79.5	**75.9**	83.1
SetCon-8B [52]	**78.8**	–	–	**80.8**	–	–
BiSeek-2B	76.5	**73.8**	**79.1**	77.5	72.6	82.4
BiSeek-4B	**77.9**	**75.4**	**80.5**	79.9	75.7	**84.1**

[A] Additional supervised fine-tuning. [B] Post-optimization.

Table S12. Long-RVOS validation results (%). Temporal and video IoU are computed from the predicted mask sequences.

Model	Subset	Expressions	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	tIoU	vIoU
BiSeek-2B	Static	460	42.3	44.0	43.1	34.5	21.0
BiSeek-2B	Dynamic	427	44.2	46.1	45.1	41.1	26.3
BiSeek-2B	Hybrid	439	45.7	47.8	46.7	43.5	28.2
BiSeek-2B	Overall	1,326	44.0	45.9	45.0	39.6	25.1
BiSeek-4B	Overall	1,326	54.0	55.8	54.9	–	–

Table S13. Long-RVOS validation comparison (%). External validation results follow supplementary Table 9 of Long-RVOS [24].

Method	$\mathcal{J}\&\mathcal{F}$	tIoU	vIoU
SAMWISE	41.2	63.5	31.2
ReferDINO	49.6	**74.7**	**45.6**
ReferMo	**53.7**	**74.3**	**46.8**
BiSeek-2B	45.0	39.6	25.1
BiSeek-4B	**54.9**	–	–

The evaluation covers 1,326 expressions; temporal and video IoU are computed from mask sequences.

Table S14. Question-type breakdown on GroundMoRe test (%). Earlier baseline scores follow [8, 38].

Method	Overall			Causal			Sequential			Counterfactual			Descriptive		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Title+ReferFormer	10.1	9.7	10.5	10.5	9.9	11.1	9.1	8.6	9.6	9.4	9.1	9.8	11.4	11.2	11.6
ReferFormer [40]	12.7	11.2	14.3	12.7	11.0	14.4	10.7	9.4	12.1	12.4	11.1	13.6	15.2	13.3	17.0
SgMg	17.5	15.8	19.2	18.9	17.0	20.9	15.5	14.0	17.0	16.4	15.2	17.5	18.8	16.7	20.9
HTR	14.5	12.7	16.2	15.7	13.8	17.5	12.5	10.8	14.1	13.5	12.2	14.9	15.9	13.8	18.1
LMPM [10]	13.3	12.7	14.0	13.7	13.1	14.3	11.3	10.6	12.0	13.2	12.6	13.7	14.3	13.6	15.1
LISA-7B [20]	6.1	4.9	7.3	5.5	4.4	6.7	5.2	4.1	6.3	5.9	4.8	7.0	7.6	6.1	9.0
PixelLM-7B [31]	10.0	9.9	10.0	9.5	9.4	9.7	9.2	9.2	9.3	9.8	9.8	9.9	10.9	11.0	10.9
PG-Video-LLaVA	11.2	11.5	10.9	10.6	10.8	10.3	10.9	11.2	10.7	10.4	10.4	10.3	12.5	13.0	12.0
PG-Video-LLaVA+SAM2	11.9	11.2	12.6	13.4	12.8	13.9	3.9	2.5	5.2	15.5	15.2	15.9	12.1	11.4	12.9
VISA [42]	5.3	4.7	5.9	6.1	5.5	6.6	4.5	3.8	5.2	5.1	4.3	5.9	5.4	5.0	5.8
ViLA+ReferFormer	16.4	14.8	17.9	15.9	14.3	17.5	16.3	14.9	17.7	13.6	12.4	14.9	19.6	17.8	21.5
ViLA+SgMg	17.1	15.3	18.9	17.3	15.4	19.3	16.3	14.6	18.0	15.3	14.0	16.7	19.2	17.0	21.4
ViLA+HTR	15.7	13.9	17.5	16.5	14.6	18.3	15.1	13.4	16.8	12.6	11.2	14.0	18.4	16.1	20.6
VideoChat2+ReferFormer	15.4	13.8	16.9	14.7	13.1	16.4	14.0	12.7	15.2	13.1	11.8	14.4	19.8	17.9	21.7
VideoChat2+SgMg	16.7	14.8	18.5	16.6	14.8	18.5	14.7	13.0	16.4	15.2	13.9	16.6	20.0	17.7	22.3
VideoChat2+HTR	15.1	13.3	16.9	15.9	14.0	17.8	13.9	12.3	15.5	12.0	10.6	13.4	18.2	15.8	20.5
SeViLA+ReferFormer	20.4	19.0	21.8	21.3	19.9	22.7	18.3	17.1	19.5	17.6	16.4	18.8	24.0	22.4	25.7
SeViLA+SgMg	22.3	20.9	23.8	21.6	20.2	23.0	21.7	20.5	22.9	20.5	19.3	21.6	24.5	22.8	26.2
SeViLA+HTR	21.8	19.9	23.6	22.3	20.4	24.2	20.4	18.6	22.1	19.7	18.0	21.4	24.4	22.1	26.7
MoRA (zero-shot) [9]	23.1	22.8	23.5	21.6	21.5	21.8	22.4	22.0	22.8	26.4	25.9	26.9	22.6	22.3	23.0
VideoSEG-O3-2B [8]	29.1	27.4	30.7	30.9	29.2	32.7	27.0	25.4	28.5	22.0	20.6	23.5	35.5	33.5	37.4
VideoSEG-O3-4B [8]	32.0	30.2	33.7	34.7	32.9	36.5	29.7	28.1	31.3	25.1	23.5	26.6	37.4	35.2	39.6
Veason-R1-3B [13]	23.5	22.0	25.0	23.7	22.0	25.4	21.1	19.8	22.3	22.8	21.5	24.1	26.5	24.8	28.3
BiSeek-2B	**58.0**	**57.9**	**58.1**	**56.4**	**56.2**	**56.6**	**62.8**	**62.7**	**62.8**	**62.3**	**62.3**	**62.4**	**51.1**	**50.9**	**51.4**
BiSeek-4B	**61.9**	**61.1**	**62.7**	**62.3**	**61.4**	**63.2**	**66.7**	**66.2**	**67.2**	**62.2**	**61.5**	**62.8**	**56.5**	**55.5**	**57.6**

E Additional qualitative comparisons

We present twelve fixed qualitative cases, with four cases each for video temporal grounding (VTG), spatiotemporal video grounding (STVG), and referring video object segmentation (RVOS). The queries and six display frames per case were selected before running the public comparison methods. Every method is shown on the same source frames, and each figure retains the complete query. Case metrics describe these selected clips. The BiSeek predictions use the 4B model at update 30,210 and retain the saved final outputs.

E.1 Video temporal grounding

We compare UFVideo-7B [28], VideoLoom-4B [32], and BiSeek-4B using their saved native predictions and the same annotations. The timelines and tIoU values in these figures use continuous clip-relative seconds. The benchmark tables apply the TimeLens one-second endpoint rounding protocol, so the two presentations can yield different scores for the same prediction. BiSeek uses the fixed order-constrained MAP output throughout these comparisons.



Figure S1. A complete action chain. The query spans picking up a pillow, placing it behind a door, and closing the door. BiSeek predicts 3.20–12.23 seconds against the annotated 3.00–13.00 seconds, reaching 90.28% continuous tIoU. UFVideo begins at the clip start and extends beyond the annotated end, while VideoLoom also begins early. The remaining BiSeek end error is 0.77 seconds. This sequence illustrates the need to align the interval with several related actions as the manipulated object moves out of view.

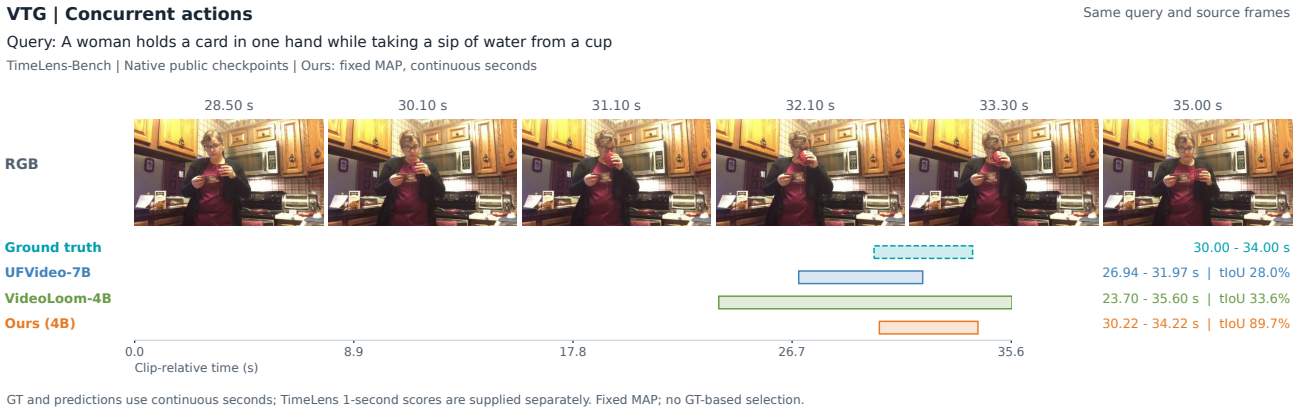

Figure S2. Concurrent actions within a short interval. The referred event requires card holding and drinking to occur together. BiSeek obtains 89.72% continuous tIoU, with start and end absolute errors of 0.22 seconds each. UFVideo and VideoLoom obtain 27.98% and 33.61%, respectively. The six common frames expose the surrounding activity and the brief co-occurrence described by the query. The remaining endpoint offsets show the temporal precision still required when a short event occupies only a small part of the clip.

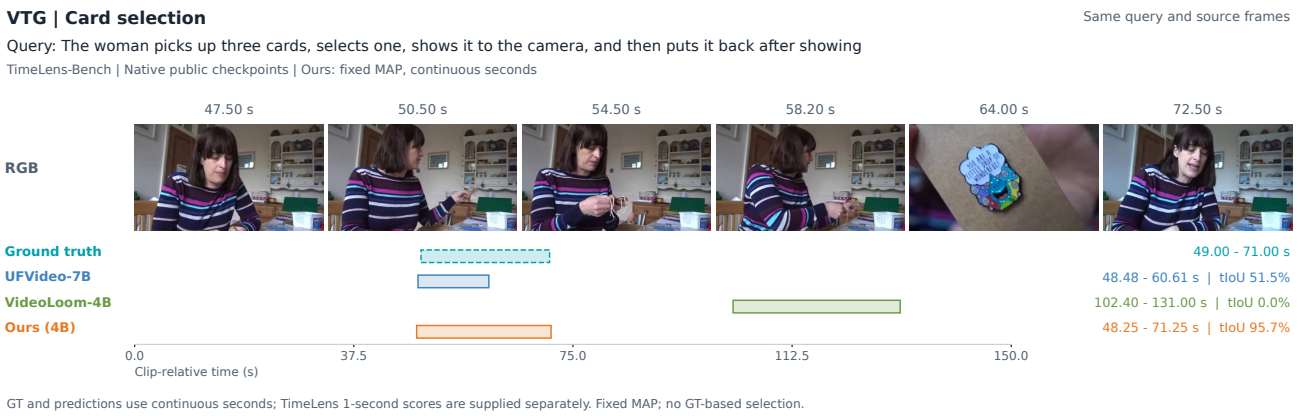

Figure S3. Selecting, presenting, and returning a card. This query combines several actions while the apparent scale of the cards changes during presentation to the camera. BiSeek reaches 95.65% continuous tIoU, with start and end errors of 0.75 and 0.25 seconds. UFVideo reaches 51.55% and has a 10.39-second end error. VideoLoom predicts a disjoint interval and receives zero overlap. The displayed frames place the selected card sequence within its surrounding activity and retain the native output intervals for all three methods.

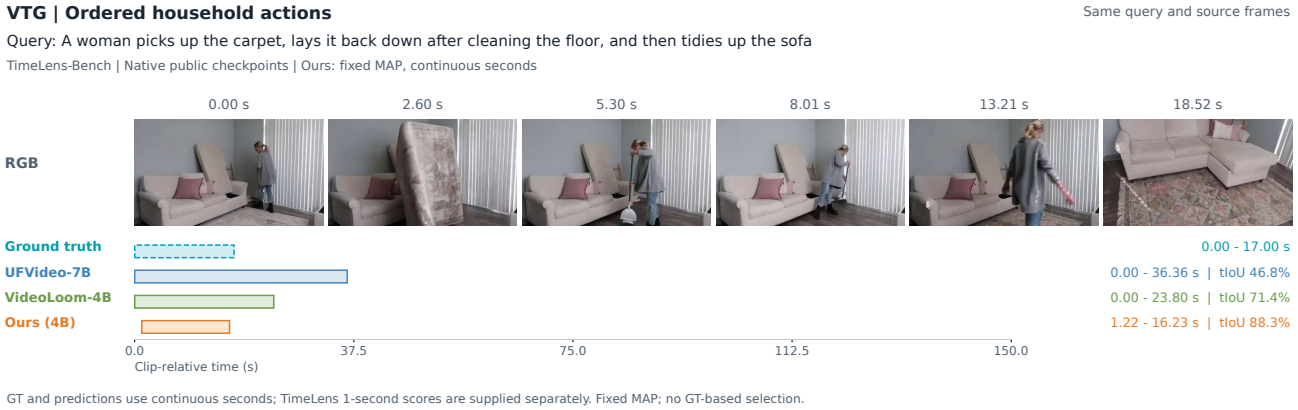

Figure S4. Connecting carpet handling with sofa tidying. The expression spans lifting a carpet, cleaning the floor, replacing the carpet, and tidying the sofa. BiSeek obtains 88.31% continuous tIoU, with start and end errors of 1.22 and 0.77 seconds. UFVideo and VideoLoom obtain 46.75% and 71.43%, with end errors of 19.36 and 6.80 seconds. Their exact starting boundaries still accompany substantially different ending boundaries, illustrating how a query with several actions can remain difficult after its beginning has been localized.

E.2 Spatiotemporal video grounding

We compare STVG-R1-7B [49], DeViL-7B [11], and BiSeek-4B on four VidSTG queries. All predictions are rescored with the same grounding scorer, annotations, and evaluation frames. STVG-R1's native proposal outputs are continued within its predicted interval using its ID, IoU, and largest-box fallback rules. This export initializes from predictions alone.

DeViL retains its native 64-frame timestamp labels and is mapped to common frames by the nearest label; a sampled frame with no box preserves that missing output. The resulting case scores use this shared frame protocol. Both temporal and spatial errors are visible in the figures.

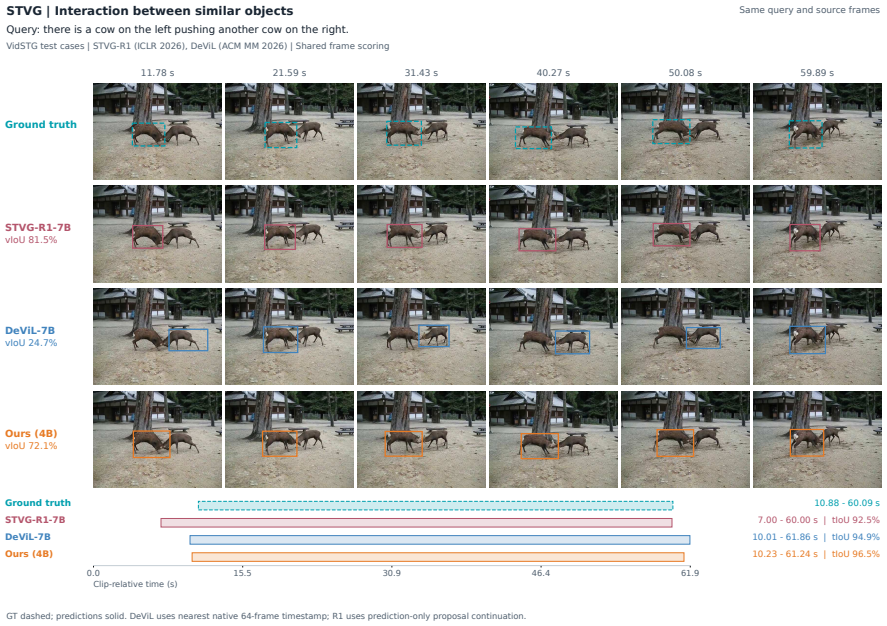

Figure S5. Interaction roles between similar animals. The query identifies the animal on the left that pushes its neighbor. BiSeek follows the annotated individual across the displayed frames and obtains 96.47% tIoU and 72.10% vIoU. STVG-R1 has a higher vIoU of 81.53%, together with 92.52% tIoU. DeViL reaches 94.92% tIoU and 24.68% vIoU while its displayed boxes frequently select the neighboring animal. The case retains the spatial advantage of STVG-R1 and BiSeek's residual box extent errors despite close temporal boundaries.

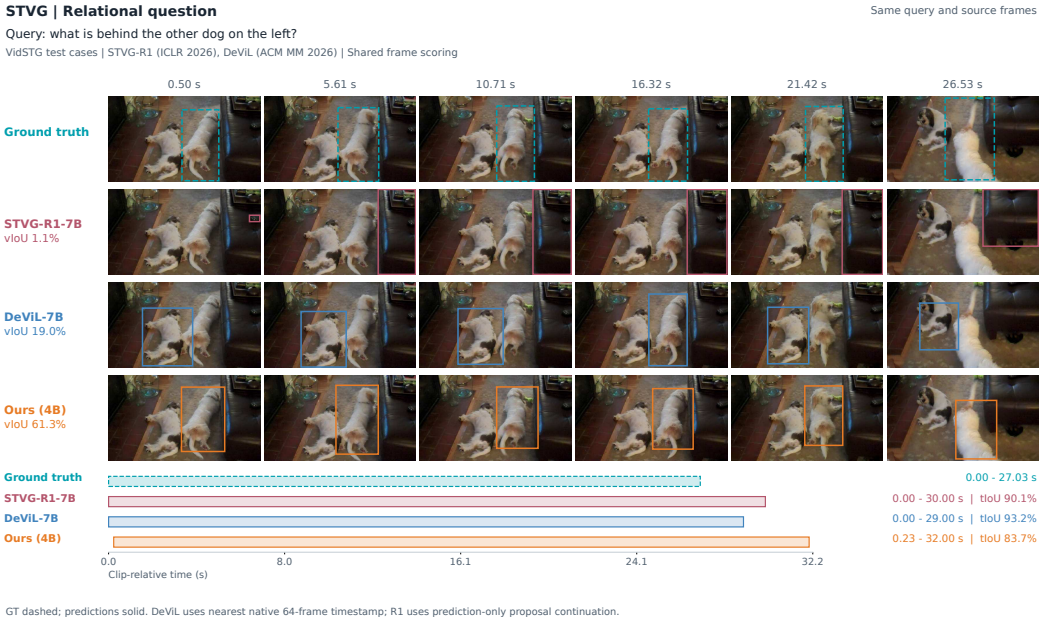

Figure S6. A relational query with temporal overprediction. The two dogs change their relative positions as the scene progresses. BiSeek obtains 61.31% vIoU, compared with 1.06% for STVG-R1 and 18.99% for DeViL. BiSeek retains the target through much of the sequence, while the final displayed box covers only part of its body. Its interval extends to 32.00 seconds against an annotated end of 27.03 seconds, producing a 4.97-second end error and 83.73% tIoU. Both comparison methods have higher temporal overlap in this example.



Figure S7. Foreground identity under partial occlusion. The expression combines green clothing with the person's position in front of a white-clothed adult. The foreground person's hand-held sheet intermittently occludes the face and upper body. BiSeek obtains 91.11% tIoU and 88.66% vIoU, while DeViL reaches 91.20% and 92.70%, respectively. STVG-R1 starts its interval substantially earlier and obtains 29.66% vIoU. The retained DeViL advantage shows that successful referent selection and close temporal boundaries can still leave differences in box tightness.



Figure S8. Hat attributes and a neighboring person. Several adults appear together, and the query combines the target's gray hat with its position next to a white-hatted person. BiSeek obtains 98.92% tIoU and 80.65% vIoU. STVG-R1 obtains 98.15% and 80.00%, while DeViL obtains 98.41% and 73.14%. All three temporal intervals closely cover the annotation. The remaining differences arise in the spatial predictions, where similar clothing, adjacent bodies, and changing posture affect the extent of the selected person's box.

E.3 Referring video object segmentation

We compare UniPixel-3B [27], UFVideo-7B [28], VideoLoom-4B [32], and BiSeek-4B on four ReasonVOS queries. Saved masks are independently rescored against the same full sequences using common region and boundary metrics. The four sequences contain 205, 55, 30, and 193 frames, yielding 483 evaluated masks per method and 1,932 masks across the four methods.

Each figure displays six fixed frames, while its scores use every original sequence frame. The hand-transfer case uses synchronized context and identical zoom windows for all methods.



Figure S9. A small object transferred between hands. The full query identifies a hand-held object through its change of hand. Identical crops magnify the existing image and mask pixels, exposing local shape errors and partial occlusion. BiSeek obtains 78.40% $\mathcal{J}\&\mathcal{F}$ over all 205 frames, compared with 67.79% for UniPixel, 59.41% for UFVideo, and 4.00% for VideoLoom. The tiny visible target at display index 5 is missed by BiSeek, preserving a difficult occlusion example within the otherwise higher-scoring sequence.

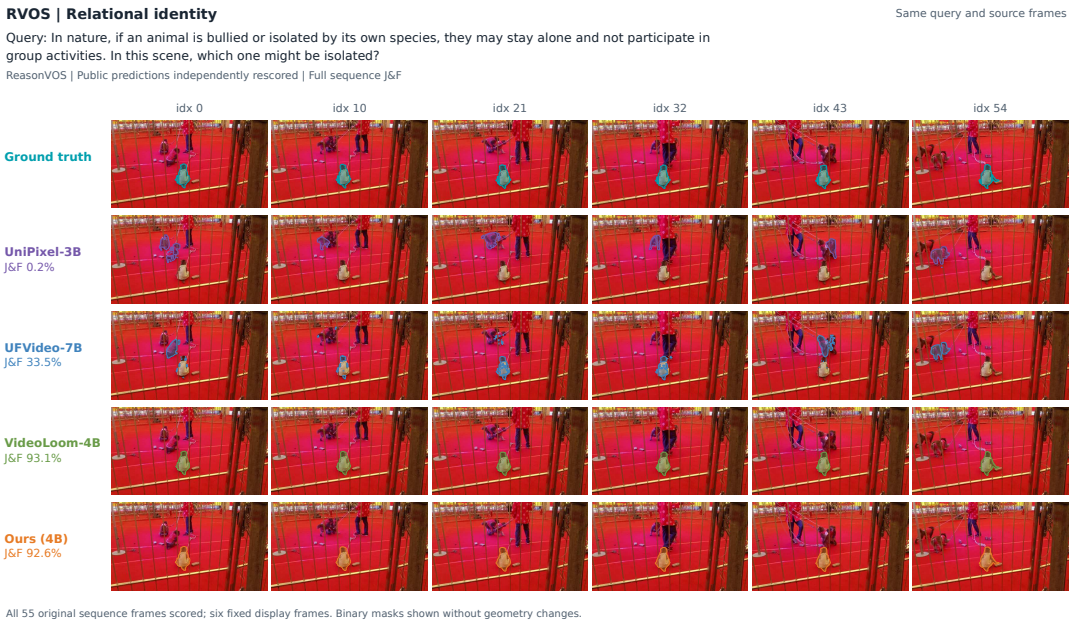

Figure S10. Identifying an isolated animal from group behavior. The query relates an individual's separation from group activity to its identity among similar animals, with foreground bars crossing the scene. VideoLoom reaches 93.12% $\mathcal{J}\&\mathcal{F}$ and exceeds BiSeek's 92.61%. BiSeek has higher region IoU, 88.47% against 87.05%, while VideoLoom has higher boundary F, 99.18% against 96.76%. UniPixel and UFVideo reach 0.17% and 33.54% $\mathcal{J}\&\mathcal{F}$. All 55 frames contribute to these scores, retaining the complementary region and boundary behavior of the two strongest predictions.

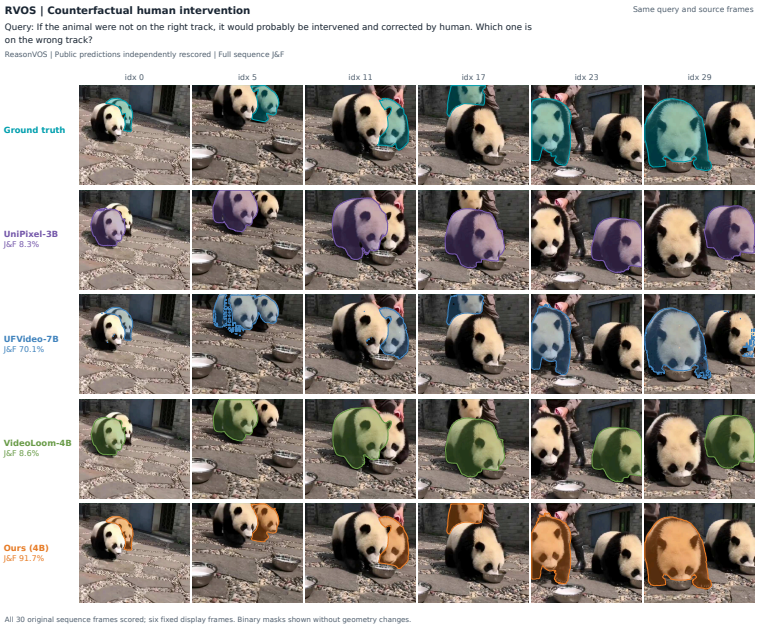

Figure S11. A conditional query resolved through human intervention. The query refers to the panda whose behavior prompts a person to intervene and correct it. The displayed sequence exposes that interaction together with nearby animals that provide competing referents. BiSeek obtains 91.71% $\mathcal{J}\&\mathcal{F}$ over the complete 30-frame sequence, with region IoU of 93.03% and boundary F of 90.38%. UFVideo obtains 70.06%, while UniPixel and VideoLoom obtain 8.27% and 8.56%. The masks illustrate how the selected referent is maintained as the intervention unfolds.



Figure S12. A blanket identified through its warming function. The query connects feeling cold after watching television with an object that can provide warmth. The blanket changes its visible area and shape as it is handled and unfolded. BiSeek obtains 85.69% $\mathcal{J}\&\mathcal{F}$ across all 193 frames, with region IoU of 83.70% and boundary F of 87.69%. VideoLoom obtains 19.10%, UFVideo 1.13%, and UniPixel 0.00%. The common frames expose both the initially compact visible region and the later extended fabric, including residual mask-boundary differences.

F Attention visualization and controlled inspection

Figures S13 and S14 extend the query-attention examples to all four native temporal units. Both 2B checkpoints receive the same eight frames and instruction, with matching input-token and pixel hashes, visual grids, and readout positions. Each pair is encoded once.

Maps average all 16 heads at intermediate layer 14, on a 14×26 post-merge spatial grid per unit. Each row shows the first of its two input frames. Dividing attention by each unit's spatial mean makes uniform attention equal to one. Both models, both cases, and all four units share the main paper's linear color scale, with limits 0 and 21.5395336. The upper 0.5% of this pooled density distribution is clipped. The first two units reproduce the main-paper comparison; the remaining units extend the same visualization protocol.

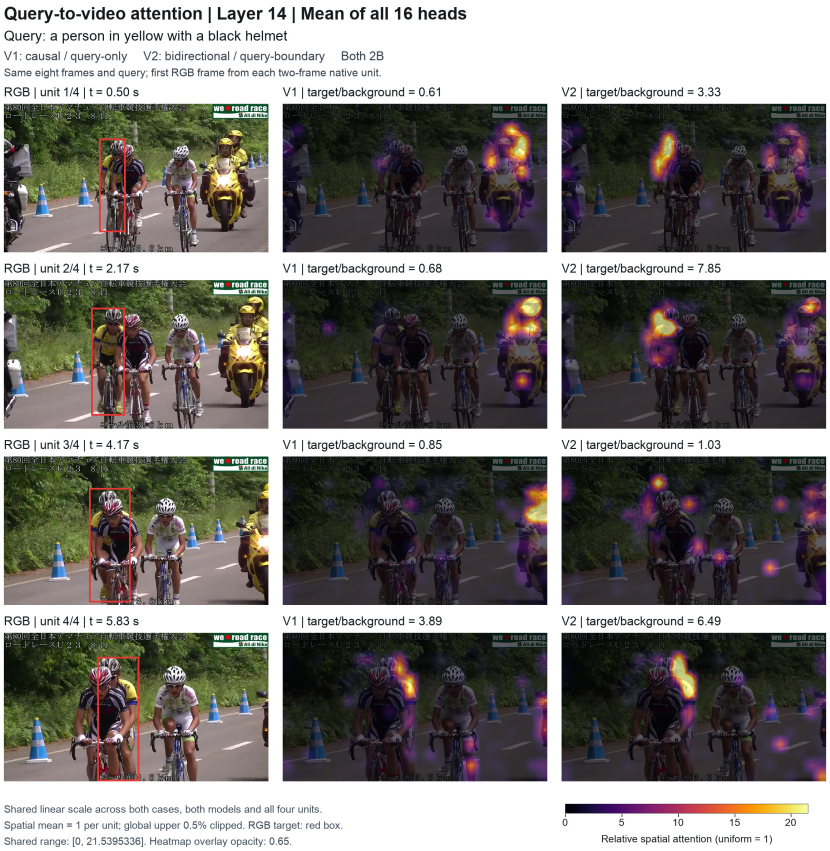

Figure S13. Black-helmet cyclist, Ref-YouTube-VOS video `547416bda1`, expression 4. V1 is causal/query-only; V2 is bidirectional/query–boundary. The first two units' target/background attention-density ratios change from 0.61/0.68 to 3.33/7.85. The third unit shows a smaller change, 0.85 to 1.03. Red boxes identify the annotated target in the RGB frames. Target masks are used for the post-hoc density statistic.

Matched-frame mask inspection. Independent SAM 3 decoding of the same eight images, using cached conditions and confidence threshold 0.5, yields 58.94 → 76.54 $\mathcal{J}\&\mathcal{F}$ for the black-helmet case and 65.35 → 95.36 for the yellow-jacket case. This diagnostic uses independent image predictions and evaluates those eight images. The benchmark tables use the full-video evaluation protocol.

The checkpoints differ in both contextualization and conditioning, with one projected query row in V1 and three query–boundary rows in V2. Their training histories also differ. These comparisons therefore measure the complete variants, with the main paper's ablations addressing individual architectural choices.



Figure S14. Yellow-jacket cyclist, Ref-YouTube-VOS video `3674b2c70a`, expression 2. V1 is causal/query-only; V2 is bidirectional/query–boundary. Both variants attend to the target. The first two units' target/background ratios increase from 2.39/2.11 to 10.03/8.59, with higher target concentration also visible in the later units. The color scale is shared with the black-helmet case and the main-paper attention figure.

G Limitations and failure analysis

Selected examples and remaining failures. The attention examples were selected from 32 candidates. Across the eight-case shortlist evaluated with identical eight-frame independent decoding, the phone example `reasonvos_6FV5P_2` scores 0.00 $\mathcal{J}\&\mathcal{F}$ for both variants. The cow example `refyt_3dd327ab4e_2` decreases from 61.26 to 54.31. These cases show that the complete bidirectional variant can still fail to recover the referred target or yield weaker masks. In the black-helmet example, the third unit's attention-density ratio is close to one for both variants. Relative spatial normalization describes concentration within a unit. Its color intensity alone cannot establish how much total probability the query assigns to the video.

Offline context and temporal sampling. BiSeek assumes that the complete sampled clip and instruction are available before prediction. Its bidirectional encoding uses future sampled context, which makes the present formulation an offline system. RVOS inference represents the clip with 32 uniformly sampled context frames. Brief actions or identity cues occurring between these frames may be absent from the semantic conditions. Independent image features supply spatial detail for decoded frames, while their language conditions retain the temporal resolution of the sampled native units. Long clips can therefore place widely separated moments under nearby or shared conditions.

Output scope. The temporal branch selects one interval, and the box branch predicts a single-target tube. Queries requiring multiple disjoint event intervals or multiple independent box trajectories require an extended output formulation. The mask branch supports target sets, whose success still depends on selecting the intended instances and maintaining their identities over time. These architectural limits define the scope of the current evaluation.