SEEKSPAN: CALIBRATING TEMPORAL EXTENT FOR VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video temporal grounding predicts the position (the midpoint) and the duration (extent) of an event in a video. Existing training-free verification methods replace the generated interval with one built from clip or window scores, even when the generated interval is well placed and only its duration is wrong. We propose SEEKSPAN, a training-free method that corrects the duration of a predicted interval around its original midpoint. A frozen video-language model, the verifier, views the full video and scores each temporal window by the yes–no log-probability margin for the queried event occurring within it, and the scores across windows estimate how long the event lasts. We take the geometric mean of this estimate and the predicted duration and place the start and end times around the original midpoint, adjusting at video endpoints when needed. With one frozen verifier shared across generator families, SEEKSPAN improves 12 of 13 configurations on Charades-STA, including all six thinking-mode configurations. The gains exceed one mIoU point in 11 configurations and reach 9.2 points. With a frozen Qwen3-VL-8B generator, SEEKSPAN reaches 57.8 mIoU on the full Charades-STA test set, above every published training-free method and above the re-measured systems trained without Charades-STA annotations.

1 INTRODUCTION

Video temporal grounding localizes the time interval that corresponds to a natural-language query in an untrimmed video (Gao et al., 2017b; Hendricks et al., 2017). Recent video-language models generate the start and end timestamps in their answers. Temporal instruction tuning and reinforcement-learning post-training improve their grounding (Huang et al., 2024; Qian et al., 2024; Wang et al., 2025b). A well-placed interval can still have the wrong length. For example, a short interval inside the queried event covers only part of it and overlaps little with the annotation (Figure 1). Position and duration are therefore separate sources of grounding error.

Training-free approaches extract more temporal evidence from frozen models (Zheng et al., 2024). REZE, UG-Ground, and FV-Action build the interval from per-clip or per-window verification scores (Li et al., 2026a; Kim et al., 2026; Huang et al., 2026). In these methods, the verification signal also sets the location, even when a generated prediction already places the event well. We use verification only to correct the duration and keep the generated position (Figure 1; Table 2).

Thinking mode in Qwen3.5-4B (Team, 2026) shows this separation on Charades-STA. On 3,322 queries where both modes return a valid interval, enabling thinking shortens the median predicted duration from 7.0 s to 5.0 s and lowers mIoU by 6.6 points. The center error changes only from 3.27 s to 3.32 s. Giving the thinking answer the direct-mode duration recovers 6.7 points. Replacing its center instead changes mIoU by -1.1 (Section 4.3). Direct mode also underestimates duration. For Qwen3-VL-8B, InternVL3.5-8B, and MiMo-VL-7B (Bai et al., 2025a; Wang et al., 2025a; Team et al., 2025), 57–62% of direct predictions are shorter than their annotations.

We introduce SEEKSPAN, a training-free method that calibrates the duration of an existing temporal-grounding prediction. A frozen verifier scores temporal windows over the full video, and these scores give a second duration estimate. SEEKSPAN combines this estimate with the generated duration and rebuilds the interval around the generated midpoint. Both models stay frozen, and one verifier is shared across generators.

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

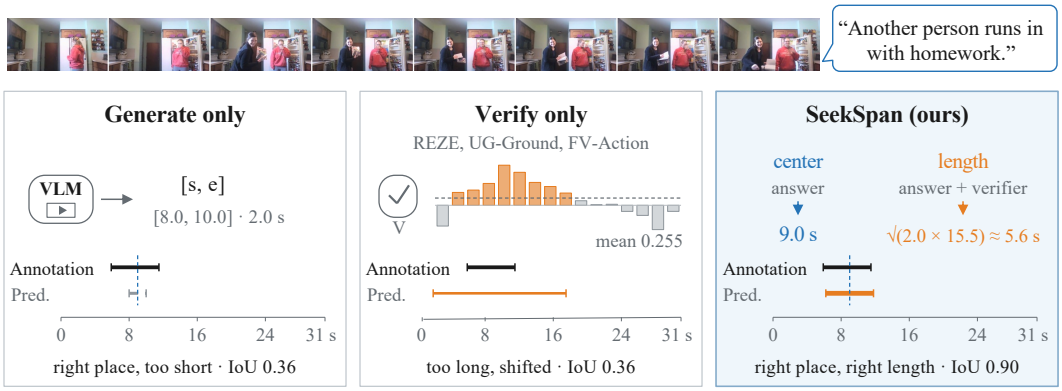


Figure 1: **Generation alone, verification alone, and SEEKSPAN on one query.** Charades-STA query 434 with Qwen3.5-4B thinking; frames every 3.9 s. The generated answer is well placed but too short (2.0 s for a 5.6 s event). The middle panel builds the interval from the verifier scores alone, as score-based methods such as REZE and UG-Ground do. Taking the contiguous above-mean region gives an interval that is too long and shifted (15.5 s). SEEKSPAN keeps the position of the answer and takes the geometric mean of the two lengths (5.6 s).

We evaluate SEEKSPAN on seven generators and thirteen model–mode configurations with a shared Qwen3.5-4B verifier and a fixed fusion weight. On the Charades-STA test set, SEEKSPAN improves twelve of the thirteen configurations, including all six thinking configurations. With a frozen Qwen3-VL-8B generator, SEEKSPAN raises mIoU from 52.9 to 57.2 in direct mode and from 55.0 to 57.8 with thinking, above the 55.7 that REZE reports with the same backbone and every other published training-free method (Table 1). Ablations of the duration source and the center source favor fusing the two duration estimates and keeping the generated position. The largest gains come from predictions that underestimate duration. Calibration transfers to other datasets where the generated center is useful and the verification duration is informative. On HiREST, for example, SEEKSPAN improves all four evaluated configurations.

Our contributions are:

- We identify a position–extent separation in temporal grounding, in which a prediction places the event well but misjudges its duration.
- We propose SEEKSPAN, a training-free extent-calibration method that keeps the generated position and shares one frozen verifier across generators.
- With one shared verifier, SEEKSPAN improves 12 of 13 generator–mode configurations on Charades-STA by up to 9.2 mIoU. When the published verification rules are applied to the same verification curve, SEEKSPAN outperforms all three in five of six configurations.

2 RELATED WORK

Video temporal grounding with video-language models. Early temporal grounding models learn video–language alignment to score candidate moments and refine boundaries (Gao et al., 2017b; Hendricks et al., 2017). Recent video-language models add grounding through temporal instruction tuning (Huang et al., 2024; Qian et al., 2024; Wang et al., 2024), explicit temporal representations (Chen et al., 2024), and structured event modeling (Guo et al., 2025). Reinforcement fine-tuning with task-specific rewards and sampling strategies improves grounding further (Wang et al., 2025b; Li et al., 2025a;b). EvoGround learns from unlabeled videos through interacting proposer and solver agents (Jung et al., 2026), and Video-OPD uses on-policy distillation for dense per-token supervision (Li et al., 2026b). These methods improve grounding through training. SEEKSPAN takes an already generated interval and leaves the generator's parameters and decoding unchanged.

Training-free grounding and temporal verification. Training-free methods extract temporal evidence from pretrained models at inference time (Table 2). Some change the generator's input or de-

coding. DART routes between fast and slow paths (Zhang et al., 2026b), Know-When crops or masks the video according to attention cues before regenerating timestamps (Du et al., 2026), and MarkIt adds query-conditioned visual markers and frame indices to the frames (Fang et al., 2026). Others assemble the interval from scores. TFVTG combines sub-event proposals (Zheng et al., 2024), DSE-VTG combines per-frame and per-clip similarities with test-time adaptation (Cao et al., 2026), and UG-Ground and REZE use per-window or per-clip yes–no confidences (Kim et al., 2026; Li et al., 2026a). FV-Action places a fixed- or adaptive-width span at a verification peak (Huang et al., 2026), and ZEAL does the same for zero-shot action localization (Aklilu et al., 2025). In these score-based methods, the verification signal determines where the final interval lies. SEEKSPAN keeps the generated position and uses verification only for the temporal extent.

Reasoning and inference-time computation. TimeLens applies thinking-free reinforcement learning with verifiable rewards to temporal grounding (Zhang et al., 2026a). TaRO trains time-aware reasoning with visually grounded reasoning paths and a temporal-sensitivity reward (Zheng et al., 2026), and VideoAuto-R1 studies confidence-based selective reasoning (Liu et al., 2026). Diagnostic work finds inconsistencies between grounding predictions and temporal judgments about the same event (Jung et al., 2025). Our analysis separates how thinking changes the position of a prediction from how it changes the duration. Self-consistency samples and aggregates multiple reasoning paths (Wang et al., 2023), and video test-time scaling methods add frames when sampled answers disagree (Wang et al., 2025c) or revisit predicted spatiotemporal regions through iterative perception (Yan et al., 2025). SEEKSPAN corrects only the extent of a single generated prediction, without decoding additional answers or reasoning traces.

3 SEEKSPAN

SEEKSPAN starts from an existing temporal-grounding prediction, keeps its midpoint, and adjusts only its duration (Figure 2). A verifier scores temporal windows over the full video, and these scores give a second estimate of the event duration. We combine this estimate with the generated duration and reconstruct the interval around the original midpoint. The generator and the verifier both remain frozen. We use *position* for the interval midpoint and w for its duration.

3.1 POSITION AS THE ANCHOR

For a video of duration D and query q, let $[s, e]$ be the generated interval, with midpoint $c = (s+e)/2$ and width $w = e - s > 0$, and let c^* and w^* denote the annotated midpoint and width. When the generated and annotated intervals have the same midpoint, $c = c^*$, temporal IoU depends only on their widths:

$$\mathrm{IoU}([s, e], [s^*, e^*]) = \frac{\min(w, w^*)}{\max(w, w^*)} = \exp(-\,|\log w - \log w^*|)\,. \tag{1}$$

With an aligned midpoint, overestimating or underestimating the width by the same factor gives the same IoU, so we combine duration estimates in log-width space. When the midpoint is displaced, width and position interact (Section 3.4).

3.2 VERIFICATION-DERIVED DURATION

We divide the video timeline into B uniform windows. For each window, we give the verifier the full video, the query, and the time range of the window, and ask whether the queried event occurs within that range. Input frames are sampled at 1 fps, with at most 128 per video.

Let $\mathcal{Y}$ and $\mathcal{N}$ denote the one-token variants of "yes" and "no," including capitalization and leading-space variants. For window b, we use the next-token log-probability margin

$$v_b = \log \sum_{y \in \mathcal{Y}} p_V(y \mid \text{video}, q, b) - \log \sum_{n \in \mathcal{N}} p_V(n \mid \text{video}, q, b). \tag{2}$$

The probabilities are read at the answer position in one forward pass, without decoding an answer or a reasoning trace. A positive value means that the verifier places more probability mass on the yes

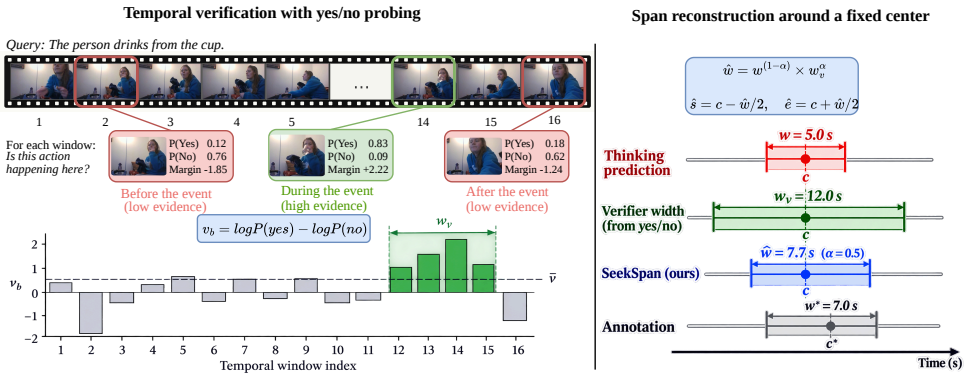


Figure 2: **SEEKSPAN combines generated and verification widths.** The verifier scores B temporal windows, and the contiguous above-mean region around the peak defines w_v. SEEKSPAN discards the region's location, combines w_v with the generated width w, and reconstructs the interval around the generated center c.

variants than on the no variants. The softmax normalizer cancels because both terms use the same output distribution. Across windows, the input differs only in the queried time range.

The resulting scores form a verification curve $v \in \mathbb{R}^B$ with peak $b^* = \arg\max_b v_b$ and mean $\bar{v} = \frac{1}{B}\sum_b v_b$. Starting from b^*, we extend a region in both directions as long as the neighboring scores stay at or above $\bar{v}$. The result is a contiguous region $[s_v, e_v]$ with verification width $w_v = e_v - s_v$. Keeping only the connected component that contains the peak prevents separate high-scoring regions from merging across low-scoring windows. We use this region only as a duration estimate and discard its location. Scaling the curve by a positive factor or shifting it by a constant leaves w_v unchanged (Appendix A).

The temporal resolution of w_v is D/B. We choose B so that an event of typical relative duration spans about four windows. With $\rho = \text{median}(w^*/D)$ on the training split, we set $B = 4/\rho$, rounded to the nearest power of two and clipped to $[8, 64]$. This gives 16 windows for Charades-STA and ActivityNet, 32 for QVHighlights and DiDeMo, 64 for TACoS, and 8 for HiREST. TVGBench has no training split and uses 16 windows. No evaluation annotation is used to choose B. Unless stated otherwise, all generators share Qwen3.5-4B as the verifier.

3.3 FUSION AND INTERVAL RECONSTRUCTION

We combine the generated width w and verification width w_v in log space. For interpolation weight α,

$$\hat{w} = w^{1-\alpha} w_v^{\alpha}, \qquad \hat{s} = c - \hat{w}/2, \qquad \hat{e} = c + \hat{w}/2, \qquad \alpha \in [0, 1]. \tag{3}$$

Equivalently,

$$\log \hat{w} = (1 - \alpha) \log w + \alpha \log w_v, \tag{4}$$

so for $0 < \alpha < 1$ the calibrated width lies between the two estimates. Interpolating in log space treats proportional expansion and contraction symmetrically, as IoU does when the midpoints are aligned (Equation 1).

We use $\alpha = 0.5$, giving $\hat{w} = \sqrt{w w_v}$. On the 223 Charades-STA development queries with Qwen3.5-4B in direct mode, we selected the geometric mean from seven constructions fixed in advance. The other six were the generated width, a training-set prior, three mixtures of the two, and the verification width. We kept this choice for all subsequent evaluations.

The reconstructed interval is centered at the generated midpoint c. If $[\hat{s}, \hat{e}]$ crosses a video boundary, we shift it back into $[0, D]$ and keep the width $\hat{w}$; if $\hat{w} > D$, we return the full video. Per query, SEEKSPAN adds B verifier forward passes to the original generation and decodes no additional tokens.

3.4 Effect of center error on width calibration

The generated width comes from decoded timestamps, and the verification width comes from yes/no token probabilities. Prior work finds discrepancies between internal temporal evidence and generated timestamps (Du et al., 2026), and binary-response scores expose temporal information through a different output interface (Li et al., 2026a; Huang et al., 2026). Neither estimate is more accurate on every query, so we combine the two. Let $J_d(u)$ denote the IoU between an interval of width u centered at the generated midpoint c and an annotation of width w^* centered at c^*, with $d = |c - c^*|$. Before video-boundary adjustment,

$$J_d(u) = \min\left\{ \frac{\min(u, w^*)}{\max(u, w^*)}, \frac{(u + w^* - 2d)_+}{u + w^* + 2d} \right\}, \tag{5}$$

where $(x)_+ = \max(x, 0)$; the first term accounts for the width mismatch and the second for the center displacement. For $d = 0$ it reduces to Equation 1, and the interpolation in Equation 3 satisfies

$$J_0(\hat{w}) \geq J_0(w)^{1-\alpha} J_0(w_v)^{\alpha} \tag{6}$$

by the triangle inequality in log-width space. When w and w_v lie on opposite sides of w^*, their signed log-width errors partly cancel, so the fused width can be more accurate than either estimate. For fixed $d > 0$, Equation 5 is maximized at $u = w^* + 2d$. A wider interval partly compensates for a displaced center, so moving the predicted duration toward the annotated duration can lower IoU. SeekSpan keeps the generated position, so its width calibration relies on that position being informative. Appendix A gives proofs and further properties of the calibration rule.

4 Experiments

4.1 Experimental setup

Data and evaluation. Charades-STA is our primary benchmark (Gao et al., 2017a). Table 1 evaluates all 3,720 test queries, as published systems do. Because the fusion weight was selected on 69 development videos (223 queries), Appendix D.1 repeats every paired comparison on the remaining 3,497 queries. On this subset, each change stays within 0.2 mIoU points of its full-set value and remains significant. We report R@0.3, R@0.5, R@0.7, and mIoU, and use mIoU for the compact comparisons below. Primary results assign IoU zero to unparsed answers. Unless stated otherwise, analyses that need valid intervals use the parsed queries outside the development videos. The direct–thinking diagnostics also require both modes to parse. Exploratory transfer evaluations use DiDeMo-derived, QVHighlights, ActivityNet-Captions, HiREST, TACoS, and related subsets (Hendricks et al., 2017; Lei et al., 2021; Krishna et al., 2017; Zala et al., 2023; Regneri et al., 2013; Wang et al., 2025b).

Generators and controls. We evaluate Qwen3.5-2B, -4B, and -9B, Qwen3-VL-8B, Qwen2.5-VL-7B, InternVL3.5-8B, and MiMo-VL-7B-RL (Team, 2026; Bai et al., 2025a;b; Wang et al., 2025a; Team et al., 2025). The primary comparison uses one shared Qwen3.5-4B verifier and the fusion weight $\alpha = 0.5$ selected on the development split. Each Base–SeekSpan pair uses the same generator, answer mode, and REZE timestamp-only prompt (Li et al., 2026a). Calibration changes neither the model parameters nor the decoding. Qwen3-VL provides its two modes as separate Instruct and Thinking checkpoints.

Comparison methods. Table 1 compares SeekSpan with published systems that are fine-tuned on the Charades-STA training split, trained on other grounding data or unlabeled videos, or training-free. We re-measure VideoChat-R1, TempSamp-R1, Time-R1, and EvoGround on the same 3,720 queries with their released checkpoints, prompts, frame budgets, decoding settings, and answer parsers. REZE is re-measured with the prompt, clip budget, and interval decoder given in its paper. Appendix E lists the reported values next to our measurements. The remaining rows are copied from the papers. The Base columns report our own runs of general video-language models that generate timestamps with one prompt.

Uncertainty. All differences are paired by query. Because several queries share a video, the 95% intervals come from 2,000 bootstrap resamples of videos. We apply Holm correction within each comparison family (Holm, 1979). Changes are computed before rounding, in mIoU points.

Table 1: **Charades-STA test set (3,720 queries): published systems and SEEKSPAN.** R@1 at IoU 0.3/0.5/0.7 and mIoU in %; unparsed answers score 0. Top: published systems grouped by training data, with SEEKSPAN in shaded rows. † marks systems re-measured on these queries with their released checkpoints and evaluation code, or for REZE with the prompt and decoder of its paper; Appendix E gives their reported values. Other rows are copied from the papers. Bold marks the best training-free mIoU. Bottom: mIoU before and after SEEKSPAN with the shared Qwen3.5-4B verifier for all thirteen configurations (R@k in Appendix D.1); bold marks the better of each pair.

Method	Backbone	Training data	R@0.3	R@0.5	R@0.7	mIoU
VideoChat-R1.5 (Yan et al., 2025)	Qwen2.5-VL-7B	Charades-STA	82.8	71.6	48.3	60.6
VideoChat-R1† (Li et al., 2025a)	Qwen2.5-VL-7B	Charades-STA	80.9	69.1	46.4	58.8
TempSamp-R1† (Li et al., 2025b)	Qwen2.5-VL-7B	Charades-STA	82.2	71.0	47.9	59.9
TimeMarker (Chen et al., 2024)	LLaMA3-8B	grounding data	73.5	51.9	26.9	48.4
EvoGround† (Jung et al., 2026)	Qwen2.5-VL-7B	unlabeled videos	78.5	59.9	34.0	53.5
Time-R1† (Wang et al., 2025b)	Qwen2.5-VL-7B	grounding data	79.0	62.2	34.9	54.2
TFVTG (Zheng et al., 2024)	BLIP-2 + GPT-4 Turbo	none	67.0	50.0	24.3	44.5
DART (Zhang et al., 2026b)	LLaVA-1.6-7B	none	71.0	52.0	29.4	48.9
UG-Ground (Kim et al., 2026)	InternVideo2.5-8B	none	74.7	52.5	32.6	51.0
DSE-VTG (Cao et al., 2026)	Qwen3-VL-Embedding-8B	none	75.7	58.6	32.2	51.3
UG-Ground (Kim et al., 2026)	Qwen2.5-VL-7B	none	70.1	51.0	30.1	48.4
FV-Action (Huang et al., 2026)	Qwen2.5-VL-7B	none	70.1	56.8	29.7	–
REZE† (Li et al., 2026a)	Qwen2.5-VL-7B	none	75.5	56.5	32.1	51.7
SEEKSPAN (ours)	Qwen2.5-VL-7B	none	74.7	54.0	27.0	49.6
REZE† (Li et al., 2026a)	Qwen3-VL-8B	none	77.4	59.3	33.0	52.6
SEEKSPAN (ours)	Qwen3-VL-8B	none	80.2	67.0	40.7	57.2
SEEKSPAN (ours)	Qwen3-VL-8B-Thinking	none	81.0	69.0	42.7	**57.8**

	Direct			Thinking		
Generator	Base	SEEKSPAN	Δ	Base	SEEKSPAN	Δ
Qwen3.5-2B	52.0	**54.6**	+2.62	37.5	**46.6**	+9.05
Qwen3.5-4B	55.0	**55.8**	+0.82	46.6	**54.3**	+7.71
Qwen3.5-9B	**57.8**	57.2	−0.61	52.1	**55.3**	+3.19
Qwen3-VL-8B	52.9	**57.2**	+4.26	55.0	**57.8**	+2.88
Qwen2.5-VL-7B	46.1	**49.6**	+3.48	—	—	—
InternVL3.5-8B	36.4	**45.6**	+9.25	35.5	**43.1**	+7.58
MiMo-VL-7B	46.0	**51.7**	+5.72	34.8	**41.0**	+6.24
Configurations improved		**6 / 7**			**6 / 6**	

4.2 CALIBRATION ACROSS GENERATORS AND ANSWER MODES

With the same verifier and fusion weight for every generator, SEEKSPAN improves **12 of 13 configurations**, six of seven direct and all six thinking configurations (Table 1). All twelve gains remain significant after correction for multiple comparisons. Direct-mode gains reach +9.2 points for InternVL3.5-8B, and five of the six positive direct-mode changes exceed two points. Thinking-mode gains range from +2.9 to +9.0 points.

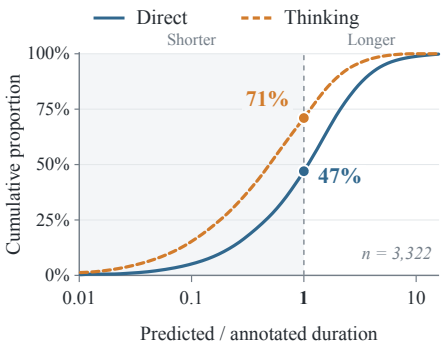
The one decrease is Qwen3.5-9B direct (−0.6 points), whose generated intervals are already wide. For every generator except InternVL3.5-8B, which gains on both, SEEKSPAN gains on predictions shorter than their annotations and gives back part of that gain on wider ones (Table 12).

Comparison with published systems. In Table 1, only the three systems fine-tuned on the Charades-STA training split reach higher mIoU than SEEKSPAN with Qwen3-VL-8B. Without training, SEEKSPAN raises the Qwen3-VL-8B thinking checkpoint from 55.0 to 57.8 mIoU and closes more than half of its gap to TempSamp-R1 (59.9). At R@0.5 it matches VideoChat-R1 (69.0 versus 69.1). With SEEKSPAN, this thinking checkpoint scores higher than every published training-free method on all four metrics. In direct mode, SEEKSPAN exceeds REZE on the same backbone (57.2 versus 52.6 re-measured and 55.7 reported). It also exceeds Time-R1 (54.2) and EvoGround (53.5), both trained without Charades-STA annotations and re-measured with their released checkpoints and evaluation scripts. The Time-R1 paper reports 58.1, 3.9 points above our re-measurement (Appendix E). On Qwen2.5-VL-7B, REZE and FV-Action score higher than SEEKSPAN at R@0.5

Table 2: **How inference-time temporal grounding methods build the output interval.** Among these methods, only SEEKSPAN keeps the position of an existing answer and changes its duration. Its verifier can be the generator itself (Table 4). Shaded rows are compared numerically in this section, with self-consistency implemented as the IoU medoid of its samples; "scores": an interval assembled from per-clip, per-window, or per-frame scores; *regenerated only when a confidence gate fails; black box: uses only the timestamps of one existing answer (–: no generated answer is used); ≤1 decode: no decoding beyond one answer.

Method	Position from	Duration from	Models	Cost per query	Black box	≤1 decode
Self-consistency (Wang et al., 2023)	samples	samples	generator	K sampled decodes	✗	✗
Know-When (Du et al., 2026)	generator*	generator*	generator	1–2 decodes, blank pass	✗	✗
MarkIt (Fang et al., 2026)	generator	generator	gen., LLM, segmenter	tags, masks, decode	✗	✗
DART (Zhang et al., 2026b)	generator	generator	generator	keyframes, decode	✗	✓
TFVTG (Zheng et al., 2024)	scores	scores	BLIP-2, LLM	LLM, frame scores	–	✗
DSE-VTG (Cao et al., 2026)	scores	scores	embedding model	embeddings, offset	–	✓
UG-Ground (Kim et al., 2026)	scores	scores	VLM	1 pass per window	–	✓
REZE (Li et al., 2026a)	scores	scores	VLM	1 pass per 3-s clip	–	✓
FV-Action (Huang et al., 2026)	score peak	fixed (8 s)	VLM	36 passes	–	✓
SEEKSPAN (ours)	generator	gen. ⊕ verifier	gen., VLM	decode, 16 passes	✓	✓

(a) Duration ratio distribution



(b) Exchanging one coordinate between answer modes

Generator	n	mIoU Direct	mIoU Thinking	Change to T D dur.	Change to T D ctr.
Qwen3.5-2B	3,139	53.3	42.0	+7.9	+1.2
Qwen3.5-4B	3,322	55.6	49.0	+6.7	−1.1
Qwen3.5-9B	3,331	58.5	54.9	+2.4	+0.1
MiMo-VL-7B	3,248	46.7	37.4	+3.1	+1.2
Qwen3-VL-8B‡	3,449	53.2	55.9	−1.3	−3.2
InternVL3.5-8B	3,426	36.2	36.1	−0.2	+0.1

‡Separate Instruct and Thinking checkpoints.

Figure 3: **Useful temporal positions coexist with inaccurate durations.** (a) For Qwen3.5-4B on 3,322 jointly parsed queries outside the development videos, thinking shifts the ratio of predicted to annotated duration toward shorter intervals. (b) mIoU of the direct and thinking answers on each generator's subset selected as in (a), and the change in the thinking answer (T) when it takes its duration (D dur.) or its center (D ctr.) from the direct answer to the same query. Appendix D.2 gives intervals and the full factorial.

and R@0.7. On this backbone, per-clip recognition locates events 5.6 mIoU points better than the model's own timestamps (Appendix E). Both methods replace the generated interval. SEEKSPAN instead keeps its position and calibrates only the duration. Table 2 summarizes where each inference-time method obtains the position and the duration.

4.3 EVIDENCE FOR SEPARATING POSITION AND DURATION

Direct–thinking comparison. For Qwen3.5-4B, thinking changes the predicted duration much more than the predicted position. On 3,322 jointly parsed queries, enabling thinking shortens the median predicted duration from 7.0 to 5.0 s but changes the center error only from 3.27 to 3.32 s. Figure 3(b) exchanges one coordinate at a time between the direct-mode and thinking-mode answers to the same query. Taking the duration from the direct answer raises the thinking answer by 6.7 points (95% interval $[+6.0, +7.3]$) and closes the whole 6.6-point gap. Taking the center instead changes the thinking answer by -1.1 ($[-1.8, -0.4]$). For Qwen3.5-2B and -9B, the duration from the direct answer likewise recovers 70% and 65% of the gap. With thinking disabled, inserting unrelated text reproduces 89% of the Qwen3.5-4B accuracy loss, so the narrowing does not require query-relevant reasoning (Appendix D.2).

Table 3: **Duration fusion is the best duration source in all six comparisons and outperforms construction-matched published rules on five.** mIoU on Charades-STA queries outside the development videos with valid intervals; each generator is its own verifier, and all columns that use verification share one verification curve. The four duration sources keep the generated center. Region takes the above-mean verification region as the interval, as score-based methods such as REZE and UG-Ground build intervals from verification scores. The two Peak variants center a width on the verification peak, as FV-Action does with a fixed 8 s width. These three constructions also move the center. Bold marks the best value per row; Δ is SEEKSPAN minus the generated interval.

	Duration source, generated center			Published-style constructions				
Generator	Generated	Prior	Verification	Region	Peak + prior	Peak + gen.	SEEKSPAN	Δ
Direct								
Qwen3.5-4B	55.05	51.71	50.73	52.08	53.27	55.42	**55.82**	+0.78
Qwen3.5-2B	52.23	49.93	50.86	50.42	49.50	50.33	**54.48**	+2.26
Qwen3-VL-8B	52.96	52.13	52.90	53.25	52.02	52.07	**56.51**	+3.55
Qwen2.5-VL-7B	46.39	44.68	44.13	42.23	40.89	41.53	**48.22**	+1.83
Thinking								
Qwen3.5-4B	49.01	52.48	51.19	52.67	54.09	49.10	**57.04**	+8.03
Qwen3.5-2B	42.02	48.13	49.29	**51.48**	50.44	42.13	50.77	+8.75

Width-source ablation. Table 3 compares four duration sources at the generated center with three construction-matched variants of the published rules that also move the center. One variant takes the verified region as the interval, and the other two center the training-prior or generated width on the verification peak. Each generator is its own verifier on its parsed analysis subset. Fusion is the best duration source in all six configurations, and the three published-style constructions trail SEEKSPAN in five of six. Fusing the generated width with the training-set prior at the same weight stays 0.9–4.1 points below SEEKSPAN in all six configurations, so the verifier adds information beyond a typical event length. The prior or the verification duration alone improves the two thinking configurations but lowers all four direct ones. For Qwen3.5-4B, the verification duration alone is 5.1 points below fusion in direct mode and 5.9 in thinking mode.

Center-source ablation. With the SEEKSPAN duration fixed, replacing the generated center with the center of the verified region lowers mIoU by 3–6 points in all six ablation configurations, and moving the center a quarter of the way toward it gives no consistent improvement.

Robustness checks. Table 4 varies the two choices that SEEKSPAN fixes in advance. The preset $\alpha = 0.5$ is the best weight in four of six configurations and within 1.1 points of the best in the other two. The shared Qwen3.5-4B verifier is the best of four verifiers in five of six configurations. Every verifier raises every configuration above its base except the Qwen2.5-VL-7B curve on Qwen3.5-4B direct. Using 8, 32, or 64 verification windows on Charades-STA keeps mIoU within 0.35 points of the sixteen-window result. On 400 queries, two paraphrased verifier questions give curves with median Spearman correlations of 0.89 and 0.95 with the default and leave accuracy unchanged within their confidence intervals. A verifier prefix that follows the chat template exactly changes each primary gain by at most 0.09 points (Tables 9–11).

4.4 INFERENCE COST AND TRANSFER

Inference cost. We compare SEEKSPAN with consensus over $K = 16$ sampled answers, which returns the medoid by pairwise IoU. For Qwen3.5-4B thinking on 374 paired queries, SEEKSPAN matches consensus ($+0.41$ points, 95% interval $[-2.07, +2.73]$) with an order of magnitude less time. In a separate 200-query timing run on an RTX A6000, one greedy thinking generation takes a median 20.54 s and sixteen verifier forwards take 6.82–9.56 s. Generation followed by SEEKSPAN takes 27.4–30.1 s, against 329 s for sixteen sequential generations (Appendix D.4). SEEKSPAN also improves thinking-mode consensus at every tested size from $K = 1$ to 16.

Table 4: **Sensitivity to the fusion weight and to the verifier.** mIoU on the analysis subsets of Table 3; Base is the generated interval. Left: fusion weight α, with each generator as its own verifier; $\alpha = 0.5$ was fixed in advance. Right: verifier, with $\alpha = 0.5$; Qwen3.5-4B is the shared verifier of Table 1. Bold marks the best value within each group.

		Fusion weight α			Verifier ($\alpha = 0.5$)			
Generator	Base	0.25	0.5	0.75	Qwen3.5-4B	Qwen3.5-2B	Qwen3-VL-8B	Qwen2.5-VL-7B
Direct								
Qwen3.5-4B	55.05	**56.46**	55.82	53.73	55.82	55.97	**55.98**	54.01
Qwen3.5-2B	52.23	54.12	**54.48**	53.19	**54.76**	54.48	54.41	52.89
Qwen3-VL-8B	52.96	55.91	**56.51**	55.33	**57.17**	56.48	56.51	54.79
Qwen2.5-VL-7B	46.39	48.12	**48.22**	46.77	**49.69**	49.30	49.15	48.22
Thinking								
Qwen3.5-4B	49.01	54.35	**57.04**	55.75	**57.04**	56.26	56.17	55.10
Qwen3.5-2B	42.02	46.73	50.77	**51.85**	**51.98**	50.77	50.75	50.07

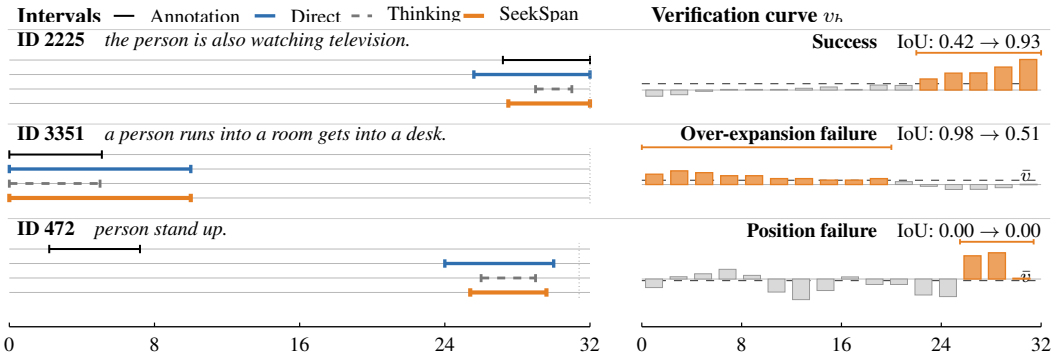


Figure 4: **Duration correction helps when the generated position and verification extent are informative.** Qwen3.5-4B thinking on Charades-STA (selection criteria in Appendix D.6). Each row shows the annotation, the direct and thinking predictions, the SEEKSPAN output, the verification curve, and the IoU of the thinking prediction before and after SEEKSPAN. From top: a narrow prediction widened in place; a well-localized prediction over-expanded by an inaccurate verification extent; a wrong generated position that no duration change can fix.

Cross-dataset transfer. Transfer requires an accurate generated position and an informative verification duration. SEEKSPAN improves all four evaluated configurations on HiREST, by 3.3–4.3 points. On QVHighlights, TACoS, and TVGBench, inaccurate verification widths increase the width error. On ActivityNet-A, calibration moves the median predicted duration from 13.4 s toward the annotated 22.8 s, yet mIoU falls by 0.2–2.7 points in all four configurations. There the mean center error is 19–22 s, as long as the events themselves, so a better duration does not recover the overlap (Section 3.4; Appendix D.5).

5 CONCLUSION

We introduced SEEKSPAN, a training-free method that keeps the generated position and fuses the generated duration with a second estimate from a frozen verifier. With one shared verifier, it improves 12 of 13 configurations on the Charades-STA test set. On a frozen Qwen3-VL-8B generator, it reaches 57.2 mIoU in direct mode and 57.8 with thinking, above every published training-free method. Generation supplies the position and verification a second duration estimate, so calibration improves grounding on Charades-STA and HiREST without retraining or extra decoding.

AI USE STATEMENT

We used ChatGPT and Claude to polish the writing of this manuscript, including grammar, clarity, and concision. The use of pretrained models for image generation, editing, verification, and preference judgment is part of our research methodology and is documented in the paper. The authors are responsible for the research design, implementation, experimental analysis, and conclusions. We reviewed all AI-assisted text before incorporating it and remain responsible for the final manuscript.

ETHICS STATEMENT

This work evaluates publicly released model checkpoints on established video temporal-grounding benchmarks and collects no new data or human-subject annotations. We do not redistribute source videos. The method operates only at inference time and does not update the evaluated model parameters.

REPRODUCIBILITY STATEMENT

Section 4.1 and Appendix B describe the evaluation subsets, inference settings, parsing procedure, and analysis protocol. Appendix C reports the statistical procedures, and Appendix G provides the prompts and verifier templates. Analyses added after inspection of earlier results are identified as exploratory.

REFERENCES

Josiah Aklilu, Xiaohan Wang, and Serena Yeung-Levy. Zero-shot action localization via the confidence of large vision-language models, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b.

Zhuo Cao, Bingqing Zhang, Sen Wang, and Xue Li. DSE-VTG: Dual-side enhancement for training-free video temporal grounding. *arXiv preprint arXiv:2609.08850*, 2026.

Shimin Chen, Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. TimeMarker: A versatile video-LLM for long and short video understanding with superior temporal localization ability. *arXiv preprint arXiv:2411.18211*, 2024.

Dazhao Du, Liao Duan, Jian Liu, Tao Han, Yujia Zhang, Eric Liu, Xi Chen, and Song Guo. Mllms know when before speaking: Revealing and recovering temporal grounding via attention cues, 2026.

Pengcheng Fang, Yuxia Chen, and Xiao-Hao Cai. Markit: Training-free visual markers for precise video temporal grounding. *ArXiv*, abs/2604.25886, 2026.

Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, Yan Feng, Peng Pei, Xunliang Cai, and Xiangyu Yue. OneThinker: All-in-one reasoning model for image and video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *IEEE International Conference on Computer Vision (ICCV)*, 2017a.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017b.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. TRACE: Temporal grounding video LLM via causal event modeling. In *International Conference on Learning Representations (ICLR)*, 2025.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 5804–5813, 2017. doi: 10.1109/ICCV.2017.618.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280, 2024.

Ji Huang, Barry Devereux, and Hui Wang. Your VLM already knows when: Training-free temporal grounding by asking yes or no. *arXiv preprint arXiv:2608.08315*, 2026.

Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. On the consistency of video large language models in temporal comprehension. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13713–13722, 2025.

Minjoon Jung, Byoung-Tak Zhang, and Lorenzo Torresani. EvoGround: Self-evolving video agents for video temporal grounding. *arXiv preprint arXiv:2605.13803*, 2026.

Sanghwan Kim, Rui Xiao, Stephan Alaniz, Yongqin Xian, and Zeynep Akata. Training-free uncertainty guidance for complex visual tasks with MLLMs. In *European Conference on Computer Vision (ECCV)*, volume 17006 of *Lecture Notes in Computer Science*, pp. 606–624. Springer, 2026. doi: 10.1007/978-3-032-37020-4_33.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 706–715, 2017. doi: 10.1109/ICCV.2017.83.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, volume 7, pp. 231–238, 1994.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Boyang Li, Chenhui Gou, and Jianfei Cai. Reze: Recognition-based zero-shot extraction for video temporal grounding, 2026a.

Jiaze Li, Hao Yin, Haoran Xu, Boshen Xu, Wenhui Tan, Zewen He, Jianzhong Ju, Zhenbo Luo, and Jian Luan. Video-OPD: Efficient post-training of multimodal large language models for temporal video grounding via on-policy distillation. In *International Conference on Machine Learning (ICML)*, 2026b.

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*, 2025a.

Yunheng Li, Jing Cheng, Shaoyong Jia, Hangyi Kuang, Shaohui Jiao, Qibin Hou, and Ming-Ming Cheng. TempSamp-R1: Effective temporal sampling with reinforcement fine-tuning for video LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025b. doi: 10.52202/085713-1358.

Shuming Liu, Mingchen Zhuge, Changsheng Zhao, Jun Chen, Lemeng Wu, Zechun Liu, Chenchen Zhu, Zhipeng Cai, Chong Zhou, Haozhe Liu, Ernie Chang, Saksham Suri, Hongyu Xu, Qi Qian, Wei Wen, Balakrishnan Varadarajan, Zhuang Liu, Hu Xu, Florian Bordes, Raghuraman Krishnamoorthi, Bernard Ghanem, Vikas Chandra, and Yunyang Xiong. VideoAuto-R1: Video auto reasoning via thinking once, answering twice. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32829–32839, 2026.

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41340–41356. PMLR, 2024.

Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025.

Qwen Team. Qwen3.5-omni technical report, 2026. URL https://arxiv.org/abs/2604.15804.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025a.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2023.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025b. doi: 10.52202/085713-2793.

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. HawkEye: Training video-text LLMs for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024.

Ziyang Wang, Jaehong Yoon, Shoubin Yu, Md Mohaiminul Islam, Gedas Bertasius, and Mohit Bansal. Video-RTS: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose,

and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28126–28140, Suzhou, China, November 2025c. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1428. URL https://aclanthology.org/2025.emnlp-main.1428/.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. VideoChat-R1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oğuz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10419–10429, 2026a.

Zhengbo Zhang, Mark He Huang, Zhigang Tu, and Ming-Hsuan Yang. DART: Difficulty-adaptive routing for zero-shot video temporal grounding. In *European Conference on Computer Vision (ECCV)*, volume 17061 of *Lecture Notes in Computer Science*, pp. 73–91. Springer, 2026b. doi: 10.1007/978-3-032-37095-2_5.

Minghang Zheng, Xinhao Cai, Qingchao Chen, Yuxin Peng, and Yang Liu. Training-free video temporal grounding using large-scale pre-trained models. In *European Conference on Computer Vision*, 2024.

Minghang Zheng, Zihao Yin, Yi Yang, Yuxin Peng, and Yang Liu. Temporal-aware reasoning optimization for video temporal grounding. In *International Conference on Machine Learning (ICML)*, 2026.

A Properties of Extent Calibration

We analyze the specified inference rule without assuming unbiased or independent generator and verifier errors. Unless stated otherwise, intervals are reconstructed on the real line before video-boundary adjustment. Let the annotation have width $w^* > 0$ and center c^*, and fix the generated center c with error $d = |c - c^*|$.

Relative-score invariance. For $v'_b = av_b + k$ with common $a > 0$ and k, the mean becomes $\bar{v}' = a\bar{v} + k$. Therefore $v'_b \geq \bar{v}'$ if and only if $v_b \geq \bar{v}$, and the set of maximizing windows is unchanged. With the same tie-breaking rule, the peak-connected region and its width are identical. This invariance does not extend to arbitrary monotone transformations or window-dependent biases. A constant curve, for example, returns the full video with the non-strict threshold.

Exact interval geometry. For a candidate width $u > 0$, the intersection length is

$$L_d(u) = \max\left\{0, \min\left(u, w^*, \frac{u + w^*}{2} - d\right)\right\}. \tag{7}$$

Its IoU is

$$J_d(u) = \frac{L_d(u)}{u + w^* - L_d(u)}.$$

If $2d \leq |u - w^*|$, one interval contains the other and

$$J_d(u) = \frac{\min(u, w^*)}{\max(u, w^*)}.$$

If $|u - w^*| < 2d < u + w^*$, the intervals partially overlap and

$$J_d(u) = \frac{u + w^* - 2d}{u + w^* + 2d}.$$

If $2d \geq u + w^*$, the IoU is zero. These cases give Equation 5.

A geometric lower bound at aligned centers. Let

$$\epsilon_g = \log(w/w^*), \qquad \epsilon_v = \log(w_v/w^*).$$

The fused log-width error is

$$\epsilon_\alpha = (1 - \alpha)\epsilon_g + \alpha\epsilon_v.$$

The triangle inequality gives

$$|\epsilon_\alpha| \leq (1 - \alpha)|\epsilon_g| + \alpha|\epsilon_v|. \tag{8}$$

Applying $\exp(-\cdot)$ yields Equation 6. Equality holds when the two errors have the same sign. With opposite signs and $0 < \alpha < 1$, the errors cancel at least in part and the fused IoU exceeds the bound. For example, if the generated and verification widths are $w^*/2$ and $2w^*$, their equal-weight geometric fusion recovers w^* exactly. If instead $w = w^*$ and $w_v = 4w^*$, equal-weight fusion attains the bound and has aligned-center IoU $1/2$, against 1 for the generated width.

Conservative interpolation with an arbitrary fixed center. The function $J_d(u)$ is nondecreasing up to $u = w^* + 2d$ and nonincreasing thereafter. In the lower-width containment case, $J_d(u) = u/w^*$. In the positive partial-overlap case,

$$\frac{d}{du}\frac{u + w^* - 2d}{u + w^* + 2d} = \frac{4d}{(u + w^* + 2d)^2} \geq 0. \tag{9}$$

Once $u \geq w^* + 2d$, the candidate contains the annotation and $J_d(u) = w^*/u$ decreases. An initial disjoint region may form a zero plateau. The optimum is therefore $u = w^* + 2d$, and every width $\tilde{w}$ between w and w_v satisfies

$$J_d(\tilde{w}) \geq \min\{J_d(w), J_d(w_v)\}. \tag{10}$$

Because $\hat{w} = w^{1-\alpha}w_v^\alpha$ lies between w and w_v for every $\alpha \in [0, 1]$, the calibrated width satisfies this bound. Both endpoint candidates are centered at c, so $J_d(w_v)$ is the IoU of the verification width placed at the generated center, wherever $[s_v, e_v]$ lies.

Effect of video-boundary adjustment. Assume the annotation lies within $[0, D]$. Take a candidate of width at most D that crosses the left endpoint. Shifting it until its start is zero cannot decrease its intersection with the annotation. The right-endpoint case is symmetric. The interval length remains unchanged, so IoU cannot decrease. An interval entirely outside the video initially has zero IoU. For $u > D$, any unadjusted IoU is at most $w^*/u < w^*/D$, whereas returning the full video attains w^*/D. Thus the final adjustment cannot reduce IoU relative to the unadjusted candidate.

Error complementarity for squared log-width loss. Applying the squared-error ambiguity decomposition (Krogh & Vedelsby, 1994) to log widths gives a complementary diagnostic:

$$\epsilon_\alpha^2 = (1-\alpha)\epsilon_g^2 + \alpha\epsilon_v^2 - \alpha(1-\alpha)(\epsilon_g - \epsilon_v)^2. \tag{11}$$

This identity is exact and requires neither independence nor zero bias. With finite second moments, define

$$M_g = \mathbb{E}[\epsilon_g^2], \qquad M_v = \mathbb{E}[\epsilon_v^2], \qquad C = \mathbb{E}[\epsilon_g \epsilon_v].$$

For equal-weight fusion,

$$\mathbb{E}[\epsilon_{1/2}^2] - M_g = \frac{M_v + 2C - 3M_g}{4}. \tag{12}$$

Fusion therefore reduces squared log-width error relative to generation when $M_v + 2C < 3M_g$. The cross-moment C is the covariance plus the product of the mean errors.

B Evaluation Protocol and Reproducibility

Analysis design. We distinguish analyses specified before evaluation from analyses added after inspecting earlier results. The primary Charades-STA comparisons, thinking-mode comparison, boundary interventions, width-calibration test, and sampling comparison were specified before evaluation and use fixed evaluation subsets and metrics. Cross-dataset evaluation, video-context regression, per-query variants, annotation-based width analyses, the direct-prompt paraphrases, and the later fusion-weight sweep are exploratory.

Evaluation sets. Charades-STA contains 3,720 test queries over 1,334 videos. We reserve 69 videos containing 223 queries for development. Table 1 evaluates all 3,720 test queries, and Appendix D.1 repeats every paired comparison on the remaining 3,497 queries.

The DiDeMo-derived subset uses the OneThinker SFT queries whose videos were available. It has 497 queries over 384 videos. Videos longer than 30 s are truncated to the first 30 s, and each query has one integer-second annotated moment. This subset therefore differs from the official DiDeMo protocol. Charades-train contains 807 queries over 749 videos, none of which is a Charades-STA test video. QVHighlights uses the full validation split of 1,550 queries and is scored by maximum IoU over its annotated moments.

The mixed temporal-grounding set contains 800 queries: 138 from Charades, 168 from ActivityNet-Captions, 158 from HiREST, 168 from TACoS, and 168 from TVGBench. The Charades portion overlaps the primary benchmark, so we treat it as an in-domain replication. The TACoS portion contains five videos, so its confidence intervals use five video clusters.

We also evaluate ActivityNet-A (500 queries), ActivityNet-B (500 queries), and ActivityNet-2000 (2,000 queries), all drawn from the ActivityNet-Captions val_2 split. We drop 9 queries in ActivityNet-B and 32 in ActivityNet-2000 because their videos cannot be decoded. ActivityNet-2000 contains ActivityNet-B and 38 queries of ActivityNet-A, and the mixed-set ActivityNet portion shares videos with all three.

Inference settings. For Qwen-family models, videos are sampled at 1 fps with at most 128 frames and at most 262,144 pixels per frame. Direct generation uses the REZE prompt with at most 128 new tokens. Thinking mode uses the same prompt with the thinking channel enabled, greedy decoding, and at most 4,096 new tokens. A separate thinking-stability diagnostic uses the model-card sampling setting with a 32,768-token budget.

For thinking-mode outputs, the parser reads only text after the closing thinking tag, so numbers in the reasoning trace are ignored. For direct outputs, it reads the full response. After mm:ss normalization,

the first two numbers are interpreted as start and end times. Missing, equal, non-finite, or non-positive spans count as unparsed. So do thinking outputs that do not close the thinking tag within the token budget. When an evaluation includes unparsed outputs, they receive IoU 0.

The verifier uses B uniform temporal windows and input frames sampled at 1 fps. We use 16 windows on Charades-STA, Charades-train, the Charades, ActivityNet, and TVGBench portions of the mixed set, and the ActivityNet subsets; 32 on QVHighlights and DiDeMo; 64 on TACoS; and 8 on HiREST. For each window, we compute the log-sum-exp over six yes token forms and six no token forms at the answer position. All twelve forms are single tokens in the Qwen3.5 tokenizer. For verifiers with other tokenizers, we omit any form that is not a single token and merge duplicate token ids. SEEKSPAN uses $\alpha = 0.5$ and the above-mean contiguous region containing the verification peak.

Sampling consensus draws $K = 16$ answers using the model-card sampling settings for each mode and returns the medoid by pairwise IoU among parsed samples. Direct Qwen-family sampling uses temperature 0.7 and top-p 0.8; thinking uses temperature 1.0 and top-p 0.95. Both use top-k 20 and presence penalty 1.5. For the consensus experiment, thinking samples are capped at 4,096 tokens. We sample InternVL3.5-8B at its model-card temperature of 0.6.

InternVL3.5-8B uses 32 frames, one tile, scaled-dot-product attention in its language model, and at most 128 new tokens in direct mode and 4,096 in thinking mode. Its released frame prefix numbers the frames but gives no timestamps. The timestamped variant adds the video duration and the time of each frame. Primary comparisons use the released model-card setting. MiMo-VL-7B-RL uses the same video input and query prompt as the Qwen models. Only its direct mode appends the `/no_think` suffix to the prompt. Bootstrap intervals use 2,000 video-cluster resamples with seed 20260813.

Checkpoints and software. The evaluated checkpoints are Qwen3.5-2B `15852e8c`, Qwen3.5-4B `851bf6e8` (also used as the shared verifier), Qwen3.5-9B `c2022362`, Qwen3-VL-8B-Instruct `0c351dd0`, Qwen3-VL-8B-Thinking `92f3c4b4`, Qwen2.5-VL-7B-Instruct `cc594898`, InternVL3.5-8B `9bb6a56a`, and MiMo-VL-7B-RL-2508 `4bfb2707`. The inference environment at submission uses Python 3.11.15, PyTorch 2.11.0 with CUDA 13.0, Transformers 5.8.0, and vLLM 0.20.2.

C STATISTICAL ANALYSIS

Paired uncertainty estimates. All differences are paired by query. Because queries from the same video are dependent, we bootstrap over videos. We use 2,000 resamples with replacement and report the 2.5 and 97.5 percentiles of the resampled mean difference. We call a difference significant when its confidence interval excludes zero.

Multiple comparisons. For each designated comparison family, we compute two-sided bootstrap p-values and apply Holm correction at a family-wise error rate of 0.05 (Holm, 1979). In the analyses reported here, the corrected families are the 13 primary SEEKSPAN comparisons (one family per query set), 7 direct-versus-thinking comparisons, 24 per-dataset thinking losses, and 40 cross-dataset calibration results. All 13 primary SEEKSPAN differences, including the Qwen3.5-9B direct-mode decrease, remain significant after correction.

Determinism. Rerunning 620 Qwen3.5-4B thinking-mode queries reproduces all raw outputs and parsed spans exactly.

D ADDITIONAL EXPERIMENTAL RESULTS

D.1 PRIMARY CONFIDENCE INTERVALS AND COMPARISON PROTOCOLS

Table 5 reports all four metrics for the thirteen paired configurations on the full test set. Table 6 gives the paired mIoU changes with confidence intervals on all 3,720 queries and on the 3,497 queries outside the 69 development videos. We apply Holm correction separately to the 13 generator–mode comparisons of each query set. After correction, all twelve gains and the Qwen3.5-9B direct

decrease remain significant on both query sets. No change differs by more than 0.2 points between the two sets.

Table 5: **All four metrics for the thirteen paired configurations on the full Charades-STA test set (3,720 queries).** Unparsed answers score 0. All configurations use the shared Qwen3.5-4B verifier and $\alpha = 0.5$. Qwen3-VL-8B uses separate Instruct and Thinking checkpoints; InternVL3.5-8B uses its model-card setting.

Generator	Mode	R@0.3		R@0.5		R@0.7		mIoU		Δ mIoU
		Base	SeekSpan	Base	SeekSpan	Base	SeekSpan	Base	SeekSpan	
Qwen3.5-2B	Direct	75.7	79.2	59.1	63.3	33.4	36.1	52.0	54.6	+2.62
Qwen3.5-4B	Direct	77.9	80.7	62.6	65.5	39.3	37.7	55.0	55.8	+0.82
Qwen3.5-9B	Direct	81.9	83.3	67.7	67.8	41.6	37.7	57.8	57.2	−0.61
Qwen3-VL-8B	Direct	77.2	80.2	60.5	67.0	34.4	40.7	52.9	57.2	+4.26
Qwen2.5-VL-7B	Direct	69.7	74.7	48.6	54.0	25.1	27.0	46.1	49.6	+3.48
InternVL3.5-8B	Direct	55.4	68.2	35.9	48.0	16.6	25.6	36.4	45.6	+9.25
MiMo-VL-7B	Direct	70.6	75.7	50.6	58.5	21.8	32.8	46.0	51.7	+5.72
Qwen3.5-2B	Thinking	59.0	67.1	35.0	52.2	15.9	30.4	37.5	46.6	+9.05
Qwen3.5-4B	Thinking	71.2	76.2	50.7	63.8	24.4	38.9	46.6	54.3	+7.71
Qwen3.5-9B	Thinking	76.1	78.1	59.7	65.1	34.0	39.7	52.1	55.3	+3.19
Qwen3-VL-8B	Thinking	78.7	81.0	64.0	69.0	38.3	42.7	55.0	57.8	+2.88
InternVL3.5-8B	Thinking	53.5	64.6	34.8	44.9	16.0	22.3	35.5	43.1	+7.58
MiMo-VL-7B	Thinking	52.6	59.2	38.3	45.6	17.6	26.2	34.8	41.0	+6.24

Table 6: **Paired uncertainty for the primary results on both query sets.** mIoU changes with 95% video-cluster bootstrap confidence intervals on all 3,720 Charades-STA test queries (Table 1) and on the 3,497 queries outside the 69 development videos. Each configuration uses the shared Qwen3.5-4B verifier. All changes remain significant after Holm correction within each 13-comparison family.

		All 3,720 queries		3,497 outside development videos	
Generator	Mode	Δ	95% CI	Δ	95% CI
Qwen3.5-2B	Direct	+2.62	[+2.10, +3.15]	+2.53	[+1.97, +3.09]
Qwen3.5-4B	Direct	+0.82	[+0.30, +1.32]	+0.78	[+0.23, +1.32]
Qwen3.5-9B	Direct	−0.61	[-1.13, -0.13]	−0.75	[-1.28, -0.23]
Qwen3-VL-8B	Direct	+4.26	[+3.68, +4.84]	+4.21	[+3.58, +4.79]
Qwen2.5-VL-7B	Direct	+3.48	[+2.99, +3.95]	+3.30	[+2.78, +3.81]
InternVL3.5-8B	Direct	+9.25	[+8.37, +10.13]	+9.22	[+8.28, +10.16]
MiMo-VL-7B	Direct	+5.72	[+5.15, +6.33]	+5.62	[+5.03, +6.25]
Qwen3.5-2B	Thinking	+9.05	[+8.48, +9.61]	+8.94	[+8.32, +9.53]
Qwen3.5-4B	Thinking	+7.71	[+7.11, +8.28]	+7.63	[+7.03, +8.23]
Qwen3.5-9B	Thinking	+3.19	[+2.64, +3.71]	+3.06	[+2.50, +3.59]
Qwen3-VL-8B	Thinking	+2.88	[+2.34, +3.43]	+2.90	[+2.31, +3.45]
InternVL3.5-8B	Thinking	+7.58	[+6.85, +8.32]	+7.67	[+6.86, +8.48]
MiMo-VL-7B	Thinking	+6.24	[+5.70, +6.78]	+6.23	[+5.70, +6.83]

Distinct evaluation subsets. The primary comparison evaluates every test query and assigns zero to unparsed answers. The width-source ablation uses valid-interval analysis subsets and each generator as its own verifier. The direct–thinking diagnosis further requires both modes to parse on the same query. Scores therefore differ across these tables. Before calibration, Qwen3.5-4B thinking scores 46.6 in the primary comparison and 49.01 in the width-source analysis.

D.2 Extent-error diagnostics

The four checkpoints in Table 7 produce shorter median intervals with thinking. Duration and center error separate most cleanly for Qwen3.5-4B, and this motivates the coordinate substitution in

Table 7: **Duration and position change differently across checkpoints.** Direct → thinking on jointly parsed Charades-STA queries outside the development videos. Duration is the median predicted width; "Shorter, %" is the fraction strictly shorter than the annotation. Center errors are in seconds. All four mIoU decreases have 95% video-cluster bootstrap intervals excluding zero.

Generator	n	mIoU	Duration, s	Shorter, %	Center error, s
Qwen3.5-2B	3,139	53.3 → 42.0	7.2 → 5.0	50 → 69	3.44 → 3.73
Qwen3.5-4B	3,322	55.6 → 49.0	7.0 → 5.0	47 → 71	3.27 → 3.32
Qwen3.5-9B	3,331	58.5 → 54.9	8.0 → 6.6	38 → 58	3.00 → 3.21
MiMo-VL-7B	3,248	46.7 → 37.4	7.0 → 5.0	58 → 72	3.93 → 5.60

Figure 3. MiMo-VL also changes its position substantially. The separately trained Qwen3-VL-8B-Thinking checkpoint gains 2.7 mIoU points over Instruct on paired queries, with little width change and a more accurate center. InternVL3.5-8B changes mIoU on paired queries by about 0.1 point when thinking is enabled.

Full coordinate factorial. Table 8 gives the complete center–duration factorial behind Figure 3(b) for all six generators with two answer modes, together with paired confidence intervals against the thinking-mode output. Its two annotation-based oracles combine the thinking-mode center with the annotated duration and the annotated center with the thinking-mode duration. For the four generators that narrow with thinking, substituting the thinking-mode duration into the direct-mode answer reproduces most or all of the loss. Substituting the thinking-mode center instead changes the score by at most 1.3 points for Qwen3.5-4B and -9B. Its effect is larger for Qwen3.5-2B and largest for MiMo-VL.

Table 8: **Complete center–duration factorial for the six generators with two answer modes.** mIoU on jointly parsed Charades-STA queries outside the development videos. Each column label lists the center source and then the duration source. D and T denote the direct and thinking answers to the same query. Intervals are clipped to the video. Parenthesized values are paired changes against T+T with 95% video-cluster bootstrap intervals. The two oracle columns take one coordinate from the annotation. They combine the thinking-mode center with the annotated duration and the annotated center with the thinking-mode duration. Center error is the mean absolute center error in seconds for the direct and thinking answers.

Generator	n	D+D	T+T	T+D	D+T	T + annot. W	annot. C + T	Center err. D/T (s)
Qwen3.5-2B	3,139	53.3	42.0	49.9 (+7.9 [+7.2, +8.7])	43.2 (+1.2 [+0.4, +2.0])	54.3 (+12.2 [+11.4, +13.1])	60.5 (+18.5 [+17.5, +19.6])	3.44/3.73
Qwen3.5-4B	3,322	55.6	49.0	55.7 (+6.7 [+6.0, +7.3])	47.9 (-1.1 [-1.8, -0.4])	59.7 (+10.7 [+9.9, +11.5])	66.8 (+17.8 [+16.7, +18.9])	3.27/3.32
Qwen3.5-9B	3,331	58.5	54.9	57.3 (+2.4 [+1.9, +2.9])	55.0 (+0.1 [-0.6, +0.8])	60.3 (+5.4 [+4.9, +6.0])	73.7 (+18.8 [+17.8, +19.9])	3.00/3.21
MiMo-VL-7B	3,248	46.7	37.4	40.4 (+3.1 [+2.4, +3.7])	38.5 (+1.2 [+0.5, +1.9])	46.0 (+8.6 [+7.9, +9.3])	57.3 (+19.9 [+18.9, +20.9])	3.93/5.60
Qwen3-VL-8B[‡]	3,449	53.2	55.9	54.7 (-1.3 [-1.8, -0.8])	52.8 (-3.2 [-4.1, -2.3])	62.0 (+6.1 [+5.6, +6.6])	75.9 (+19.9 [+18.8, +21.1])	3.42/3.10
InternVL3.5-8B	3,426	36.2	36.1	36.0 (-0.2 [-0.5, +0.2])	36.2 (+0.1 [-0.8, +1.0])	40.0 (+3.8 [+3.3, +4.4])	70.6 (+34.5 [+33.1, +35.9])	5.54/6.89

[‡] Separate Instruct and Thinking checkpoints.

Stability across decoding settings. For Qwen3.5-2B and -4B, the direct–thinking difference remains similar with greedy decoding on the full Charades-STA analysis subset, with model-card sampling and a 32k-token budget, and with two additional seeds on 400-query subsets. Across the Qwen3.5 family, the direct–thinking loss also decreases with model size, from 11.3 points for 2B to 6.6 for 4B and 3.7 for 9B.

Inserted-text controls. The narrowing can be reproduced without query-specific reasoning content. With thinking disabled, an unrelated prefix reproduces 89% of the Qwen3.5-4B accuracy loss and 64% of the Qwen3.5-2B loss. A reasoning trace copied from another query produces a similar or larger loss, and listing temporal windows increases the loss further. The effect persists across forced thinking budgets from 64 to 1,024 tokens and without forced early closure. All tested inserted-text controls also reduce predicted duration. The narrowing is therefore sensitive to text that precedes the final answer.

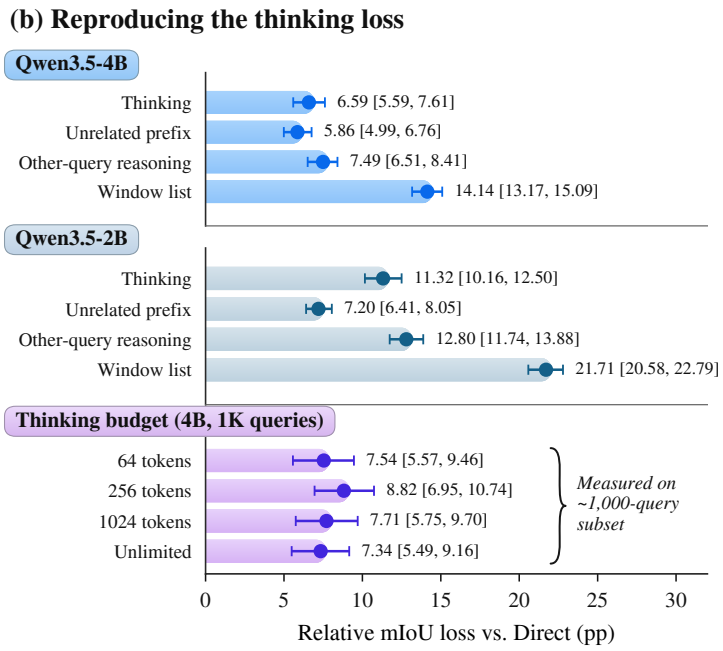


Figure 5: **Narrowing-related accuracy losses persist with inserted text and across thinking budgets.** Direct-minus-control mIoU in percentage points with 95% confidence intervals. The "Unlimited" budget applies no forced early closure and keeps the 4,096-token generation cap.

Boundary behavior. With thinking, both boundaries move inward, and the start moves more. The narrowing therefore does not come mainly from an earlier predicted end. Separate controls find less reliable end estimates, and neither asking for the end first nor supplying the model's own start removes the error. Keeping the video duration fixed and replacing the frames after the annotated end with unrelated frames reduces Qwen3.5-4B thinking-mode end error from 4.0 to 2.9 s and Qwen3.5-2B end error from 4.7 to 3.2 s. A balanced yes/no probe also finds lower end accuracy in four of five checkpoints across three model families. These controls distinguish end-boundary reliability from the broader narrowing effect.

Dependence on surrounding video context. A regression across five evaluation groups associates shorter-video settings with larger thinking losses (Figure 6). This regression uses 11,425 paired queries. Unlike the primary evaluation, its Charades-STA group includes parsed queries from development videos. ActivityNet-A is the only group whose leave-one-group-out prediction has the wrong sign. In a separate 2,000-query ActivityNet evaluation, thinking improves mIoU by 4.0 points for Qwen3.5-2B and 1.4 points for Qwen3.5-4B. Only the 2B improvement remains significant after correction.

Padding the same events into videos four times as long changes the thinking loss from $+7.7$ to -1.4 for Qwen3.5-4B and from $+11.7$ to $+0.8$ for Qwen3.5-2B. Both confidence intervals include zero after padding. Padding also changes the background content and the relative event position, so it perturbs the surrounding context as a whole. The effect of thinking therefore depends on the video setting.

D.3 Calibration controls and robustness

Width and center controls. The training-set prior in Table 3 is

$$w_{\text{prior}} = D \exp(\mu),$$

where $\mu = -1.3818$ is the mean log annotated width-to-video ratio on Charades-train. This prior differs from the median ratio used for window selection. The table compares prior-only replacement, verification-only replacement, and the fixed SeekSpan fusion. Table 9 sweeps the fusion weight.

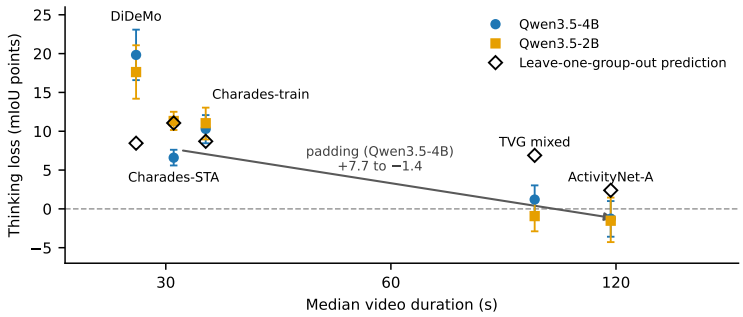


Figure 6: **The effect of thinking varies across video settings.** Thinking loss versus log median video duration for five regression groups; positive values favor direct mode. Error bars show 95% video-cluster bootstrap intervals, open markers are leave-one-group-out predictions, and the arrow denotes the Charades-STA padding intervention. Groups near 30 s are horizontally offset for legibility.

The preselected $\alpha = 0.5$ is the best value on four of the six configurations and within 1.1 points of the best on the other two. The verification width alone ($\alpha = 1$) is the worst setting for all four direct configurations.

Table 9: **Sensitivity to the fusion weight.** mIoU for $\alpha \in \{0, 0.25, 0.5, 0.75, 1\}$ on the analysis subsets of Table 3, with each generator as its own verifier. Setting $\alpha = 0$ reconstructs the generated width around the generated center with boundary adjustment, and $\alpha = 1$ uses the verification width alone. SEEKSPAN fixes $\alpha = 0.5$ in advance; bold marks the best value in each row.

Generator	n	$\alpha = 0.0$	$\alpha = 0.25$	$\alpha = 0.5$	$\alpha = 0.75$	$\alpha = 1.0$
Direct						
Qwen3.5-4B	3,497	55.26	**56.46**	55.82	53.73	50.73
Qwen3.5-2B	3,497	52.28	54.12	**54.48**	53.19	50.86
Qwen3-VL-8B	3,497	53.33	55.91	**56.51**	55.33	52.90
Qwen2.5-VL-7B	3,497	46.49	48.12	**48.22**	46.77	44.13
Thinking						
Qwen3.5-4B	3,322	49.12	54.35	**57.04**	55.75	51.19
Qwen3.5-2B	3,139	42.08	46.73	50.77	**51.85**	49.29

The center-source controls keep the SEEKSPAN width and move the center toward the verified region. A one-quarter shift improves Qwen3.5-4B direct by 0.5 points and gives no consistent improvement on the other configurations. A full move to the verified-region center lowers mIoU by 3–6 points. On five of six configurations, SEEKSPAN scores above the interval spanning the complete verified region.

Verifier and temporal resolution. We apply verification curves from four models to each generator (Table 10). The shared Qwen3.5-4B verifier is the best verifier on five of six configurations. On all four configurations with a generator other than Qwen3.5-4B, it also scores above the generator's own verifier. Every curve raises every configuration above its base except the Qwen2.5-VL-7B curve on Qwen3.5-4B direct, and that curve scores lowest on all six configurations. On Charades-STA, all tested window counts in $\{8, 16, 32, 64\}$ are within 0.35 mIoU points of the sixteen-window result (Table 11). Coarser windows perform better on ActivityNet-A and finer windows on QVHighlights. Fixed two-second windows let the window count vary per query. They stay within 0.2 mIoU points of the sixteen-window result on Charades-STA and require more verifier calls on the longer ActivityNet-A and QVHighlights videos.

Prompt and prefix sensitivity. Two verifier-question paraphrases on a 400-query subset have median Spearman correlations of 0.89 and 0.95 with the reported verification curves. The corresponding changes in SEEKSPAN accuracy have confidence intervals including zero. A verifier prefix that

Table 10: **Verification curves transfer across generators.** mIoU after SEEKSPAN with $\alpha = 0.5$ on the analysis subsets of Table 3 when the verification curve comes from each of four verifiers. Base is the generated interval. Every entry differs from its base with a 95% video-cluster bootstrap interval excluding zero; the only decrease is the Qwen2.5-VL-7B curve applied to Qwen3.5-4B direct. Bold marks the best verifier in each row. The primary results use Qwen3.5-4B as the shared verifier.

			Verifier			
Generator	n	Base	Qwen3.5-4B	Qwen3.5-2B	Qwen3-VL-8B	Qwen2.5-VL-7B
Direct						
Qwen3.5-4B	3,497	55.05	55.82	55.97	**55.98**	54.01
Qwen3.5-2B	3,497	52.23	**54.76**	54.48	54.41	52.89
Qwen3-VL-8B	3,497	52.96	**57.17**	56.48	56.51	54.79
Qwen2.5-VL-7B	3,497	46.39	**49.69**	49.30	49.15	48.22
Thinking						
Qwen3.5-4B	3,322	49.01	**57.04**	56.26	56.17	55.10
Qwen3.5-2B	3,139	42.02	**51.98**	50.77	50.75	50.07

Table 11: **Sensitivity to the number of verification windows.** mIoU after SEEKSPAN on the jointly parsed queries of each dataset, for $B \in \{8, 16, 32, 64\}$ uniform windows and for fixed two-second windows (2 s). Base is the generated interval. SEEKSPAN uses the training-split rule of Section 3.2 (16 windows on Charades-STA and ActivityNet, 32 on QVHighlights); bold marks the best setting in each row.

Dataset	Generator	n	Base	$B = 8$	$B = 16$	$B = 32$	$B = 64$	2 s
Charades-STA	Qwen3.5-4B, direct	3,322	55.60	56.15	**56.34**	56.27	56.11	56.19
Charades-STA	Qwen3.5-4B, thinking	3,322	49.01	**57.17**	57.04	57.05	56.89	57.02
Charades-STA	Qwen3.5-2B, direct	3,139	53.34	55.69	**55.81**	55.69	55.61	55.66
Charades-STA	Qwen3.5-2B, thinking	3,139	42.02	**52.20**	51.98	51.84	51.64	51.99
ActivityNet-A	Qwen3.5-4B, direct	440	42.39	**41.59**	40.32	39.30	39.23	39.42
ActivityNet-A	Qwen3.5-4B, thinking	440	43.65	**42.24**	40.99	40.35	40.23	40.61
ActivityNet-A	Qwen3.5-2B, direct	404	38.94	**39.92**	38.70	37.90	37.61	37.54
ActivityNet-A	Qwen3.5-2B, thinking	404	40.46	**40.50**	39.24	38.88	38.06	38.60
QVHighlights	Qwen3.5-4B, direct	439	67.95	57.36	59.30	60.77	61.45	**61.54**
QVHighlights	Qwen3.5-4B, thinking	439	68.10	57.91	59.95	61.61	62.09	**62.25**
QVHighlights	Qwen3.5-2B, direct	434	56.12	50.23	51.54	52.43	53.15	**53.41**
QVHighlights	Qwen3.5-2B, thinking	434	63.70	55.39	57.04	58.14	58.67	**58.79**

follows the chat template exactly changes each primary calibration gain by at most 0.09 points, and no paired prefix difference has a confidence interval excluding zero.

Dependence on the generated duration. Gains are concentrated on predictions shorter than their annotations (Table 12). InternVL3.5-8B has the largest center errors in Table 8 and also improves on wider predictions. With a displaced center, the IoU-maximizing width exceeds the annotated width (Section 3.4). On Charades-STA, fusion substantially reduces absolute log-width error for thinking outputs. For direct outputs, the generated duration is already more accurate than the verification duration.

Table 12: **Gains are concentrated on predictions shorter than their annotations.** Paired mIoU change of SEEKSPAN with the shared Qwen3.5-4B verifier on Charades-STA queries outside the development videos, split by whether the generated interval is narrower or wider than the annotation. * marks a 95% video-cluster bootstrap interval excluding zero.

		Narrower than annotation				Wider than annotation			
Generator	Mode	n	Base	SEEKSPAN	Δ	n	Base	SEEKSPAN	Δ
Qwen3.5-2B	Direct	1,762	54.1	61.6	+7.54*	1,735	50.4	47.8	−2.56*
Qwen3.5-4B	Direct	1,636	53.4	59.5	+6.16*	1,861	56.5	52.6	−3.96*
Qwen3.5-9B	Direct	1,329	58.7	63.1	+4.39*	2,168	57.6	53.7	−3.90*
Qwen3-VL-8B	Direct	2,158	55.2	64.0	+8.81*	1,339	49.3	46.1	−3.20*
Qwen2.5-VL-7B	Direct	1,713	48.1	56.3	+8.16*	1,784	44.7	43.4	−1.37*
InternVL3.5-8B	Direct	2,039	36.2	48.5	+12.37*	1,455	36.4	41.2	+4.83*
MiMo-VL-7B	Direct	1,972	50.9	62.0	+11.08*	1,504	40.8	39.3	−1.46*
Qwen3.5-2B	Thinking	2,155	42.8	58.3	+15.48*	984	40.3	38.2	−2.14*
Qwen3.5-4B	Thinking	2,365	50.5	63.0	+12.58*	957	45.4	42.2	−3.20*
Qwen3.5-9B	Thinking	1,919	58.0	66.3	+8.24*	1,412	50.6	46.9	−3.63*
Qwen3-VL-8B	Thinking	2,220	58.1	65.5	+7.47*	1,229	52.1	46.9	−5.25*
InternVL3.5-8B	Thinking	1,966	36.2	47.3	+11.12*	1,439	36.5	40.0	+3.44*
MiMo-VL-7B	Thinking	2,343	34.9	45.1	+10.23*	926	43.4	41.0	−2.34*

D.4 SAMPLING AND COMPUTATIONAL COST

Consensus samples $K = 16$ answers and selects the interval with the highest average pairwise IoU to the others. The thinking-mode accuracy comparison uses 374 queries, and the timing run uses a separate set of 200 queries. On the larger analysis subset of Table 3, SEEKSPAN raises Qwen3.5-4B thinking from 49.0 to 57.0 mIoU. Consensus takes its position from the samples. On the 3,497 direct-mode analysis queries of Table 3, it raises Qwen3.5-4B from 55.0 to 58.1 mIoU, 2.3 points above SEEKSPAN (55.8). Calibrating the consensus answer lowers mIoU by 1.3 points.

On one RTX A6000, the median greedy thinking-generation time is 20.54 s. Sixteen verifier forwards take 9.56 s when video preparation is repeated for every window; a cached-input proxy gives 6.82 s for the verifier component. The 27.4 s estimate sums generation and cached verification; repeated preparation instead gives 30.1 s. The 329 s consensus estimate is 16×20.54 s.

D.5 CROSS-DATASET TRANSFER

The cross-dataset evaluation covers 40 model–mode configurations. Fixed calibration improves 17 of them and lowers 23. HiREST is the only dataset outside Charades on which all four evaluated configurations improve, by 3.3–4.3 mIoU points. On QVHighlights, TACoS, and TVGBench, inaccurate verification widths increase the width error. On ActivityNet, calibration can reduce the width error and still lower IoU because the generated position is inaccurate. On ActivityNet-A, the generated intervals of Qwen3.5-4B and -2B have a mean absolute center error of 19–22 s against a median annotated width of 22–23 s. On Charades-STA, the center error is 3.3–3.7 s against a median annotated width of 7.0 s.

Table 14 gives per-configuration results on five of the evaluated datasets. Its Charades-STA values come from an analysis subset of the test queries, so they differ from the all-query values in Table 1.

Table 13: **Components of the inference-cost estimate.** Qwen3.5-4B thinking, timed on one RTX A6000 over a separate set of 200 queries. Totals are sums or multiples of the measured components.

Quantity	Seconds/query	Measurement or construction
One thinking generation	20.54	Measured median
Sixteen verifier forwards	9.56	Repeated video preparation
Cached verification	6.82	Cached-input proxy
SEEKSPAN total	27.4	Generation + cached proxy
SEEKSPAN total, repeated preparation	30.1	Generation + verifier forwards
Sixteen-sample consensus	329	16× generation median

Table 14: **Selected cross-dataset calibration results.** Paired mIoU changes from SEEKSPAN calibration applied to every query. 4B and 2B denote Qwen3.5 generators. * marks a 95% bootstrap interval excluding zero; † marks significance after correction.

	Qwen3.5-4B		Qwen3.5-2B	
Dataset	Direct	Thinking	Direct	Thinking
Charades-train	-0.7	$+9.9^{*\dagger}$	$+2.5^{*\dagger}$	$+10.9^{*\dagger}$
Charades-STA	$+0.7^{*}$	$+8.0^{*\dagger}$	$+2.5^{*\dagger}$	$+10.0^{*\dagger}$
HiREST	$+3.3^{*}$	$+3.9^{*}$	$+4.3^{*\dagger}$	$+3.6^{*\dagger}$
TVG-ANet	-3.1	-3.0	-1.6	-2.1
ActivityNet-A	-2.1^{*}	-2.7^{*}	-0.2	-1.2

D.6 QUALITATIVE EXAMPLE SELECTION

Figure 4 uses examples selected from the evaluation outputs by fixed criteria. Candidates exclude development videos, have annotated widths of 4–10 s, and come from videos no longer than 32 s. Success cases require direct IoU above 0.6, thinking width below 0.6 of the annotated width, thinking-center error below 1.5 s, and SEEKSPAN IoU above 0.7. Figure 4 shows the third-largest qualifying gain and Figure 1 the second-largest. The two diagnostic failure examples are the largest qualifying IoU decrease from over-expansion and the largest qualifying thinking-center error with final IoU below 0.3.

E BASELINE CONSTRUCTION AND COMPARISON METHODS

Sensitivity to the direct prompt. The timestamp-only REZE prompt improves the four tested backbones by 1.2–17.4 mIoU points over the narrative query template of the OneThinker evaluation items. Two paraphrases with the same output format score 0.5–4.2 points below the REZE wording. Each Base–SEEKSPAN pair uses the same prompt, so the prompt choice does not contribute to the calibration gain.

With the same prompt, we compare generated intervals with two verification-based interval estimates. Neither verification estimate improves over the generated interval for the Qwen3-family models. On Qwen2.5-VL-7B, the REZE isolated-clip estimate improves mIoU by 5.6 points.

Reproduction of released methods. We re-evaluate each released method with its released checkpoint, prompt, frame budget, decoding procedure, and parser (Table 15). Table 1 lists Time-R1, TempSamp-R1, and EvoGround as run by their released evaluation scripts. With REZE's prompt wording and pixel budget, our REZE results are 2.5–3.1 mIoU points below the reported values. We run Time-R1 both in the thinking mode described in its paper and with the no-think suffix of its released evaluation script. The paper reports 58.1 mIoU for the 7B model (Table 7 of the arXiv version), 3.9 points above our re-measurement with the released script. The reported R@0.5 (60.8) lies between our two settings (59.8 and 62.2). VideoChat-R1 reaches 58.8 mIoU in our evaluation and 60.8 in its paper. TempSamp-R1 reaches 59.9 with the prompt of its released evaluation script and 59.4 with the corresponding thinking prompt. Its paper reports 61.7 and 62.1. With its released

script (its prompt with the no-think suffix, 2 frames per second, greedy decoding, at most 200 new tokens), EvoGround reaches 53.5 mIoU against a reported 53.1. Our R@0.3 is above the reported value, and our R@0.5 and R@0.7 are below. Without the suffix, the same prompt asks for a thinking segment and gives 51.8 mIoU with 12 unparsed answers.

DSE-VTG, posted on 8 September 2026, reports 51.3 mIoU on Charades-STA using a frozen Qwen3-VL-Embedding-8B and similarity-based interval decoding (Cao et al., 2026). Its code is not released, so we report its published numbers. Table 16 lists every published Charades-STA result we compare with, including methods omitted from Table 1.

Table 15: **Reproduction of released methods on all 3,720 Charades-STA test queries.** Each grounding-trained method uses its released checkpoint and inference settings, and REZE uses the prompt and decoder of its paper. Unparsed answers receive IoU 0. Left: measured here. Right: reported in the paper (–: not reported).

Method	Backbone	R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
TempSamp-R1 (released evaluation script, no thinking; Table 1)	Qwen2.5-VL-7B	82.2	71.0	47.9	59.9	83.3	73.6	52.2	61.7
TempSamp-R1 (thinking prompt of the same script)	Qwen2.5-VL-7B	82.1	70.1	46.7	59.4	83.6	74.1	52.9	62.1
VideoChat-R1 (released evaluation script; Table 1)	Qwen2.5-VL-7B	80.9	69.1	46.4	58.8	–	71.7	50.2	60.8
EvoGround (released evaluation script, no-think suffix; Table 1)	Qwen2.5-VL-7B	78.5	59.9	34.0	53.5	77.2	60.5	35.5	53.1
EvoGround (thinking, no-think suffix removed)	Qwen2.5-VL-7B	76.8	57.6	31.8	51.8	77.2	60.5	35.5	53.1
Time-R1 (released evaluation script, no-think suffix; Table 1)	Qwen2.5-VL-7B	79.0	62.2	34.9	54.2	78.1	60.8	35.3	58.1
Time-R1 (thinking, as the paper describes)	Qwen2.5-VL-7B	78.9	59.8	32.7	53.5	78.1	60.8	35.3	58.1
REZE (verbatim setting; Table 1)	Qwen2.5-VL-7B	75.5	56.5	32.1	51.7	79.8	59.9	35.0	54.2
REZE (verbatim setting; Table 1)	Qwen3-VL-8B	77.4	59.3	33.0	52.6	80.6	64.9	37.8	55.7

Table 16: **Published Charades-STA results used for comparison (full list).** R@1 at IoU 0.3/0.5/0.7 and mIoU (%) are copied from each paper's own results table and not re-run here (–: not reported). Base-model rows reported by other papers are omitted because they vary by more than ten points across prompts and frame budgets.

Method	Backbone	arXiv	R@0.3	R@0.5	R@0.7	mIoU
Fine-tuned on the Charades-STA training split						
HawkEye (FT)	Vicuna-7B	2403.10228	72.5	58.3	28.8	49.3
TRACE-7B (FT)	Mistral-7B	2410.05643	–	61.7	41.4	–
TimeSuite (FT)	Mistral-7B	2410.19702	79.4	67.1	43.0	–
VideoChat-R1	Qwen2.5-VL-7B	2504.06958	–	71.7	50.2	60.8
VideoChat-R1.5 (FT)	Qwen2.5-VL-7B	2509.21100	82.8	71.6	48.3	60.6
Time-R1 (FT)	Qwen2.5-VL-7B	2503.13377	82.8	72.2	50.1	60.9
TempSamp-R1	Qwen2.5-VL-7B	2509.18056	83.6	74.1	52.9	62.1
TempSamp-R1 (no CoT)	Qwen2.5-VL-7B	2509.18056	83.3	73.6	52.2	61.7
Trained on other grounding data or unlabeled videos						
TimeChat	LLaMA-2-7B	2312.02051	–	32.2	13.4	–
VTimeLLM-7B	Vicuna-7B	2311.18445	51.0	27.5	11.4	31.2
Momentor-7B	LLaMA-7B	2402.11435	42.6	26.6	11.6	28.5
HawkEye	Vicuna-7B	2403.10228	50.6	31.4	14.5	33.7
TRACE-7B	Mistral-7B	2410.05643	–	40.3	19.4	38.7
TimeSuite	Mistral-7B	2410.19702	69.9	48.7	24.0	–
TimeMarker-8B	LLaMA3-8B	2411.18211	73.5	51.9	26.9	48.4
Time-R1	Qwen2.5-VL-7B	2503.13377	78.1	60.8	35.3	58.1
EvoGround	Qwen2.5-VL-7B	2605.13803	77.2	60.5	35.5	53.1
Training-free						
TFVTG	BLIP-2 + GPT-4 Turbo	2408.16219	67.0	50.0	24.3	44.5
UG-Ground	Qwen2.5-VL-7B	2510.00705	70.1	51.0	30.1	48.4
UG-Ground	InternVideo2.5-8B	2510.00705	74.7	52.5	32.6	51.0
REZE	Qwen2.5-VL-7B	2608.04480	79.8	59.9	35.0	54.2
REZE	Qwen3-VL-8B	2608.04480	80.6	64.9	37.8	55.7
DART	LLaVA-1.6-7B	2607.00672	71.0	52.0	29.4	48.9
DSE-VTG	Qwen3-VL-Embedding-8B	2609.08850	75.7	58.6	32.2	51.3
FV-Action	Qwen2.5-VL-7B	2608.08315	70.1	56.8	29.7	–
Training-free, test split not stated						
MarkIt	Qwen2-VL-7B	2604.25886	62.3	38.8	16.6	41.1

Published verification-based methods. FV-Action and UG-Ground use Qwen2.5-VL-7B with their own frame sampling and verification procedures (Huang et al., 2026; Kim et al., 2026). FV-Action reports R@0.3/0.5/0.7 of 70.1/56.8/29.7; UG-Ground reports 70.1/51.0/30.1 with 48.4 mIoU. Our corresponding SEEKSPAN configuration uses Qwen2.5-VL-7B to generate the interval and the shared Qwen3.5-4B model to estimate its duration. It reaches 74.7/54.0/27.0 and 49.6 mIoU, with the highest R@0.3 of the three, higher R@0.5 and mIoU than UG-Ground, and lower R@0.7 than both.

SEEKSPAN makes sixteen verifier calls per query after generation and FV-Action makes 36, though the two use different models and visual inputs. Construction-matched comparisons using the same verifier and temporal windows are reported in Section 4.3 and Appendix D.3.

Grounding-trained systems. The grounding-trained systems in Table 1 are trained for temporal grounding, and their predicted durations are already calibrated. On Charades-STA, their median predicted-to-video duration ratios are 0.27–0.31, close to the 0.26 of the training annotations. SEEKSPAN targets generators whose predicted durations are too short (Table 12), so these systems lie outside the setting it is designed for.

F ADDITIONAL CALIBRATION ANALYSES

Verification-curve decoding. SEEKSPAN uses the above-mean contiguous run through the verification peak. We compare this rule with maximum-sum, two-means, largest-mass, and smoothed alternatives. With Qwen3.5-4B, the above-mean run performs best among the tested decoding rules.

Coordinate substitutions. The annotation-based oracle analysis replaces either the generated width or the generated center and keeps the other coordinate unchanged. Because IoU depends jointly on position and width, the two substitutions do not form an additive error decomposition. Thinking outputs have the larger extent error, and correcting the width recovers a larger fraction of the available oracle improvement for them than for direct outputs.

Fusion weight across datasets. The weight $\alpha = 0.5$ is selected once on the 223 Charades-STA development queries and remains fixed in all reported evaluations. A later sweep evaluates other weights across datasets. Where SEEKSPAN helps, $\alpha = 0.5$ is usually the best value or close to it. Where calibration lowers accuracy, the loss grows as more weight goes to the verification estimate.

G PROMPTS AND TEMPLATES

The core prompts and tested paraphrases are reproduced below. `{query}` denotes the query sentence, and `{t0}` and `{t1}` denote temporal-window boundaries.

Direct prompt.

```
Please find the visual event described by the sentence
'{query}', determining its starting and ending times.  The
format should be:  'The event happens in <start time> - <end
time> seconds'.
```

Thinking mode uses the same query prompt with the corresponding thinking channel enabled. MiMo-VL-7B direct mode appends the `/no_think` suffix, while thinking mode omits it. InternVL3.5-8B uses its released R1-style system prompt.

Direct-prompt paraphrases. The output format is unchanged.

```
p2:  Find when the event '{query}' happens in the video,
giving its start and end times.  The format should be:  'The
event happens in <start time> - <end time> seconds'.
p3:  During which part of the video does '{query}' occur?
State the starting and ending times.  The format should be:
'The event happens in <start time> - <end time> seconds'.
```

Narrative template of the OneThinker evaluation items.

```
Give the query: '{query}', when does the described content
occur in the video?
```

The OneThinker evaluation item stores its reference answer inside `<answer>` tags as a JSON object of the form `{"time": [start, end]}` (Feng et al., 2026).

Verifier question. For each temporal window, the verifier views the full video and receives the following yes/no question. We read next-token probabilities at the answer position without autoregressive generation. Times are printed to one decimal place.

```
Watch the full video. Between {t0} seconds and {t1}
seconds, is this happening: "{query}"? Answer yes or no.
Please answer this question based on the visual
content.Provide your thinking process between the <think>
and </think> tags, and then give your final answer between
the <answer> and </answer> tags.At the end, you must output
the final answer in the format:
<answer><your_answer_here></answer>
```

The instruction text is the released OneThinker evaluation template (Feng et al., 2026). We keep it unchanged in verifier runs, down to the missing spaces after `content.` and `tags.` in the original.

By default, the verifier uses the assistant prefill `<think>\n </think>\n\n <answer>`, and reads the next-token distribution immediately after `<answer>`. We also evaluate a prefix that follows the chat template exactly. The two prefixes produce highly correlated curves (mean Spearman correlation 0.97), and replacing the default prefix changes every main SEEKSPAN gain by at most 0.09 mIoU points. No paired prefix difference has a confidence interval excluding zero.

Verifier-question paraphrases. The following alternatives change only the temporal-window question and retain the remaining instruction text and assistant prefix.

```
p1: Watch the full video. Does the event "{query}" occur
between {t0} s and {t1} s? Answer yes or no.
p2: Watch the full video. Consider the interval from {t0}
to {t1} seconds. Is "{query}" happening there? Answer yes
or no.
```

Cross-task boundary probe. We ask six balanced yes/no questions per query, with thinking disabled, greedy decoding, and at most 8 new tokens. The queried time t is placed inside or outside the annotated event according to the probe type.

```
State: At {t} seconds into the video, is this happening:
'{query}'? Answer yes or no.

Onset: By {t} seconds into the video, has this already
started: '{query}'? Answer yes or no.

Offset: By {t} seconds into the video, has this already
ended: '{query}'? Answer yes or no.
```

For the state question, t is the annotated midpoint or 3 s outside the span. For onset, t is 2 s after or 3 s before the annotated start. For offset, t is 2 s after or 3 s before the annotated end.

Boundary interventions.

```
I-1 (end first): Please find the visual event described by
the sentence '{query}', determining its ending and starting
times. The format should be: 'The event ends at <end time>
seconds and starts at <start time> seconds'.
```

```
I-2 (own start given): Please find the visual event
described by the sentence '{query}'. The event starts at
{start} seconds. Determine when it ends. The format should
be: 'The event ends at <end time> seconds'.
```

For I-2, the supplied start is the model's own direct-mode start for that query. The unrelated-text control inserts approximately 380 tokens of video-unrelated text between the video and the query with thinking disabled. The intervention scripts contain the complete inserted text and the code that builds the remaining controls.