IN-CONTEXT SEARCH WITH MASKED RESAMPLING FOR LONG-VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large language models (Video LLMs) localize moments in short videos with high accuracy, yet struggle when the target lasts a few seconds within hours of video. Existing multi-stage methods for long videos, with or without Video LLMs, search from coarse to fine, but refine within a single contiguous window at a time; if that window misses the target, the refinement is wasted. We instead bridge the coarse and fine stages with an in-context search that carries several candidates from one stage to the next. Since a Video LLM outputs only the window it is most confident about, we propose *masked resampling*, which hides the frames of each predicted window from the model and lets it answer again, so that a single pass over the video yields several distinct candidates. The candidates are joined into a *montage*, a sequence of short clips shown with their original timestamps, on which we find that Video LLMs ground without any adaptation. Montages also remove the need to feed a uniformly downsampled whole video to the Video LLM as its coarse input: an embedder instead filters out clearly irrelevant content, and the Video LLM searches a montage of the most relevant clips, which retains the target in most cases at a higher per-frame resolution. The resulting method, MaRS, requires no training and improves three Video LLMs on four long-video benchmarks at the same visual budget, including models already trained on long videos. When long-video training data are available, fine-tuning on montages reduces long-video training to training on short inputs and brings further gains: with MaRS, a fine-tuned 4B model outperforms larger fine-tuned Video LLMs.

1 INTRODUCTION

Temporal video grounding aims to localize the moment in a video that corresponds to a natural-language query. Video large language models (Video LLMs) have recently advanced this task considerably (Zhang et al., 2026; Li et al., 2025; 2026c; Zhu et al., 2026), yet their success has largely been confined to short videos. In long videos, where the queried moment may last only a few seconds within hours of footage, their accuracy degrades sharply (Zhu et al., 2026). A long video yields far more visual tokens than the context window of an LLM can hold, so it must be sampled more sparsely or at a lower resolution, and the moment of interest is easily lost among the surrounding content. Even when the tokens fit, redundant content can distract the model: token reduction and frame selection methods find that Video LLMs perform as well or better with far fewer visual tokens (Huang et al., 2025; Park et al., 2026; Wang et al., 2025a; Qin et al., 2025).

Existing approaches to long-video grounding follow two main directions. One line of work trains Video LLMs directly on long videos (Li et al., 2026c; Zhu et al., 2026). Besides being costly, this does not remove the underlying difficulty, since the model must still locate the target in a single pass over a sparsely sampled input. The other line searches from coarse to fine, a strategy inherited from earlier long-video methods, in which a coarse stage selects contiguous windows and the final prediction is made within one of them (Hou et al., 2023; Pan et al., 2023; Hannan et al., 2024). Video LLM-based methods may refine over several stages, each on a shorter window and thus at a higher per-frame resolution under a fixed visual token budget, but each stage only narrows down the window chosen in the stage before (Li et al., 2025; Hannan et al., 2025). In both cases, if the very first window misses the target, every later step is wasted.

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.

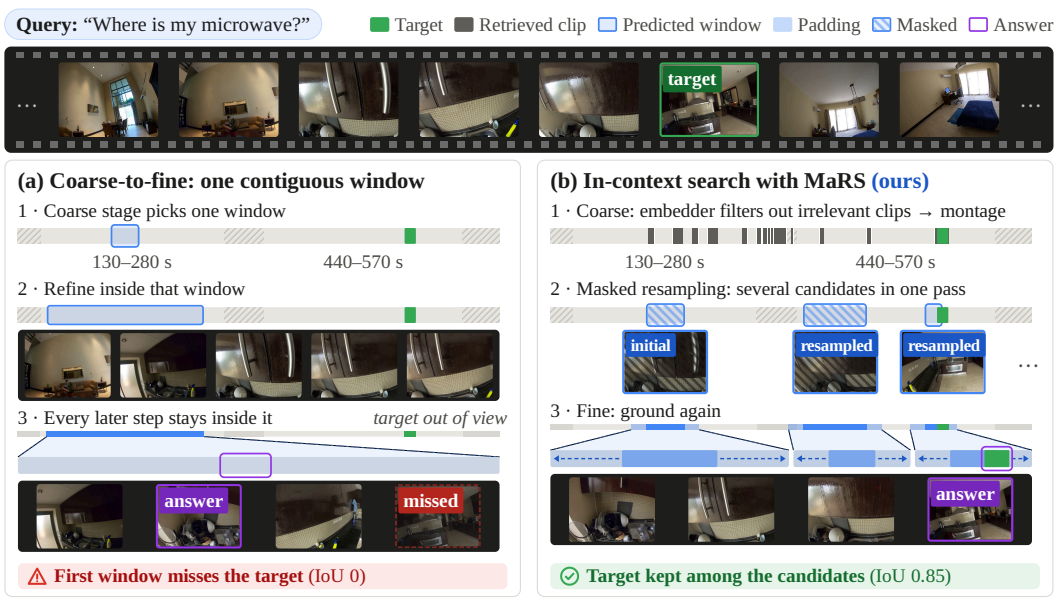


Figure 1: Long-video grounding with a single contiguous window vs. MaRS. (a) Coarse-to-fine methods refine within one contiguous window at a time, so the target is lost when that window misses it. (b) MaRS obtains several candidates from a montage of retrieved clips by masked resampling and grounds again on a montage of the candidates.

We propose to bridge the coarse and fine stages with an in-context search over several candidates (Figure 1). Instead of passing a single window from one stage to the next, the Video LLM proposes several candidate windows, which are joined into one input that is shorter than the coarse one, so that it grounds the query again at a higher per-frame resolution. The target then only needs to be among the candidates, so a single early mistake no longer decides the outcome. Such a search, however, requires the Video LLM to give not only its most confident answer but also other plausible ones, whereas it is trained to output only the single window it considers most likely.

Unlike an embedder, which assigns a score to every clip, a single timestamp response provides no ranking of alternative locations. In our experiments, sampling at a higher temperature yields little candidate diversity (Table 1), consistent with the entropy collapse reported in reinforcement learning for reasoning LLMs (Cui et al., 2025; Yu et al., 2025). We therefore propose *masked resampling*: after each prediction, the frames of the predicted window are masked out of the key–value cache, and the model reads the query again and gives a new answer. Each new answer is thus drawn to a different part of the video, and since the rest of the cache is reused, each additional candidate costs only a short decoding step.

The candidates obtained in this way usually lie far apart in the video, so we join them into a *montage*: a sequence of short clips arranged in temporal order and shown with their original timestamps. Since Video LLMs are commonly trained on short videos or on uniformly downsampled long videos, whose timestamps are evenly spaced, existing coarse-to-fine methods follow this convention and feed the whole downsampled video to the coarse stage (Li et al., 2025; Liu et al., 2026; Hannan et al., 2025). We find, however, that Video LLMs ground on montages directly, without any adaptation, despite their irregularly spaced timestamps. We believe this is because grounding only requires the model to find the relevant frames and read off their timestamps, which does not depend on regular spacing. This compatibility opens a more effective option for the coarse stage: an embedder first filters out clearly irrelevant clips, and the Video LLM then searches a montage of the most relevant ones. Every later round is then left entirely to the in-context search of the Video LLM: masked resampling yields several candidates, which are joined into a new montage for a closer look in the next round, until the final answer is given.

The framework above requires no long-video training and applies even to Video LLMs trained only on short videos. When long-video training data are available, however, the question becomes how

to exploit them without the cost of processing entire long videos, and in a way that is consistent with how the model is used within our framework. We address this with *montage-based fine-tuning*, in which each training sample is a montage of clips that the embedder retrieves from a long video. Such a sample then has the input length of a short video yet carries the supervision of a long one, and the model is trained on the same kind of input it receives at test time. Since the embedder has already removed the clearly irrelevant content, training also concentrates on the hard part of the task: telling the target apart from the most semantically similar content in the video.

We evaluate our in-context **Ma**sked **R**esampling **S**earch framework, *MaRS*, on four long-video grounding benchmarks. Without any training and at the same visual budget, it enables a Video LLM trained on short videos to rival those trained specifically on long videos, and further improves those trained on long videos. Moreover, montage-based fine-tuning trains Video LLMs on long-video grounding data at the input length of short videos, enabling a 4B model to outperform larger fine-tuned Video LLMs. Our contributions are as follows:

- We identify that coarse-to-fine methods for long-video grounding refine within a single contiguous window at a time, and show that Video LLMs ground on montages of short clips without adaptation, which lets them search beyond a single window.
- We propose masked resampling, which obtains several distinct candidates from a single encoding of the video by hiding earlier answers, and build on it MaRS, a multi-round search that carries several candidates from one round to the next.
- We propose montage-based fine-tuning, which provides long-video supervision at the input length of a short video and matches the input the model receives during the search.

2 Related Work

Temporal grounding with Video LLMs. Temporal grounding was long addressed with specialized models that score or regress candidate moments from video features (Zhang et al., 2020b;a; Lei et al., 2021). Video LLMs instead learn temporal grounding through instruction tuning (Ren et al., 2024; Huang et al., 2024; Pang et al., 2026) and, more recently, reinforcement learning with temporal IoU rewards (Wang et al., 2025c; Zhang et al., 2026; Zhu et al., 2026), mostly by generating the start and end times of the queried moment as text. These models ground short videos accurately but degrade as videos grow longer (Zhu et al., 2026). Rather than training another such model, we ask how an existing one can search within a long video in context; MaRS therefore applies to these models without any training. When long-video data are available, montage-based fine-tuning turns long-video training into training on short montages, which focuses the model on telling the target apart from similar content.

Long-video temporal grounding. Benchmarks such as MAD (Soldan et al., 2022), Ego4D-NLQ (Grauman et al., 2022), MomentSeeker (Yuan et al., 2025), and VUE-TR (Vidi Team et al., 2025a;b) ask for moments of a few seconds within videos of up to several hours. Earlier methods without Video LLMs select contiguous windows in a coarse step and make the final prediction within one of them (Hou et al., 2023; Pan et al., 2023; Hannan et al., 2024). Video LLMs either train on long videos directly (Zhu et al., 2026; Li et al., 2026c) or search over several stages: UniTime (Li et al., 2025) narrows the video down to ever shorter windows, and VideoMind (Liu et al., 2026) verifies several proposed windows one by one before keeping one. Methods that pass several windows beyond the coarse stage obtain them from a separately trained proposer (Liu et al., 2026) or from a fixed partition of the video (Li et al., 2025; Hannan et al., 2025); none obtains them from an off-the-shelf Video LLM without training, as masked resampling does.

Retrieval for long videos. Retrieval-augmented generation (Lewis et al., 2020) has been adapted to long videos, mostly for question answering (Luo et al., 2025; Wang et al., 2025d). For grounding, a recent pipeline retrieves a few long windows with an image embedder and grounds once on their union (Seo & Kim, 2026), so that the retriever alone decides what the Video LLM sees. In MaRS, the embedder only removes clearly irrelevant clips, and the Video LLM makes the decisive choices over several stages. Masked resampling also relates to reranking with language models (Nogueira et al., 2020; Qin et al., 2024; Sun et al., 2023; Zhuang et al., 2024), as it reads a ranking of candidate windows directly out of the grounding model.

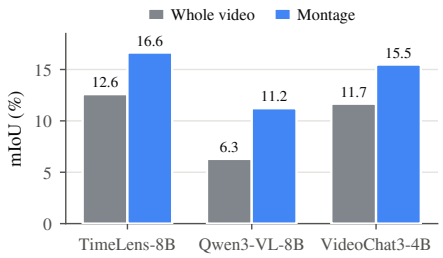


Figure 2: Grounding on the whole video vs. on a montage of the top-32 clips (4 s each) from Qwen3-VL-Embedding-2B (Li et al., 2026a), at the same visual budget (Ego4D-NLQ (Grauman et al., 2022)).

Table 1: Eight candidates from one montage (Ego4D-NLQ (Grauman et al., 2022)). *Distinct*: number of distinct candidates; *Recall*: share of queries (%) for which some candidate overlaps the target. *Temperature*: sampling at T=1; *List*: prompting for up to eight windows; *Masked*: masked resampling.

	TimeLens-8B		Qwen3-VL-8B	
Method	Distinct	Recall	Distinct	Recall
Single answer	1.0	34.1	1.0	20.4
Temperature	2.2	53.5	3.5	43.2
List	1.5	34.3	**5.1**	26.1
Masked	**5.4**	**69.5**	4.8	**48.9**

3 METHOD

Given a long video V and a natural-language query q, our goal is to localize the moment $[t_s, t_e]$ that q describes. MaRS searches for this moment from coarse to fine without settling on a single window before its final answer: at every stage, the Video LLM proposes several candidate windows, which are joined into a shorter input for the next stage. Section 3.1 introduces the montage, the input on which the Video LLM grounds at every stage, and describes how an embedder builds the first one. Section 3.2 presents masked resampling, which obtains several candidates from a single pass over a montage, and Section 3.3 combines the two into a multi-stage search. Section 3.4 shows how the same montages allow the Video LLM to be fine-tuned on long-video data.

3.1 GROUNDING ON MONTAGES

Setup. A Video LLM grounds a query by reading a sequence of sampled frames, each preceded by a textual timestamp (Zhang et al., 2026; Bai et al., 2025a; Li et al., 2026c), and generating the start and end times of the queried moment as text. Let frames x_τ be sampled from V at times $\tau \in \mathcal{T}$, and let T be the duration of V. For a time interval $c \subseteq [0, T]$, we define its *frame sequence* $X(c)$ as the frames sampled within c, in temporal order, each preceded by its timestamp $\langle\tau\rangle$ written as text. The model receives a frame sequence together with q and decodes a window $w = [s, e]$.

Candidate clips. We divide V into non-overlapping clips $\{c_j\}$ of length ℓ and score each clip against the query with an embedder whose visual and text encoders are f_v and f_t:

$$r_j = \frac{1}{|\mathcal{U}_j|} \sum_{u \in \mathcal{U}_j} \cos\big(f_v(u),\, f_t(q)\big), \tag{1}$$

where $\mathcal{U}_j$ contains the visual inputs that f_v encodes: the clip itself, $\mathcal{U}_j = \{c_j\}$, for a video embedder (Li et al., 2026a), and its frames, $\mathcal{U}_j = \{x_\tau\}_{\tau \in c_j}$, for an image embedder (Tschannen et al., 2025; Radford et al., 2021). We keep the k highest-scoring clips. The budget $k\ell$ is deliberately generous: the embedder only needs to discard content that is clearly irrelevant, not to locate the target. When the video is shorter than $k\ell$, all clips are kept. We use $\ell = 4$ s.

Montage. Given disjoint time intervals $c_1, \ldots, c_m$ in temporal order, we concatenate their frame sequences into a single input that we call a *montage*:

$$M = X(c_1) \oplus X(c_2) \oplus \cdots \oplus X(c_m), \tag{2}$$

where $\oplus$ denotes concatenation. The first montage is built from the embedder's clips; later stages build montages from the model's own candidate windows. Video LLMs are typically trained on uniformly sampled videos, and a montage departs from this setting. Nevertheless, they ground on montages directly, without any adaptation, and more accurately than on the whole video sampled uniformly at the same visual budget (Figure 2). We believe that montages pose no difficulty because the model only needs to identify the relevant frames and report their timestamps, a skill that does not depend on how evenly the timestamps are spaced.

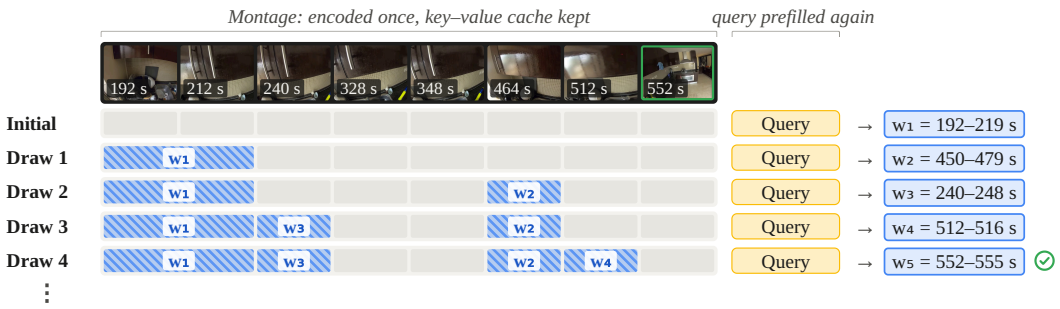


Figure 3: Masked resampling. The montage is encoded once and its key–value cache is kept. After each answer, the frames and timestamps of the answered window are hidden from the query, which is prefilled again, and the next window is decoded; the masks accumulate over draws. Draws are shown until the target (552 s) is found.

3.2 Masked Resampling

MaRS passes several candidates from one stage to the next, yet a Video LLM answers with only one window per pass. Grounding models are typically trained to output the window they are most confident about, and unlike an embedder, which ranks every clip, they leave the remaining candidates unranked. When the top answer is wrong, the correct moment may well be among these alternatives, but a single pass offers no access to them. To recover them, we propose *masked resampling* (Figure 3): once the model has named a window, we hide it and let the model answer again.

Let $w_1 = [s_1, e_1]$ be the initial prediction on a montage M. Before the k-th decoding, we collect the previously predicted windows:

$$H_k = \bigcup_{i<k} [\, s_i,\ e_i \,]. \tag{3}$$

For every $\tau \in H_k$, we prevent the query and answer tokens from attending to the cached timestamp and frame tokens $(\langle\tau\rangle, x_\tau)$ of M. Since the montage precedes the query in the input, its key–value cache from the initial pass does not depend on the query and is reused as is. The query tokens, in contrast, attended to the hidden frames during the initial pass, so we prefill them again under an attention mask that excludes the hidden tokens, and decode w_k under the same mask. Each additional candidate thus costs only the prefill of the short query and the decoding of a window. Some models may output several windows in one answer, in which case all of these windows enter H_k, so the next decoding looks beyond every moment the model has already found.

Why masking. Two simpler ways to obtain several candidates exist, but neither works well in practice (Table 1). Temperature sampling draws further windows from the same output distribution: with eight draws, TimeLens-8B produces barely two distinct windows on average. This low diversity is consistent with the entropy collapse reported during reinforcement learning for reasoning LLMs (Cui et al., 2025; Yu et al., 2025). Prompting the model to list several windows fares no better. TimeLens-8B, trained to output one window, nearly always still returns one, whereas the base Qwen3-VL-8B tends to split a single region into adjacent pieces of equal length. Masking avoids both problems: by removing the evidence for the windows already found, it forces each new draw toward a different part of the montage, yielding more distinct candidates and a higher chance that one of them overlaps the target.

3.3 Multi-Stage Search

Search in rounds. Montages and masked resampling together yield a search from coarse to fine that proceeds in rounds. The first round grounds on the montage of the clips selected by the embedder, and masked resampling returns several candidate windows. These candidates are padded and joined into a new montage of fixed length, on which the next round grounds and draws candidates again. When the new montage is shorter than the previous one, the fixed visual token budget lets the model see it at a higher per-frame resolution, with finer spatial detail; when it keeps the same duration, it instead preserves the temporal reach needed for long target moments, so that both of

their boundaries remain in view. Finally, the model grounds once more on the montage composed from the last round's candidates to give the final prediction. Details are in Appendices A.1 and A.2.

Two search regimes. How much of the search is left to the Video LLM depends on how far the embedder can be trusted, and we therefore instantiate the procedure in two regimes. FOCUS places more trust in the embedder: it starts from a compact selection of clips, e.g., 128 s, which a single round of search narrows before the final prediction. EXPLORE shifts more of the search to the Video LLM: it starts from a selection as long as current single-pass Video LLMs can take in before they subsample the video, e.g., 448–768 s (Zhang et al., 2026; Bai et al., 2025a), and narrows it over two rounds of search. The former is cheaper, whereas the latter is less limited by the embedder.

3.4 MONTAGE-BASED FINE-TUNING

MaRS requires no training, but when long-video training data are available, they can be used to fine-tune the Video LLM for the input it receives during the search. However, fine-tuning on entire long videos is costly, and most of what the model would see is unrelated to the query. We instead fine-tune on montages: the embedder removes clearly irrelevant content in advance, so that training focuses on telling the target apart from the most semantically similar content in the video.

For each training query, we build a montage from the long training video in the same way as at inference, so that each sample has the input length of a short video but the supervision of a long one. We fine-tune the model on these montages with Group Relative Policy Optimization (GRPO) (Shao et al., 2024), following recent work on temporal grounding (Wang et al., 2025c; Zhang et al., 2026). For a query q and its montage M, GRPO samples G responses $\{o_1, \ldots, o_G\}$ from the policy π_θ and scores each with a reward $r(o_i)$: the temporal IoU between the predicted and the ground-truth window, or zero if the response cannot be parsed. Each response receives the group-relative advantage

$$\hat{A}_i = \frac{r(o_i) - \mathrm{mean}\big(\{r(o_j)\}_{j=1}^{G}\big)}{\mathrm{std}\big(\{r(o_j)\}_{j=1}^{G}\big)}, \tag{4}$$

and the model is trained to minimize

$$\mathcal{L}_{\mathrm{GRPO}}(\theta) = -\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\frac{\pi_\theta(o_{i,t} \mid M, q, o_{i,<t})}{\pi_{\theta_{\mathrm{old}}}(o_{i,t} \mid M, q, o_{i,<t})}\,\hat{A}_i, \tag{5}$$

where $o_{i,t}$ is the t-th token of o_i and $\pi_{\theta_{\mathrm{old}}}$ is the policy that generated the group; following previous practice, we omit the clipping operation and the KL term of the original objective (Wang et al., 2025c; Zhang et al., 2026; Hu et al., 2025; Liu et al., 2025).

4 EXPERIMENTS

4.1 EVALUATION

Benchmarks and metrics. We evaluate on four benchmarks with videos of several minutes to two hours: the full validation split of Ego4D-NLQ (Grauman et al., 2022), the text-to-video subset of MomentSeeker (Yuan et al., 2025), and the visual queries of VUE-TR (Vidi Team et al., 2025a) and VUE-TR-V2 (Vidi Team et al., 2025b) (512 and 691 queries whose videos were available). For fine-tuning, we also use TACoS (Regneri et al., 2013) and evaluate on its test split. We report R1@m, the percentage of queries whose prediction reaches a temporal IoU of at least m with the ground truth, and the mean IoU (mIoU); to analyze the search, we also report *coverage*, the fraction of the ground-truth moment inside a montage.

4.2 IMPLEMENTATION DETAILS

Models and search. We apply MaRS to TimeLens-8B (Zhang et al., 2026), TimeLens2-8B (Zhu et al., 2026), and VideoChat3 (Li et al., 2026c). Every grounding pass receives 28,672 visual tokens (14,336 in the ablations), and clips of 4 s are scored with Qwen3-VL-Embedding-2B (Li et al., 2026a) by default, and with SigLIP2 (Tschannen et al., 2025) and CLIP (Radford et al., 2021) in the ablations. FOCUS starts from the top-32 clips (128 s) and searches one round with two additional

Table 2: Temporal grounding results (R1@0.5 and mIoU, %). Group labels give the visual token budget: 128K rows are reported by Zhu et al. (2026) under its evaluation protocol; 28K rows are our evaluation with each model's official prompt and 28,672 tokens per pass. Green: gain over one pass over the whole video by the same model. R1@0.3 and R1@0.7 for the models we evaluated are given in Appendix B.1. †: VUE-TR vision subsets; ‡: MomentSeeker text-only queries. Gray: TimeLens2 models are fine-tuned on Ego4D-NLQ, so their results are not zero-shot there.

Model	Ego4D-NLQ		MomentSeeker‡		VUE-TR†		VUE-TR-V2†	
	R1@0.5	mIoU	R1@0.5	mIoU	R1@0.5	mIoU	R1@0.5	mIoU
4B (128K)								
InternVL3.5-4B (Wang et al., 2025b)	0.7	2.0	2.1	3.2	14.1	16.5	6.0	8.5
Qwen3-VL-4B (Bai et al., 2025a)	10.6	11.4	13.5	15.3	33.0	33.6	19.9	20.3
Molmo2-4B (Clark et al., 2026)	7.7	10.0	14.7	19.5	42.5	43.1	30.9	32.1
TimeLens2-4B (Zhu et al., 2026)	18.1	18.6	27.0	27.9	52.4	53.2	49.5	48.1
VideoChat3-4B (Li et al., 2026c)	13.9	14.6	24.4	25.9	47.8	47.9	40.4	40.2
4B (28K)								
VideoChat3-4B (Li et al., 2026c)	13.9	14.5	24.1	25.4	42.0	40.8	38.5	36.6
+ **MaRS-FOCUS**	17.0$_{+3.1}$	17.6$_{+3.1}$	25.3$_{+1.2}$	26.9$_{+1.5}$	49.8$_{+7.8}$	48.8$_{+8.0}$	49.8$_{+11.3}$	46.8$_{+10.2}$
+ **MaRS-EXPLORE**	17.0$_{+3.1}$	17.5$_{+3.0}$	26.5$_{+2.4}$	27.3$_{+1.9}$	50.8$_{+8.8}$	49.4$_{+8.6}$	51.8$_{+13.3}$	48.6$_{+12.0}$
7B+ (128K)								
Qwen3.5-397B-A17B (Qwen Team, 2026)	13.3	14.5	22.3	23.3	42.7	42.2	35.5	34.5
Qwen3-VL-235B-A22B (Bai et al., 2025a)	11.5	12.7	21.2	23.7	42.5	43.1	34.5	34.1
Qwen3.5-35B-A3B (Qwen Team, 2026)	12.0	13.0	20.2	22.6	46.3	44.6	37.1	35.3
Qwen3-VL-30B-A3B (Bai et al., 2025a)	10.4	11.5	19.6	22.6	40.4	42.5	32.3	32.6
Vidi-1.5-9B (Vidi Team et al., 2025a)	6.5	8.4	14.8	18.6	45.0	47.0	31.9	34.3
LLaVA-OneVision-2-8B (An et al., 2026)	12.0	12.8	18.0	19.6	40.8	41.3	35.0	34.5
Qwen3-VL-8B (Bai et al., 2025a)	4.8	5.3	9.4	9.8	23.8	25.5	13.1	14.0
InternVideo3-8B (Yan et al., 2026)	7.3	8.6	17.9	19.3	39.4	41.1	32.4	32.6
TimeLens-8B (Zhang et al., 2026)	14.6	15.7	20.4	21.9	42.9	43.6	34.4	34.0
Video-o3-7B (Zeng et al., 2026)	1.9	3.1	9.8	13.0	22.1	25.1	12.9	15.3
VideoChat-Flash-7B (Li et al., 2026b)	1.5	2.5	5.9	7.2	15.4	17.3	9.5	10.1
MiMo-VL-7B (Xiaomi LLM-Core Team, 2025)	0.4	1.0	3.8	6.0	17.5	19.1	10.1	10.4
TimeSuite-7B (Zeng et al., 2025)	0.6	1.4	3.3	5.9	10.3	13.2	5.3	7.3
TimeLens2-8B (Zhu et al., 2026)	18.4	19.0	27.3	28.5	51.6	53.5	49.1	47.7
7B+ (28K)								
TimeLens-8B (Zhang et al., 2026)	14.4	15.2	24.2	26.2	42.6	42.0	38.4	36.4
+ **MaRS-FOCUS**	19.1$_{+4.7}$	20.1$_{+4.9}$	25.9$_{+1.7}$	27.2$_{+1.0}$	54.3$_{+11.7}$	51.9$_{+9.9}$	55.6$_{+17.2}$	51.8$_{+15.4}$
+ **MaRS-EXPLORE**	18.8$_{+4.4}$	19.8$_{+4.6}$	26.0$_{+1.8}$	27.5$_{+1.3}$	52.3$_{+9.7}$	51.5$_{+9.5}$	54.7$_{+16.3}$	50.6$_{+14.2}$
TimeLens2-8B (Zhu et al., 2026)	16.3	16.8	25.9	27.1	46.3	46.5	43.4	41.2
+ **MaRS-FOCUS**	20.9$_{+4.6}$	21.5$_{+4.7}$	25.9$_{+0.0}$	28.5$_{+1.4}$	53.7$_{+7.4}$	52.7$_{+6.2}$	58.2$_{+14.8}$	54.0$_{+12.8}$
+ **MaRS-EXPLORE**	21.0$_{+4.7}$	21.6$_{+4.8}$	26.5$_{+0.6}$	28.7$_{+1.6}$	53.7$_{+7.4}$	52.8$_{+6.3}$	58.6$_{+15.2}$	54.7$_{+13.5}$

candidates; EXPLORE starts from the top-128 clips (512 s) and searches two rounds, drawing eight and then two candidates. Further details are given in Appendix A.

Fine-tuning. We fine-tune Qwen3-VL-4B-Instruct (Bai et al., 2025a) with supervised fine-tuning followed by GRPO, both with LoRA adapters (Hu et al., 2022). GRPO trains separate models for Ego4D-NLQ and TACoS on montages of the top-32 clips from their training splits, matching the input of FOCUS. Hyperparameters are given in Appendix A.

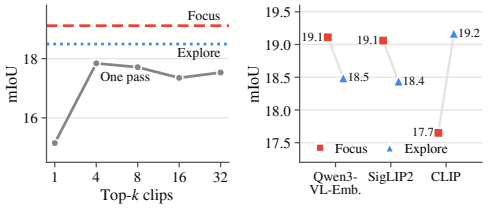
4.3 MAIN RESULTS

Training-free results. At the same budget of 28K tokens per pass, MaRS improves all three Video LLMs in mIoU on every benchmark (Table 2), benefiting both from the montage built by the embedder (Figure 2) and from the in-context search of the Video LLM; Appendix B.1 separates the two by reporting one pass on the montage. With MaRS, TimeLens-8B, which is trained only on short videos, outperforms all baselines other than TimeLens2 and even surpasses TimeLens2 on Ego4D-NLQ and VUE-TR-V2. FOCUS suits TimeLens-8B, which is trained on short videos only, whereas TimeLens2-8B and VideoChat3, which are trained on long videos, benefit more from EXPLORE; MaRS thus bridges a strong retrieval model and a Video LLM for temporal grounding.

Montage-based fine-tuning. Table 3 compares fine-tuned models, each evaluated with its own inference procedure. After montage-based fine-tuning, Qwen3-VL-4B with FOCUS outperforms all previous methods on both benchmarks, by 2.4 to 3.6 points over UniTime-Full, which is trained on a

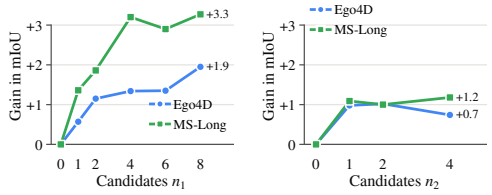
Table 3: Results after fine-tuning on long-video benchmarks. Baseline numbers are taken from Li et al. (2025), and those of TimeLens2 from Zhu et al. (2026). Each method uses its own inference procedure, e.g., multi-stage refinement for UniTime and a single pass for TimeLens2, which is fine-tuned on Ego4D-NLQ but not on TACoS. SP: dataset-specific fine-tuning; Full: universal pre-training. MaRS-FOCUS-4B: Qwen3-VL-4B after montage-based fine-tuning (Section 3.4).

Model	Ego4D-NLQ			TACoS		
	R1@0.3	R1@0.5	mIoU	R1@0.3	R1@0.5	mIoU
2D-TAN (Zhang et al., 2020b)	5.0	2.0	3.5	45.6	35.8	–
VSLNet (Zhang et al., 2020a)	10.8	6.8	8.8	35.5	23.5	25.0
RGNet (Hannan et al., 2024)	18.3	12.0	12.4	–	–	–
SnAG (Mu et al., 2024)	15.9	11.3	–	56.4	44.9	–
SG-DETR (Gordeev et al., 2026)	–	–	–	56.7	44.7	40.9
LD-DETR (Zhao et al., 2025)	–	–	–	57.6	44.3	40.3
UniVTG (Lin et al., 2023)	7.3	4.0	4.9	51.4	35.0	33.6
UniVTG w/ PT	11.7	7.5	7.9	56.1	43.4	38.6
Qwen2-VL-7B (Wang et al., 2024)	0.5	0.2	0.5	4.1	1.3	2.6
Qwen2.5-VL-7B (Bai et al., 2025b)	1.1	0.5	0.8	7.7	3.4	5.3
UniTime-SP-7B (Li et al., 2025)	24.8	16.8	17.3	61.2	48.3	45.0
UniTime-Full-7B (Li et al., 2025)	27.1	18.4	18.8	66.9	55.1	50.0
TimeLens2-8B (Zhu et al., 2026)	–	18.4	19.0	–	–	–
MaRS-FOCUS-4B	**30.2**	**21.1**	**21.2**	**69.3**	**58.7**	**53.2**



(a) Search vs. one pass. (b) Across embedders.

Figure 4: Search beyond the embedder (Ego4D-NLQ). (a) One pass on the top-k clips (expanded to 128 s) vs. FOCUS and EXPLORE. (b) Both regimes with three embedders.



(a) First round (n_1). (b) Last round (n_2).

Figure 5: Number of candidates drawn by masked resampling in the first and last rounds: gain in final mIoU over drawing none.

large collection of grounding data and refines its predictions over multiple stages, and by more than 2 points over TimeLens2-8B on Ego4D-NLQ, although TimeLens2 is twice as large and trained on long videos including Ego4D-NLQ.

4.4 ABLATION STUDIES

Unless stated otherwise, the ablations use TimeLens-8B with a visual budget of 14,336 tokens per pass and Qwen3-VL-Embedding-2B as the embedder. They are conducted on a subset of 1,500 queries from the Ego4D-NLQ validation split and on the MomentSeeker videos longer than twenty minutes, which we denote as MS-Long.

Search beyond the embedder. Figure 4 examines what the search adds beyond the embedder and how much it should rely on it. A single grounding pass on the embedder's top clips saturates at a few clips, and no choice of k reaches either search regime (Figure 4a). Beyond a few clips, a single pass gains nothing from additional clips, whereas FOCUS, starting from the same top-32 clips, gains 1.6 points by searching among them. Across embedders (Figure 4b), the two regimes perform on par with a strong embedder, where FOCUS is the cheaper choice. With a weaker embedder, however, FOCUS, which trusts the embedder's compact selection, loses 1.5 points, whereas EXPLORE, which keeps a broader selection and relies more on the Video LLM's own search by searching over two

Table 4: Different numbers of search rounds from the top-128 clips. *None*: one pass directly gives the final answer. *No draws*: each round passes on only its first answer.

	Ego4D		MS-Long	
Rounds	R1@0.5	mIoU	R1@0.5	mIoU
None	12.2	13.3	6.3	6.7
One	15.7	16.3	9.2	9.3
Two, no draws	15.9	16.3	7.8	7.7
Two	17.9	18.5	10.7	11.1

Table 5: Coverage (%) of the ground truth by a 128 s montage. TL: TimeLens.

Montage	Ego4D	MS-Long
First answer	54.8	25.8
+ 8 random	44.3	27.0
+ 8 draws (TL-8B)	65.2	38.7
+ 8 draws (TL2-8B)	71.4	52.2
Embedder's top-32 clips		
Qwen3-VL-Emb.	70.1	45.6
SigLIP2	63.9	38.4
CLIP	56.4	35.5

Table 6: Montage-based fine-tuning with different training inputs. See main text for explanations of abbreviated names.

	Ego4D		TACoS	
Input	R1@0.5	mIoU	R1@0.5	mIoU
T-DS	5.6	9.2	35.0	34.9
S-DS	6.6	9.1	37.6	36.9
Crop	10.1	12.1	37.2	37.4
Montage	10.6	12.5	39.1	39.2

rounds, varies by less than one point across embedders. Leaving more of the search to the Video LLM thus makes the result less dependent on the quality of the embedder.

Number of search rounds. Table 4 varies the number of search rounds, starting from the same selection as EXPLORE, the top-128 clips. Grounding once on this selection performs poorly, and each round of search improves the result, with two rounds performing best on both benchmarks. The third row passes only the model's first answer from one round to the next, as single-window coarse-to-fine methods do. The second round then brings no gain over a single round on Ego4D-NLQ and even lowers the result on MS-Long. An additional round thus pays off only when several candidates are carried into it.

Number of candidates. Figure 5 varies the number of candidates drawn by masked resampling in each round of search: n_1 in the first round, which starts from the top-128 clips, and n_2 in the last round, which FOCUS and EXPLORE share. n_2 is ablated with a single round of search from the top-32 clips. Accuracy rises with n_1 up to eight candidates on Ego4D-NLQ and levels off after four on MS-Long. Since the first round searches a long selection in which the target is easily missed, we use $n_1 = 8$, which generalizes better across different kinds of data. In the last round, one or two candidates already bring the full gain and further candidates add nothing, so we select $n_2 = 2$.

Coverage of the drawn candidates. Table 5 measures how much of the ground truth the drawn candidates cover, with every selection expanded to 128 s. With TimeLens-8B, eight draws raise coverage over the first answer alone, whereas eight random windows do not, so the draws point to relevant parts of the video. Their coverage lies between that of the strongest and the weaker embedders, and with TimeLens2-8B it exceeds even the strongest embedder, showing that a strong Video LLM has the potential to search its input more effectively than the embedder.

Training inputs. To study the effect of the training input, we train Qwen2.5-VL-3B (Bai et al., 2025b), with interleaved textual timestamps as in TimeLens (Zhang et al., 2026), on Ego4D-NLQ and TACoS with GRPO directly from the base model. We compare four inputs of 8,192 visual tokens: the whole video downsampled in time (T-DS) or in space (S-DS), a contiguous 128 s crop containing the target (Crop), and a montage of the top-32 clips (Montage). Evaluated by grounding once on the same montages (Table 6), training on the downsampled whole video performs worst, and training on montages performs best, ahead of the crop, which does not force the model to tell the target apart from hard negatives.

5 CONCLUSION

We showed that coarse-to-fine search for long-video temporal grounding need not refine one window at a time. Video LLMs can ground on montages of clips with irregular temporal gaps, and with masked resampling they give not only their most confident answer but several candidates, which the next round refines together. Without any training, MaRS improves three Video LLMs on four benchmarks, and a 4B model fine-tuned on montages outperforms larger fine-tuned Video LLMs. On easy queries that the embedder and a single pass already solve, however, the search brings little gain. In future work, we will study how to decide adaptively when to search.

AI USE STATEMENT

In this work, we used generative AI tools to implement methods. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design research methodology or experiments, interpret results, or assist with translation, and formulating mathematical claims, proofs, and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create and edit software code, create scientific figures, suggest experimental parameters, search for and summarize literature, draft and edit parts of the paper, format references, and suggest a structure and title for the paper. We have reviewed all AI-assisted work: we checked the search code, verified the experimental results, checked all cited works against their sources, and revised all text. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses only publicly available benchmarks and models, each under its respective licence and terms of use; we collect no new data and conduct no study with human subjects. Ego4D-NLQ consists of egocentric recordings of daily life, which may capture bystanders and private settings; we access it under the Ego4D licence agreement (Grauman et al., 2022) and do not redistribute any of its videos. Temporal grounding can be applied to large collections of video, including surveillance footage, where locating specific events may raise privacy concerns. Our method does not add recognition capabilities beyond those of the underlying Video LLMs and embedders, but it makes searching long videos more effective, and we encourage its use in accordance with applicable privacy regulations.

REPRODUCIBILITY STATEMENT

Section 3 describes MaRS in full, and Section 4.2 and Appendix A give the settings used in all experiments, including the search schedule, the candidate composition rule (Appendix A.2), and the fine-tuning hyperparameters (Appendix A.3). All benchmarks, Video LLMs, and embedders we use are publicly available, and MaRS requires no training for the main results. An anonymized version of our code is provided in the supplementary material, and we will release the code, the clip scores, and the fine-tuned checkpoints upon publication.

REFERENCES

Xiang An, Yin Xie, Feilong Tang, Yunyao Yan, Huajie Tan, Didi Zhu, Changrui Chen, Xiuwei Zhao, Bin Qin, Kaicheng Yang, et al. LLaVA-OneVision-2: Towards next-generation perceptual intelligence. *arXiv preprint arXiv:2605.25979*, 2026. URL https://arxiv.org/abs/2605.25979.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, et al. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b. URL https://arxiv.org/abs/2502.13923.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Clark_Molmo2_Open_Weights_and_Data_for_Vision-Language_Models_with_Video_CVPR_2026_paper.html.

Ganqu Cui, Yuchen Zhang, Jiacheng Chen, Lifan Yuan, Zhi Wang, Yuxin Zuo, Haozhan Li, Yuchen Fan, Huayu Chen, Weize Chen, Zhiyuan Liu, Hao Peng, Lei Bai, Wanli Ouyang, Yu Cheng,

Bowen Zhou, and Ning Ding. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025. URL https://arxiv.org/abs/2505.22617.

Aleksandr Gordeev, Vladimir Dokholyan, Irina Tolstykh, and Maksim Kuprashevich. Saliency-guided DETR for moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 907–916, 2026. URL https://openaccess.thecvf.com/content/WACV2026/html/Gordeev_Saliency-Guided_DETR_for_Moment_Retrieval_and_Highlight_Detection_WACV_2026_paper.html.

Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Grauman_Ego4D_Around_the_World_in_3000_Hours_of_Egocentric_Video_CVPR_2022_paper.html.

Tanveer Hannan, Md Mohaiminul Islam, Thomas Seidl, and Gedas Bertasius. RGNet: A unified clip retrieval and grounding network for long videos. In *Proceedings of the European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/03186.pdf.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. ReVisionLLM: Recursive vision-language model for temporal grounding in hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Hannan_ReVisionLLM_Recursive_Vision-Language_Model_for_Temporal_Grounding_in_Hour-Long_Videos_CVPR_2025_paper.html.

Zhijian Hou, Wanjun Zhong, Lei Ji, Difei Gao, Kun Yan, W.K. Chan, Chong-Wah Ngo, Mike Zheng Shou, and Nan Duan. CONE: An efficient COarse-to-fiNE alignment framework for long video temporal grounding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8013–8028, 2023. URL https://aclanthology.org/2023.acl-long.445/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Jingcheng Hu, Yinmin Zhang, Qi Han, Daxin Jiang, Xiangyu Zhang, and Heung-Yeung Shum. Open-Reasoner-Zero: An open source approach to scaling up reinforcement learning on the base model. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/ed873d79e7c268c020c4b4db13a2812a-Abstract-Conference.html.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VTimeLLM_Empower_LLM_to_Grasp_Video_Moments_CVPR_2024_paper.html.

Xiaohu Huang, Hao Zhou, and Kai Han. PruneVid: Visual token pruning for efficient video large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19959–19973, 2025. URL https://aclanthology.org/2025.findings-acl.1024/.

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*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/62e0973455fd26eb03e91d5741a4a3bb-Abstract.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026a. URL https://arxiv.org/abs/2601.04720.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. VideoChat-Flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, 2026b. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b14d7175755b180dc2163e15e3110cb6-Abstract-Conference.html.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, Xinyu Chen, Haoran Chen, Zizheng Huang, Jun Zhang, Kun Ouyang, Lin Sui, Ziang Yan, Yicheng Xu, Chenting Wang, Yinan He, Hongjie Zhang, Yi Wang, Yu Qiao, Yali Wang, Ziwei Liu, Kai Chen, and Limin Wang. VideoChat3: Fully open video MLLM for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026c. URL https://arxiv.org/abs/2607.14935.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. In *Advances in Neural Information Processing Systems*, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/5d2e24df9cfaad3189833b819c40b392-Abstract-Conference.html.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. UniVTG: Towards unified video-language temporal grounding. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Lin_UniVTG_Towards_Unified_Video-Language_Temporal_Grounding_ICCV_2023_paper.html.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. VideoMind: A chain-of-LoRA agent for temporal-grounded video reasoning. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2503.13444.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-like training: A critical perspective. In *Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=5PAF7PAY2Y.

Yongdong Luo, Xiawu Zheng, Guilin Li, Shukang Yin, Haojia Lin, Chaoyou Fu, Jinfa Huang, Jiayi Ji, Fei Chao, Jiebo Luo, and Rongrong Ji. Video-RAG: Visually-aligned retrieval-augmented long video comprehension. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/f593c9c251d4d7cf14d4ab9861dfb7eb-Abstract-Conference.html.

Fangzhou Mu, Sicheng Mo, and Yin Li. SnAG: Scalable and accurate video grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Mu_SnAG_Scalable_and_Accurate_Video_Grounding_CVPR_2024_paper.html.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020. URL https://aclanthology.org/2020.findings-emnlp.63/.

Yulin Pan, Xiangteng He, Biao Gong, Yiliang Lv, Yujun Shen, Yuxin Peng, and Deli Zhao. Scanning only once: An end-to-end framework for fast temporal grounding in long videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Pan_Scanning_Only_Once_An_End-to-end_Framework_for_Fast_Temporal_Grounding_ICCV_2023_paper.html.

Zongshang Pang, Mayu Otani, and Yuta Nakashima. Measure twice, cut once: A semantic-oriented approach to video temporal localization with video LLMs. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2503.09027.

Jongwoo Park, Kanchana Ranasinghe, Kumara Kahatapitiya, Wonjeong Ryu, Donghyun Kim, and Michael S. Ryoo. Too many frames, not all useful: Efficient strategies for long-form video QA. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 3569–3588, 2026. URL https://aclanthology.org/2026.eacl-long.164/.

Jialong Qin, Xin Zou, Di Lu, Yibo Yan, and Xuming Hu. Sharp eyes and memory for VideoLLMs: Information-aware visual token pruning for efficient and reliable VideoLLM reasoning. *arXiv preprint arXiv:2511.08003*, 2025. URL https://arxiv.org/abs/2511.08003.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. Large language models are effective text rankers with pairwise ranking prompting. In *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024. URL https://aclanthology.org/2024.findings-naacl.97/.

Qwen Team. Qwen3.5: Towards native multimodal agents. Qwen Blog, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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. URL https://aclanthology.org/Q13-1003/.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ren_TimeChat_A_Time-sensitive_Multimodal_Large_Language_Model_for_Long_Video_CVPR_2024_paper.html.

Sukmin Seo and Geewook Kim. Natural-language temporal grounding in hour-long videos is a search problem: A benchmark and empirical decomposition. *arXiv preprint arXiv:2606.12300*, 2026. URL https://arxiv.org/abs/2606.12300.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Mattia Soldan, Alejandro Pardo, Juan León Alcázar, Fabian Caba, Chen Zhao, Silvio Giancola, and Bernard Ghanem. MAD: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Soldan_MAD_A_Scalable_Dataset_for_Language_Grounding_in_Videos_From_CVPR_2022_paper.html.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://aclanthology.org/2023.emnlp-main.923/.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. URL https://arxiv.org/abs/2502.14786.

Vidi Team, Celong Liu, Chia-Wen Kuo, Dawei Du, Fan Chen, Guang Chen, Jiamin Yuan, Lingxi Zhang, Lu Guo, Lusha Li, Longyin Wen, Qingyu Chen, et al. Vidi: Large multimodal models for video understanding and editing. *arXiv preprint arXiv:2504.15681*, 2025a. URL https://arxiv.org/abs/2504.15681.

Vidi Team, Celong Liu, Chia-Wen Kuo, Chuang Huang, Dawei Du, Fan Chen, Guang Chen, Haoji Zhang, Haojun Zhao, Lingxi Zhang, et al. Vidi2: Large multimodal models for video understanding and creation. *arXiv preprint arXiv:2511.19529v1*, 2025b. URL https://arxiv.org/abs/2511.19529v1.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. URL https://arxiv.org/abs/2409.12191.

Shaoguang Wang, Weiyu Guo, Ziyang Chen, Yijie Xu, Xuming Hu, and Hui Xiong. Less is more: Token-efficient video-QA via adaptive frame-pruning and semantic graph integration. *arXiv preprint arXiv:2508.03337*, 2025a. URL https://arxiv.org/abs/2508.03337.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. URL https://arxiv.org/abs/2508.18265.

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*, volume 38, 2025c. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7801b29c93b599b8d0c44138596bdeed-Abstract-Conference.html.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. VideoTree: Adaptive tree-based video representation for LLM reasoning on long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025d. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_VideoTree_Adaptive_Tree-based_Video_Representation_for_LLM_Reasoning_on_Long_CVPR_2025_paper.html.

Xiaomi LLM-Core Team. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025. URL https://arxiv.org/abs/2506.03569.

Ziang Yan, Sheng Xia, Jiashuo Yu, Yue Wu, Tianxiang Jiang, Songze Li, Kanghui Tian, Yicheng Xu, Yinan He, Kai Chen, Limin Wang, Yu Qiao, and Yi Wang. InternVideo3: Agentify foundation models with multimodal contextual reasoning. *arXiv preprint arXiv:2606.12195*, 2026. URL https://arxiv.org/abs/2606.12195.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, Lingjun Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a4277440d50f1f15d2cb4c14f7e0c0d2-Abstract-Conference.html.

Huaying Yuan, Jian Ni, Zheng Liu, Yueze Wang, Junjie Zhou, Zhengyang Liang, Bo Zhao, Zhao Cao, Ji-Rong Wen, and Zhicheng Dou. MomentSeeker: A task-oriented benchmark for long-video moment retrieval. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/281e0b9142763f2b6c944fedb8550ba9-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, et al. TimeSuite: Improving MLLMs for long video understanding via grounded tuning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5e8309c9ca683e11672e3dbcd4b87776-Abstract-Conference.html.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, Xinhao Li, Zikang Wang, Changlian Ma, Qingyu Zhang, Zizheng Huang, Kun Ouyang, Tianxiang Jiang, Ziang Yan, Yi Wang, Hongjie Zhang, Yali Wang, and Limin Wang. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2601.23224.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Annual Meeting of the Association for Computational Linguistics*, 2020a. URL https://aclanthology.org/2020.acl-main.585/.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10419–10429, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_TimeLens_Rethinking_Video_Temporal_Grounding_with_Multimodal_LLMs_CVPR_2026_paper.html.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D temporal adjacent networks for moment localization with natural language. In *AAAI Conference on Artificial Intelligence*, 2020b. URL https://ojs.aaai.org/index.php/AAAI/article/view/6984.

Pengcheng Zhao, Zhixian He, Fuwei Zhang, Shujin Lin, and Fan Zhou. LD-DETR: Loop decoder DEtection TRansformer for video moment retrieval and highlight detection. *arXiv preprint arXiv:2501.10787*, 2025. URL https://arxiv.org/abs/2501.10787.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. TimeLens2: Generalist video temporal grounding with multimodal LLMs. *arXiv preprint arXiv:2607.17423*, 2026. URL https://arxiv.org/abs/2607.17423.

Shengyao Zhuang, Honglei Zhuang, Bevan Koopman, and Guido Zuccon. A setwise approach for effective and highly efficient zero-shot ranking with large language models. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024. URL https://arxiv.org/abs/2310.09497.

A Implementation Details

A.1 Inputs and Search

Frames are sampled at 2 fps and interleaved with their timestamps. Clips of 4 s are scored as in Eq. (1). FOCUS draws two candidates in addition to the model's first answer and composes them into the final montage. EXPLORE draws eight additional candidates in its first round and composes them into a 128 s montage; its second round draws two and composes them into the final montage. The final montage lasts 64 s on Ego4D-NLQ and MomentSeeker and 128 s on TACoS and the VUE benchmarks. In masked resampling, the first answer is decoded greedily and the additional candidates are sampled at temperature 1. Candidates of 64 s or longer are discarded before composing, and if the first answer of the first round is 64 s or longer, the search stops and returns it as the final prediction.

A.2 Composing Candidates

A predicted window may extend over parts of the video that the montage does not show. We keep only its visible parts, each of which becomes a separate subwindow. To build the next montage, we pad the windows equally on both sides until together they fill the duration set for the next round. The padding restores context around each candidate, so that a later round can recover boundaries that the montage has cut.

A.3 Fine-Tuning

Supervised fine-tuning uses 36K samples: 32K from TimeLens-100K (Zhang et al., 2026), which consists mostly of short videos, and 4.5K montages of 128 s built from the Ego4D-NLQ training split. It runs for one epoch with a learning rate of 2×10^{-5} under a cosine schedule. GRPO then trains separate models for Ego4D-NLQ and TACoS, on montages built from 1.5K and 5.6K training queries, respectively; each montage consists of the top-32 clips (4 s each) retrieved by Qwen3-VL-Embedding (Li et al., 2026a). We do not train on the longer selections of EXPLORE, since videos in Ego4D-NLQ and TACoS rarely exceed 512 s. GRPO uses a group size of 8, a constant learning rate of 10^{-5}, a sampling temperature of 1, and one epoch. Both stages train LoRA adapters (Hu et al., 2022) on all attention and feed-forward projections of the language model, with rank 64 and $\alpha = 128$ for supervised fine-tuning and rank 32 and $\alpha = 64$ for GRPO, and a dropout of 0.05; all other parameters, including the vision encoder, remain frozen.

B Additional Results

B.1 Full Metrics

Table 7 reports all metrics for the models that we evaluated ourselves. All runs use 28,672 visual tokens per pass. One pass grounds either the whole video, sampled at 2 fps, or the montage of the top-32 or top-128 clips, which are the inputs of FOCUS and EXPLORE, respectively. Of the vision-only queries of VUE-TR and VUE-TR-V2, we use the 512 and 691 whose videos could be downloaded; Zhu et al. (2026) reports 525 and 715 queries, so the 128K rows in Table 2 are computed on slightly different subsets.

Table 7: All metrics (%) at 28,672 visual tokens per pass. Gray: TimeLens2 is fine-tuned on Ego4D-NLQ.

	Ego4D-NLQ				MomentSeeker				VUE-TR				VUE-TR-V2			
Method	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU	R1@0.3	R1@0.5	R1@0.7	mIoU
VideoChat3-4B																
One pass, whole video	20.0	13.9	7.6	14.5	37.9	24.1	10.7	25.4	52.0	42.0	31.4	40.8	47.3	38.5	25.3	36.6
Top-32, one pass	22.5	15.6	8.4	16.2	40.4	26.5	11.6	27.0	61.3	47.9	35.0	46.9	57.5	45.7	30.1	43.4
Top-128, one pass	20.4	14.1	7.9	14.7	38.6	25.1	11.1	26.0	59.2	48.6	37.1	47.0	59.5	49.5	35.3	46.5
MaRS-FOCUS	24.6	17.0	9.4	17.6	39.9	25.3	11.5	26.9	62.7	49.8	37.3	48.8	61.4	49.8	34.0	46.8
MaRS-EXPLORE	24.9	17.0	9.0	17.5	40.1	26.5	13.0	27.3	62.9	50.8	38.5	49.4	63.1	51.8	37.3	48.6
TimeLens-8B																
One pass, whole video	21.2	14.4	8.1	15.2	40.1	24.2	12.3	26.2	51.4	42.6	32.4	42.0	46.6	38.4	27.2	36.4
Top-32, one pass	26.1	17.4	9.3	18.6	38.8	25.2	12.0	26.5	63.1	53.1	39.3	50.4	61.9	51.1	36.5	48.2
Top-128, one pass	21.4	14.6	8.2	15.4	40.4	25.5	12.9	26.9	60.2	48.8	38.1	48.2	60.1	48.6	33.3	45.8
MaRS-FOCUS	28.3	19.1	10.8	20.1	40.4	25.9	11.9	27.2	64.8	54.3	41.2	51.9	66.7	55.6	40.1	51.8
MaRS-EXPLORE	27.8	18.8	10.7	19.8	41.1	26.0	12.2	27.5	65.0	52.3	40.4	51.5	65.4	54.7	38.8	50.6
TimeLens2-8B																
One pass, whole video	23.5	16.3	8.8	16.8	38.2	25.9	13.2	27.1	58.4	46.3	35.5	46.5	53.1	43.4	30.5	41.2
Top-32, one pass	28.0	19.1	9.9	19.8	40.4	26.5	11.8	27.6	61.9	52.0	41.2	51.0	65.8	55.3	41.1	51.8
Top-128, one pass	24.3	16.7	9.2	17.3	39.0	26.5	13.2	27.8	63.3	51.4	41.8	51.6	65.6	54.3	39.8	51.4
MaRS-FOCUS	30.4	20.9	11.2	21.5	41.3	25.9	12.6	28.5	64.6	53.7	42.8	52.7	69.2	58.2	43.1	54.0
MaRS-EXPLORE	30.7	21.0	11.7	21.6	40.1	26.5	13.5	28.7	63.9	53.7	43.0	52.8	68.7	58.6	43.6	54.7