VideoSlate: STATEFUL COMMIT CONTROL FOR TEMPORALLY GROUNDED AGENTIC VIDEO REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic video reasoning systems gather and assess video evidence across stages, selecting relevant spans as visual evidence for answering questions. New candidates and observations prompt reassessment of the current span, but revisions can correct localization errors or discard useful evidence already present in the adopted span. Systems therefore need to connect evidence organization across stages with temporal state control, allowing accumulated information to jointly inform revision assessment while explicit adoption decisions govern how assessment outcomes affect downstream visual evidence. We introduce VideoSlate, a stateful commit-control framework built around shared evidence. A typed evidence blackboard links candidate spans, provenance, construction, and verifier observations, enabling an incumbent-relative certifier to assess revision support from the candidate, current span, and related proposals. Based on certification, the controller authorizes a revision or retains the current span, separating evidence accumulation from state updates and recording each decision and its evidence on the blackboard. Extensive evaluations across video grounding and grounded question-answering benchmarks show that VideoSlate can improve temporal localization through selective revision.

1 INTRODUCTION

Agentic video reasoning methods use planning, tool invocation, or role coordination to gather video evidence over multiple steps and form judgments from it (Gao et al., 2023; Fan et al., 2024; Liu et al., 2026). To use this evidence, the system must account for temporal relations: visually similar content may occur at different times, but only some occurrences satisfy the query's temporal conditions (Wang et al., 2024b). Subsequent judgments therefore need to relate new observations to the evidence and interpretations already available. If these connections no longer inform later judgments, recorded observations may remain accessible without constraining the interpretation of new evidence. Agentic video reasoning thus needs to organize and maintain evidence so that accumulated temporal context continues to guide the interpretation of new observations and the formation of subsequent judgments.

Memory and retrieval help maintain these connections. VideoAgent uses structured event and object memories with tool queries, while LangRepo organizes descriptions at multiple temporal scales (Fan et al., 2024; Kahatapitiya et al., 2024). DrVideo iteratively retrieves and augments a video document, and more recent agents maintain hierarchical or multimodal memories for adaptive tool use and retrieval (Ma et al., 2025; Yin et al., 2026; Yeo et al., 2026). These systems make prior information available for later reasoning, often through active reuse. However, storing or retrieving evidence does not by itself determine how to compare observations that support different temporal interpretations. The remaining question is what each observation supports and how conflicting support should affect the judgment carried forward.

When later reasoning relies on the currently adopted interval, resolving such disagreements requires determining whether the available evidence is sufficient to change the incumbent and the visual content it supplies downstream. A candidate's match to the query characterizes support for that candidate in isolation; deciding whether to revise the current temporal judgment also requires comparing that support with the evidence for the existing span. VideoMind provides one concrete instance of the former: its verifier scores each candidate independently according to whether it accurately

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.

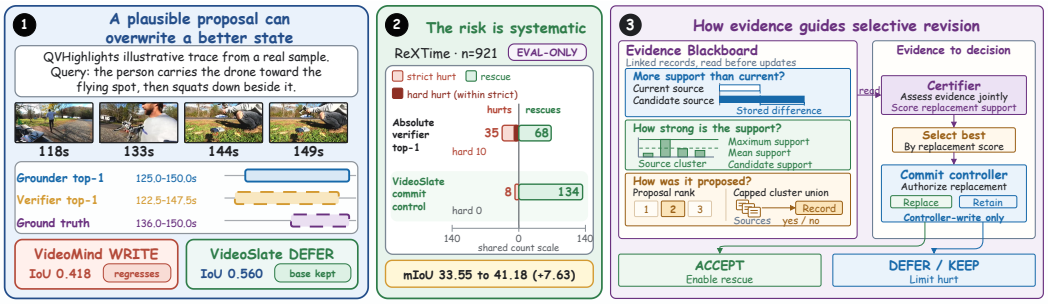


Figure 1: (a) A challenger selected by VideoMind's verifier can overwrite a stronger incumbent; deferral preserves it. (b) Selective commits retain beneficial revisions while suppressing harmful writes. (c) The certifier jointly assesses relative support, source support, and proposal context from the evidence blackboard. The controller uses the best candidate's replacement score to authorize replacement or retain the current result.

covers the queried event and selects the highest-scoring proposal (Liu et al., 2026). Figure 1(a) illustrates the resulting risk: the verifier prefers a shifted candidate that has less overlap with the reference span than the initial localization. This comparison is further complicated by candidate construction: combined or modified intervals can inherit support assessed for source intervals with different temporal boundaries.

To make information from different reasoning stages jointly inform temporal judgments, we connect candidate generation and evaluation through shared evidence and use explicit state control to coordinate how evaluation results affect subsequent reasoning. Shared evidence preserves links among candidates, observations, and their sources, allowing revision assessment to combine support for the candidate, related proposals, and the incumbent. This evidence jointly informs whether to revise the adopted span, while the resulting recommendation requires a separate commit decision before changing the temporal state used downstream. The aim is to decouple evidence accumulation from state updates, with the former expanding the basis for comparison and the latter controlling which revisions are adopted. Through this division of responsibility, candidate generation, evidence assessment, and result adoption are connected through shared evidence. Information from each stage contributes to temporal judgments, while explicit state-update rules control its effect on the downstream visual context.

Guided by these principles, we introduce VideoSlate, a stateful commit-control framework. The framework incorporates a typed evidence blackboard that stores the incumbent alongside candidate intervals, linking each candidate to its provenance, construction, and verifier observations. A learned incumbent-relative certifier scores each proposed replacement based on features from these records, combining verifier support for the candidate and incumbent, aggregated support from related proposals, and candidate provenance. A dedicated controller selects the candidate with the highest certifier score and authorizes replacement only when that score meets the acceptance threshold; otherwise, it retains the incumbent. The controller records each decision and its supporting evidence on the blackboard, allowing the returned interval to be traced to the evidence and commitment rule that produced it. Candidate records remain separate from the committed span: tools append candidates and observations, while only the controller can change the interval used downstream.

The contributions are:

- We introduce VideoSlate, a framework that models temporal refinement as state updates under asymmetric replacement risk, explicitly decoupling proposal, certification, and commitment to improve localization through beneficial revisions and provide more precise visual evidence downstream.
- We design a typed, append-only evidence blackboard that ensures traceable replacement decisions by explicitly linking proposals, provenance, verifier observations, and traces while keeping candidate records separate from the committed span.
- We propose an incumbent-relative certifier that learns replacement value directly from pairs, paired with a controller that dynamically commits revisions based on runtime evidence or defaults to the incumbent when support is insufficient.

2 Related Work

Temporally Grounded Video Understanding. Temporal grounding localizes query-relevant events; grounded VideoQA also requires evidence-supported answers. TALL/CTRL refines windows (Gao et al., 2017); ActivityNet Captions supplies localized descriptions (Krishna et al., 2017). QVHighlights and UMT address moment retrieval and highlight detection (Lei et al., 2021; Liu et al., 2022). Unified supervision, image–language tuning, and diffusion further extend grounding (Lin et al., 2023b; Liu et al., 2024c; Li et al., 2023). Timestamp-aware models support localization (Huang et al., 2024a; Ren et al., 2024; Huang et al., 2024b; Chen et al., 2024c), with extensions using generation, experts, reinforcement learning, and curated data (Li et al., 2025; Yang et al., 2025; Wang et al., 2025; Zhang et al., 2026). NExT-GQA, ReXTime, CG-Bench, and ETBench evaluate grounded video reasoning (Xiao et al., 2024; Chen et al., 2024a; 2025; Liu et al., 2024d). We study whether generated candidates should replace the retained span.

Agentic Video Systems and State Control. Agentic video systems compose tools (Surís et al., 2023; Yang et al., 2023; Gao et al., 2023) and organize evidence through textual repositories or memory (Zhang et al., 2024a; Kahatapitiya et al., 2024; Wang et al., 2024b; Fan et al., 2024). SeViLA couples localization and answering (Yu et al., 2023). VideoMind passes the verifier's top proposal to the answerer (Liu et al., 2026). REVISOR uses initial reasoning to identify a video interval for denser resampling, then combines the new visual evidence with the initial reasoning trace to refine its answer (Li et al., 2026). Beyond video, bMAS/LbMAS coordinates agents through shared blackboard messages; its control unit uses the query, accumulated messages, and agent capabilities to select participants for each round (Han & Zhang, 2025). VideoSlate certifies revisions against the incumbent and separately authorizes changes to the span used downstream.

Candidate Verification and Selective Prediction. Verification assesses quality; selective prediction governs reliance on outputs. Verifiers, process supervision, and search evaluate solutions (Cobbe et al., 2021; Lightman et al., 2024; Wang et al., 2024c); pairwise learning-to-rank estimates preferences (Burges et al., 2005). Selective models trade coverage for risk (Geifman & El-Yaniv, 2017; 2019); deferral routes cases to experts (Madras et al., 2018; Mozannar & Sontag, 2020). VideoSlate applies selective prediction to revision, retaining the incumbent when evidence is insufficient and logging the decision.

3 VideoSlate: Selective Temporal State Revision

3.1 Problem Formulation and Overview

Given a video V of duration T and a query q, temporal grounding provides a span whose visual content can support downstream reasoning. Once this span is in use, a proposed revision must be assessed relative to what it would replace. We denote the currently retained span by $b_t = [s_t, e_t]$, the incumbent, and describe one revision step indexed by t. A challenger is a candidate proposed to replace this span. The task is to determine whether the available evidence supports a replacement or retaining the incumbent. The resulting span b_{t+1} supplies the visual evidence used downstream.

Making this decision requires keeping the incumbent available as a comparison reference while gathering evidence for alternative spans. We organize this information in a typed evidence blackboard $\mathcal{B}_t$, which stores the incumbent alongside candidate records, verifier observations, and decision traces. Candidate identifiers link each proposed interval to its source proposals and observations. Tools append proposals and observations while the incumbent remains protected; only the controller can replace it. This separation lets evidence accumulate without changing the committed span. Figure 2 organizes this step into Ground & Propose, Record & Certify, and Decide & Commit.

3.2 Ground & Propose: Constructing Revision Candidates

A grounder's proposed interval may omit part of the queried event or include unrelated context. Revision therefore needs alternatives that can reduce or extend its temporal coverage. Given V and q, the frozen VideoMind grounder (Liu et al., 2026) produces ranked temporal spans. The primary configuration uses the first-ranked output as the initial incumbent span b_0 and retains the others as raw

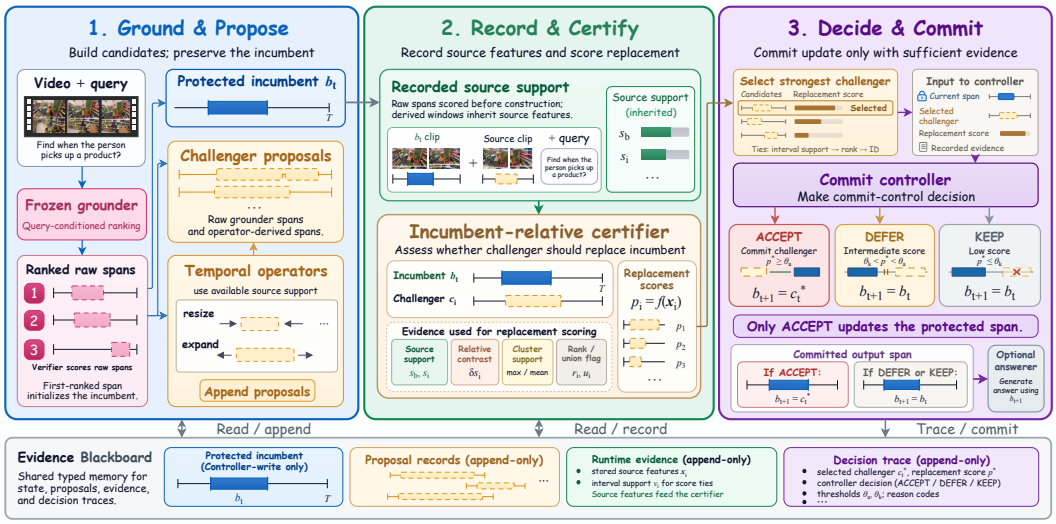


Figure 2: VideoSlate's selective temporal state revision. Ground & Propose obtains raw spans and their verifier support before constructing additional candidates. Record & Certify links candidates to their sources and scores replacement using stored source features. Decide & Commit selects the strongest challenger, using separate interval observations to resolve score ties when available. Only acceptance changes the committed span; both retention actions preserve it with distinct traces. The blackboard records proposals, evidence, and decisions throughout the step. Score bars are schematic.

proposals. The verifier first scores the incumbent and raw proposals against q. These raw proposals and their source-support scores provide the inputs for constructing alternative spans.

Using these inputs, temporal operators construct candidates that preserve, narrow, or broaden the proposed coverage. Direct candidates retain the raw proposal boundaries. Nonempty intersections between the incumbent and a raw proposal restrict coverage to their shared region, while duration edits expand a proposal to include additional context. Union operators combine the incumbent with a raw proposal or combine proposals within a verifier-supported cluster. For cluster construction, the raw proposal with the highest source support serves as the anchor; sufficiently supported proposals join when they overlap the anchor or lie near its center.

These constructions can yield repeated intervals, which we consolidate in the evaluation collection while retaining their source information. Each derived candidate inherits source-support features and retains links to the proposals used to construct it. Such inherited support does not directly verify the boundaries of the resulting interval. Merging duplicate intervals combines these links and takes the maximum of each stored numeric feature, allowing one candidate to carry evidence from several constructions. The resulting raw and derived candidates form the challenger set $\mathcal{C}_t$ for certification; the incumbent is excluded and remains available as the fallback output.

3.3 Record & Certify: Incumbent-Relative Revision Evidence

Once the candidate set is constructed, certification assesses whether each challenger should replace the incumbent. A candidate's absolute support does not establish improvement over the incumbent. We therefore compare the two spans using the stored source features linked on the blackboard, whose support scores originate from the frozen VideoMind verifier (Liu et al., 2026).

To represent replacement evidence, we combine incumbent support as the comparison baseline with candidate and cluster support for the alternative. The stored support difference adds an explicit relative comparison, while rank and a cluster-union indicator encode proposal order and construction type. For challenger c_i, these features give the seven-dimensional input

$$\mathbf{x}_i = [s_b,\ r_i,\ s_i,\ s_i^{\max},\ \bar{s}_i,\ u_i,\ \delta s_i]^\top. \tag{1}$$

In Equation 1, s_b and s_i are the stored incumbent and candidate support scores. The maximum $s_i^{\max}$ and mean $\bar{s}_i$ summarize the source cluster, which contains a single member for an ordinary

proposal. The coordinate r_i stores proposal rank, and u_i indicates whether a capped verifier-cluster union contributed to the record. The stored difference δs_i is computed as candidate minus incumbent support before feature inheritance and record merging.

A logistic certifier maps this vector to a replacement score p_i by applying a sigmoid to a learned linear combination with an intercept. Training preferences are defined by overlap improvement relative to the incumbent, so larger scores indicate stronger learned support for replacement. The score encodes this replacement preference; it does not estimate the magnitude of an IoU gain.

To distinguish challengers with equal certifier scores, we retain a separate interval-support value v_i. When available, this value is the frozen verifier's yes/no observation of the candidate's own window. Such observations cover a subset of candidate intervals; for a missing observation, v_i falls back to the maximum available stored verifier-support or source-cluster value for that candidate. This observation or fallback provides a tie-breaking input to the controller without changing the source features used to compute p_i.

3.4 Decide & Commit: Selective State Transition

The certifier scores rank challengers by replacement support, but even the highest-scoring candidate may lack enough support to replace the incumbent. The controller therefore first identifies the preferred challenger, then checks whether its score warrants a state change. For a nonempty, scorable challenger set, it selects the preferred challenger using the deterministic lexicographic order

$$i^* = \underset{i:\, c_i \in \mathcal{C}_t}{\arg\max} \big(p_i,\ v_i,\ -r_i,\ \mathrm{id}(c_i)\big). \tag{2}$$

Here i^* indexes the selected challenger. The order first maximizes p_i, using higher interval support v_i only to break score ties. Remaining ties favor a lower stored rank and, finally, the lexicographically larger stable identifier. Selection uses no reference-dependent information.

The selected score then determines whether the incumbent changes. With frozen thresholds $\theta_{\mathrm{a}} > \theta_{\mathrm{k}}$, Accept applies when $p_{i^*} \geq \theta_{\mathrm{a}}$ and sets $b_{t+1} = c_{i^*}$. Otherwise, the controller retains the incumbent, returning $b_{t+1} = b_t$. It records Defer for a score between the thresholds and Keep for a score at or below θ_{k}, distinguishing intermediate from low replacement support while returning the same span. An empty or unscorable candidate set also retains the incumbent and records Keep with a fallback reason.

For each such transition, the controller appends a decision trace to the blackboard, linking the returned span to the evidence and rule that produced it. Each trace stores the selected challenger, when available, its score, feature snapshot, thresholds, action, and any fallback reason. For a retained span, these records show whether the controller applied the score thresholds or used a fallback.

Answer generation. The resulting span b_{t+1} is the localization output of this revision step. For grounded question answering, this span is cropped and passed with q to an answerer outside the replacement policy. Both an accepted challenger and a retained incumbent supply this same downstream interface.

3.5 Certifier Training and Calibration

We learn replacement scores from each incumbent–challenger pair (b_i, c_i) using its overlap change against reference span g_i:

$$\Delta_i = \mathrm{IoU}(c_i, g_i) - \mathrm{IoU}(b_i, g_i). \tag{3}$$

Using Equation 3, pairs with $\Delta_i \geq 0.05$ receive $y_i = 1$, and those with $\Delta_i \leq -0.05$ receive $y_i = 0$; ties are excluded from fitting but retained during calibration.

With the grounder and verifier frozen, we fit certifier weights $\mathbf{w}$ and intercept β on N labeled pairs:

$$\mathcal{L}(\mathbf{w}, \beta) = \sum_{i=1}^{N} \alpha_{y_i}\, \ell_{\mathrm{BCE}}(y_i, p_i) + \frac{\|\mathbf{w}\|_2^2}{2C}. \tag{4}$$

In Equation 4, ℓ_{BCE} is binary cross-entropy, $\alpha_y = N/(2N_y)$ balances classes with label counts N_y, and $C = 1$ controls L2 regularization.

Five video-grouped folds provide out-of-fold (OOF) scores. Since each query returns one span, calibration evaluates outputs after candidate selection. Eligible thresholds improve R@0.5 under harm and retention constraints; selection prioritizes localization gain. Primary OOF score ties use training-candidate IoU and operator-row identifiers; evaluation follows the annotation-free order in Equation 2. We select $\theta_{\mathrm{a}} = 0.55$ and $\theta_{\mathrm{k}} = 0.20$; the latter follows the grid tie convention. The refitted certifier and thresholds are frozen for evaluation.

4 EXPERIMENTS

We evaluate grounded VideoQA and temporal grounding. Benchmark comparisons assess localization and joint answer–localization quality; controlled comparisons examine how evidence and commitment rules affect revisions.

4.1 EXPERIMENTAL SETTINGS

Setup and scope. We evaluate REXTIME validation (921 examples), CG-Bench Mini (3,000 questions), NExT-GQA official test (5,553 queries), Charades-STA official test (3,720 queries), and ActivityNet-Captions (17,031 queries). The first three evaluate grounded QA; the last two isolate localization. The primary REXTIME configuration uses the frozen VideoMind-7B grounder/verifier. Its offline evaluation replays one state transition per query using frozen candidates and observations. Other benchmarks use dataset-specific policies. In VideoSlate rows, Size denotes the 7B localization backbone; Qwen3-VL-8B (Bai et al., 2025) answers from the selected temporal context. Published baselines retain their own answering configurations. CG-Bench baseline results follow Table 2 of Liu et al. (2026).

Metrics. We report R@0.3, R@0.5, mIoU, and R@0.7 where available. Acc measures answer accuracy; Tables 1–3 define the benchmark-specific joint metrics. At IoU 0.5, a rescue changes a miss into a hit and a strict hurt changes a hit into a miss; a hard hurt replaces an incumbent at or above 0.5 IoU with an output below 0.3.

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Grounded VideoQA. VideoSlate leads the reported metrics on REXTIME and CG-Bench (Tables 1 and 2). On REXTIME, joint accuracy improves markedly over VideoMind despite similar answer accuracy. Expanded answering windows may cover similar answer-relevant cues even when predicted boundaries differ, while the joint gains highlight VideoSlate's improved temporal grounding of correct responses. CG-Bench provides complementary evidence: among localization hits pooled across thresholds, VideoSlate also answers a larger share correctly than VideoMind.

On NExT-GQA (Table 3), VideoSlate achieves the highest reported Acc@GQA, mIoU, and mIoP. Relative to VideoMind, its localization advantage is clearer at the stricter IoU cutoff: coarse recall is slightly lower and mIoU is nearly tied. At the stricter IoP cutoff, recall is close to but slightly below LLoVi, while higher mIoP and Acc@GQA indicate better average temporal overlap and joint QA performance at a comparable localization hit rate.

Video Temporal Grounding. Using candidates from the frozen VideoMind grounder, VideoSlate improves all four metrics over VideoMind on Charades-STA and ActivityNet-Captions (Table 4). It leads the reported ActivityNet-Captions metrics and achieves higher recall than UniTime at medium and high IoU thresholds on Charades-STA, although UniTime retains higher mIoU. UniTime uses adaptive frame scaling for boundary prediction (Li et al., 2025), while residual candidate-selection errors may limit further mIoU gains in our pipeline.

4.3 CONTROLLED COMPARISONS AND ABLATIONS

Framework ablations and controls. Table 5 presents ablations and controls across VideoSlate's three runtime stages, examining how candidate construction, certification features, and commit rules affect localization. Further candidate controls show that direct proposals retain the gains, whereas

Table 1: Grounded VideoQA on REXTIME (Chen et al., 2024a). Acc@IoU requires a correct answer and temporal IoU ≥ 0.5. Bold marks the best results; underlining marks the second best.

Method	Size	R@0.3	R@0.5	mIoU	Acc	Acc@IoU
VTimeLLM (Huang et al., 2024a)	7B	28.84	17.41	20.14	36.16	–
TimeChat (Ren et al., 2024)	7B	14.42	7.61	11.65	40.04	–
LITA (Huang et al., 2024b)	13B	29.49	16.29	21.49	34.44	–
VTimeLLM (FT) (Huang et al., 2024a)	7B	<u>43.69</u>	26.13	29.92	57.58	17.13
TimeChat (FT) (Ren et al., 2024)	7B	40.13	21.42	26.29	49.46	10.92
TOGA (Gupta et al., 2025)	7B	29.91	19.79	25.53	–	–
SynRL (Jiang et al., 2026)	4B	38.80	26.20	28.90	–	–
Divid (Tang et al., 2026)	7B	42.56	<u>31.05</u>	<u>35.78</u>	–	<u>22.26</u>
VideoMind (Liu et al., 2026)	7B	38.22	25.52	27.61	<u>74.59</u>	20.20
VideoSlate (Ours)	7B	**58.85**	**38.98**	**41.18**	**75.46**	**31.60**

Table 2: Grounded VideoQA on the 3,000-question CG-Bench Mini evaluation (Chen et al., 2025). *long-acc.* is answer accuracy; *rec.@IoU* and *acc.@IoU* average localization recall and answer-correct hits over IoU <u>thresholds 0.1–0.5</u>.

Method	Size	long-acc.	mIoU	rec.@IoU	acc.@IoU
Closed-source systems					
GPT-4o (OpenAI, 2024)	–	<u>45.2</u>	5.62	8.30	4.38
Gemini-1.5-Pro (Reid et al., 2024)	–	37.2	3.95	5.81	2.53
Claude-3.5-Sonnet (Anthropic, 2024)	–	40.5	3.99	5.67	2.79
Open-source systems					
Video-LLaVA (Lin et al., 2023a)	7B	16.2	1.13	1.96	0.59
VideoLLaMA (Zhang et al., 2023)	7B	18.4	1.21	1.87	0.84
VideoChat2 (Li et al., 2024b)	7B	19.3	1.28	1.98	0.94
ST-LLM (Liu et al., 2024b)	7B	23.8	2.23	2.86	1.13
ShareGPT4Video (Chen et al., 2024b)	16B	26.7	1.85	2.65	1.01
Chat-UniVi-v1.5 (Jin et al., 2023)	13B	25.9	2.07	2.53	1.21
VILA (Lin et al., 2024)	8B	28.7	1.56	2.89	1.35
LongVA (Zhang et al., 2024b)	7B	28.7	2.94	3.86	1.78
LLaVA-OneVision (Li et al., 2024a)	7B	31.1	1.63	1.78	1.08
Video-CCAM (Fei et al., 2024)	14B	29.7	2.63	3.48	1.83
Kangaroo (Liu et al., 2024a)	8B	30.2	2.56	2.81	1.94
VITA (Fu et al., 2024)	8×7B	33.3	3.06	3.53	2.06
Qwen2-VL (Wang et al., 2024a)	72B	41.3	3.58	5.32	3.31
InternVL2 (OpenGVLab, 2024)	78B	42.2	3.91	5.05	2.64
VideoMind (Liu et al., 2026)	7B	38.4	<u>7.10</u>	<u>9.93</u>	<u>4.67</u>
VideoSlate (Ours)	7B	**46.6**	**7.98**	**12.51**	**7.45**

including merge candidates adds few rescues but more strict hurts. Useful revisions therefore already exist among the grounder's outputs, and expanding the set does not ensure that they will be selected.

The feature controls separate candidate availability from the evidence needed to adopt a revision. Removing cluster-support terms or the union indicator preserves candidates but lowers scores; with remaining coefficients and the threshold fixed, useful alternatives go unselected. Thus stored evidence affects whether an existing revision opportunity becomes an actual update.

The decision controls further show that preserving correct incumbents can coexist with recovering more errors. The full policy improves both outcomes over unconditional verifier selection, while the tested margin gate falls short in localization. These findings support assessing replacement value from candidate and incumbent evidence before changing the committed span.

Allocation under a fixed commit budget. Table 6 examines how scoring rules prioritize updates when commit counts are matched. Each query's incumbent and challenger are fixed; only which queries commit changes. Source-support ranking trails random allocation, indicating that candidate support alone is insufficient to prioritize useful replacements. A well-supported candidate may add little when the incumbent is already adequate. Source-margin ranking improves localization, while the certifier gives the best results at 50%. These results show that jointly using candidate, incumbent, and construction evidence enables more effective update allocation within the shared pool.

Table 3: Localization and grounded-QA performance on the NExT-GQA official test (Xiao et al., 2024). IoP divides temporal intersection by predicted duration; Acc@GQA requires a correct answer and IoP ≥ 0.5.

Method	Size	IoU			IoP			Acc@GQA
		R@0.3	R@0.5	mIoU	R@0.3	R@0.5	mIoP	
FrozenBiLM NG+ (Yang et al., 2022)	890M	13.50	6.10	9.60	28.50	23.70	24.20	17.50
SeViLA (Yu et al., 2023)	4B	29.20	13.80	21.70	34.70	22.90	29.50	16.60
LangRepo (Kahatapitiya et al., 2024)	8×7B	–	12.20	18.50	–	28.70	31.30	17.10
VideoStreaming (Qian et al., 2024b)	8.3B	–	13.30	19.30	–	31.00	32.20	17.80
LLoVi (Zhang et al., 2024a)	1.8T	–	15.30	20.00	–	**36.90**	37.30	24.30
HawkEye (Wang et al., 2024d)	7B	37.00	19.50	25.70	–	–	–	–
VideoChat-TPO (Yan et al., 2024)	7B	41.20	23.40	27.70	47.50	32.80	35.60	25.50
VideoMind (Liu et al., 2026)	7B	**50.20**	25.80	31.40	**56.00**	35.30	39.00	28.20
VideoSlate (Ours)	7B	49.47	**26.54**	**31.48**	55.70	36.81	**39.73**	**30.43**

Table 4: Video temporal grounding on the Charades-STA official test and ActivityNet-Captions.

Method	Size	Charades-STA				ActivityNet-Captions			
		R@0.3	R@0.5	R@0.7	mIoU	R@0.3	R@0.5	R@0.7	mIoU
VTimeLLM (Huang et al., 2024a)	7B	51.00	27.50	11.40	31.20	44.00	27.80	14.30	30.40
TimeChat (Ren et al., 2024)	7B	51.50	32.20	13.40	–	–	–	–	–
Momentor (Qian et al., 2024a)	7B	42.60	26.60	11.60	28.50	42.90	23.00	12.40	29.30
ChatVTG (Qu et al., 2024)	7B	52.70	33.00	15.90	34.90	40.70	22.50	9.40	27.20
E.T. Chat (Liu et al., 2024d)	4B	65.70	45.90	20.00	42.30	24.10	12.80	6.10	18.90
UniTime (Li et al., 2025)	7B	–	59.10	31.90	**52.20**	–	22.80	14.10	27.30
ED-VTG (Pramanick et al., 2025)	–	–	–	–	–	52.10	33.10	16.00	35.20
VideoChat-R1.5 (Yan et al., 2025)	7B	–	–	–	–	52.40	32.30	16.80	35.50
VideoMind (Liu et al., 2026)	7B	73.50	59.10	31.20	50.20	48.40	30.30	15.70	33.30
VideoSlate (Ours)	7B	**74.81**	**61.16**	**33.20**	51.19	**55.31**	**35.95**	**18.00**	**37.68**

Table 5: Module-aligned controls and diagnostics on the REXTIME replay. Rows follow the three runtime stages in Figure 2. Feature and candidate interventions retain the remaining coefficients and thresholds.

Stage and intervention	R@0.3	R@0.5	mIoU	Acc	Acc@IoU
Ground & Propose					
Cluster-union proposals only	41.04	25.41	28.32	74.27	19.98
Record & Certify					
w/o cluster-support terms	42.13	26.60	29.46	74.48	21.17
w/o union-indicator term	41.37	25.84	28.61	74.48	20.63
Decide & Commit					
KEEP only (protected incumbent)	40.28	25.30	27.98	74.59	20.20
Unconditional verifier top-1	46.91	28.88	33.55	73.62	22.37
Fixed verifier-margin gate	42.78	27.90	31.36	74.05	22.26
VideoSlate (full policy)	**58.85**	**38.98**	**41.18**	**75.46**	**31.60**

Candidate selection and commitment. Table 7 compares candidate selection and commitment on shared frozen Charades-STA candidates. Proposal ranking replaces VideoSlate's candidate selection with the highest-probability candidate while retaining its commit rule. It outperforms Proposal-prob. gate, which selects the same candidate but uses a probability threshold. VideoSlate further improves all four metrics over Proposal ranking. Since the candidate pool is shared, these gains reflect better use of available alternatives. Higher recall across the reported thresholds and higher mean overlap indicate gains beyond a single overlap cutoff. The commit rule evaluates only the selected challenger, so effective revision depends on both choosing an alternative and deciding whether to adopt it.

4.4 VISUALIZATION

Figure 3 illustrates why a relevant candidate need not improve the current span. In the rescue case, the incumbent already covers the activity, but a tighter challenger removes excess context and improves boundary alignment. Conversely, the deferred merge inherits strong source support while

Table 6: Commit allocation on REXTIME with a 50% budget and one shared challenger per query. Unselected queries retain their incumbents. Ties use a fixed order; random allocation uses one run. Localization metrics (%) cover all queries.

Allocation rule	R@0.3	R@0.5	mIoU
Random allocation	48.97	31.49	34.22
Source-support ranking	48.21	30.40	33.34
Source-margin ranking	50.81	32.14	35.44
Certifier-score ranking	**51.68**	**34.20**	**36.37**

Table 7: Post-hoc Charades-STA controls on shared frozen candidates. The proposal controls share probability ranking but use a probability threshold or VideoSlate's gate, respectively. Commit counts differ; metrics in %.

Commit rule	R@0.3	R@0.5	R@0.7	mIoU
KEEP only	73.71	58.76	29.92	49.72
Proposal-prob. gate	73.82	58.92	30.16	49.83
Proposal ranking	74.35	60.27	32.42	50.78
VideoSlate	**74.81**	**61.16**	**33.20**	**51.19**

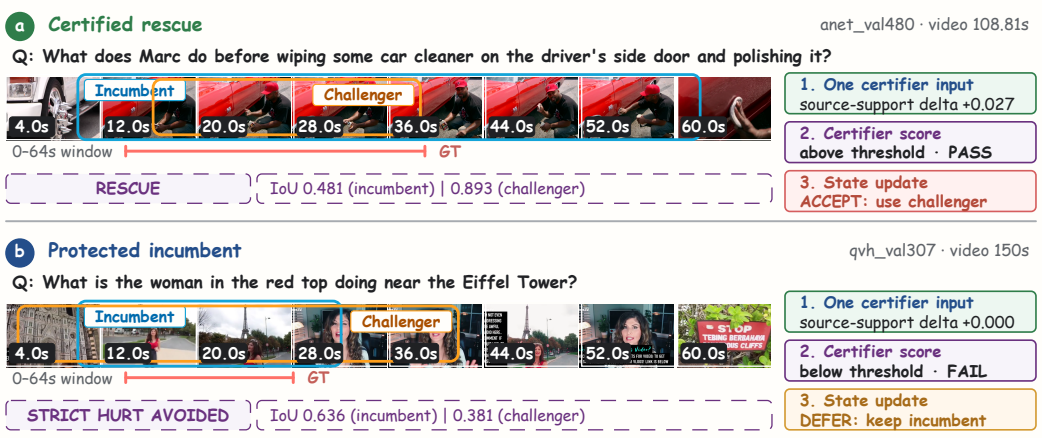


Figure 3: Two outcomes of selective temporal revision. Blue and orange outlines mark the incumbent and challenger; coral reference spans and IoU values are evaluation-only overlays. Source-support difference is one certifier input; the replacement score determines whether the acceptance threshold is met. Top: accepting a challenger yields an R@0.5 rescue. Bottom: deferral preserves the incumbent and avoids a harmful expansion.

adding only time outside the reference event. This expansion preserves event coverage but reduces temporal precision by including more surrounding context. Retaining the incumbent preserves useful localization despite another relevant interval being available. Both incumbents already cover the reference event, showing that event coverage alone cannot determine whether a revision is useful. The incumbent provides a reference for distinguishing improved boundary precision from expansion that only adds excess context. High source support therefore does not, by itself, establish that a derived candidate is a better replacement. The certifier evaluates candidate and incumbent evidence jointly to assess whether the proposed boundary change warrants replacing the current span.

5 CONCLUSION

We presented VideoSlate, a framework for agentic video reasoning that separates proposing temporal revisions from deciding whether to replace the current span. Its evidence blackboard links candidates to their provenance and verifier observations, supplying evidence for incumbent-relative comparison. The certifier assesses replacement value from this evidence, and the controller authorizes supported updates while retaining the incumbent otherwise. Evaluations show gains in temporal localization and joint answer-localization quality. In the primary controlled evaluation, the full policy corrects more errors with fewer harmful replacements than unconditional verifier selection. These findings highlight the value of judging a candidate against what it would replace. Keeping the incumbent as both a comparison reference and an available output allows VideoSlate to refine temporal evidence selectively for downstream reasoning. Each returned span remains traceable to the evidence and decision rule recorded on the blackboard.

AI USE STATEMENT

AI tools were used to assist writing, polish wording, check terminology consistency, draft sections of the supplementary material, and support literature retrieval and discovery of related work. They also assisted with experimental code development, LaTeX formatting, and figure and table layout adjustments. The authors are responsible for the scientific claims, interpretation of results, citations, and final content.

REPRODUCIBILITY STATEMENT

Our code is available at `https://anonymous.4open.science/r/VideoSlate/`. The repository includes the implementation, configuration files, and instructions for training and evaluation. We also provide tests covering candidate construction, commit decisions, and evaluation behavior.

REFERENCES

Anthropic. Introducing Claude 3.5 Sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

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

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd International Conference on Machine Learning*, 2005. URL https://www.microsoft.com/en-us/research/publication/learning-to-rank-using-gradient-descent/.

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. CG-Bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/70fa5df8e3300dc30bf19bee44a56155-Abstract-Conference.html.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang Frank Wang. ReXTime: A benchmark suite for reasoning-across-time in videos. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, volume 37, 2024a. doi: 10.52202/079017-0900. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/32683193e1d0e7a5795b073acecb3549-Abstract-Datasets_and_Benchmarks_Track.html.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, et al. ShareGPT4Video: Improving video understanding and generation with better captions. In *Advances in Neural Information Processing Systems*, pp. 19472–19495, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/22a7476e4fd36818777c47e666f61a41-Abstract-Datasets_and_Benchmarks_Track.html.

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*, 2024c. URL https://arxiv.org/abs/2411.18211.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *Computer Vision – ECCV 2024*,

pp. 75–92, 2024. doi: 10.1007/978-3-031-72670-5_5. URL https://doi.org/10.1007/978-3-031-72670-5_5.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-CCAM: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024. URL https://arxiv.org/abs/2408.14023.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. VITA: Towards open-source interactive omni multimodal LLM. *arXiv preprint arXiv:2408.05211*, 2024. URL https://arxiv.org/abs/2408.05211.

Difei Gao, Lei Ji, Luowei Zhou, Kevin Qinghong Lin, Joya Chen, Zihan Fan, and Mike Zheng Shou. AssistGPT: A general multi-modal assistant that can plan, execute, inspect, and learn. *arXiv preprint arXiv:2306.08640*, 2023. URL https://arxiv.org/abs/2306.08640.

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*, pp. 5267–5275, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Gao_TALL_Temporal_Activity_ICCV_2017_paper.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Ayush Gupta, Anirban Roy, Rama Chellappa, Nathaniel D. Bastian, Alvaro Velasquez, and Susmit Jha. TOGA: Temporally grounded open-ended video QA with weak supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23593–23603, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Gupta_TOGA_Temporally_Grounded_Open-Ended_Video_QA_with_Weak_Supervision_ICCV_2025_paper.html.

Bochen Han and Songmao Zhang. Exploring advanced LLM multi-agent systems based on blackboard architecture. *arXiv preprint arXiv:2507.01701*, 2025. URL https://arxiv.org/abs/2507.01701.

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*, pp. 14271–14280, 2024a. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_VTimeLLM_Empower_LLM_to_Grasp_Video_Moments_CVPR_2024_paper.html.

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. LITA: Language instructed temporal-localization assistant. *arXiv preprint arXiv:2403.19046*, 2024b. URL https://arxiv.org/abs/2403.19046.

Songtao Jiang, Sibo Song, Chenyi Zhou, Yuan Wang, Ruizhe Chen, Tongkun Guan, Ruilin Luo, Yan Zhang, Zhihang Tang, Yuchong Sun, Hang Zhang, Zhibo Yang, Shuai Bai, Junyang Lin, and Zuozhu Liu. Learning transferable temporal primitives for video reasoning via synthetic videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/papers/Jiang_Learning_Transferable_Temporal_Primitives_for_Video_Reasoning_via_Synthetic_Videos_CVPR_2026_paper.pdf.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. Chat-UniVi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023. URL https://arxiv.org/abs/2311.08046.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S. Ryoo. Language repository for long video understanding. *arXiv preprint arXiv:2403.14622*, 2024. URL https://arxiv.org/abs/2403.14622.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 706–715, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Krishna_Dense-Captioning_Events_in_ICCV_2017_paper.html.

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*, volume 34, 2021. URL https://papers.nips.cc/paper/2021/hash/62e0973455fd26eb03e91d5741a4a3bb-Abstract.html.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a. URL https://arxiv.org/abs/2408.03326.

Jiaze Li, Hao Yin, Wenhui Tan, Jingyang Chen, Boshen Xu, Yuxun Qu, Yijing Chen, Jianzhong Ju, Zhenbo Luo, and Jian Luan. REVISOR: Beyond textual reflection, towards multimodal introspective reasoning in long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5059–5069, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_REVISOR_Beyond_Textual_Reflection_Towards_Multimodal_Introspective_Reasoning_in_Long-Form_CVPR_2026_paper.html.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Li_MVBench_A_Comprehensive_Multi-modal_Video_Understanding_Benchmark_CVPR_2024_paper.html.

Pandeng Li, Chen-Wei Xie, Hongtao Xie, Liming Zhao, Lei Zhang, Yun Zheng, Deli Zhao, and Yongdong Zhang. MomentDiff: Generative video moment retrieval from random to real. In *Advances in Neural Information Processing Systems*, volume 36, pp. 65948–65966, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/d01bda31bbcd780774ff15b534e03c40-Abstract-Conference.html.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. *arXiv preprint arXiv:2506.18883*, 2025. URL https://arxiv.org/abs/2506.18883.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/aca97732e30bcf1303bc22ac3924fd16-Abstract-Conference.html.

Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023a. URL https://arxiv.org/abs/2311.10122v1.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. VILA: On pre-training for visual language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26689–26699, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Lin_VILA_On_Pre-training_for_Visual_Language_Models_CVPR_2024_paper.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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2794–2804, 2023b. URL https://openaccess.thecvf.com/content/ICCV2023/html/Lin_UniVTG_Towards_Unified_Video-Language_Temporal_Grounding_ICCV_2023_paper.html.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024a. URL https://arxiv.org/abs/2408.15542.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. ST-LLM: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pp. 1–18, 2024b. URL https://doi.org/10.1007/978-3-031-72998-0_1.

Ye Liu, Siyuan Li, Yang Wu, Chang Wen Chen, Ying Shan, and Xiaohu Qie. UMT: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3042–3051, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Liu_UMT_Unified_Multi-Modal_Transformers_for_Joint_Video_Moment_Retrieval_and_CVPR_2022_paper.html.

Ye Liu, Jixuan He, Wanhua Li, Junsik Kim, Donglai Wei, Hanspeter Pfister, and Chang Wen Chen. R^2-Tuning: Efficient image-to-video transfer learning for video temporal grounding. In *European Conference on Computer Vision*, 2024c. URL https://arxiv.org/abs/2404.00801.

Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. E.T. Bench: Towards open-ended event-level video-language understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 32076–32110, 2024d. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/38ace89a980ead30c6be2b46afc1a5bd-Abstract-Datasets_and_Benchmarks_Track.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://proceedings.iclr.cc/paper_files/paper/2026/hash/5e84e4413268b713f0d4a1b23a9dae57-Abstract-Conference.html.

Ziyu Ma, Chenhui Gou, Hengcan Shi, Bin Sun, Shutao Li, Hamid Rezatofighi, and Jianfei Cai. DrVideo: Document retrieval based long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18936–18946, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Ma_DrVideo_Document_Retrieval_Based_Long_Video_Understanding_CVPR_2025_paper.html.

David Madras, Toniann Pitassi, and Richard Zemel. Predict responsibly: Improving fairness and accuracy by learning to defer. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/09d37c08f7b129e96277388757530c72-Abstract.html.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

OpenAI. GPT-4o system card, 2024. URL https://openai.com/index/gpt-4o-system-card/.

OpenGVLab. InternVL2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy, 2024. URL https://internvl.github.io/blog/2024-07-02-InternVL-2.0/.

Shraman Pramanick, Effrosyni Mavroudi, Yale Song, Rama Chellappa, Lorenzo Torresani, and Triantafyllos Afouras. Enrich and detect: Video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24297–24308, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Pramanick_Enrich_and_Detect_Video_Temporal_Grounding_with_Multimodal_LLMs_ICCV_2025_paper.html.

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, 2024a. URL https://proceedings.mlr.press/v235/qian24a.html.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. In *Advances in Neural Information Processing Systems*, pp. 119336–119360, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d7ce06e9293c3d8e6cb3f80b4157f875-Abstract-Conference.html.

Mengxue Qu, Xiaodong Chen, Wu Liu, Alicia Li, and Yao Zhao. ChatVTG: Video temporal grounding via chat with video dialogue large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1847–1856, 2024. URL https://openaccess.thecvf.com/content/CVPR2024W/PVUW/html/Qu_ChatVTG_Video_Temporal_Grounding_via_Chat_with_Video_Dialogue_Large_CVPRW_2024_paper.html.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024. URL https://arxiv.org/abs/2403.05530v1.

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*, pp. 14313–14323, 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.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Suris_ViperGPT_Visual_Inference_via_Python_Execution_for_Reasoning_ICCV_2023_paper.html.

Yepeng Tang, Weining Wang, Longteng Guo, Tongtian Yue, Wenxuan Wang, Chunjie Zhang, and Jing Liu. Divid: Disentangled spatial-temporal modeling within LLMs for temporally grounded video understanding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1c85c302ece39939c1b334c78f7ee1b8-Abstract-Conference.html.

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*, 2024a. URL https://arxiv.org/abs/2409.12191.

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *Computer Vision – ECCV 2024*, pp. 58–76, 2024b. doi: 10.1007/978-3-031-72989-8_4. URL https://doi.org/10.1007/978-3-031-72989-8_4.

Xiyao Wang, Linfeng Song, Ye Tian, Dian Yu, Baolin Peng, Haitao Mi, Furong Huang, and Dong Yu. Towards self-improvement of LLMs via MCTS: Leveraging stepwise knowledge with curriculum

preference learning. *arXiv preprint arXiv:2410.06508*, 2024c. URL https://arxiv.org/abs/2410.06508.

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

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*, 2024d. URL https://arxiv.org/abs/2403.10228.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xiao_Can_I_Trust_Your_Answer_Visually_Grounded_Video_Question_Answering_CVPR_2024_paper.html.

Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, et al. Task preference optimization: Improving multimodal large language models with vision task alignment. *arXiv preprint arXiv:2412.19326*, 2024. URL https://arxiv.org/abs/2412.19326.

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*, volume 38, pp. 119152–119184, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/acb94e709f02895fd98b5867f0b184f3-Abstract-Conference.html.

Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 124–141, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/00d1f03b87a401b1c7957e0cc785d0bc-Abstract-Conference.html.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. MM-ReAct: Prompting ChatGPT for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023. URL https://arxiv.org/abs/2303.11381.

Zuhao Yang, Yingchen Yu, Yunqing Zhao, Shijian Lu, and Song Bai. TimeExpert: An expert-guided video LLM for video temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24286–24296, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Yang_TimeExpert_An_Expert-Guided_Video_LLM_for_Video_Temporal_Grounding_ICCV_2025_paper.html.

Woongyeong Yeo, Kangsan Kim, Jaehong Yoon, and Sung Ju Hwang. WorldMM: Dynamic multimodal memory agent for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25599–25609, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yeo_WorldMM_Dynamic_Multimodal_Memory_Agent_for_Long_Video_Reasoning_CVPR_2026_paper.html.

Yufei Yin, Qianke Meng, Minghao Chen, Jiajun Ding, Zhenwei Shao, and Zhou Yu. VideoARM: Agentic reasoning over hierarchical memory for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24042–24051, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yin_VideoARM_Agentic_Reasoning_over_Hierarchical_Memory_for_Long-Form_Video_Understanding_CVPR_2026_paper.html.

Shoubin Yu, Jaemin Cho, Prateek Yadav, and Mohit Bansal. Self-chained image-language model for video localization and question answering. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/f22a9af8dbb348952b08bd58d4734b50-Abstract-Conference.html.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. A simple LLM framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21715–21737, 2024a. doi: 10.18653/v1/2024.emnlp-main.1209. URL https://aclanthology.org/2024.emnlp-main.1209/.

Hang Zhang, Xin Li, and Lidong Bing. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023. URL https://arxiv.org/abs/2306.02858.

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.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024b. URL https://arxiv.org/abs/2406.16852.