TempCloze: Can Video-LLMs Identify the Missing Middle?

Wenqi Pei[1,*] **Henry Hengyuan Zhao**[2,*] **Yilai Liu**[1,*] **Jiahao Meng**[3]
Han Chen[2] **Ziyu Wang**[3] **Hongyang Du**[1,†]

[1]The University of Hong Kong [2]National University of Singapore [3]Peking University

Abstract

Temporal reasoning benchmarks for Video-LLMs are often mediated by language, leaving room for linguistic shortcuts from option wording, answer correlations, or language priors. To reduce such shortcuts, we introduce **TempCloze**[1], a video cloze benchmark for evaluating visual temporal reasoning in Video-LLMs. Given the beginning and ending clips of a video, models must identify the true missing middle from four candidates. TempCloze contains 1,521 carefully filtered videos from seven sources, mainly long-take and egocentric videos. We construct same-source distractors along three dimensions: *Semantic* asks what event should happen, *Alignment* probes when it should occur, and *Progression* tests how it should unfold, while shared scenes and objects reduce appearance cues. Our evaluation of 10 proprietary and 21 open-source Video-LLMs reveals Alignment as the primary bottleneck: models often recognize plausible semantic content and local event progression but struggle with temporal alignment. We further conduct error pattern and behavioral sensitivity analyses on TempCloze-Mixed and TempCloze-Hard with four representative models to examine where errors arise and how candidate order, context direction, visible span, frame density, and test-time scaling influence model choices.

1 Introduction

Video Large Language Models (Video-LLMs) have become increasingly capable of video temporal reasoning with advances in multimodal architectures and model scaling (Li et al., 2023; Zhao et al., 2025; Meng et al., 2025, 2026a). Though Video-LLMs have achieved strong performance on temporal reasoning benchmarks like TempCompass and TVBench (Liu et al., 2024; Cores et al., 2025), the benchmarks are moving beyond conventional VideoQA to cover longer contexts and more fine-grained questions (Fu et al., 2025; Zhang et al., 2025a; Meng et al., 2026b; Xu et al., 2026).

However, these evaluations are still mediated by language. Models need to choose among textual options or produce detailed captions (Lei et al., 2019; Maaz et al., 2024). Such formats can introduce linguistic shortcuts, allowing models to exploit option wording, answer correlations, or language priors to obtain competitive scores (Xiao et al., 2024; Balepur et al., 2024). For example, pretrained language models can exceed the random baseline by more than 25% without multimodal context (Yang et al., 2020). Accordingly, a more direct evaluation should ask models to compare visual evidence in video segments rather than textual descriptions. This motivates a benchmark for the evaluation of visual temporal reasoning in Video-LLMs.

To address the gap, we introduce TempCloze, a video cloze benchmark centered on the question: *Can Video-LLMs Identify the Missing Middle?* Given the beginning and ending clips of a video, models are required to identify the true missing middle from a set of candidates. The correct choice is not merely visually related to the video, but occupies the proper temporal position between the observed clips. Identifying the middle requires models to infer what happens, when it occurs, and how it unfolds. This setup encourages models to reason over time from what they see, reducing the influence of linguistic shortcuts in evaluation.

To evaluate distinct aspects of visual temporal reasoning, TempCloze constructs multi-dimension distractors from the same source. They share scenes and objects with the original video. This reduces appearance cues and encourages temporal comparison rather than visual matching. Moreover, distractors cover three dimensions: **Semantic (S)** tests what event should happen by contrasting the correct event with other non-overlapping intervals, **Alignment (A)** probes when the event should occur

[*] Equal contribution. [†] Corresponding author.
[1] https://github.com/CedricPei/Temporal-Cloze

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 13676–13693
October 24-29, 2026 ©2026 Association for Computational Linguistics

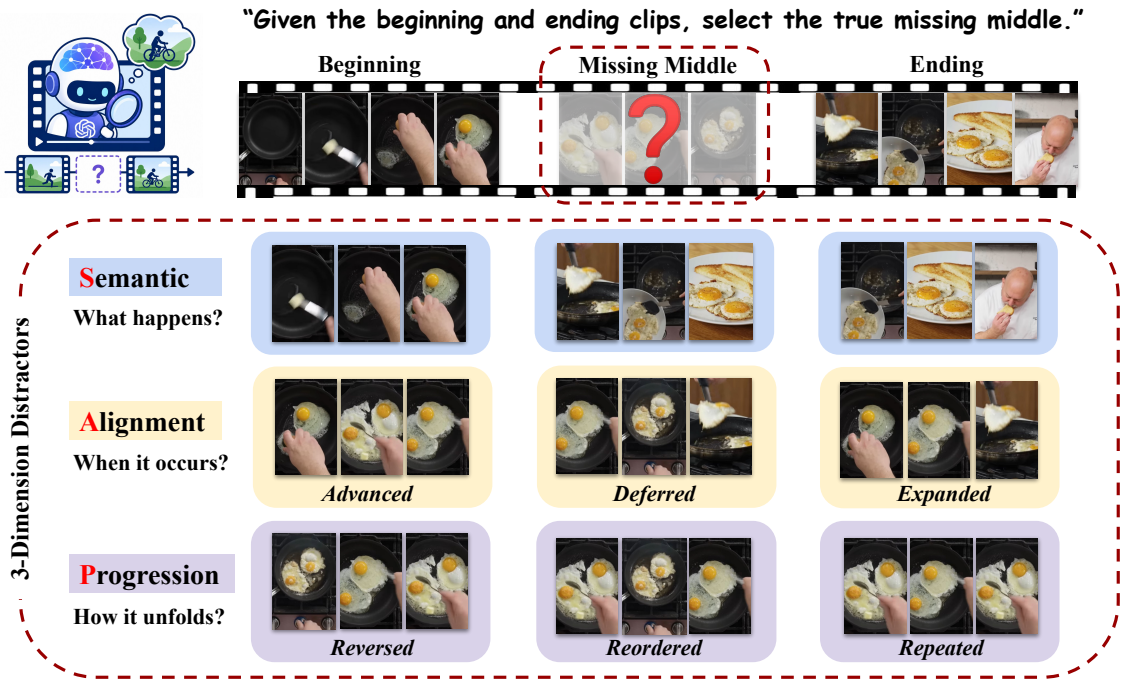


Figure 1: Overview of TempCloze. Given the beginning and ending clips, models are asked to identify the true missing middle. Distractors are constructed along three dimensions: Semantic, Alignment and Progression.

by shifting or expanding the ground-truth span, and **Progression (P)** evaluates how the event should develop by reversing, reordering, or repeating the segment. Collectively, we assess event semantics, temporal alignment, and temporal progression.

We construct TempCloze from seven sources covering long-take (Xiong et al., 2024), egocentric (Yang et al., 2026), and fine-grained motion (Tu et al., 2025) videos. After filtering out overlength and low-quality videos, we employ a reasoning LLM to examine their captions and exclude those unsuitable for cloze formatting. We then use optical flow to remove largely static videos (Farnebäck, 2003). The resulting benchmark contains 1,521 videos, each instantiated along three dimensions. We evaluate 10 proprietary and 21 open-source Video-LLMs and observe a consistent bottleneck: while models can identify plausible event content and recognize its progression, they struggle substantially with temporal alignment. Additionally, we construct two auxiliary subsets: TempCloze-Mixed, which mixes distractors across temporal dimensions, and TempCloze-Hard, which consists of the most widely misclassified instances.

We further conduct Error Pattern Analysis and Behavioral Sensitivity Analysis to examine where errors arise and what factors affect model choices. Dimension-specific errors show that Alignment failures are dominated by Expanded, while Progression failures are marked by Reversed. Mixed-dimension errors show that Alignment is the most misleading alternative. We also find that selections are unstable under candidate reordering but with relatively stable performance. Furthermore, models rely more on beginning context than ending. They depend strongly on endpoint information, and adding frame density can dilute the information. Test-time scaling brings model-dependent gains, but does not change the relative ordering among dimensions. These analyses offer broader insights into the performance and limitations of visual temporal reasoning in existing Video-LLMs.

Our contributions are summarized as follows:

- We introduce TempCloze, a video cloze benchmark comprising 1,521 videos to evaluate visual temporal reasoning in Video-LLMs. Each video is instantiated with three dimensions: Semantic, Alignment, and Progression.
- We evaluate 10 proprietary and 21 open-source Video-LLMs and identify the Alignment dimension as the main bottleneck.
- We conduct Error Pattern Analysis and Behavioral Sensitivity Analysis, revealing where errors arise and how some factors influence model choices.

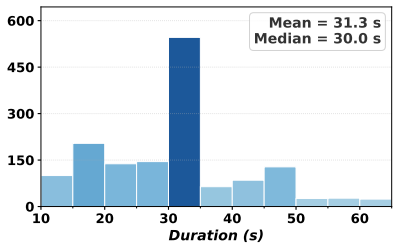
2 Related Work

2.1 Temporal Reasoning Benchmarks

Temporal reasoning benchmarks for video have grown increasingly complex in recent years. One

Table 1: Composition over 1,521 videos. Left: videos per source; Right: duration distribution in seconds.

Source	#
LVD-2M	515
EgoLife	437
MiraData	198
FAVOR	145
CaReBench	94
Video-TT	89
Daily-Omni	43



line comes from general video understanding benchmarks that include temporal reasoning as part of a broader evaluation. Early VideoQA datasets such as TGIF-QA (Jang et al., 2017), NExT-QA (Xiao et al., 2021) and ActivityNet-QA (Yu et al., 2019) established this setting with questions about explicit actions and spatio-temporal relations. Recent long-context benchmarks such as LongVideoBench (Wu et al., 2024), MVBench (Li et al., 2024) and VideoMME (Fu et al., 2025) broaden coverage across diverse tasks, requiring evidence aggregation over extended dynamics rather than isolated cues. Another line focuses more directly on temporal structure, with TemporalBench (Cai et al., 2024), TVBench (Cores et al., 2025), and TempCompass (Liu et al., 2024) testing event ordering, moment localization, and fine-grained motion. However, these benchmarks are still language-mediated. This leaves room for linguistic shortcuts, where models can benefit from surface regularities (Ma et al., 2024; Zhong et al., 2022). TempCloze asks models to identify the missing middle from video candidates, shifting evaluation toward visual temporal reasoning.

2.2 Cloze for Video Understanding

Cloze tasks remove part of a context and infer the missing content (Taylor, 1953). MovieFIB (Maharaj et al., 2017) asks models to fill blanks in movie descriptions and FIBER (Castro et al., 2022) extends this with constituent-level blanks and multiple valid answers. While they evaluate video understanding through textual completion, cloze-style objectives have also been used for video representation learning. VideoBERT (Sun et al., 2019) is an early pre-training method that applies masked prediction to learn temporal correspondences between video and text. VCP (Luo et al., 2020) learns by predicting transformations applied to withheld video clips. MaskFeat (Wei et al., 2023) masks parts of the video input and predicts features of the regions, while VideoMAE (Tong et al., 2022) reconstructs masked video patches for pre-training. These works demonstrate the value of cloze formatting, but they use it as a training objective. TempCloze instead formulates video cloze for the evaluation of temporal reasoning in Video-LLMs.

3 TempCloze

3.1 Overview

A video V is defined over the time interval $(0,T)$. We use $C_{a,b}$ to denote the clip spanning interval (a,b). For a target interval (s,e), the video can be decomposed into the beginning $B = C_{0,s}$, the middle $M = C_{s,e}$, and the ending $E = C_{e,T}$:

$$V = [C_{0,s} \mid C_{s,e} \mid C_{e,T}] = [B \mid M \mid E].$$

TempCloze instantiates the missing-middle video cloze task. Given the visible context (B,E) and a candidate set $\mathcal{Y}_d$, the model must select the clip that correctly fills the missing interval. For each dimension d, the candidate set contains the middle M and three distractors:

$$\mathcal{Y}_d = \{M, D_d^1, D_d^2, D_d^3\}, \quad d \in \{S, A, P\}.$$

Distractors are designed around three dimensions: Semantic (S), Alignment (A), and Progression (P).

3.2 Data Sources

We build TempCloze from seven public sources: CaReBench (Xu et al., 2025), Daily-Omni (Zhou et al., 2026), EgoLife (Yang et al., 2026), FAVOR-Bench (Tu et al., 2025), LVD-2M (Xiong et al., 2024), MiraData (Ju et al., 2024), and Video Thinking Test (Zhang et al., 2025b). Long-take sources such as MiraData and LVD-2M and egocentric sources including EgoLife and FAVOR-Bench provide temporally continuous videos that reduce scene-change cues, while the remaining sources broaden the activity domains. Detailed dataset descriptions are provided in Appendix D.1.

3.3 Dataset Construction

We construct TempCloze with a pipeline that filters videos efficiently before synthesizing distractors.

3.3.1 Video Filtering

Duration and LLM Filtering. We first rely on source metadata to retain videos between 12 and 90 seconds. We leverage temporally dense captions provided by sources such as LVD-2M and MiraData to screen videos before loading the raw files. We then employ the reasoning model GPT-o3 (OpenAI, 2025) to assess whether the events are

Table 2: Main results on TempCloze. Dimension Accuracy reports Semantic, Alignment, and Progression accuracy; Cumulative Accuracy reports the fraction of videos with at least one, at least two, or all three dimensions correct. Peach, blue, and sage mark the top three; **Bold** and underlined values mark the best and second-best. Human* Baseline is measured on 100 randomly selected videos, and the detailed protocol is provided in Appendix C.

Model		Dimension Accuracy (%)				Cumulative Accuracy (%)		
Name	Think	S↑	A↑	P↑	Mean↑	≥1 ↑	≥2 ↑	3/3 ↑
Proprietary Models								
Seed1.8	✓	95.07	**76.92**	**93.75**	**88.58**	**99.54**	**95.40**	**70.81**
Qwen3.5-Plus	✓	90.72	76.32	90.20	85.74	98.75	91.25	67.22
Seed1.8	✗	**96.25**	61.93	92.11	83.43	99.35	94.02	56.94
Gemini2.5-Pro	✓	90.33	66.67	67.78	74.92	97.44	83.49	43.95
Gemini2.5-Flash	✗	82.18	59.83	74.82	72.28	94.61	78.17	44.05
GPT5.4	✗	68.11	37.41	65.15	56.89	89.35	60.16	21.17
Claude4.6-Sonnet	✗	55.69	28.80	70.94	51.81	86.12	52.66	16.63
Claude4.6-Opus	✓	49.47	39.34	62.89	50.57	81.97	50.85	18.88
Gemini3-Flash	✗	61.08	40.37	48.13	49.86	84.09	49.57	15.91
Seed1.6	✗	64.59	17.83	55.40	45.93	81.65	46.68	9.38
Grok4.1	✗	24.52	24.06	23.73	24.11	57.53	14.07	0.72
Avg.		70.73	48.13	67.72	62.19	88.22	65.12	33.24
Open-source Models								
Qwen3.5-397B-A17B	✓	**75.94**	50.62	**78.24**	**68.27**	**94.54**	**74.49**	**35.77**
Qwen3.5-35B-A3B	✓	75.10	**51.55**	69.96	65.54	92.33	70.84	33.80
KimiK2.5	✗	71.27	41.95	65.02	59.41	89.94	62.79	25.51
Qwen3VL-32B-I	✗	51.35	10.91	63.25	41.84	78.70	41.36	5.46
InternVL3.5-38B	✗	37.81	32.63	46.23	38.96	70.43	36.54	11.35
Qwen3VL-8B-I	✗	31.14	15.06	45.46	30.55	64.49	23.71	3.49
InternVL3.5-8B	✗	26.30	28.34	29.52	28.05	60.15	20.18	3.81
InternVL3-38B	✗	30.31	20.12	32.81	27.74	59.04	21.24	2.96
KimiVL-A3B-T	✓	30.17	25.52	24.98	26.91	62.17	16.23	2.17
Qwen3VL-32B-T	✓	25.05	25.38	25.72	25.38	58.76	15.40	1.91
LLaVA-CriticR1-7B	✓	23.81	26.90	25.36	25.36	58.22	16.17	1.73
Qwen2.5VL-7B-I	✗	27.22	26.22	24.17	25.87	57.93	17.70	2.11
GLM4.6V-Flash	✗	21.78	14.71	39.58	25.36	55.64	17.36	3.04
Qwen3VL-4B-T	✓	27.81	24.85	22.95	25.20	58.38	15.58	1.64
Qwen3VL-8B-T	✓	24.65	25.38	24.59	24.87	57.53	15.45	1.64
ThinkLiteVL-7B	✓	24.36	25.79	23.58	24.58	57.80	14.60	1.33
MiMoVL-7B-RL	✗	20.84	23.73	28.80	24.46	53.91	16.70	2.76
Qwen3VL-4B-I	✗	20.38	19.72	32.48	24.19	54.30	15.51	2.76
KimiVL-A3B-I	✗	21.09	26.78	23.13	23.67	52.23	15.60	2.75
MiMoVL-7B-SFT	✗	23.41	20.51	25.77	23.23	52.86	14.60	2.24
Molmo2-8B	✗	24.19	20.70	24.73	23.21	53.50	14.25	1.88
Avg.		34.00	26.54	36.97	32.51	63.95	26.49	7.15
Human*	–	**96.00**	**98.00**	**97.00**	**97.00**	**100.00**	**99.00**	**92.00**
Random	–	25.00	25.00	25.00	25.00	57.81	15.62	1.56

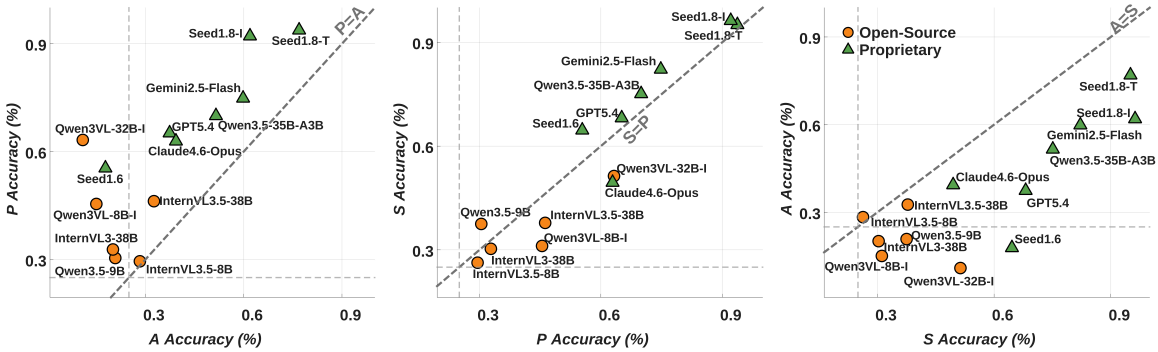
temporally coherent and whether the surrounding context can constrain a missing middle. Videos that are unsuitable for cloze formatting are excluded.

Quality and Motion Filtering. We remove clips with a bitrate below 200 kbps or sharpness below 30. For videos that pass quality checks, we sample a gap from the central 50% of the timeline, with length set to 20%–40% of the full duration. This leaves context on both sides and avoids trivial and underconstrained boundary cases. We validate each gap with Farnebäck optical flow (Farnebäck, 2003), requiring average flow magnitude above 1.0 and resampling up to three times.

These stages preserve video quality while reducing unnecessary loading. After this pipeline, TempCloze retains 1,521 videos. Table 1 summarizes the source and duration statistics. Appendix D.2 provides detailed filtering statistics, and Appendix D.3 reports human validation of filtered-video quality.

3.3.2 Distractor Generation

We construct three distractors per dimension. Formal mathematical definitions are in Appendix A. **Semantic** focuses on the specific content that the missing clip needed to connect. Each distractor is a same-duration clip, $D_S = C_{u,u+\ell}$, where $(u,u+\ell)$ does not overlap (s,e). Distinguishing them requires identifying the event implied by the context and the core activity of the video.



(a) Progression vs. Alignment (b) Semantic vs. Progression (c) Alignment vs. Semantic

Figure 2: Pairwise comparison of dimension-level accuracy on TempCloze across leading models.

Alignment focuses on the temporal compatibility of the missing clip when the content remains plausible. Each distractor is a nearby clip, $D_A = C_{s',e'}$, where (s',e') shifts or expands (s,e). Advanced and Deferred place the candidate before or after the target, while Expanded widens it to include adjacent context from both. Distinguishing them requires sensitivity to temporal boundaries.

Progression focuses on event development of the missing clip when both content and coarse temporal compatibility remain plausible. We use Reversed, Reordered, and Repeated variants: Reversed plays the answer segment backward, Reordered disrupts the local order of subevents, and Repeated duplicates a short subsegment. Distinguishing them requires fine-grained perception of event direction, subevent order, and action continuity.

4 Experiments

4.1 Setup

4.1.1 Models

We evaluate 10 proprietary and 21 open-source models. For models with different versions, we use the suffix T for thinking and I for instruct. Proprietary models are Qwen3.5-Plus (Qwen Team, 2026), Seed1.8 (ByteDance Seed, 2025b), Gemini2.5-Pro, Gemini2.5-Flash (Comanici et al., 2025), GPT5.4 (OpenAI, 2026), Claude4.6-Sonnet, Claude4.6-Opus (Anthropic, 2026), Gemini3-Flash (Google, 2025), Seed1.6 (ByteDance Seed, 2025a), and Grok4.1 (xAI, 2025). Open-source models are Qwen3.5-397B-A17B, Qwen3.5-35B-A3B (Qwen Team, 2026), KimiK2.5 (Team et al., 2026), Qwen3VL-32B-I, Qwen3VL-32B-T, Qwen3VL-8B-T, Qwen3VL-8B-I, Qwen3VL-4B-T, Qwen3VL-4B-I (Bai et al., 2025a), InternVL3.5-38B, InternVL3.5-8B (Wang et al., 2025a), InternVL3-38B (Zhu et al., 2025), KimiVL-A3B-T, KimiVL-A3B-I (Kimi Team, 2025), LLaVA-CriticR1-7B (Wang et al., 2025b), Qwen2.5VL-7B-I (Bai et al., 2025b), GLM4.6V-Flash (Z.ai, 2025), ThinkLiteVL-7B (Wang et al., 2025c), MiMoVL-7B-RL, MiMoVL-7B-SFT (Yue et al., 2025), and Molmo2-8B (Clark et al., 2026).

4.1.2 Configurations

By default, we uniformly sample 16 frames from each clip, yielding 96 frames per dimension across six clips. Appendix B.1 details the sampling rule. To avoid exact boundary-frame shortcuts, we also conduct a boundary sanity check in Appendix B.2. Open-source models are deployed on A6000 GPUs using vLLM (Kwon et al., 2023).

4.1.3 Representative Models and Subsets

For in-depth analysis, we use four representative models: Gemini2.5-Pro, Seed1.6, Qwen3.5-397B-A17B, and Qwen3VL-8B-I. We also construct two auxiliary subsets. TempCloze-Mixed contains a randomly shuffled subset of 300 videos, each with candidates from the ground-truth and S, A, and P. TempCloze-Hard contains the 150 instances with the largest number of errors across models.

4.2 Main Results

Table 2 presents the main results for both proprietary and open-source models. We provide the full leaderboard in Figure 8 and report source-level statistics for Gemini2.5-Pro and Qwen3.5-35B-A3B in Appendix E.

Dimension accuracy. On Semantic, proprietary models average 70.73%, led by Seed1.8 at 96.25%. Open-source models average 34.00%. Nevertheless, Qwen3.5-397B reaches 75.94%, showing that the strongest open-source systems can approach proprietary-level Semantic performance. A similar group-level gap appears on Progression, where pro-

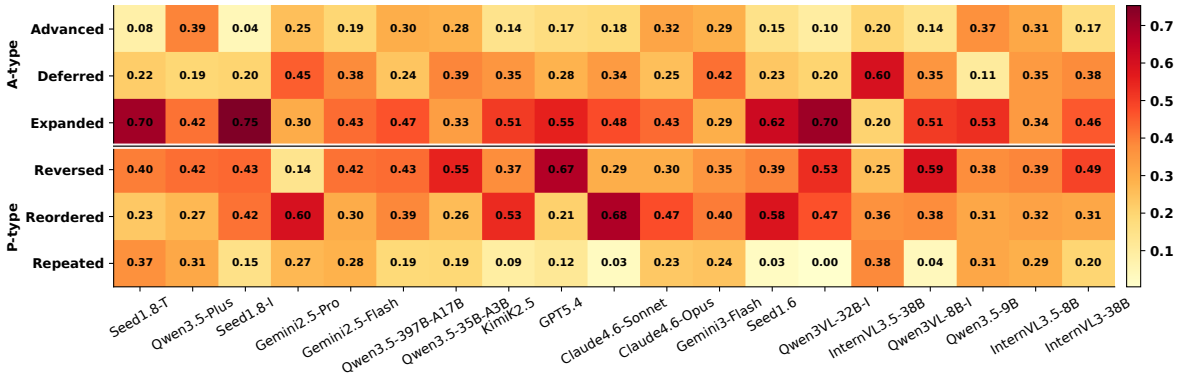


Figure 3: Error proportions for alignment and progression distractors. Darker cells indicate higher error rates.

Table 3: TempCloze-Mixed error attribution (%). The Alignment and Progression options are each randomly sampled from one of three designed variants.

Model	Acc.	S	A	P
Gemini2.5-Pro	76.3	2.7	**19.7**	1.3
Seed1.6	68.9	7.0	**14.7**	9.4
Qwen3.5-397B-A17B	62.0	9.0	**20.3**	8.7
Qwen3VL-8B-I	24.0	29.3	**37.3**	9.3

prietary models average 67.72% and open-source models average 36.97%. Qwen3.5-397B reaches 78.24%, indicating that stronger open-source models can identify local event order. Alignment is substantially harder. Proprietary models drop to 48.13% on average, while open-source models drop to 26.54%. Even the strongest scores, 76.92% for Seed1.8 and 51.55% for Qwen3.5-35B, are well below the 98.00% Human Baseline. ***Alignment is the primary bottleneck for visual temporal reasoning in current Video-LLMs.*** Models are better at recognizing plausible event content and local progression than at judging exactly when an event should occur relative to the observed endpoints.

To further examine dimension-level behavior, Figure 2 compares leading models across pairs. Two panels involving Alignment reinforce the pattern raised, as most models score lower on Alignment than the paired dimension. A notable observation emerges from the S–P panel. Open-source models have low absolute scores but perform better on Progression than Semantic, suggesting that local motion continuity can support a basic Progression score. By contrast, proprietary models sit at higher scores and favor Semantic over Progression, indicating that high Progression accuracy requires stronger temporal understanding. We report joint accuracy for dimension pairs in Appendix F.1.

Cumulative accuracy. Cumulative results shift the view from dimensions to full-video understanding. We interpret ≥ 1 as limited understanding, ≥ 2 as sufficient understanding, and 3/3 as complete understanding. Proprietary models show a meaningful level of sufficient understanding, averaging 65.12% on ≥ 2 and reaching 95.40% for the best model. However, complete understanding is harder. The proprietary average drops to 33.24% on 3/3, and even the best model reaches only 70.81%, far below the 92.00% Human Baseline. Open-source models show a more limited level of sufficient understanding. Even the strongest one reaches only 74.49% on ≥ 2 and 35.77% on 3/3, indicating that open-source models have not yet reached sufficient understanding. Overall, current Video-LLMs remain short of complete video understanding that requires robust visual temporal reasoning. We provide a video-level breakdown in Appendix F.2.

Table 4: Effect of candidate order permutation on TempCloze-Hard across representative models. Disp. denotes accuracy dispersion. FR denotes Flip Rate, and CFR denotes Clip Flip Rate. All values are percentages.

Model	Dim.	Mean	Range	Disp.	FR	CFR
Gemini2.5-Pro	S	70.0	9.3	3.9	30.2	36.9
	A	41.2	6.0	2.6	41.9	64.6
	P	49.7	10.0	3.7	48.4	64.8
	All	**53.6**	**3.8**	**1.5**	**40.2**	**55.4**
Seed1.6	S	76.8	6.0	2.4	18.8	22.0
	A	33.7	1.3	0.6	34.9	48.4
	P	81.8	6.7	2.4	24.3	26.8
	All	**64.1**	**2.7**	**0.9**	**26.0**	**32.4**
Qwen3.5-397B-A17B	S	61.1	8.8	3.2	31.2	40.2
	A	39.1	10.1	3.9	36.6	56.9
	P	64.3	10.1	4.1	36.4	44.2
	All	**54.8**	**3.1**	**1.2**	**34.7**	**47.2**
Qwen3VL-8B-I	S	7.4	4.7	1.7	11.5	51.2
	A	11.7	6.0	2.3	20.6	66.1
	P	37.4	4.7	1.9	45.5	64.8
	All	**18.9**	**1.8**	**0.8**	**25.9**	**60.7**

4.3 Error Pattern Analysis

4.3.1 Dimension-Specific Errors

Because Alignment and Progression use controlled temporal edits rather than random clips, Figure 3 analyzes their error patterns. The resulting errors

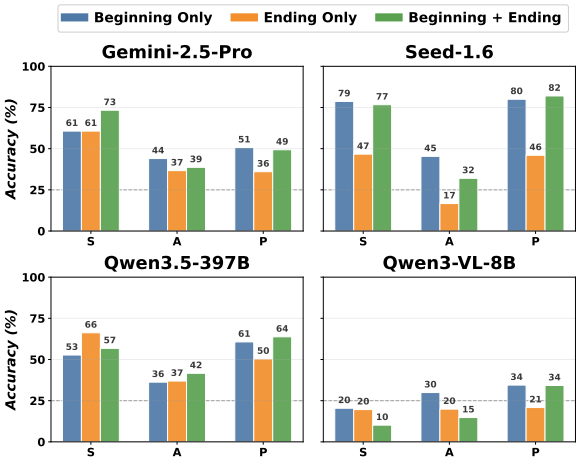


Figure 4: Effect of context direction ablation on TempCloze-Hard across representative models.

are structured rather than random. For Alignment, Expanded is the most prominent signature. The selected clip contains the correct event content but extends both before and after the target interval. For example, Seed1.8-I selects Expanded in 75% of its Alignment errors. For Progression, Reversed is the strongest failure mode. GPT-5.4 selects Reversed in 67% of its Progression errors. These two errors share visual plausibility, as models select clips whose content still appears relevant in the surrounding video while failing at temporal boundaries or at the directional order of event development.

4.3.2 Mixed-Dimension Errors

Table 3 reports accuracy together with the share of each distractor dimension on TempCloze-Mixed. Across models, Alignment receives the highest error share. This makes Alignment the most distracting dimension, echoing the finding raised in Section 4.2 that Alignment is the primary bottleneck. A counterintuitive pattern is that Semantic can be more distracting than Progression, even though models generally achieve slightly higher Semantic accuracy than Progression accuracy in Table 2. This suggests that incorrect progression is easier to reject than an incorrect event. Beyond dimension shares, the accuracy drop indicates that competition across dimensions is harder than evaluating a single dimension. Although this setting reduces the diagnostic separation within individual dimensions, it shows that models have substantial room to improve in visual temporal reasoning, and reveals failures hidden by marginal accuracy alone.

4.4 Behavioral Sensitivity Analysis

On TempCloze-Hard, we examine how multiple factors affect representative models.

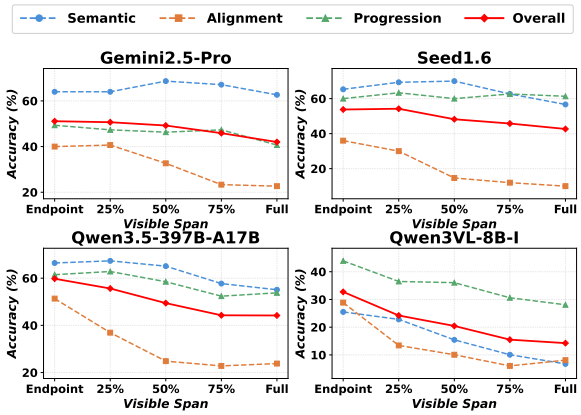


Figure 5: Effect of visible span scaling on TempCloze-Hard across representative models.

4.4.1 Candidate Order Permutation

As shown in Table 4, we evaluate the effect of candidate order by testing four candidate permutations per instance at temperature 0. FR is the share of instances for which the correctness changes at least once and CFR is the share for which the selected clip changes at least once. Mean accuracy changes little, with variance below 4 points for every model, but both instability measures remain substantial: CFR ranges from 32.4% to 60.7%, and FR ranges from 25.9% to 40.2%. This indicates that stable accuracy can hide unstable choices. Another consistent pattern is that Alignment and Progression are generally less stable than Semantic, suggesting that visual temporal reasoning is sensitive to candidate order when candidates are visually similar but temporally distinct. This is expected because content differences are salient, whereas timing and progression differences are subtler.

4.4.2 Context Direction Ablation

Figure 4 examines which part of the visible context plays a stronger role in model decisions. We observe that ending-only is often the weakest, while beginning-only frequently approaches or exceeds the performance of both contexts. For example, Seed-1.6 reaches 80% on Progression with beginning-only, close to 82% with both, but drops to 46% with ending-only. The asymmetry suggests that models rely more on forward continuation cues than backward inference. This may reflect forward ordered pretraining traces, which bias inference in the same direction. Consequently, jointly considering information from both directions remains a difficulty for visual temporal reasoning.

4.4.3 Visible Span Scaling

Figure 5 studies how performance changes as the visible span expands from endpoint frames to full

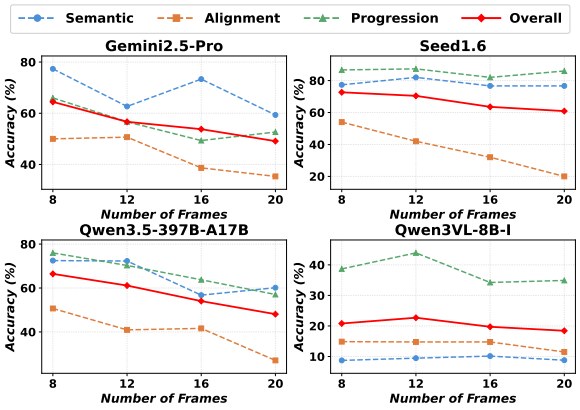


Figure 6: Effect of sampling density scaling on TempCloze-Hard across representative models.

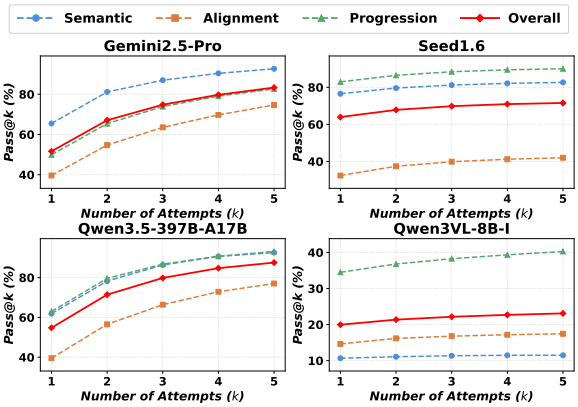


Figure 7: Effect of test-time scaling on TempCloze-Hard across representative models.

context. Alignment drops sharply for all models, since its evidence lies near the missing gap boundaries and longer spans dilute these cues. For Semantic and Progression, stronger models stay stable or improve slightly at short spans before declining mildly, while weaker models degrade much more clearly. This pattern suggests that longer context is double-edged: it adds temporal evidence for event-level reasoning but can distract models from the most diagnostic frames. Current Video-LLMs still need stronger mechanisms for integrating long-range context and selecting critical frames.

4.4.4 Sampling Density Scaling

Figure 6 studies how performance changes as sampling density increases from 8 to 20 frames per clip. The pattern is similar to visible span scaling: stronger models are more likely to preserve or slightly improve Semantic and Progression, while Alignment declines across models. This suggests that higher sampling density does not automatically provide richer evidence for Video-LLMs. Instead, dense similar frames can add confusion and reduce attention to temporally decisive cues. Exploring video native architectures that do not treat videos as sequences of images may be a promising direction for exploiting dense temporal evidence.

4.4.5 Test-Time Scaling

Figure 7 explores the test-time performance upper bound at temperature 0.7. Pass@k counts an instance as solved if any of k sampled attempts is correct. All models improve with more attempts, but the gain varies substantially. Gemini2.5-Pro and Qwen3.5-397B-A17B show the largest gains, with Overall pass@k rising by roughly 30 points from $k = 1$ to $k = 5$. This suggests that their predictions are more unstable, consistent with Table 4, where their Disp. values are higher (1.5 and 1.2) than those of Seed1.6 and Qwen3VL-8B-I (0.9 and 0.8). However, extra attempts do not change the relative ordering among dimensions: Alignment remains the lowest as the bottleneck while Semantic and Progression vary by model.

5 Conclusion

TempCloze evaluates visual temporal reasoning by requiring models to identify the missing middle from surrounding clips, instead of reasoning over natural-language descriptions. Across 1,521 videos, it builds same-source candidates along three dimensions: *Semantic* targets what event should happen, *Alignment* probes when it should occur, and *Progression* tests how it should unfold. Because candidates share scenes and objects, TempCloze limits appearance shortcuts and tests temporal evidence among visually similar options.

Results expose an uneven boundary in Video-LLMs. Many models identify Semantic content and event Progression, but temporal Alignment remains the bottleneck: models struggle to place the missing middle precisely between endpoints, while humans solve the task reliably. Error and sensitivity analyses reveal a finer-grained picture. Within dimensions, failures often choose clips with plausible content but incompatible temporal structure. Across dimensions, Alignment is the most confusing. Models remain sensitive to presentation order and context direction, showing that current gains do not reflect stable reasoning. Even with test-time scaling,the relative ordering is largely unchanged and keeps Alignment as the hardest.

Looking forward, Video-LLMs need video native mechanisms that integrate both temporal directions, select key evidence, and represent temporal structure explicitly. Future benchmarks should focus on visual temporal reasoning in longer videos while controlling linguistic shortcuts.

Limitations

TempCloze is a controlled diagnostic benchmark for missing-middle identification, rather than a complete measure of all video understanding abilities. Its design focuses on visually grounded temporal fit between the observed beginning and ending clips. It therefore does not directly cover open-ended generation, dialogue-based narrative understanding, audio-grounded inference, or unconstrained future prediction. The fixed candidate format also means that performance should be interpreted with respect to the designed alternatives, rather than as a general measure of free-form video reasoning.

All candidates are drawn from the same source video to reduce appearance shortcuts and keep alternatives visually plausible. This choice makes the task stricter: TempCloze measures exact temporal compatibility, not only broad event plausibility. In particular, Alignment distractors may contain relevant content while still failing to match the removed interval. Such cases are intentional, as they test whether models can identify the correct temporal span instead of selecting a generally plausible continuation.

Our model results should be viewed as a snapshot of current Video-LLMs. APIs, decoding behavior, and video input limits can change over time. TempCloze also draws from public sources that emphasize long-take, egocentric, and fine-grained motion videos, which supports temporal comparison but cannot cover every domain or camera style. Future extensions can broaden video domains and modalities while preserving the same visual candidate-comparison principle.

LLM Usage Statement

Central to our methodology, multiple Large Language Models (LLMs) served as the subjects of our experiments to test our proposed benchmark. We explicitly state that we have never relied on LLMs to generate core research ideas, methodologies, experimental designs, or conclusions. All technical contributions and analyses presented are the original work of the authors.

References

Anthropic. 2026. Introducing Claude Opus 4.6.

Shuai Bai, Yuxuan Cai, and 1 others. 2025a. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, and 1 others. 2025b. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. Artifacts or abduction: How do llms answer multiple-choice questions without the question? *Preprint*, arXiv:2402.12483.

ByteDance Seed. 2025a. Introduction to techniques used in Seed1.6.

ByteDance Seed. 2025b. Official release of Seed1.8: A generalized agentic model. Accessed 2026-03-21.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *Preprint*, arXiv:2410.10818.

Santiago Castro, Ruoyao Wang, Pingxuan Huang, Ian Stewart, Oana Ignat, Nan Liu, Jonathan C. Stroud, and Rada Mihalcea. 2022. Fiber: Fill-in-the-blanks as a challenging video understanding evaluation framework. *Preprint*, arXiv:2104.04182.

Christopher Clark, Jieyu Zhang, and 1 others. 2026. Molmo2: Open weights and data for vision-language models with video understanding and grounding. *arXiv preprint arXiv:2601.10611*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. 2025. Lost in time: A new temporal benchmark for videollms. *Preprint*, arXiv:2410.07752.

Gunnar Farnebäck. 2003. Two-frame motion estimation based on polynomial expansion. In *Scandinavian Conference on Image Analysis*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Google. 2025. Gemini 3 Flash Preview.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. *Preprint*, arXiv:1704.04497.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. 2024. Miradata: A large-scale video dataset with long durations and structured captions. *Preprint*, arXiv:2407.06358.

Kimi Team. 2025. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. 2019. Tvqa: Localized, compositional video question answering. *Preprint*, arXiv:1809.01696.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *Preprint*, arXiv:2403.00476.

Dezhao Luo, Chang Liu, Yu Zhou, Dongbao Yang, Can Ma, Qixiang Ye, and Weiping Wang. 2020. Video cloze procedure for self-supervised spatio-temporal learning. *Preprint*, arXiv:2001.00294.

Jie Ma, Pinghui Wang, Dechen Kong, Zewei Wang, Jun Liu, Hongbin Pei, and Junzhou Zhao. 2024. Robust visual question answering: Datasets, methods, and future challenges. *Preprint*, arXiv:2307.11471.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Preprint*, arXiv:2306.05424.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. *Preprint*, arXiv:1611.07810.

Jiahao Meng, Xiangtai Li, Haochen Wang, Yue Tan, Tao Zhang, Lingdong Kong, Yunhai Tong, Anran Wang, Zhiyang Teng, Yujing Wang, and 1 others. 2025. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*.

Jiahao Meng, Yue Tan, Qi Xu, Kuan Gao, Weisong Liu, Yanwei Li, Jason Li, Lingdong Kong, Haochen Wang, Qianyu Zhou, and 1 others. 2026a. Watch, remember, reason: Human-view video understanding with mllms. *arXiv preprint arXiv:2606.07433*.

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. 2026b. Videozerobench: Probing the limits of video mllms with spatio-temporal evidence verification. *arXiv preprint arXiv:2604.01569*.

OpenAI. 2025. Introducing openai o3 and o4-mini. *Preprint*.

OpenAI. 2026. Introducing GPT-5.4.

Qwen Team. 2026. Qwen3.5: Towards native multimodal agents.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. *Preprint*, arXiv:1904.01766.

Wilson L. Taylor. 1953. "cloze procedure": A new tool for measuring readability. *Journalism & Mass Communication Quarterly*, 30:415 – 433.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 307 others. 2026. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Preprint*, arXiv:2203.12602.

Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. 2025. Favor-bench: A comprehensive benchmark for fine-grained video motion understanding. *Preprint*, arXiv:2503.14935.

Weiyun Wang, Zhangwei Gao, and 1 others. 2025a. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xiyao Wang, Chunyuan Li, Jianwei Yang, Kai Zhang, Bo Liu, Tianyi Xiong, and Furong Huang. 2025b. LLaVA-Critic-R1: Your critic model is secretly a strong policy model. *arXiv preprint arXiv:2509.00676*.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025c. Sota with less: MCTS-guided sample selection for data-efficient visual reasoning self-improvement. *arXiv preprint arXiv:2504.07934*.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. 2023. Masked feature prediction for self-supervised visual pre-training. *Preprint*, arXiv:2112.09133.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Preprint*, arXiv:2407.15754.

xAI. 2025. Grok 4.1 model card. Technical report, xAI.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa:next phase of question-answering to explaining temporal actions. *Preprint*, arXiv:2105.08276.

Junbin Xiao, Angela Yao, Yicong Li, and Tat Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. *Preprint*, arXiv:2309.01327.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. 2024. Lvd-2m: A long-take video dataset with temporally dense captions. *Preprint*, arXiv:2410.10816.

Qi Xu, Yue Tan, Shihao Chen, Jiahao Meng, Anna Wang, Shunping Ji, Hao Fei, and Jason Li. 2026. Towards one-to-many temporal grounding. *arXiv preprint arXiv:2606.06294*.

Yifan Xu, Xinhao Li, Yichun Yang, Desen Meng, Rui Huang, and Limin Wang. 2025. Carebench: A fine-grained benchmark for video captioning and retrieval. *Preprint*, arXiv:2501.00513.

Jianing Yang, Yuying Zhu, Yongxin Wang, Ruitao Yi, Amir Zadeh, and Louis-Philippe Morency. 2020. What gives the answer away? question answering bias analysis on video qa datasets. *Preprint*, arXiv:2007.03626.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhongang Cai, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, Francesco Gringoli, and 3 others. 2026. Egolife: Towards egocentric life assistant. *Preprint*, arXiv:2503.03803.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. *Preprint*, arXiv:1906.02467.

Zihao Yue and 1 others. 2025. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*.

Z.ai. 2025. GLM-4.6V.

Xuan Zhang, Ziyan Jiang, Rui Meng, Yifei Leng, Zhenbang Xiao, Zora Zhiruo Wang, Yanyi Shang, and Dehan Kong. 2025a. Universal retrieval for multimodal trajectory modeling. *Preprint*, arXiv:2506.22056.

Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. 2025b. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. *Preprint*, arXiv:2507.15028.

Henry Hengyuan Zhao, Wenqi Pei, Yifei Tao, Haiyang Mei, and Mike Zheng Shou. 2025. Interfeedback: Unveiling interactive intelligence of large multimodal models via human feedback. *Preprint*, arXiv:2502.15027.

Yaoyao Zhong, Wei Ji, Junbin Xiao, Yicong Li, Wei Deng, and Tat seng Chua. 2022. Video question answering: Datasets, algorithms and challenges.

Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. 2026. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *Preprint*, arXiv:2505.17862.

Jinguo Zhu, Weiyun Wang, and 1 others. 2025. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

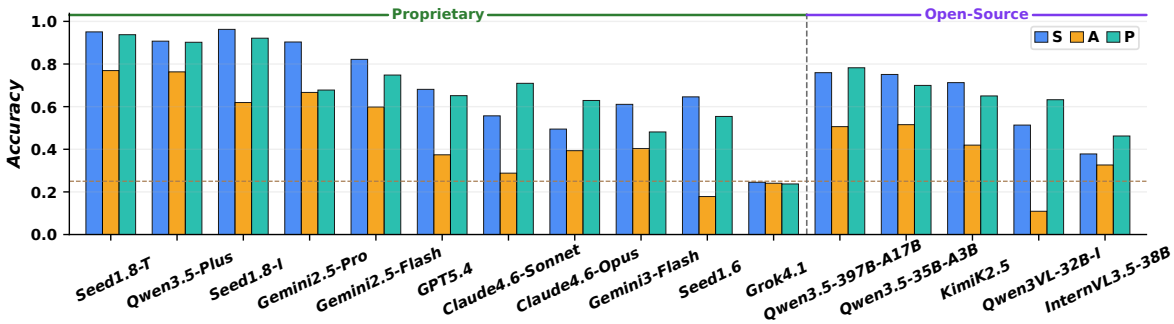


Figure 8: Top metrics of proprietary and open-source models on TempCloze.

A Distractor Formalization

Using the notation in Section 3, we define the distractors below. All intervals are valid source-video subclips. The ground-truth interval (s, e) lies in the central 50% of the video and has duration $\ell = e - s \in [0.2T, 0.4T]$. Advanced and Deferred shift it by $\ell/2$, while Expanded extends it by $\ell/2$ on each side; hence $\delta = \delta_1 = \delta_2 = \ell/2 \in [0.1T, 0.2T]$, preserving context on both sides.

Semantic distractors are same-duration clips outside the missing interval:

$$\begin{aligned} D_S^i &= C_{u_i, u_i+\ell}, \\ (u_i, u_i+\ell) \cap (s, e) &= \emptyset. \end{aligned}$$

Alignment distractors stay near the target but perturb its boundaries:

$$\begin{aligned} D_A^{\text{advanced}} &= C_{s-\delta, e-\delta}, \\ D_A^{\text{deferred}} &= C_{s+\delta, e+\delta}, \\ D_A^{\text{expanded}} &= C_{s-\delta_1, e+\delta_2}. \end{aligned}$$

The sampling procedure retains only perturbations whose intervals remain inside the source video.

Progression distractors alter the target span's internal order. For reorder, $J_1, \ldots, J_m$ partitions (s, e) and π permutes $\{1, \ldots, m\}$; for repeat, $J \subset (s, e)$ is a short subinterval:

$$\begin{aligned} D_P^{\text{rev}} &= \text{reverse}(C_{s,e}), \\ D_P^{\text{reorder}} &= \text{concat}_{k=1}^{m} C_{J_{\pi(k)}}, \\ D_P^{\text{repeat}} &= \text{trim}_\ell(C_J \mid C_J \mid \cdots). \end{aligned}$$

B Frame Sampling

B.1 Bin-Centered Sampling

For each input clip, we divide the interval (a, b) into n equal bins and sample each bin center:

$$t_i = a + \left(i + \frac{1}{2}\right) \frac{b-a}{n}, \quad i = 0, \ldots, n-1.$$

Decoded frames are selected from these timestamps. Unlike endpoint-inclusive sampling, this keeps the first and last samples inside the clip rather than exactly at a or b. This reduces boundary-frame leakage, since visible context clips meet the missing middle at exact temporal boundaries and some candidates touch or overlap those boundaries.

B.2 Boundary Sanity Checks

B.2.1 Exact Boundary-Frame Matches

We test whether each candidate's first or last 1–3 sampled frames exactly match the adjacent context boundary. Across 1,521 videos per dimension and 18,252 candidate clips, Table 5 shows zero matches.

Table 5: Exact boundary-frame matches in sampled candidates.

Dimension	Clips	Begin-side			End-side		
		1f	2f	3f	1f	2f	3f
S	6084	0	0	0	0	0	0
A	6084	0	0	0	0	0	0
P	6084	0	0	0	0	0	0
Overall	18252	0	0	0	0	0	0

B.2.2 Edge Baseline

Edge-1 uses each candidate's first and last frame; Edge-4 uses its first and last four. We average DINOv2 similarities between candidate edges and adjacent context frames, then normalize the four scores into probabilities.

Table 7 reports mean ground-truth probabilities for Semantic, Alignment, and Progression; Mean averages them and is not top-1 accuracy. Values only slightly above the uniform 25% show that edges can occasionally be decisive but are not substantial overall.

B.2.3 Single-Frame Baseline

For Seed1.8 and Gemini2.5-Pro, we represent every beginning, ending, and candidate clip by its

Table 6: Joint accuracy and video-level breakdown on TempCloze. Joint columns count videos where the listed dimensions are all correct; 0/3–3/3 count exact numbers of correct dimensions. The $S+A+P$ column equals 3/3.

Model		Joint Accuracy (%)				Video-Level Breakdown (%)			
Name	Think	$S+A\uparrow$	$S+P\uparrow$	$A+P\uparrow$	$S+A+P\uparrow$	0/3	1/3	2/3	3/3
Proprietary Models									
Seed1.8	✓	**74.88**	**89.55**	**72.58**	**70.81**	0.46	4.14	24.59	**70.81**
Qwen3.5-Plus	✓	71.23	82.88	71.56	67.22	1.25	7.50	24.03	67.22
Seed1.8	✗	60.82	89.15	57.92	56.94	0.66	5.33	37.08	56.94
Gemini2.5-Pro	✓	62.76	62.17	46.45	43.95	2.57	13.95	39.54	43.95
Gemini2.5-Flash	✗	52.99	64.89	48.39	44.05	5.39	16.44	34.12	44.05
GPT5.4	✗	29.85	47.01	25.64	21.17	10.65	29.19	38.99	21.17
Claude4.6-Sonnet	✗	21.04	42.21	22.68	16.63	13.87	33.46	36.03	16.63
Claude4.6-Opus	✓	25.79	35.39	27.43	18.88	18.03	31.12	31.97	18.88
Gemini3-Flash	✗	27.74	31.89	21.76	15.91	15.91	34.52	33.66	15.91
Seed1.6	✗	13.63	40.16	11.64	9.38	18.35	34.97	37.30	9.38
Grok4.1	✗	5.19	5.06	5.26	0.72	42.47	43.46	13.35	0.72
Open-source Models									
Qwen3.5-397B-A17B	✓	41.62	**62.26**	**42.14**	**35.77**	5.46	20.05	38.72	**35.77**
Qwen3.5-35B-A3B	✓	**42.46**	57.08	38.89	33.80	7.67	21.49	37.04	33.80
KimiK2.5	✗	33.53	49.77	30.51	25.51	10.06	27.15	37.28	25.51
Qwen3VL-32B-I	✗	7.30	37.34	7.63	5.46	21.30	37.34	35.90	5.46
InternVL3.5-38B	✗	16.40	22.62	20.22	11.35	29.58	33.89	25.19	11.35
Qwen3VL-8B-I	✗	6.92	16.21	7.58	3.49	35.51	40.78	20.22	3.49
InternVL3.5-8B	✗	8.15	9.60	10.06	3.81	39.84	39.97	16.37	3.81
InternVL3-38B	✗	8.81	10.91	7.43	2.96	40.96	37.80	18.28	2.96
KimiVL-A3B-T	✓	6.96	6.67	6.96	2.17	37.83	45.94	14.06	2.17
Qwen3VL-32B-T	✓	5.92	6.58	6.71	1.91	41.25	43.36	13.49	1.91
LLaVACriticR1-7B	✓	6.99	5.79	6.85	1.73	41.78	42.05	14.44	1.73
Qwen2.5VL-7B-I	✗	6.47	8.03	7.42	2.11	42.07	40.23	15.59	2.11
GLM4.6V-Flash	✗	5.02	10.43	7.99	3.04	44.36	38.28	14.32	3.04
Qwen3VL-4B-T	✓	7.23	5.79	5.85	1.64	41.62	42.80	13.94	1.64
Qwen3VL-8B-T	✓	6.84	5.65	6.25	1.64	42.47	42.08	13.81	1.64
ThinkLiteVL-7B	✓	6.04	5.37	5.84	1.33	42.20	43.20	13.27	1.33
MiMoVL-7B-RL	✗	7.43	6.38	8.42	2.76	46.09	37.21	13.94	2.76
Qwen3VL-4B-I	✗	6.05	7.36	7.63	2.76	45.69	38.79	12.75	2.76
KimiVL-A3B-I	✗	8.25	5.98	6.87	2.75	47.77	36.63	12.85	2.75
MiMoVL-7B-SFT	✗	6.05	6.57	6.44	2.24	47.14	38.26	12.36	2.24
Molmo2-8B	✗	6.18	6.18	5.65	1.88	46.51	39.25	12.37	1.88

Table 7: Edge baseline, measured by the mean probability (%) assigned to the ground-truth candidate.

Setting	S	A	P	Mean
Edge-1	26.69	26.06	26.29	26.35
Edge-4	25.91	24.91	25.66	25.49

midpoint frame. Table 8 compares this setting with the original multi-frame evaluation.

Table 8: Original multi-frame evaluation versus the single-frame baseline. All values are percentages.

Name	S↑	A↑	P↑	Mean↑	≥1↑	≥2↑	3/3↑
Seed1.8	96.25	61.93	92.11	83.43	99.35	94.02	56.94
Seed1.8 (Single)	**38.72**	**6.97**	**6.25**	**17.31**	**44.64**	**6.77**	**0.53**
Gemini2.5-Pro	90.33	66.67	67.78	74.92	97.44	83.49	43.95
Gemini2.5-Pro (Single)	**40.50**	**8.81**	**7.89**	**19.07**	**48.13**	**8.55**	**0.53**

C Human Baseline

We estimate a Human Baseline on 100 randomly selected videos from TempCloze. Five annotators completed the task after a short briefing. A Streamlit interface showed the same inputs as model evaluation: beginning clip, ending clip, and four candidate middle clips (Figure 11).

Each annotator answered 120 questions. Every question received two independent judgments. We count a question as correct only if both annotators selected the ground-truth clip; all disagreements count as incorrect.

Under this strict rule, humans solve 96/100 Semantic, 98/100 Alignment, and 97/100 Progression questions, for 291/300 = 97.0% mean dimension accuracy. At the video level, humans solve at least one dimension for 100/100 videos, at least two for 99/100, and all three for 92/100.

D Source Datasets

D.1 Dataset Overview

LVD-2M (Xiong et al., 2024) is a large-scale open-domain video dataset collected from the web. It contains approximately 2 million long-take videos, each lasting more than 10 seconds with continuous motion. The dataset provides temporally

Table 9: Source-level filtering statistics.

Source	Input	Duration		LLM		Quality		Motion		Filtered (%)
		Count	Share (%)	Count	Share (%)	Count	Share (%)	Count	Share (%)	
CaReBench	1,000	437	43.70	73	12.97	86	17.55	233	57.67	90.60
Daily-Omni	653	0	0.00	222	34.00	154	35.73	161	58.12	93.42
EgoLife	2,347	1	0.04	1,391	59.29	0	0.00	265	27.87	81.38
FAVOR	1,776	503	28.32	503	39.51	104	13.63	414	62.82	91.84
LVD-2M	82,196	75,252	91.55	3,843	55.34	156	5.03	192	6.52	99.37
MiraData	1,972	–	–	705	35.75	31	2.49	4	0.33	89.96
Video-TT	791	–	–	265	33.50	151	35.53	144	52.55	88.75

dense captions generated through a hierarchical annotation pipeline, describing events at fine-grained intervals. Videos cover varied scenes, camera motion, and durations, supporting models that learn extended spatiotemporal dependencies.

Video-TT (Zhang et al., 2025b) contains 1,000 short video clips sourced from YouTube Shorts. Each video is paired with one open-ended question and four natural adversarial variants for visual and narrative comprehension. The dataset spans daily life, comedy, sports, and 18 question types covering element attributes, event sequencing, and plot understanding. Videos were filtered for length (under 65 seconds), content safety, and frame sampling quality, with rationales curated to support reasoning under question variation.

FAVOR-Bench (Tu et al., 2025) focuses on fine-grained actions and temporally sensitive visual states. It contains short to medium-length clips that capture changes in object states or human motions. Videos are annotated for temporal dynamics, scene transitions, and event variations, supporting analysis of temporal reasoning.

CaReBench (Xu et al., 2025) is a curated video dataset with temporal descriptions. It consists of moderate-length videos depicting indoor and outdoor everyday activities. Each video includes annotations for scene content and event transitions, with structured temporal information separating spatial and temporal aspects of the scene. The dataset supports fine-grained evaluation of video comprehension and retrieval.

Daily-Omni (Zhou et al., 2026) consists of daily-life videos paired with captions describing human activities and object interactions. Videos are recorded across different environments and scenarios, capturing common daily behaviors. The annotations are aligned to observable actions and interactions.

EgoLife (Yang et al., 2026) is a dataset of first-person life-logging videos recorded from head-mounted or body-mounted cameras. It contains continuous footage of everyday activities across multiple participants, capturing egocentric perspectives, motion patterns, and interactions. Audio is included in the original recordings, but the dataset focuses mainly on visually grounded action annotations.

MiraData (Ju et al., 2024) provides densely captioned videos with event descriptions. Clips vary in length and are filtered and deduplicated for quality and continuity. Structured captions describe objects, actions, background, and motion patterns at multiple granularities. The dataset contains hundreds of thousands of videos, making it suitable for temporal reasoning and continuous video understanding.

D.2 Filtering Statistics

Table 9 reports the filtering ratio at each stage. For sources without captions, we first generate fine-grained captions with Gemini-2.5-Pro, then use the available or generated captions for LLM suitability screening. Overall, duration filtering removes the largest number of videos, mainly because long web-scale sources contain many clips outside our target length range. Among later stages, LLM suitability and motion filtering are the most selective.

Retained videos then enter the same gap-generation stage: sample a central missing span, validate it with Farnebäck optical flow, and allow up to three resampling attempts. Stage percentages are filter-conditional; the total column is relative to the original source input.

D.3 Human Validation

We conduct a small-scale human validation on the filtered videos. Five annotators each review 60 videos, for 300 videos in total. Each annotator

Table 10: Gemini2.5-Pro statistics by source.

Source	Dimension Accuracy (%)				Cumulative Accuracy (%)		
	S↑	A↑	P↑	Mean↑	≥ 1 ↑	≥ 2 ↑	3/3↑
CaReBench	91.49	64.89	62.77	73.05	94.68	81.91	42.55
Daily-Omni	93.02	79.07	60.47	77.52	97.67	83.72	51.16
EgoLife	83.52	59.73	66.13	69.79	95.88	78.03	35.47
FAVOR-Bench	95.17	64.83	75.86	78.62	99.31	84.83	51.72
LVD-2M	93.19	68.54	69.71	77.15	98.44	87.35	45.91
MiraData	94.95	78.28	69.70	80.98	98.99	90.40	53.54
Video-TT	86.52	62.92	56.18	68.54	95.51	71.91	38.20
Overall	**90.33**	**66.67**	**67.78**	**74.92**	**97.43**	**83.49**	**43.95**

Table 11: Qwen3.5-35B-A3B statistics by source.

Source	Dimension Accuracy (%)				Cumulative Accuracy (%)		
	S↑	A↑	P↑	Mean↑	≥ 1 ↑	≥ 2 ↑	3/3↑
CaReBench	61.70	33.33	51.06	48.70	80.65	50.54	15.05
Daily-Omni	86.05	62.79	81.40	76.75	95.35	83.72	51.16
EgoLife	79.86	53.09	71.10	68.02	94.95	73.62	35.55
FAVOR-Bench	76.92	51.75	75.17	67.95	93.66	75.35	36.62
LVD-2M	70.23	47.96	65.24	61.14	89.88	64.40	29.38
MiraData	76.26	61.42	74.49	70.72	94.36	78.97	39.49
Video-TT	83.15	56.18	87.64	75.66	97.75	84.27	44.94
Overall	**75.10**	**51.55**	**69.96**	**65.54**	**92.33**	**70.83**	**33.80**

rates whether a video is suitable for the cloze task on a five-point scale: 1 = Unusable, 2 = Somewhat unsuitable, 3 = Acceptable, 4 = Somewhat suitable, and 5 = Very suitable. The Streamlit front end used for this review is shown in Figure 12.

Table 12: Human validation of filtered video suitability.

Annotator	1	2	3	4	5	Total	Avg.
No.1	0	2	10	14	34	60	4.33
No.2	0	5	17	27	11	60	3.73
No.3	0	2	3	34	21	60	4.23
No.4	0	4	4	12	40	60	4.47
No.5	0	3	16	25	16	60	3.90
Overall	**0**	**16**	**50**	**112**	**122**	**300**	**4.13**
Share (%)	0.0	5.3	16.7	37.3	40.7	100.0	–

Table 12 shows that the filtered videos are generally suitable for the cloze task. The overall mean rating is 4.13 out of 5, with 234/300 ratings (78.0%) in the two positive categories and 284/300 ratings (94.7%) at least acceptable. Importantly, rating 1 (Unusable) occurs 0 times, indicating that the filtering pipeline removes clearly invalid cases.

E Statistics by Sources

Tables 10 and 11 report source-level results for two strong models. Metrics follow Table 2: S/A/P are dimension accuracies, Mean averages them, and cumulative columns count dimensions solved on the same video.

Gemini2.5-Pro is consistently strong on Semantic, while Alignment remains lower. MiraData gives its best 3/3 score.

Qwen3.5-35B-A3B varies more by source and is strongest on Daily-Omni and Video-TT. For both models, high ≥ 1 accuracy with much lower 3/3 accuracy shows that failures are often partial.

F Additional Analysis on Main Results

F.1 Joint Accuracy

Pairs involving Alignment are consistently weaker than $S{+}P$. Proprietary models average 40.54% on $S{+}A$ and 37.39% on $A{+}P$, versus 53.67% on $S{+}P$; open-source models show the same pattern at lower levels (11.93%, 12.06%, and 16.79%). Seed 1.8 (thinking) leads proprietary joint accuracy, while Qwen3.5-35B-A3B and Qwen3.5-397B-A17B lead open-source models.

F.2 Video-Level Breakdown

The proprietary–open-source gap is larger at the video level: 1.9× in mean dimension accuracy (62.19% vs. 32.51%) but 4.6× in 3/3 accuracy (33.24% vs. 7.15%). Proprietary failures are often narrow 2/3 misses (31.88% vs. 19.34%); open-source failures concentrate in 0/3 and 1/3 cases. Qwen3.5-397B-A17B reaches 38.72% in 2/3 but 35.77% in 3/3, while the best proprietary models reach 67.22%–70.81% in 3/3. For Seed 1.8, thinking mainly shifts videos from 2/3 to 3/3 (37.08% to 24.59%, 56.94% to 70.81%).

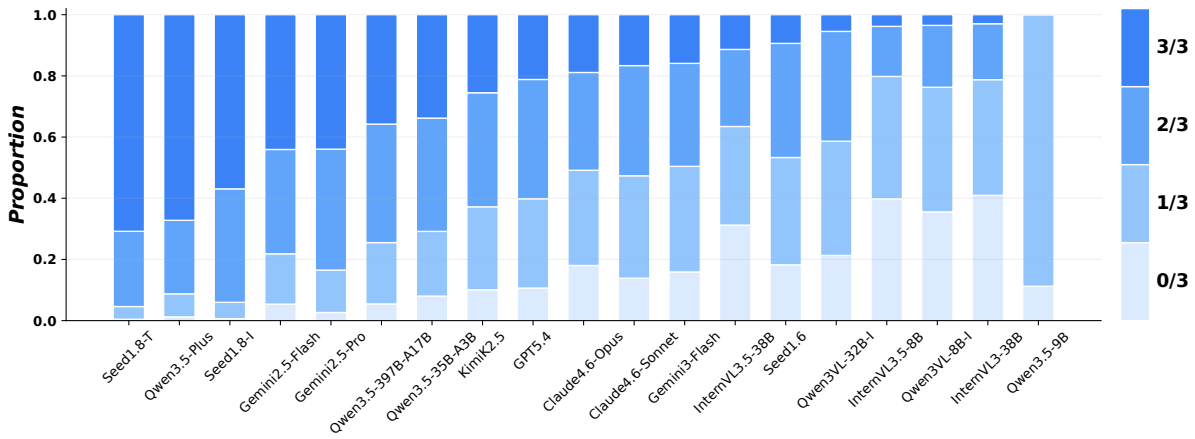
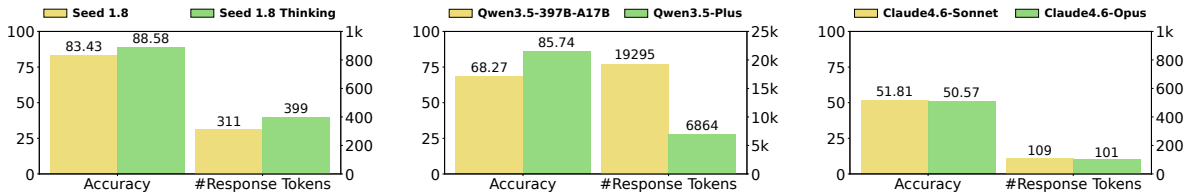


Figure 9: Video-level breakdown across models.



Seed with and without thinking | Qwen open-weight and hosted | Claude Sonnet & Opus

Figure 10: Thinking and mode comparisons: mean accuracy vs. average response tokens.

Table 13: Effect of thinking variants.

Pair	ΔS	ΔA	ΔP	ΔMean
Seed1.8	-1.18	+14.99	+1.64	+5.15
Qwen3VL-32B	-26.30	+14.47	-37.53	-16.46
Qwen3VL-8B	-6.49	+10.32	-20.87	-5.68
Qwen3VL-4B	+7.43	+5.13	-9.53	+1.01
KimiVL-A3B	+9.08	-1.26	+1.85	+3.24

F.3 Effect of Thinking

The Think column in Table 2 mixes proprietary reasoning modes and open-source thinking checkpoints; we use matched pairs where possible. Table 13 reports percentage-point changes from non-thinking or instruct baselines, and Figure 10 compares representative accuracy–token tradeoffs.

Thinking helps most when a strong base model stays visually grounded. Seed 1.8 gains 5.15 mean-accuracy points and 13.87 full-joint points, mainly from a 14.99-point Alignment gain; Semantic changes little.

Open-source pairs are mixed. Qwen3VL thinking checkpoints improve Alignment at all matched sizes but often lose more on Semantic and Progression. Qwen3VL-32B-T, for example, gains 14.47 Alignment points but drops 26.30 on Semantic and 37.53 on Progression relative to Qwen3VL-32B-I.

KimiVL-A3B is milder: thinking improves mean accuracy by 3.24 points, mainly through Semantic, but slightly lowers full joint accuracy. Overall, thinking is not uniformly beneficial on TempCloze; it helps only when extra reasoning improves temporal timing without losing visual or progression constraints.

G Prompts

G.1 Temporal Cloze Suitability Prompt

```
O3_PROMPT = """[Temporal Cloze Task Suitability]

You are judging whether the video described by
     the caption is suitable for a temporal cloze
      benchmark. In this benchmark, a video is
     split into three ordered visual segments:
- BEGINNING: the context shown before the
     missing middle.
- ANSWER: the missing middle segment that a
     model should choose or predict.
- ENDING: the context shown after the missing
     middle.

A good example lets a viewer infer a plausible
     ANSWER from BEGINNING and ENDING through
     visual temporal and causal continuity. Judge
      only what can be supported by visible
     actions or events in the caption. Ignore
     audio-only information, speech or dialogue
     content, subtitles, narration, background
     music, creator commentary, and any purely
     emotional or aesthetic judgment unless it is
      visually grounded.

Please consider these quality aspects:
- Scene transition: whether BEGINNING, ANSWER,
     and ENDING appear to belong to a coherent
```

```
    visual scene or a justifiable scene
    transition.
- Subject consistency: whether the main person,
    object, or environment remains consistent
    across the three segments.
- Action consistency: whether the motion or
    activity in ANSWER reasonably continues from
     BEGINNING and leads toward ENDING.
- Logical consistency: whether the full sequence
     makes temporal and causal sense as a middle
    -completion example.

PASS only if:
- The caption describes a concrete visual
    process, action, interaction, transformation
    , or event with ordered steps.
- The likely ANSWER segment would be visually
    distinguishable and meaningfully constrained
     by BEGINNING and ENDING.
- BEGINNING, ANSWER, and ENDING would form one
    coherent temporal sequence with stable
    subjects and a reasonable causal or physical
     progression.

REJECT if any of these apply:
- The caption is mostly static description,
    scenery, appearance, atmosphere, or a
    montage without a clear ordered visual
    action.
- The event depends mainly on dialogue, audio,
    text, subtitles, narration, or hidden intent
     rather than visible motion.
- The middle could be almost anything, is too
    ambiguous, or cannot be inferred from before
    /after visual context.
- There are abrupt unrelated scene cuts,
    inconsistent subjects or environments, or
    disconnected actions that would make
    BEGINNING, ANSWER, and ENDING incoherent.
- The video only shows repeated motion with no
    meaningful progress, or a single state with
    little temporal change.

Caption:
{caption}

Return valid JSON only, using one of these
    shapes:
{{"pass": true, "reason": "one or two concise
    sentences naming the decisive quality
    aspects"}}
{{"pass": false, "reason": "one or two concise
    sentences naming the decisive quality
    aspects"}}
"""
```

G.2 Temporal Cloze Evaluation Prompt

```
eval_prompt = """
You are given:
- **BEGINNING**: the first part of a video (
    frames in temporal order).
- **END**: the last part of the same video (
    frames in temporal order).
The middle segment between BEGINNING
and END was removed.

You will then see four candidate middle
segments, labeled A, B, C, D.
Each candidate is a short clip;
exactly one is the true middle that
connects BEGINNING to END.
The others are wrong.

Task: Which candidate is the correct middle?
    Choose one of A, B, C, D.

**Output JSON only** : {"answer": "<A, B, C, or
    D>", "reason": "<one or two sentences>"}
"""
```

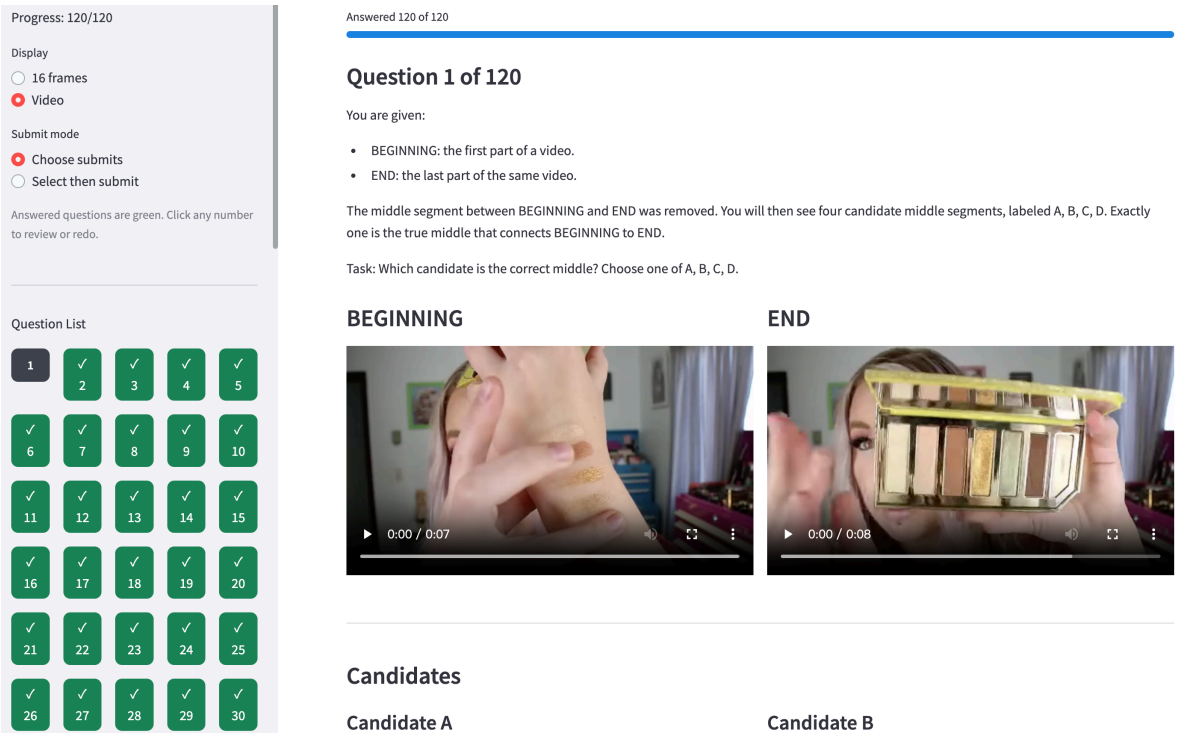


Figure 11: Human annotation interface. Annotators viewed the same clips as the models.

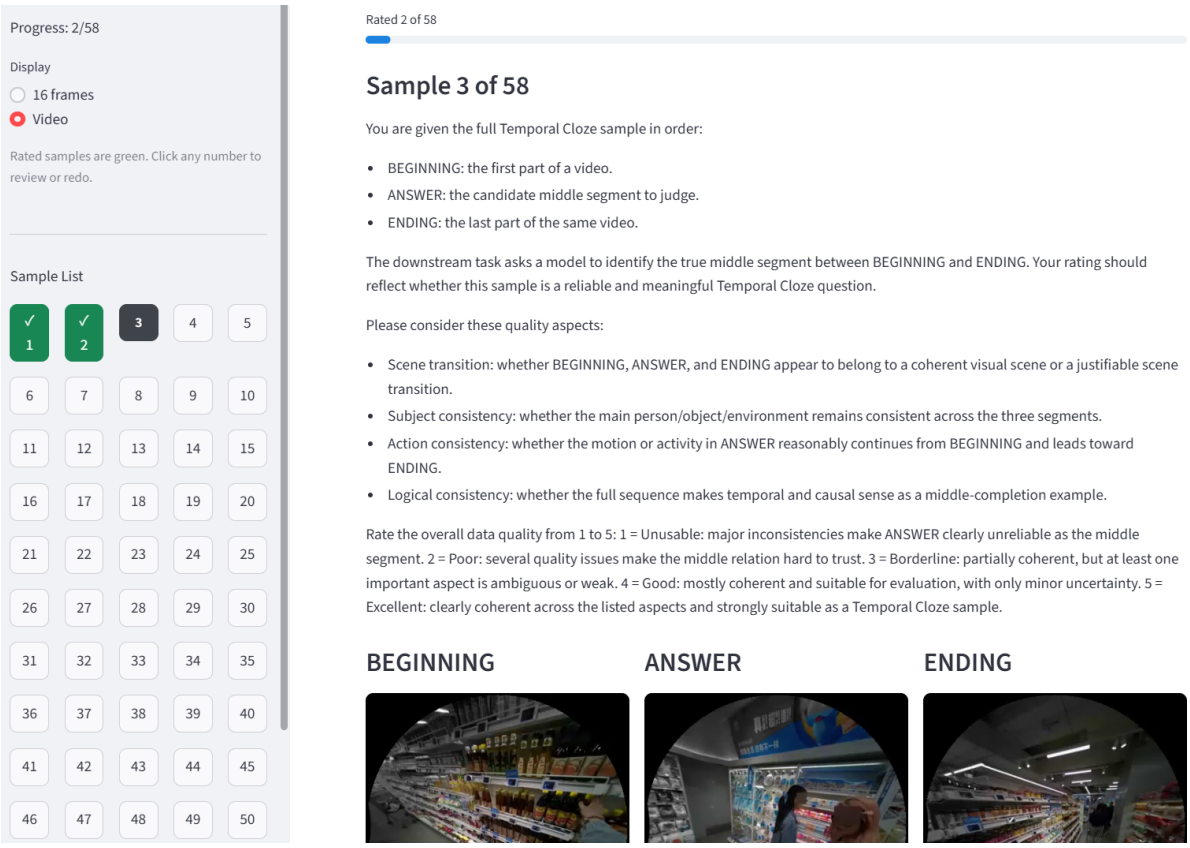


Figure 12: Streamlit interface for filtered-video quality review.