PERSISTQA: DIAGNOSING IDENTITY ASSOCIATION IN LONG-VIDEO QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

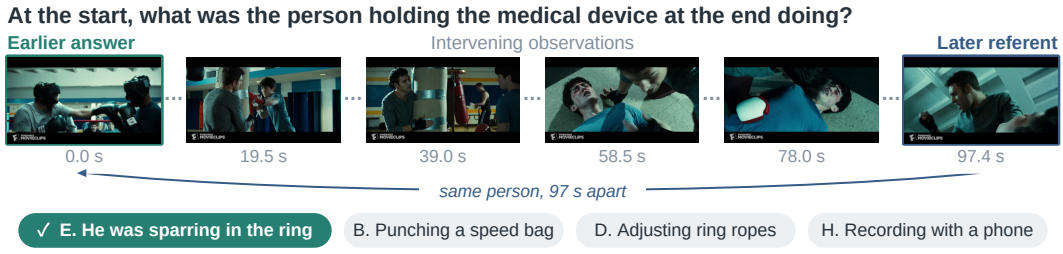

Figure 1: A PersistQA question. The later referent identifies the person, whose earlier activity is the answer. Four of eight options are shown, shortened, with the designated answer checked.

ABSTRACT

Long videos often contain people who reappear across different scenes and moments. Answering a question about one person therefore requires more than finding a relevant event. A model must recognize the intended person across time and associate the correct observations with that person. However, long-video question answering is typically evaluated using aggregate answer accuracy, which does not reveal where this process fails. We introduce PersistQA, a diagnostic benchmark for studying identity grounding in long-video question answering. PersistQA contains 3,299 eight-option questions over 409 videos. Each question identifies a person through visual or relational cues and asks about the same individual at another moment. Across 27 models, the best open-weight and closed-source systems achieve only 27.2% and 29.3% accuracy, compared with 60% for a human no-look-back baseline. Our diagnostics reveal several sources of error. Frozen CLIP features distinguish same-identity from different-identity pairs with 85.2% accuracy, while VLMs explicitly judging identity reach only 50.6–62.6%. Eight uniformly sampled frames miss approximately 53% of candidate-answer windows, and replacing one frame with candidate evidence improves accuracy by 4.3–9.7 percentage points. Even when candidate evidence is directly provided, 61–78% of questions remain incorrect. These results highlight identity association, evidence access, and evidence use as distinct challenges for long-video question answering.

1 INTRODUCTION

Long videos often contain people who disappear and reappear across scenes (Song et al., 2024, Rawal et al., 2024). A question may identify a person at one moment and ask about that same person much later. Finding a relevant event is therefore not enough. A model must also determine whether that event concerns the intended individual, which relates to person re-identification (Zheng et al., 2015, Ye et al., 2022). We refer to this ability as *identity grounding* (Figure 1).

This distinction is difficult to study with standard long-video question answering evaluations (Fu et al., 2025, Zhou et al., 2025, Wu et al., 2024). Aggregate answer accuracy tells us whether a model produced the correct answer, but not whether it connected the right observations to the right person. A model may miss the relevant moment. It may observe the moment but associate it with another person. It may also have access to useful evidence and still answer incorrectly. These behaviors lead

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.

to the same final error, yet they suggest different limitations. Telling these failures apart requires diagnostics that look beyond aggregate QA accuracy.

We introduce PersistQA, a diagnostic benchmark for studying identity grounding in long-video question answering. PersistQA contains 3,299 questions over 409 videos. Each question identifies an individual through visual or relational cues and asks about that same individual's property or activity at another, possibly distant, moment. To reduce language-only shortcuts (Goyal et al., 2017), questions are filtered using multiple language models. Nearly all questions also include manually verified candidate-answer windows. They are withheld during standard evaluation and instead enable controlled analyses of identity correspondence and access to relevant visual evidence. Table 1 compares this design with existing long-video benchmarks.

We evaluate 27 models and find that identity-grounded QA remains difficult. The strongest open-weight or dedicated long-video model reaches 27.2% accuracy, while the strongest closed-source model reaches 29.3%. A human no-look-back baseline reaches 60%. Our diagnostics provide additional context for this gap. Frozen CLIP (Radford et al., 2021) features distinguish same-identity from different-identity pairs with 85.2% accuracy, while explicit identity judgments from VLMs reach only 50.6% to 62.6%. Thus, visual correspondence can be recoverable even when explicit identity judgments remain unreliable. Evidence access also matters. Eight uniformly sampled frames miss about 53% of candidate-answer windows. Replacing one sampled frame with candidate evidence improves QA accuracy by 4.3 to 9.7 percentage points at the same frame budget. Yet substantial errors remain when inputs are restricted to candidate evidence. Across the evaluated models, 61% to 78% of questions are still answered incorrectly under this condition. Together, these results suggest that longer inputs (Song et al., 2024, He et al., 2024, Shen et al., 2025, Liu et al., 2025) or recoverable visual correspondence alone do not resolve identity-grounded QA. Models must reliably connect observations across time to the described individual while accessing and using relevant evidence. PersistQA is designed to measure these behaviors directly.

Our main contributions are: 1) **Identity grounding benchmark.** We introduce PersistQA, a diagnostic benchmark that tests whether long-video QA models can connect observations about the same individual across time, with 3,299 questions over 409 videos and verified candidate-answer windows; 2) **Targeted diagnostics.** We complement answer accuracy with identity correspondence probes and controlled frame interventions that examine how identity judgments and access to candidate evidence relate to QA performance; and 3) **Empirical diagnosis.** Across 27 models, we find a large gap from human performance and show that recoverable visual correspondence and increased evidence access do not by themselves yield reliable identity-grounded answers.

2 Related Work

Table 1: Benchmark design and diagnostics. I: QA models judge whether two crops show the same individual. R: explicit referring design, with descriptive reference as a task or construction objective. T: annotated evidence is inserted or removed at a fixed frame budget. ✗ means not reported rather than absent. Counts cover all MLVU tasks and NarrativeTrack entity clips.

	Benchmark size		Design		
Benchmark	Videos	Questions	I	R	T
Video-MME	900	2,700	✗	✗	✗
MLVU	1,730	3,102	✗	✓	✗
LongVideoBench	3,763	6,678	✗	✓	✗
CG-Bench	1,219	12,129	✗	✗	✗
Neptune	2,405	3,268	✗	✗	✗
NarrativeTrack	406	1,006	✗	✓	✗
Video-MME-v2	800	3,200	✗	✗	✗
PersistQA (ours)	409	3,299	✓	✓	✓

Benchmarks for long-video understanding. Video-MME, MLVU, and Neptune evaluate diverse perception and temporal reasoning (Fu et al., 2025, Zhou et al., 2025, Nagrani et al., 2025). Video-MME includes cross-appearance person tracking, and Video-MME-v2 adds grouped questions on temporal dynamics and entity persistence (Fu et al., 2026). LongVideoBench adds referring reasoning and object tracking, and NarrativeTrack evaluates entity continuity, changes, and ambiguity (Wu

et al., 2024, Ha et al., 2026). CG-Bench evaluates answer-supporting temporal clues, while KFS-Bench annotates disjoint scenes to evaluate sampling (Chen et al., 2024, Li et al., 2026). Table 1 compares explicit referring-query design and standalone identity and temporal-evidence diagnostics. PersistQA combines identity-grounded referent resolution with correspondence probes and controlled frame interventions to test how identity and evidence access affect answering.

Identity-conditioned video understanding. IDA-VLM and StoryTeller incorporate character identity into reasoning and description (Ji et al., 2025, He et al., 2025). ForeSea combines tracking and retrieval for person-centric QA and studies the role of scene context (Park et al., 2026). In contrast, PersistQA supplies no reference crop or identity track. The model must establish the described referent within the video and link that individual to answer evidence while retaining scene context.

Evidence selection and temporal grounding. Memory and compression extend long-video input (Song et al., 2024, He et al., 2024, Shen et al., 2025, Liu et al., 2025, Chen et al., 2025). Query-conditioned selectors balance relevance, coverage, and diversity (Tang et al., 2025, Zhang et al., 2025, Sun et al., 2025b), while F2C retains local clip context (Sun et al., 2025a). QSVideo, DAFS, and GCR use semantic retrieval, attention, and grounding-based replacement (Ao et al., 2026, Wang et al., 2026, Wei et al., 2026). SeViLA and R-VLM learn localization or retrieval with QA (Yu et al., 2023, Xu et al., 2026b). LongVideoAgent coordinates temporal grounding and observation, and TimeLens2 and SportsTime predict intervals or connect reasoning to temporal evidence (Liu et al., 2026, Zhu et al., 2026, Cao et al., 2026). ViTL and LeAdQA combine query-guided localization with answering, while REVEAL checks evidence sufficiency (Wang et al., 2025, Dong et al., 2025, Yan et al., 2026). A controlled allocation study separates selection from compression and reinvestment (Khatri, 2026). Our interventions test supplied observations at fixed budgets, and our exploratory selectors study allocation with frozen answerers and benchmark-informed development.

Person and video re-identification. Re-identification matches a query person to gallery images or tracks (Zheng et al., 2015, Ye et al., 2022). VCLIP learns temporal fusion and pedestrian-attribute features for video matching under clothing changes (Hill et al., 2025). DIFFER uses semantic descriptions and adversarial feature disentanglement to learn clothing-robust identity representations (Liang & Rawat, 2025). DisenQ uses structured language supervision to separate biometric, motion, and appearance features for identification across activities (Azad & Rawat, 2025). CSCI separates color-related appearance cues from identity features for image and video re-identification (Pathak & Rawat, 2025a). RLQ combines coarse attribute prediction with task-agnostic distillation to improve clothes-changing ReID under image degradation (Pathak & Rawat, 2025b). These methods improve correspondence representations. PersistQA instead tests whether a model resolves a linguistically described referent and answers from that individual's observations. Its pair probes and QA interventions distinguish recoverable identity correspondence from its use during answering.

3 The PersistQA Benchmark



Figure 2: Benchmark construction, from source videos to verified temporal windows.

We describe the task, question construction (Figure 2), and annotations, then summarize the composition. Given a video V, question q, and eight options $\mathcal{A}$, a model predicts $\hat{y} = f(V, q, \mathcal{A})$. The question identifies an individual through visual or relational cues and asks about that individual's property or activity. The observations establishing the referent and supplying the answer can occur at different times. The model receives neither annotations nor a reference crop.

Collection and question construction. We use footage from CinePile, MovieChat-1k, and LVU (Rawal et al., 2024, Song et al., 2024, Wu & Krähenbühl, 2021), resolving source identifiers before counting videos. Model-assisted construction specifies a visual referent, a queried property,

and a designated answer for each question, and it records a textual temporal anchor such as a scene transition or a preceding event. A separate prompt generates seven distractors matched in style and length. Text processing removes duplicate options and redistributes designated-answer positions. A twelve-model language-only committee removes items when at least five models select the designated answer or at least eight select the same wrong option. Its members come from model families that are also evaluated, so post-filter committee accuracy reflects the construction procedure rather than independent resistance to language-only shortcuts. The archived construction report retains 3,299 of 7,390 candidates, removing 4,091 (55.3%). Appendix A gives construction details, the committee, and the threshold sweep. Table 1 compares evaluation designs.

Annotations. Each item receives capability, referral, difficulty, and challenge labels. A separate stage uses the question, designated answer, and 32 uniform frames to mark candidate answer observations, yielding manually verified temporal windows for 3,245 questions. These coarse windows support diagnostics but may omit context needed to establish the referent. Their timestamps are withheld from standard evaluation inputs. Appendix A distinguishes these annotations from the textual temporal-anchor metadata used in selector controls.

Composition. The videos last 56–603 seconds, with a median of 198 seconds. Figure 3 summarizes video structure and question design. Automatically detected shot changes count visual discontinuities, with a median of 58 per video. *Maximum concurrent detections* is the largest number of detected people in any of eight uniformly sampled frames (median two). It describes sampled within-frame crowding and does not count distinct individuals or cover the whole video. Five capability categories describe *what is asked*, with a relatively balanced distribution in the middle panel. Referral labels describe *how the individual is identified*, and the most frequent strategies combine several cues or use activity and appearance. A question can therefore refer through one property while asking about another. Ten challenge categories include 987 cross-scene, 442 multi-hop, 406 appearance-change, and 456 single-shot questions. These are construction categories, not verified requirements for every answer. Full definitions and counts appear in Appendix A.

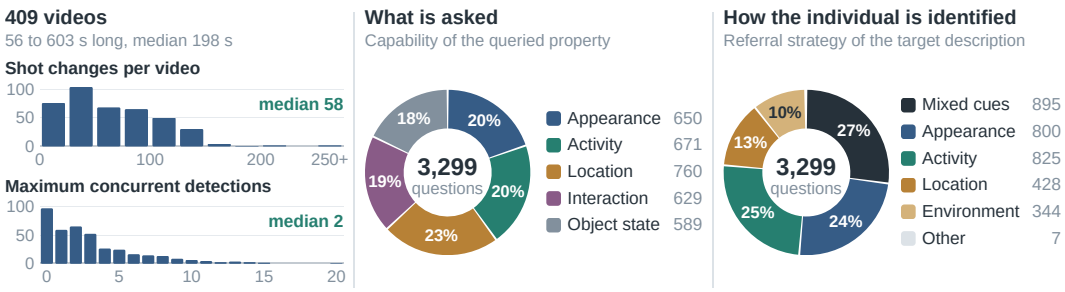

Figure 3: Benchmark composition. Other combines the interaction and object-state referral labels.

4 Evaluation on PersistQA

We evaluate 27 models, including fifteen open-weight models with eight-frame inputs, seven dedicated long-video models, and five closed-source models with eight-frame inputs. Counts include sizes from the same family. Groups describe input handling, and some eight-frame models also compress visual tokens. Appendix B gives the model families, input adapters, and response parsers. Eight frames provide a common answering budget, and larger-budget results are reported in Appendix I. Accuracy is the fraction of designated-option predictions that are correct, with absent or invalid responses counted as incorrect. Chance is 12.5%. Table 2 reports overall and five capability scores for 25 configurations. Two additional configurations appear in Table 7 in Appendix B.

We first compare overall model accuracy on all 3,299 questions (Table 2). Eight-frame open-weight models range from 13.6% to 26.4%, led by VideoChat-Flash-7B. Dedicated long-video models range from 11.3% to 27.2%, led by TimeViper-9B. Their ranges overlap, so accepting longer inputs does not consistently correspond to higher accuracy. These differences do not isolate an architectural effect, since training, computation, preprocessing, and sampling also vary across models. Closed-source models range from 16.2% to 29.3%, led by GPT-5.6 Terra. All three groups answer

Table 2: Accuracy (%) by model group. APP, ACT, LOC, INT, and OBJ denote the appearance, activity, location, interaction, and object-state capabilities. Bold and underline mark each group's best and second-best values, and darker purple means higher accuracy.

Model	All ↑	Capability accuracy					Model	All ↑	Capability accuracy				
		APP	ACT	LOC	INT	OBJ			APP	ACT	LOC	INT	OBJ
Open-weight							**Long-video implementations**						
Video-LLaVA	13.55	12.8	14.0	13.4	13.8	13.8	MA-LMM-7B	11.34	10.3	10.0	11.2	12.9	12.6
Qwen3-VL-2B	16.55	17.4	16.8	15.9	16.5	16.1	LongVU-7B	17.94	19.5	18.2	18.3	18.0	15.4
Qwen3-VL-8B	17.37	20.2	17.0	16.7	16.9	16.1	Video-XL-7B	18.10	19.1	20.3	18.0	16.9	16.0
Qwen2.5-VL-7B	17.46	19.2	19.8	18.7	14.1	14.8	Video-XL-Pro-3B	19.25	17.2	21.8	21.2	18.1	17.3
Gemma3-4B	18.22	18.6	19.2	19.1	15.9	18.0	StreamKV-7B	22.70	25.5	23.8	21.7	20.3	22.1
InternVL3-2B	20.89	23.4	20.1	21.2	20.5	19.0	BIMBA-7B	25.07	30.6	24.4	24.2	23.7	**22.2**
Gemma3-12B	21.28	24.0	20.7	21.7	20.5	19.2	TimeViper-9B	**27.19**	**30.9**	**28.5**	**28.7**	**25.0**	22.1
Ovis2.5-2B	22.22	23.7	20.9	23.6	22.7	19.9	**Closed-source**						
Ovis2.5-9B	23.01	25.4	22.2	26.2	20.2	20.2	GPT-5.4-nano	16.22	19.4	14.6	18.2	13.5	14.9
InternVL3-8B	23.07	25.7	23.5	24.1	20.0	21.6	GPT-5.4-mini	22.95	24.2	22.7	24.3	20.7	22.6
VideoChat-Flash-2B	25.37	28.6	24.9	27.5	23.1	22.1	GPT-5.5	28.04	31.5	24.0	**31.2**	**27.2**	25.6
InternVL3-14B	25.43	**29.1**	23.1	27.5	22.7	**24.3**	Gemini-3.1 Flash-Lite	28.22	31.5	26.2	29.2	**27.2**	26.7
VideoChat-Flash-7B	**26.37**	28.2	**27.4**	**28.8**	**23.7**	22.9	GPT-5.6 Terra	**29.34**	**33.1**	**26.4**	30.8	26.6	**29.7**

fewer than one third correctly despite different input budgets and training. Appearance questions are the easiest for 16 of the 25 configurations, while interaction or object-state questions are the hardest for 18, with mean accuracies of 23.6%, 20.0%, and 19.8%, respectively. We also examine shared failures. The 26 models with predictions for more than 95% of the benchmark share 2,966 questions. GPT-5.4-mini falls outside this criterion but remains in the full leaderboard. Among the shared questions, 647 (21.8%) receive no correct answer, and none is answered correctly by every model. This identifies a common set of failures whose causes cannot be determined from aggregate accuracy alone. Appendix C reports question-level agreement.

Human baseline. A human baseline achieved 60% accuracy under a no-look-back protocol, in which the video could not be revisited while answering. This provides a complementary reference for the task with answers based on the viewer's retained observations.

Challenge types and input coverage. We next study how accuracy varies with the queried property and intended association. Figure 5c averages original QA accuracy across 27 models for each challenge. Appendix C gives capability profiles and the full breakdown. Mean accuracy is 23.0% on 987 cross-scene questions and 18.2% on 442 multi-hop questions. The lower score for reference chains suggests that crossing a scene boundary alone does not explain the observed difficulty. Occlusion recovery and role swaps have only 25 and 10 questions, respectively, so their averages do not support a precise ranking. Automatic estimates of referent and answer positions for 3,160 questions show no monotonic accuracy decrease with temporal separation or with the number of intervening cuts (Appendix H). These estimates can select an incorrect referent and do not show that distance is irrelevant under verified identities. Additional frames improve several models. InternVL3-14B rises from 25.4% at eight frames to 28.9% at thirty-two (Appendix I). Across 22 models with reported sizes, log parameter count has Pearson correlation $r = 0.16$ with accuracy. This correlation describes a heterogeneous model set, without controlling for architecture or training. We next hold the answerer and frame budget fixed to test whether the supplied observations affect its answers.

5 Discussion and Analysis

We examine identity correspondence, access to candidate answer evidence, and responses when that evidence is supplied. Each analysis targets a different part of identity-grounded answering.

Identity correspondence and explicit judgments. We compare visual correspondence with explicit identity decisions using 2,848 balanced same/different pairs from 7,055 tracked crops in 114 videos. Tracker and linker assignments provide initial pair labels, which an author manually verified. Pairs share a video and coarse entity category, with gaps from ten seconds to over two

minutes. VLMs judge side-by-side crops, while frozen CLIP uses cosine similarity and a cross-validated threshold (Radford et al., 2021). Appendix D.1 gives the protocol, retrieval comparison, and tracker-based identity conditioning. VLM agreement with verified labels ranges from 50.6% to 62.6%, versus CLIP's 85.2% (Figure 4). OSNet, a person re-identification model (Zhou et al., 2019), reaches 62.8% (65.7% on person pairs), close to the strongest VLMs. Some VLMs answer *same* for most pairs, 86% for Gemma3-12B and 93% for VideoChat-Flash-7B. Longer gaps contain more different-identity pairs, 60% beyond two minutes versus 44% at 10–30 seconds, so the time-gap panel averages accuracy over same and different pairs. On this measure the same-biased models stay near chance at every gap, and CLIP stays at 83–86%. The other VLMs lose 4–7 points from the shortest to the longest gaps, and the loss comes from same-identity pairs. InternVL3-14B recognizes 70% of them at 10–30 seconds but 53% beyond two minutes, while its accuracy on different-identity pairs stays near 60%. InternVL3-8B's own features reach 84.5% rank-1 retrieval accuracy and 74.6% mAP, while its binary judgments reach 61.6%. These tasks use different candidate sets and decision rules, and shared clothing and backgrounds can also support matching. They therefore do not quantify information lost during answering. We also rendered an estimated referent track into the input, crossing four trackers with five linking representations. Averaged over seventeen answerers on their shared subset, these inputs reach 18.5–18.9% accuracy, versus 19.7% on raw video (Table 10), although rendering also suppresses surrounding context. *Visual correspondence can be recoverable from frozen features even when explicit identity judgments remain unreliable.*

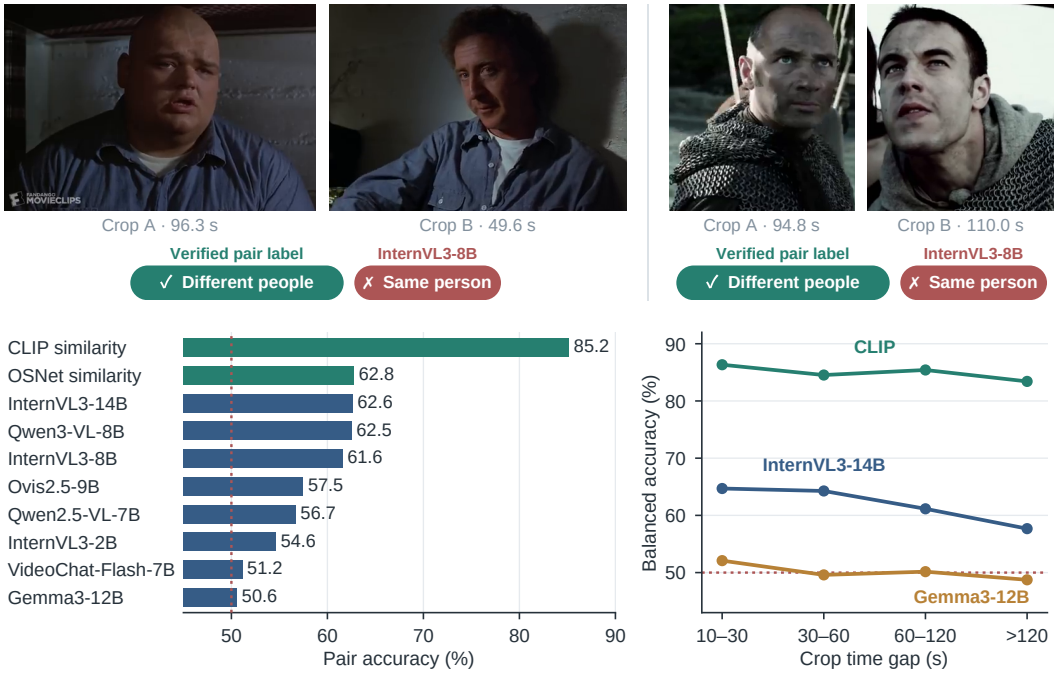

Figure 4: Identity correspondence. Top: selected different-person pairs misjudged as one person by InternVL3-8B. Bottom: accuracy on verified pairs, overall and by time gap, with chance dotted.

Temporal coverage and frame replacement. The temporal windows described in Section 3 mark candidate answer observations, with a median raw span of 4.4 seconds. Eight uniform frames miss approximately 53% of windows and sixteen miss 27%, allowing half an annotation-grid interval as tolerance (Figure 5a). Apparent saturation at 32 frames partly reflects the resolution of the annotation grid. We replace one of eight frames, fixing the question, options, answerer, and other seven frames. For missed windows, insertion swaps the nearest uniform sample for the window center. For covered windows, removal moves one covered frame outside the window. Separate subsets contain 300 paired questions each. This fixes the input budget but changes access to candidate evidence.

Insertion improves accuracy by 4.3–9.7 percentage points, while removal lowers it by 2.7–7.3 points (Figure 5b and Table 3). Two insertion effects and Ovis's removal effect remain significant after Holm correction. Different subsets and residual relevant frames can make the effects asymmetric.

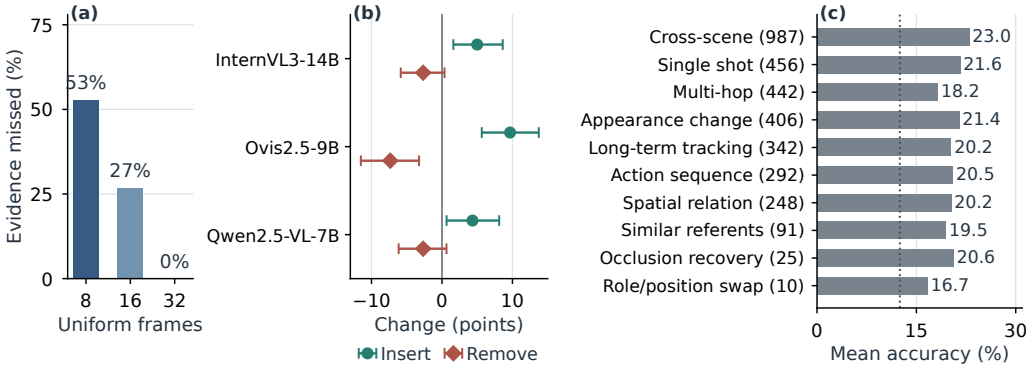

Figure 5: Evidence access and challenge profiles. (a) Missed windows. (b) One-frame insertion and removal effects with 95% intervals. (c) Mean QA accuracy by challenge, with chance dotted.

Table 3: Paired frame interventions. Δ is the accuracy change in points, with 95% video-bootstrap CIs. Bold marks effects that stay significant after Holm correction of exact McNemar tests.

Model	Improved/ regressed	Δ	95% CI	Exact p	Holm p
Evidence insertion					
InternVL3-14B	23/8	$\mathbf{+5.00}$	$[+1.62, +8.62]$	0.011	**0.043**
Ovis2.5-9B	37/8	$\mathbf{+9.67}$	$[+5.65, +13.74]$	1.54×10^{-5}	$\mathbf{9.22\times10^{-5}}$
Qwen2.5-VL-7B	21/8	$+4.33$	$[+0.66, +8.12]$	0.024	0.072
Evidence removal					
InternVL3-14B	6/14	-2.67	$[-5.84, +0.36]$	0.115	0.231
Ovis2.5-9B	8/30	$\mathbf{-7.33}$	$[-11.51, -3.25]$	4.72×10^{-4}	**0.002**
Qwen2.5-VL-7B	8/16	-2.67	$[-6.14, +0.65]$	0.152	0.231

Table 11 reports challenge-wise effects, where small slices show no consistent ordering. *Changing the supplied observations changes accuracy without increasing the frame budget.* The intervention does not separate better referent identification from improved perception or another answer cue.

Restricted inputs and internal allocation. We next test five compression-based models on 759 common questions using the full input, a timestamp hint, or only frames from the annotated window. Inputs contain RGB frames and text, with no audio stream or external subtitles. Hints add temporal instructions, while restriction concentrates all observations around the annotated candidate evidence. Hints add at most 0.9 accuracy points. Restriction gives larger gains for LongVU and VideoChat-Flash, reaching 31.9% accuracy, but 61–78% of questions remain incorrect in all three conditions (Figure 6). Restriction changes compression, distraction, and referent context together, so these responses cannot isolate a cause. Appendix D.3 and Table 12 give the protocol and scores.

InternVL3-8B's frame attention offers a complementary view (Figure 19 in Appendix F). Across twenty saved questions, mean attention favors later frame positions at both input budgets. This describes where visual attention is assigned, not whether those frames contain sufficient evidence or cause the answer. Matched-frame representations and projected option scores appear in the same appendix (Figure 21), together with normalization and cohort details. Further controls retain an estimated referent cue alongside candidate answer evidence while preserving full scenes (Appendix D.4). Their gains do not exceed a repeated answer-image control. *Candidate evidence access leaves substantial answering errors. Resolving them requires evaluating referent context and evidence use together, rather than either one in isolation.*

6 Exploratory Evidence Selection

The frame interventions motivate improving evidence access at a fixed answering budget. We consider two exploratory allocation strategies. *Frame selection* chooses individual observations, and

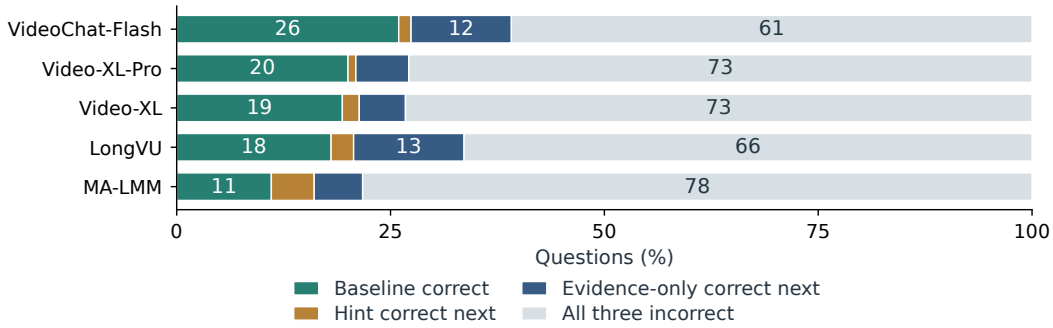

Figure 6: Restricted-input responses. Each question counts under the first condition that answers it correctly, in the order baseline, timestamp hint, and evidence-only input.

segment selection divides the budget across temporal windows. Figure 7 traces their shared scoring stage, alternative allocation branches, and frozen answerer.

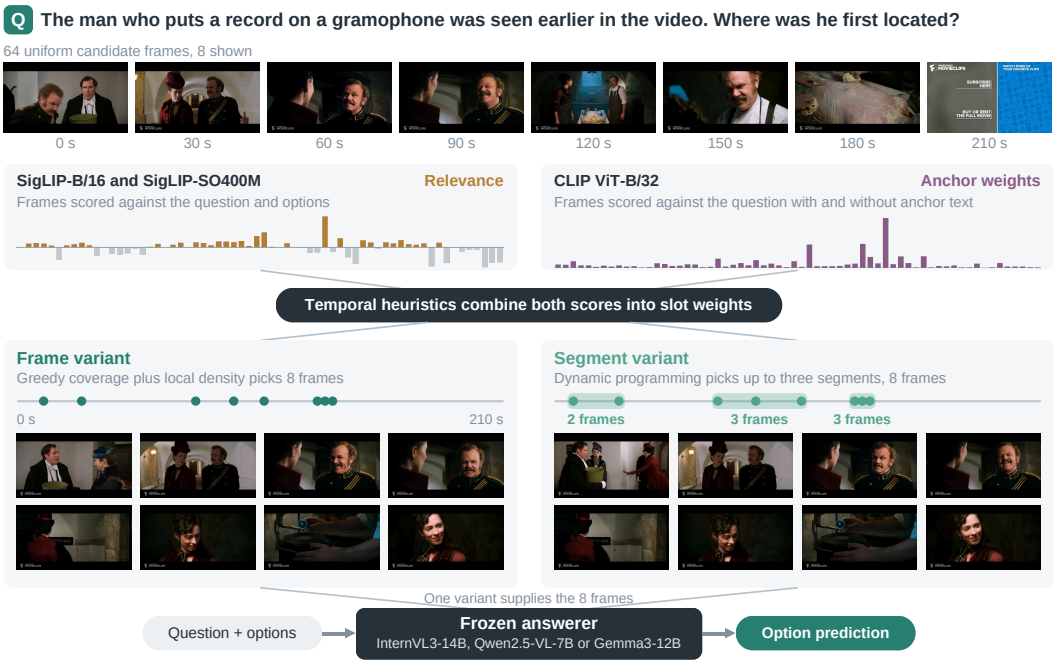

Figure 7: Exploratory selection and answering on one archived example.

We sample 64 uniform candidate frames. Frozen SigLIP-B/16 and SigLIP-SO400M score each frame against the question and its eight options, and their standardized similarities form a weighted relevance score. Frozen CLIP compares the question with and without the benchmark's additional temporal-anchor text, using score differences to weight candidate anchor moments (Zhai et al., 2023, Radford et al., 2021). This field can describe an answer moment rather than referent establishment. Three temporal heuristics combine relevance, anchor scores, and cumulative support in the question's temporal direction, producing distributions over candidate frames. The frame strategy combines these distributions through a normalized geometric mean. It greedily adds the frame that most increases an objective rewarding coverage of high-scoring moments and additional nearby observations. The segment strategy uses an arithmetic mean and dynamic programming to allocate the same total budget across at most three separated windows, sampling evenly within each. Its objective sums window values, which can count overlapping temporal support. The strategies therefore differ in both grouping and objective. One branch's selected frames reach the same frozen answerer

with the unchanged question and options. We use eight frames in the main comparisons and vary the follow-up budget from two to thirty-two. Full scoring and allocation definitions are in Appendix G (Tang et al., 2025, Sun et al., 2025b,a, Huang et al., 2026).

The *relevance* baseline uses the same SigLIP question-option scores to select high-scoring frames while suppressing nearby temporal candidates. It is the principal baseline without selector-specific training, alongside uniform sampling and three-seed Q-Frame (Zhang et al., 2025). We also compare with the temporal grounding method SeViLA (Yu et al., 2023), whose trained localizer remains frozen. Matched-input controls supply identical metadata to baseline scorers. We evaluate selection through two input protocols. The earlier rendered-clip comparison gives the frame variant 65.0% window coverage against SeViLA's 59.5%, with similar QA gains (Figure 8, with cohorts and tolerance in Appendix G). Window hits do not explain the accuracy ordering. A temporal hit measures overlap with an approximate annotation, whereas QA also requires readable answer content and context connecting it to the referent. The direct-frame comparison deterministically selects 300 questions from 300 videos independently of model outputs and excludes conflicting variants. Methods share candidate frames and each answerer's decoding path.

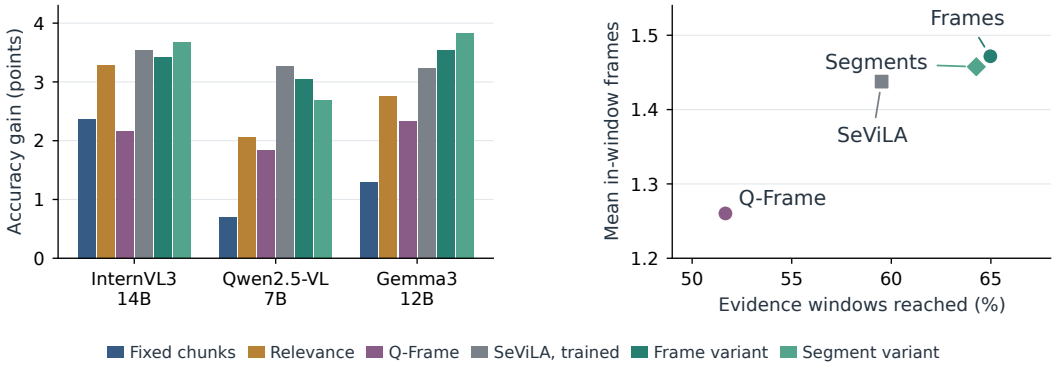

Figure 8: Exploratory selection results. Left: eight-frame accuracy gains over uniform sampling on 3,160 questions. Right: window coverage and in-window frames on 1,556 questions.

At eight frames, the frame variant gains 5.0, 7.3, and 7.3 points over uniform sampling for InternVL3, Qwen, and Gemma (Table 17). Relevance remains competitive, scoring 31.7% on Gemma against 31.3% for the full frame variant. Removing metadata changes accuracy by at most one point, with all six intervals including zero. Anchor-scoring controls show no detected gain after correction. Uncertainty, component controls, and 2–32-frame results are reported in Appendix G.4. Built with benchmark metadata and evidence labels, these exploratory selectors provide a foundation for future methods on PersistQA. Transfer to other benchmarks remains untested.

7 Conclusion

PersistQA tests whether models attribute video observations to a particular individual. Its diagnostics distinguish recoverable correspondence, evidence access, and correctness with candidate answer evidence supplied. Visual similarity remains recoverable despite explicit identity errors, and replacing one frame improves accuracy by up to 9.7 points at a fixed budget. Restricted inputs leave substantial errors, and exploratory selection increases coverage without proportional QA gains, motivating joint evaluation of localization and answering. The selectors offer exploratory starting points for future work. The controlled intervention establishes an evidence-access effect without excluding referent-grounding or perception failures. In these implementations, correspondence representations and longer inputs do not ensure that answers use observations about the intended individual.

Reproducibility Statement

The evaluation uses a fixed set of 3,299 question keys from 409 videos. Appendix B describes scoring, Appendix D.2 defines diagnostic inputs, and Appendix G specifies selector development. Saved benchmark records include questions, options, designated answers, source identifiers, construction metadata, and approximate evidence annotations. Evaluation code supplies prompts and parsers. Manifests, frame indices, predictions, and follow-up protocol hashes distinguish historical rendered-clip results from direct-frame comparisons. These records support reconstruction without embedding source videos in the benchmark annotations.

Disclosure of AI Assistance

Generative AI tools helped polish grammar, sentence structure, and terminology, supported literature retrieval and discovery, and produced preliminary section drafts from research notes and existing manuscript material. The resulting text and suggested sources were checked and revised.

Ethics Statement

The questions concern individuals in existing video collections and do not require personal names. This does not anonymize visible people or establish consent for new uses. Benchmark records reference source videos, and paper figures include selected frames for analysis. Public availability does not grant redistribution rights, which remain subject to source terms. The evaluation does not imply suitability for identifying, monitoring, or making consequential decisions about individuals.

References

Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. BoT-SORT: Robust Associations Multi-Pedestrian Tracking. *arXiv preprint arXiv:2206.14651*, 2022.

Wei Ao, Lan Wang, and Vishnu Naresh Boddeti. QSVideo: Query-Conditioned Semantic Temporal Retrieval for Video Understanding. *arXiv preprint arXiv:2607.04559*, 2026. doi: 10.48550/arXiv.2607.04559. URL https://arxiv.org/abs/2607.04559.

Shehreen Azad and Yogesh Singh Rawat. DisenQ: Disentangling Q-Former for Activity-Biometrics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13502–13512, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Azad_DisenQ_Disentangling_Q-Former_for_Activity-Biometrics_ICCV_2025_paper.html.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b.

Siyu Cao, Lu Zhang, Ruizhe Zeng, and Zhi yong Liu. Towards Temporal Compositional Reasoning in Long-Form Sports Videos. *arXiv preprint arXiv:2604.22226*, 2026.

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. *arXiv preprint arXiv:2412.12075*, 2024.

Weihua Chen, Xianzhe Xu, Jian Jia, Hao Luo, Yaohua Wang, Fan Wang, Rong Jin, and Xiuyu Sun. Beyond Appearance: A Semantic Controllable Self-Supervised Learning Framework for Human-Centric Visual Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15050–15061, 2023.

Yilong Chen, Xiang Bai, Zhibin Wang, Chengyu Bai, Yuhan Dai, Ming Lu, and Shanghang Zhang. StreamKV: Streaming Video Question-Answering with Segment-based KV Cache Retrieval and Compression. *arXiv preprint arXiv:2511.07278*, 2025.

Xinxin Dong, Baoyun Peng, Haokai Ma, Yufei Wang, Zixuan Dong, Fei Hu, and Xiaodong Wang. LeAdQA: LLM-Driven Context-Aware Temporal Grounding for Video Question Answering. *arXiv preprint arXiv:2507.14784*, 2025. doi: 10.48550/arXiv.2507.14784. URL https://arxiv.org/abs/2507.14784.

Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. StrongSORT: Make DeepSORT Great Again. *IEEE Transactions on Multimedia*, 25:8725–8737, 2023. doi: 10.1109/tmm.2023.3240881.

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, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, Yongkang Xie, Xiawu Zheng, Xue Yang, Haoyu Cao, Yunsheng Wu, Ziwei Liu, Xing Sun, Caifeng Shan, and Ran He. Video-MME-v2: Towards the Next Stage in Benchmarks for Comprehensive Video Understanding. *arXiv preprint arXiv:2604.05015*, 2026. doi: 10.48550/arXiv.2604.05015. URL https://arxiv.org/abs/2604.05015.

Gemma Team. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*, 2025.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Hyeonjeong Ha, Jinjin Ge, Bo Feng, Kaixin Ma, and Gargi Chakraborty. NarrativeTrack: Evaluating Entity-Centric Reasoning for Narrative Understanding. *arXiv preprint arXiv:2601.01095*, 2026. doi: 10.48550/arXiv.2601.01095. URL https://arxiv.org/abs/2601.01095.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13504–13514, 2024.

Shuting He, Hao Luo, Pichao Wang, Fan Wang, Hao Li, and Wei Jiang. TransReID: Transformer-Based Object Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15013–15022, 2021.

Yichen He, Yuan Lin, Jianchao Wu, Hanchong Zhang, Yuchen Zhang, and Ruicheng Le. StoryTeller: Improving Long Video Description through Global Audio-Visual Character Identification. *arXiv preprint arXiv:2411.07076v3*, 2025.

Cole Hill, Florence Yellin, Krishna Regmi, Dawei Du, and Scott McCloskey. Re-Identifying People in Video via Learned Temporal Attention and Multi-Modal Foundation Models. In *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)*, pp. 6259–6268, February 2025. URL https://openaccess.thecvf.com/content/WACV2025/html/Hill_Re-Identifying_People_in_Video_via_Learned_Temporal_Attention_and_Multi-Modal_WACV_2025_paper.html.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Yuning Huang, Xiaoyu Ji, Joseph Huang, Yichi Zhang, and Fengqing Zhu. Adaptive Greedy Frame Selection for Long Video Understanding. *arXiv preprint arXiv:2603.20180*, 2026.

Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Gedas Bertasius, and Lorenzo Torresani. BIMBA: Selective-Scan Compression for Long-Range Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29096–29107, 2025.

Yatai Ji, Shilong Zhang, Jie Wu, Peize Sun, Weifeng Chen, Xuefeng Xiao, Sidi Yang, Yujiu Yang, and Ping Luo. IDA-VLM: Towards Movie Understanding via ID-Aware Large Vision-Language Model. In *International Conference on Learning Representations*, volume 2025, pp. 52639–52652, 2025.

Prakhar Khatri. Select, Compress, Reinvest: A Controlled Study of Visual-Token Allocation in Long-Video MLLMs. *arXiv preprint arXiv:2609.03820*, 2026. doi: 10.48550/arXiv.2609.03820. URL https://arxiv.org/abs/2609.03820.

Siyuan Li, Li Sun, and Qingli Li. CLIP-ReID: Exploiting Vision-Language Model for Image Re-identification without Concrete Text Labels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1):1405–1413, 2023. doi: 10.1609/aaai.v37i1.25225.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, Yu Qiao, Yali Wang, and Limin Wang. VideoChat-Flash: Hierarchical Compression for Long-Context Video Modeling. *arXiv preprint arXiv:2501.00574v4*, 2025.

Zongyao Li, Kengo Ishida, Satoshi Yamazaki, Xiaotong Ji, and Jianquan Liu. KFS-Bench: Comprehensive Evaluation of Key Frame Sampling in Long Video Understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5643–5652, 2026.

Xin Liang and Yogesh S Rawat. DIFFER: Disentangling Identity Features via Semantic Cues for Clothes-Changing Person Re-ID. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13980–13989, June 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liang_DIFFER_Disentangling_Identity_Features_via_Semantic_Cues_for_Clothes-Changing_Person_CVPR_2025_paper.html.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5971–5984. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.342.

Runtao Liu, Ziyi Liu, Jiaqi Tang, Yue Ma, Renjie Pi, Jipeng Zhang, and Qifeng Chen. LongVideoAgent: Multi-Agent Reasoning with Long Videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 40404–40416. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1876.

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. Video-XL-Pro: Reconstructive Token Compression for Extremely Long Video Understanding. *arXiv preprint arXiv:2503.18478*, 2025.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 Technical Report. *arXiv preprint arXiv:2508.11737*, 2025.

Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep OC-Sort: Multi-Pedestrian Tracking by Adaptive Re-Identification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 3025–3029, 2023. doi: 10.1109/icip49359.2023.10222576.

Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/bf02295996.

Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, Mikhail Sirotenko, Yukun Zhu, and Tobias Weyand. Neptune: The Long Orbit to Benchmarking Long Video Understanding. *arXiv preprint arXiv:2412.09582v2*, 2025.

Hyojin Park, Yi Li, Janghoon Cho, Sungha Choi, Jungsoo Lee, Taotao Jing, Shuai Zhang, Munawar Hayat, Dashan Gao, Ning Bi, and Fatih Porikli. ForeSea: AI Forensic Search with Multi-modal Queries for Video Surveillance. *arXiv preprint arXiv:2603.22872*, 2026.

Priyank Pathak and Yogesh S. Rawat. Colors See Colors Ignore: Clothes Changing ReID with Color Disentanglement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16797–16807, October 2025a. URL https://openaccess.thecvf.com/content/ICCV2025/html/Pathak_Colors_See_Colors_Ignore_Clothes_Changing_ReID_with_Color_Disentanglement_ICCV_2025_paper.html.

Priyank Pathak and Yogesh S Rawat. Coarse Attribute Prediction with Task Agnostic Distillation for Real World Clothes Changing ReID. In *36th British Machine Vision Conference 2025 (BMVC)*. BMVA, 2025b. URL https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_346/paper.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763. PMLR, 2021.

Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. CinePile: A Long Video Question Answering Dataset and Benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, Zhuang Liu, Hu Xu, Hyunwoo J. Kim, Bilge Soran, Raghuraman Krishnamoorthi, Mohamed Elhoseiny, and Vikas Chandra. LongVU: Spatiotemporal Adaptive Compression for Long Video-Language Understanding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 54582–54599. PMLR, 2025.

Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao. Video-XL: Extra-Long Vision Language Model for Hour-Scale Video Understanding. *arXiv preprint arXiv:2409.14485*, 2024.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. MovieChat: From Dense Token to Sparse Memory for Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18221–18232, 2024.

Guangyu Sun, Archit Singhal, Burak Uzkent, Mubarak Shah, Chen Chen, and Garin Kessler. From Frames to Clips: Training-free Adaptive Key Clip Selection for Long-Form Video Understanding. *arXiv preprint arXiv:2510.02262*, 2025a.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. MDP3: A Training-free Approach for List-wise Frame Selection in Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24090–24101, 2025b.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive Keyframe Sampling for Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128, 2025.

Chendong Wang, Donglin Bai, Yifan Yang, Xiao Jin, Anlan Zhang, Rui Wang, Shiqi Jiang, Yuqing Yang, Hao Wu, Qi Dai, Chong Luo, Ting Cao, Lili Qiu, and Suman Banerjee. Video-in-the-Loop: Span-Grounded Long Video QA with Interleaved Reasoning. *arXiv preprint arXiv:2510.04022*, 2025. doi: 10.48550/arXiv.2510.04022. URL https://arxiv.org/abs/2510.04022.

Yilin Wang, Xiangxi Zheng, Dongxing Mao, Linjie Li, Zhengyuan Yang, Ping Yu, Rui Yan, Yuan Yao, and Alex Jinpeng Wang. Efficient Frame Selection for Long Videos at Test Time with Attention-Based MLLM Selectors. *arXiv preprint arXiv:2607.15689*, 2026. doi: 10.48550/arXiv.2607.15689. URL https://arxiv.org/abs/2607.15689.

Fan Wei, Siru Zhong, Runmin Dong, Miao Yang, Zhaoyang Luo, and Haohuan Fu. Ground, Cover, and Refine: Evidence-Centric Frame Selection for Long-Video Question Answering. *arXiv preprint arXiv:2608.01660*, 2026. doi: 10.48550/arXiv.2608.01660. URL https://arxiv.org/abs/2608.01660.

Chao-Yuan Wu and Philipp Krähenbühl. Towards Long-Form Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1884–1894, 2021.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A Benchmark for Long-context Interleaved Video-Language Understanding. In *Advances in Neural Information Processing Systems*, volume 37, pp. 28828–28857. Curran Associates, Inc., 2024. doi: 10.52202/079017-0907.

Boshen Xu, Zihan Xiao, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Qin Jin. TimeViper: A Hybrid Mamba-Transformer Vision-Language Model for Efficient Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40479–40493, 2026a.

Jiaqi Xu, Cuiling Lan, Wenxuan Xie, Xuejin Chen, and Yan Lu. Long Video Understanding With Learnable Retrieval in Video-Language Models. *IEEE Transactions on Multimedia*, 28:5254–5261, 2026b. doi: 10.1109/tmm.2026.3664996.

Caijun Yan, Yang Zhou, Meixing Shi, Haoran Sun, Yichen Li, Yuxiang Cai, and Yankai Jiang. REVEAL: A Rubric-Guided Agent for Explicit Evidence Sufficiency Verification in Long-Video Question Answering. *arXiv preprint arXiv:2608.08612*, 2026.

Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven C. H. Hoi. Deep Learning for Person Re-Identification: A Survey and Outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):2872–2893, 2022. doi: 10.1109/tpami.2021.3054775.

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, pp. 76749–76771. Curran Associates, Inc., 2023. doi: 10.52202/075280-3354.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-Frame: Query-aware Frame Selection and Multi-Resolution Adaptation for Video-LLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22056–22065, 2025.

Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object Tracking by Associating Every Detection Box. In *Computer Vision – ECCV 2022*, pp. 1–21. Springer Nature Switzerland, 2022.

Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable Person Re-Identification: A Benchmark. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: Benchmarking Multi-task Long Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, 2025.

Kaiyang Zhou, Yongxin Yang, Andrea Cavallaro, and Tao Xiang. Omni-Scale Feature Learning for Person Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*, 2025.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. TimeLens2: Generalist Video Temporal Grounding with Multimodal LLMs. *arXiv preprint arXiv:2607.17423*, 2026.