NARRATIVETRACK: Evaluating Entity-Centric Reasoning for Narrative Understanding

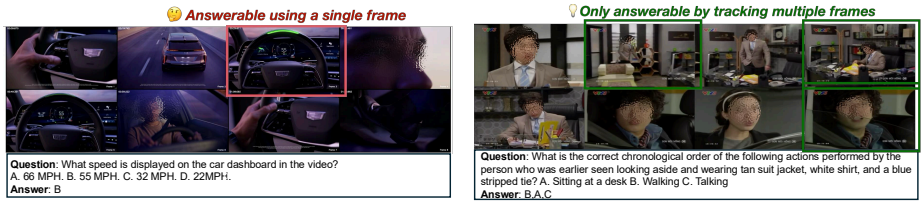
Hyeonjeong Ha[1,2†], Jinjin Ge[1], Bo Feng[1], Kaixin Ma[1], Gargi Chakraborty[1]
[1]Apple, [2]University of Illinois Urbana-Champaign
hh38@illinois.edu,
{jinjin_ge, bfeng2, kaixin_ma, g_chakraborty}@apple.com
† Work done during an internship at Apple.

Abstract. Multimodal large language models (MLLMs) have achieved impressive progress in vision-language reasoning, yet their ability to understand temporally unfolding narratives in videos remains largely underexplored. Narrative understanding requires more than recognizing isolated events: models must maintain coherent representations of who is doing what, when, and where across scene transitions and temporal gaps. We introduce NARRATIVETRACK, the first benchmark to evaluate narrative understanding in MLLMs through fine-grained entity-centric reasoning. Unlike existing benchmarks limited to short clips or coarse scene-level semantics, we decompose videos into constituent entities and evaluate models using a Compositional Reasoning Progression (CRP), a structured framework that progressively increases narrative complexity across three dimensions: entity existence, entity changes, and entity ambiguity. This progression requires models to move beyond local perception to reasoning about entities' temporal persistence, state changes, and fine-grained perceptual disambiguation. To enable scalable benchmark construction, we develop a fully automated entity-centric pipeline that extracts temporally grounded entity representations and provides the foundation for CRP. Evaluations of state-of-the-art MLLMs reveal that existing models struggle to maintain coherent entity representations under visual transitions and temporal dynamics. Open-source general-purpose MLLMs exhibit strong perceptual grounding but weak temporal continuity, while video-specialized MLLMs capture temporal context yet frequently hallucinate entities' contexts. These findings uncover a fundamental trade-off between perceptual grounding and temporal reasoning, indicating that narrative understanding emerges only from their integration. NARRATIVETRACK provides the first systematic framework to diagnose and advance temporally grounded narrative comprehension in MLLMs.

1 Introduction

Narrative understanding involves perceiving how entities evolve and interact over time to form coherent events, a fundamental aspect of human cognition. When watching a video, humans naturally build a mental model of the unfolding story by tracking who is present, what they are doing, and how their states change across scenes [49]. This process relies on maintaining entity representations: temporally grounded structures that bind an

⋆ Project Page: https://github.com/apple/ml-NarrativeTrack.
⋆ Dataset: https://huggingface.co/datasets/hjha/NarrativeTrack.



(a) Video-MME [14] **(b)** NARRATIVETRACK (Ours)

Fig. 1: Examples of existing benchmark and NARRATIVETRACK. While existing benchmarks can often be answered from a single frame, ours requires reasoning by tracking entities over time.

entity's identity and state across time, enabling coherent reasoning even under occlusion, viewpoint changes, or scene transitions. Entities thus serve as the basic units of narrative structure, organizing events into meaningful temporal and causal relationships.

Despite remarkable advances in multimodal large language models (MLLMs) across both static [26,37,38,55] and temporally evolving modalities [6,21,24,27,51], their entity-centric reasoning ability for narrative understanding remains largely unexamined. Existing benchmarks primarily assess either local recognition (e.g., object or action recognition) or global summarization (e.g., story-level understanding), leaving a critical gap in assessing whether models track entities consistently throughout a narrative. Datasets incorporating temporal information typically use short clips with minimal scene variation, emphasizing localized semantics [14,28,44,45,48] (Fig. 1a). In contrast, long-form benchmarks capture coarse global context [39,42], but rarely test long-range temporal dependencies where entities evolve and interact across disjoint scenes (Fig. 10). Consequently, many tasks can be solved using static visual cues or language priors [11], without genuine temporal reasoning, while heavy reliance on manual annotations further limits scalability and diagnostic granularity.

To address these limitations, we introduce NARRATIVETRACK, a novel benchmark designed to evaluate narrative understanding in MLLMs through an entity-centric, bottom-up formulation. Rather than assessing video understanding from top-down summaries or scene-level semantics, we decompose each video into its constituent entities, which serve as the fundamental building blocks of events, and evaluate how models reason about their temporal continuity and evolution. This design is grounded in cognitive and narratological theory [9,49]: narrative comprehension emerges from maintaining entity continuity and interpreting their evolving actions and contextual roles. Evaluating whether models can track how entities persist, change, and become confusable directly probes the ability to construct temporally grounded narrative structure.

We formalize this by defining a *Compositional Reasoning Progression* (CRP), which systematically increases narrative complexity across three interdependent reasoning dimensions: (1) **entity existence** tests basic temporal continuity, evaluating whether models can track entities persistently across time; (2) **entity changes** evaluates reasoning about evolving actions, scenes, and outfits that signal narrative development [12,13]; (3) **entity ambiguity** introduces visually similar entities to challenge fine-grained perceptual disambiguation under temporal dynamics. This progression transforms narrative reasoning from single-skill testing into a compositional diagnostic of entity-centric reasoning,

enabling a systematic evaluation of narrative understanding. To enable scalable construction of the entity representations that underpin CRP, we further present a *fully automated entity-centric pipeline* that extracts temporally grounded trajectories augmented with fine-grained attributes directly from raw videos without human supervision.

We evaluate a diverse suite of state-of-the-art MLLMs on NARRATIVETRACK, spanning open-source general-purpose, open-source video-specialized, and proprietary models. Gemini-2.5-Pro achieves the highest average accuracy of 83.80%, while open-source models exhibit substantial gaps. Notably, general-purpose MLLMs outperform video-specialized ones (e.g., Qwen2.5-VL-32B: 56.96% vs. Video-LLaMA2-72B: 49.70%), revealing a fundamental trade-off between perceptual grounding and temporal reasoning. General-purpose MLLMs show strong perceptual grounding, accurately recognizing static visual cues, yet frequently fail to maintain temporal consistency. Conversely, video-specialized MLLMs capture temporal continuity more reliably but lack perceptual robustness, often hallucinating under entity ambiguity or appearance changes. Further analysis shows that neither scaling model size nor increasing frame density improves narrative understanding, and it also exposes weaknesses in reasoning about backward temporal scenarios, highlighting a persistent difficulty in maintaining coherent entity representations. We leave architectural innovation, such as bidirectional temporal modeling and entity-centric training objectives, as future work. Overall, our results indicate that narrative understanding is a compositional capability, emerging from the integration of temporal reasoning and perceptual precision. By centering evaluation on fine-grained entity tracking, NARRATIVETRACK offers the first systematic framework to diagnose how and where MLLMs fail to maintain coherent, temporally grounded narrative structure.

2 Related Work

2.1 Multimodal Large Language Models (MLLMs)

MLLMs augment large language models with visual encoders, enabling joint reasoning over text and images for tasks such as chart understanding and visual question answering [4, 16, 26, 38, 41, 55]. Recent advances have scaled these models toward unified input-output representations, long-context reasoning, and cross-modal generalization [5, 10, 36, 40, 53], bringing them closer to general-purpose visual-language understanding. Building on these developments, recent work has extended MLLMs to the video domain, enabling them to process sequential visual inputs and reason over temporally evolving scenes [6, 7, 21, 24, 27]. However, most MLLMs remain optimized for static image understanding. Their visual encoders typically compress spatial and temporal information into coarse global embeddings, sacrificing fine-grained perceptual detail [18, 20]. Consequently, while they excel at high-level semantic alignment [17, 32], they struggle to maintain consistent reasoning about individual entities or events over time. Moreover, how current MLLMs ground their responses in fine-grained elements such as entities remains underexplored, leaving a critical gap in understanding multimodal reasoning.

2.2 Video Understanding Benchmarks

Existing video understanding benchmarks evaluate MLLMs from diverse perspectives, including global semantic comprehension [45,48], causal and temporal reasoning [21,44], action and event recognition [14], spatial reasoning [28], and long-video understanding [23,34,42]. While each benchmark targets specific reasoning skills, recent work [11] has shown that many can be solved without genuine temporal reasoning. In many cases, model performance remains largely unchanged when frame order is shuffled, suggesting reliance on language priors or static visual cues rather than true temporal reasoning. These findings raise a concern that current evaluations often measure recognition of isolated events rather than the ability to maintain coherent spatio-temporal representations over time. More recently, compositional reasoning in videos has been studied through structured evaluation protocols. VELOCITI [33] introduces a strict video-language entailment, isolating agent–action binding errors via carefully constructed positive and negative captions. This design provides a controlled testbed for event-level compositional grounding. However, its scope remains temporally localized: entities are typically continuously visible within the clip, and the primary challenge lies in correctly associating agents with actions within short temporal windows. As such, the reasoning required is largely confined to within-clip binding rather than cross-scene continuity.

In contrast, narrative understanding in long-form videos introduces a fundamentally different challenge: characters may disappear and reappear across scenes, undergo state changes, and participate in causally linked events separated by substantial temporal gaps [9,28,49]. Our benchmark moves beyond short-horizon entailment to entity-centric question answering over long videos, requiring models to maintain persistent identity and state representations across discontinuities. By explicitly evaluating long-range entity continuity under scene transitions, entity's state evolutions, and temporal gaps, our benchmark advances video understanding beyond localized grounding and static frame reasoning.

3 Evaluating VideoLLMs Beyond the Frame

Problem Statement. Each input video V is represented as a sequence of frames sampled at a fixed frame rate f, producing $N+1$ frames with timestamps $\{t_0, t_1, \ldots, t_N\}$, where $t_j = j/f$ for $j \in [0, N]$. These timestamps define the temporal axis over which visual events unfold. For each entity e_i, we define a temporally grounded *entity representation*:

$$\begin{aligned}\tau_{e_i} = \{&(t_{i0}, b_{i0}, a_{i0}, s_{i0}, o_{i0}), (t_{i1}, b_{i1}, a_{i1}, s_{i1}, o_{i1}),\\ &\ldots, (t_{iM}, b_{iM}, a_{iM}, s_{iM}, o_{iM})\},\end{aligned} \tag{1}$$

where $t_{ij} \in [t_0, t_N]$ denotes a timestamp t_j at which e_i appears, and $M \leq N$ is the total number of such observations. At each timestamp t_j, the entity representation includes: (1) b_{ij}, the bounding box of e_i; (2) a_{ij}, the action performed by the entity; (3) s_{ij}, the scene context; and (4) o_{ij}, the entity's outfit. This structured representation captures both *temporal dynamics* and *perceptual context*, serving as the foundation for evaluating whether MLLM can reason consistently about entity representation, tracking

Benchmarks	Len.(s)	#QA Pairs	Anno.	BBox	Action	Scene	Outfit	Entity-Centric
AVA [15]	900	-	M	✓	✓	✗	✗	✗
TVQA [19]	11.2	15,253	M	✗	✗	✗	✗	✗
MovieQA [35]	205.5	2,144	M	✗	✗	✗	✗	✗
How2QA [22]	15.3	2,852	M	✗	✗	✗	✗	✗
NExT-QA [44]	39.5	8,564	A	✗	✗	✗	✗	✗
Video-MME [14]	1,017.9	2,700	M	✗	✗	✗	✗	✗
PerceptionTest [31]	23.0	44,000	A & M	✓	✓	✗	✗	✗
LongVideoBench [42]	473	6,678	M	✗	✗	✗	✗	✗
LVBench [39]	4,101	1,549	M	✗	✗	✗	✗	✗
NARRATIVETRACK	55.3	1,006	A	✓	✓	✓	✓	✓

Table 1: Comparison between existing video understanding benchmarks and NARRATIVETRACK. *Len.* denotes the average duration of clips, and *Anno.* indicates the annotation type, where *A* denotes automated annotation and *M* refers to manual annotation.

their continuity, appearances, and narrative roles across scenes, to achieve coherent and fine-grained narrative understanding.

In this work, we focus exclusively on *human entities*, as they are the primary agents in narrative videos and exhibit rich temporal variations in action and appearances. This makes humans particularly representative subjects for evaluating entity-centric reasoning across temporal and perceptual grounding. In contrast, many non-human or background entities exhibit limited visual variation, which can reduce evaluation to simple identity matching, while also introducing ambiguity that may confound the assessment of compositional reasoning. Restricting the scope to visually salient human subjects enables a more controlled and challenging testbed for rigorously evaluating models' ability to maintain coherent entity-level understanding over time. Extending the framework to non-human entities remains a promising direction for future work.

3.1 Automated Entity-Centric Pipeline

Building a benchmark for evaluating narrative understanding requires rich, temporally aligned representations that capture how entities evolve over time. Existing datasets lack such structured, entity-centric annotations, as manually labeling long, unconstrained videos is costly and often inconsistent (Table 1). To overcome this limitation, we introduce a novel, fully automated pipeline that extracts structured entity representations τ_{e_i} directly from raw videos without human supervision. The pipeline comprises three stages: *entity detection*, *entity tracking*, and *contextual recognition* (Fig 2).

Entity Detection. We detect all visible entities in each frame using off-the-shelf multi-object detection models. Accurate detection is critical: missing or imprecise detections can fragment trajectories or merge distinct entities, cascading errors into tracking and contextual recognition. To improve reliability, we introduce an entity-centric ensemble detection strategy that combines predictions from Detectron2 [43] and Owlv2 [29] through spatial consensus rather than naive union. Let $\mathcal{B}^{(1)}$ and $\mathcal{B}^{(2)}$ denote the sets of bounding boxes predicted by the two detectors for a given frame. We compute pairwise IoU between boxes across detectors and consider two boxes to correspond to the same

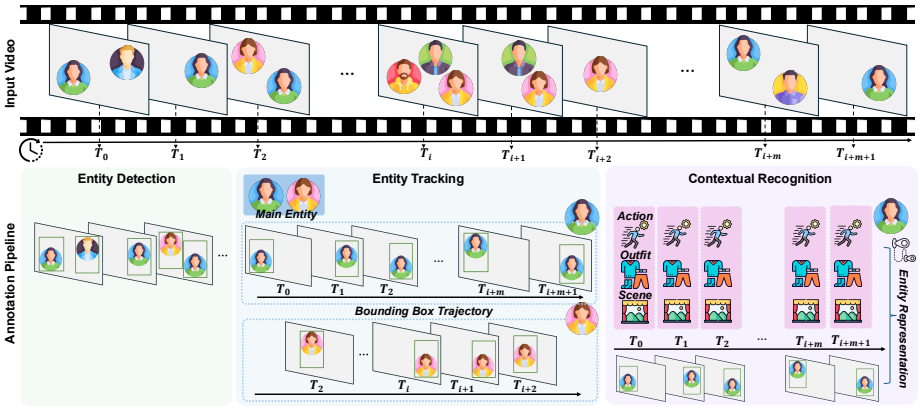


Fig. 2: Overview of Automated Entity-Centric Pipeline. Our pipeline consists of three key stages: (1) Entity Detection, (2) Entity Tracking, (3) Contextual Recognition, extracting entity representations capturing bounding-box trajectories, actions, outfits, and scene contexts associated with the target entity from raw video without human supervision.

entity if IoU ≥ 0.5, a standard threshold in detection matching and tracking that balances over-merging with duplicate retention. When two boxes meet this criterion, we keep only the higher-confidence score:

$$\mathcal{B}_{\text{final}} = \mathcal{B}^{(1)} \cup \mathcal{B}^{(2)} \setminus \{b_j^{(2)} \mid \exists\, b_i^{(1)},\ \text{IoU}(b_i^{(1)}, b_j^{(2)}) \geq 0.5\},$$

This process preserves one representative box per entity while retaining unique detections from both models, yielding comprehensive yet non-redundant coverage of visible entities in each frame. The resulting per-frame detections are subsequently linked across time to form bounding box trajectories. Each trajectory is inherently spatio-temporal: it encodes the spatial localization of an entity within individual frames as well as its temporal evolution across frames, enabling consistent entity-level tracking throughout the video.

Entity Tracking. Narrative understanding requires identifying *who persists* over time. However, naively tracking every detected entity is inefficient and error-prone, as some may appear briefly or contribute little to the narrative. To focus on salient narrative participants, we identify main characters by clustering bounding-box embeddings extracted with state-of-the-art re-identification (ReID) models [30,46,54]. We use OSNet-x1.0* as the ReID model. Each cluster corresponds to a distinct entity and is ranked by size, with the top four selected as main characters under the assumption that recurrent presence correlates with narrative centrality. Empirically, central characters tend to occupy larger spatial regions ($\geq$50% of the frame height or width) and appear consistently over time. Across 100 sampled videos, the top-four clusters satisfy this criterion for at least 50% of their instances, motivating our choice of four entities as a stable heuristic for capturing salient characters. To enhance identity consistency, we refine trajectories with

* https://kaiyangzhou.github.io/deep-person-reid/MODEL_ZOO.html

face recognition for precise alignment and apply ensemble verification across multiple MLLMs with majority voting to suppress false identities. Detailed prompt is provided in Appendix §9.2. This consensus-based strategy emulates human agreement, minimizing identity drift in a fully unsupervised setting. The resulting trajectories provide temporally consistent spatial localization for each target entity throughout the video.

Contextual Recognition. Beyond spatial localization, narrative comprehension requires understanding *what* each entity is doing, *how* and *when* it appears, and *where* it is situated. For each entity trajectory, we augment every timestep t_{ij} with contextual attributes: *actions* a_{ij}, *outfits* o_{ij}, and *scenes* s_{ij}. We use Gemini-2.5-Pro [8] to infer these contextual attributes (prompt is provided in §9.3), highlighting the target entity in overlaid clips to preserve context and ensure the model attends to the correct individual. Attributes are annotated per segment in which the entity appears in clips, and these predictions populate τ_{e_i}, yielding a structured, temporally aligned representation of how each entity evolves throughout the narrative. To support QA generation for entity changes and entity ambiguity dimensions, Gemini-2.5-Pro is used to identify videos with significant attribute changes or visually similar entities based on predicted attributes.

Pipeline Evaluation. We evaluate the quality of our detection and tracking pipeline using the AVA dataset [15], which provides frame-level bounding boxes $\mathcal{G} = \{g_1, g_2, \ldots, g_{N_G}\}$ but lacks entity identities. Since the correspondence between ground-truth and predicted boxes is unknown, each predicted box $p_i \in \mathcal{P} = \{p_1, p_2, \ldots, p_{N_P}\}$ at frame t is greedily matched to a ground-truth box:

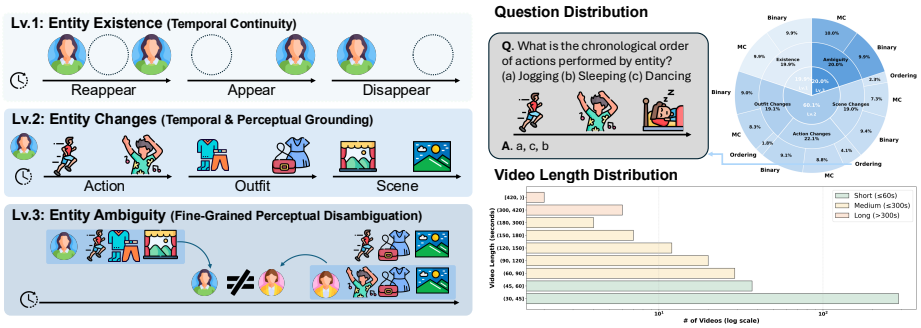
$$g^*(p_i) = \underset{g \in \mathcal{G} \setminus \mathcal{G}_{\text{matched}}}{\arg\max} \ \text{IoU}(p_i, g)$$

and if $\text{IoU}(p_i, g^*(p_i)) \geq 0.5$, we record the match, updating $\mathcal{M}$ and $\mathcal{G}_{\text{matched}}$ as: $\mathcal{M} \leftarrow \mathcal{M} \cup \{(p_i, g^*(p_i))\}$ and $\mathcal{G}_{\text{matched}} \leftarrow \mathcal{G}_{\text{matched}} \cup \{g^*(p_i)\}$. The recall is then computed as $\text{Recall} = |\mathcal{M}|/|\mathcal{G}|$. Our ensemble detection improves recall from 0.780 (Detectron2 alone) and 0.801 (OWLv2 alone) to 0.848, confirming that spatial consensus effectively reconciles missed or inconsistent detections.

To assess the reliability of our ensemble-based tracking verification, human experts from the authors label each tracked entity as *kept* if it consistently matches the same entity across frames, or *deleted* if an incorrect identity is assigned. We compare human-majority vote labels with the predictions of our model-based majority voting method on randomly sampled AVA video clips comprising 1,108 detections. Notably, two sources agree on 96.08% of cases, indicating that our consensus approach reliably filters erroneous tracks and accurately preserves coherent identities. This confirms that our entity-centric pipeline produces high-quality tracking results in a fully automated manner.

3.2 Compositional Reasoning Progression

To systematically evaluate narrative understanding, we introduce Compositional Reasoning Progression (CRP), an evaluation framework that decomposes entity-centric reasoning into progressively more complex dimensions (Fig. 3a). Each dimension isolates a distinct capability, enabling fine-grained analysis of model behavior.



(a) Compositional Reasoning Progression. **(b)** Data Statistics of NARRATIVETRACK.

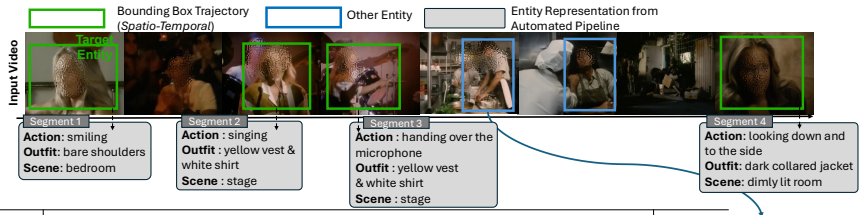
Fig. 3: Overview of NARRATIVETRACK. (a) Our benchmark is grounded in Compositional Reasoning Progression that introduces three levels of increasing complexity in entity-centric reasoning: entity existence, entity changes, and entity ambiguity. (b) The benchmark covers three question types: binary, multiple-choice (MC), and ordering, with diverse temporal scales from short to long (average of 55.3 seconds).

Entity existence evaluates *temporal continuity*: the ability to track an entity's presence across its appearances, disappearances, and reappearances. This dimension isolates the most fundamental temporal requirement of maintaining a consistent identity over time. Failures indicate breakdowns in long-range temporal tracking. **Entity changes** assess whether a model can detect and ground state transitions of a target entity. Beyond continuity, the model must integrate dynamic visual and contextual cues. We decompose this dimension into: (i) **Action changes**, testing *temporal grounding* of dynamic events; (ii) **Outfit changes**, probing *fine-grained visual grounding* of perceptual transitions; (iii) **Scene changes**, evaluating *spatial-contextual grounding* by linking entity with environmental context. The failures here reflect a weak grounding of evolving visual or contextual signals. **Entity ambiguity** introduces visually similar entities, requiring joint *temporal reasoning and fine-grained perceptual disambiguation* to ensure reliable entity tracking. Errors reveal weak compositional reasoning.

Overall, CRP progresses from entity persistence to grounded state transitions, to joint temporal–perceptual disambiguation. This design enables precise diagnosis of whether failures arise from disrupted temporal continuity, weak grounding of evolving attributes, or confusion among similar entities.

3.3 QA Generation

Building on the structured entity-centric representations produced by our automated pipeline, we construct question-answer (QA) pairs fully programmatically from extracted metadata (Fig. 4). Questions are constructed by instantiating predefined reasoning templates aligned with CRP (Table 5-9). Each representation contains temporally evolving states, including bounding box trajectories, actions, scenes, outfits, and their transitions. Template slots are deterministically filled with corresponding attributes. For example, given an entity transition ($\{a_1, o_1, s_1\} \rightarrow \{a_2, o_2, s_2\}$), we instantiate: "Is the person a_1 in s_1 with o_1 at the beginning later seen wearing o_2?" to probe entity outfit change



Dimension	Question (Fill Template with Entity Representation)	Segment
Entity Existence	Is the person with bare shoulders, smiling in the bedroom, only seen at the beginning and then gone?	Seg1
Action Changes	Identify the order of actions made by the person with bare shoulders, smiling in the bedroom at the beginning. (a) handing over the microphone (b) singing on the stage (c) looking down and to the side	Seg1 Seg2 Seg3 Seg4
Outfit Changes	Is the person with bare shoulders, smiling in the bedroom at the beginning, later seen wearing a dark collared jacket?	Seg1 Seg4
Scene Changes	In which scene does the person with bare shoulders, smiling in the bedroom at the beginning, appear later? (a) stage (b) kitchen (c) bathroom	Seg1 Seg2
Entity Ambiguity	After appearing with bare shoulders and smiling in the bedroom at the beginning, is the same person later shown cooking in the kitchen while wearing a beige shirt and an apron?	Seg1

Real Distractors: Other Entity's State in the **same** video

Synthetic Distractors: Other Entity's State in the **different** video

Fig. 4: QA Generation & Qualitative Examples. QA pairs are constructed by instantiating CRP-aligned reasoning templates using extracted entity representations. Distractors include real ones from other entities in the same video and synthetic ones from different videos.

reasoning. Ground-truth answers are directly derived from the same structured metadata to ensure semantic consistency. After automatic generation, GPT-4o is used for linguistic refinement (e.g., grammar check and fluency improvement) and to filter out invalid cases, such as questions generated from entity representations without valid state changes. The detailed prompt is provided in Appendix (§10.2).

QA Formats and Reasoning Patterns. We design three QA formats: binary, multiple-choice, and ordering. The ordering format requires a chronological arrangement of the entity's attribute transitions, explicitly evaluating fine-grained temporal compositionality beyond recognition. To probe temporal directional bias (§5), we define three reasoning patterns: (1) *forward* (tracking start $\rightarrow$ end), (2) *backward* (reasoning end $\rightarrow$ start), (3) *agnostic* (reasoning a mid-point $\rightarrow$ start & end; bidirectional), which are produced programmatically by varying temporal reference points during template instantiation.

Distractor Construction. We construct two types of distractors for binary and multiple-choice formats. *Real Distractors (Intra-Clip Negatives)* are sampled from other entities within the *same* video clip. Since these entities co-exist temporally and spatially with the target entity, the model must accurately track and distinguish the target entity. This setting tests fine-grained entity discrimination and tracking under realistic multi-entity scenarios. In contrast, *Synthetic Distractors (Cross-Clip Negatives)* are sampled from entities in *different* video clips, and thus do not appear in the queried clip. This setup evaluates hallucination robustness (§5): models should reject entities that are not present in the clip rather than relying on semantic plausibility. Intuitively, real distractors stress entity-level discrimination within context, while synthetic distractors test presence verification and hallucination avoidance.

Quality Review. To assess the quality of the automatically generated QA pairs, we randomly sample 100 items and perform a triple-annotator review (valid/invalid). A

QA pair is marked *valid* only if both the question and its ground-truth answer are correctly grounded to the target entity and its states in the video. It is marked *invalid* if either component is incorrect or misaligned, or if distractors contain obvious, non-grounded attributes that make the answer trivial. Among the sampled QA pairs, 70% are unanimously judged valid with substantial inter-annotator agreement (Fleiss' $\kappa = 0.767$). Most invalid cases arise from overly obvious synthetic distractors with implausible attribute assignments, rather than errors in entity tracking or contextual grounding. This indicates that the core entity-centric pipeline is reliable, consistent with §3.1.

To further ensure correctness, human annotators manually verify the attributes inferred by proprietary MLLMs during the contextual recognition step against the corresponding video evidence. Based on this verification, we refine all retained QA pairs by: (1) correcting inconsistent ground-truth answers; (2) revising ambiguous or weakly grounded attributes in the questions; (3) replacing trivial synthetic distractors with visually confusable alternatives. This process yields 1,006 high-quality QA pairs from 406 video clips, spanning diverse genres (e.g., documentary, news, TV drama) and temporal scales up to 659 seconds (Fig. 3b). The final distribution across CRP dimensions is balanced: entity existence (200), action changes (222), outfit changes (192), scene changes (191), and entity ambiguity (201).

To mitigate answer-choice bias, we balance the ground-truth answer distribution across candidates after manual verification. Finally, we randomly re-evaluate the cleaned benchmark with three annotators, achieving 96% average human accuracy, confirming dataset reliability and clarity.

4 Experiments

4.1 Experimental Settings

Raw Video Sources. We construct NARRATIVETRACK using raw video sources from three widely adopted video datasets: AVA [15], VideoMME [14], and LVBench [39]. To ensure suitable entity-centric evaluation, we filter out dark or low-quality clips and videos containing only a single character without meaningful attribute changes. From the selected videos, we extract entity representations using our automated pipeline (§3.1) and generate QA pairs (§3.3).

Model Baselines. We evaluate 13 open-source and 7 proprietary MLLMs on our benchmark. The open-source models include both general-purpose and video-specialized MLLMs of varying scales. We refer to open-source general-purpose MLLMs as *OGP-MLLMs* and open-source video-specialized MLLMs as *OVS-MLLMs*. The OGP-MLLMs includes Qwen-2.5-VL-7B, 32B [2] and InternVL3-8B, 38B [55]; the OVS-MLLMs include mPLUG-Owl3-7B [47], VideoChat2-7B [21], Video-LLaMA2-7B, 72B [6], Video-LLaVA-7B [24], LLaVA-Video-7B [52], LLaVA-NeXT-Video-7B, 34B [51], and VILA-8B [25]. We use Gemini-2.0-Flash, Gemini-2.5-Flash, Gemini-2.5-Pro, Gemini-3-Flash, Gemini-3.1-Pro, GPT-4o, and GPT-4.1 for a proprietary model. All open-source model evaluations are conducted by sampling 20 frames per video following prior work [50], as this setting achieves the best performance across frame densities (§5). For proprietary models, we use 128 frames per video to take advantage of its larger visual

Model	Existence (EE)		Action Changes (AC)			Outfit Changes (OC)			Scene Changes (SC)			Ambiguity (EA)		**Avg.**
	B	MC	B	MC	O	B	MC	O	B	MC	O	B	MC	
Open-Source General-Purpose MLLMs														
Qwen-2.5-VL-7B [2]	57.00	36.00	48.91	38.20	17.07	65.93	55.42	22.22	46.32	46.58	34.78	70.00	45.55	48.81
Qwen-2.5-VL-32B [2]	68.00	42.00	71.74	42.70	19.51	61.54	63.86	5.56	68.42	58.90	30.44	79.00	46.54	56.96
InternVL3-8B [55]	71.00	55.00	55.44	49.44	29.27	46.15	31.33	22.22	55.79	46.58	39.13	63.00	26.73	48.81
InternVL3-38B [55]	70.00	55.00	56.52	50.56	21.96	56.04	51.81	22.22	66.32	58.90	52.71	74.00	36.63	55.47
Open-Source Video-Specialized MLLMs														
mPLUG-Owl3-7B [47]	60.00	37.00	63.04	35.96	24.39	42.86	22.89	16.67	53.68	45.21	21.74	62.00	22.77	42.94
VideoChat2-7B [21]	45.00	35.00	60.87	38.20	12.20	58.24	30.12	5.56	57.90	35.62	47.83	60.00	21.78	42.55
Video-LLaMA2-7B [6]	53.00	42.00	58.70	35.96	12.20	50.55	28.92	16.67	55.79	52.06	26.09	51.00	25.74	43.04
Video-LLaMA2-72B [6]	59.00	46.00	52.17	49.44	19.51	42.86	42.17	11.11	62.10	57.53	43.48	71.00	36.63	49.70
Video-LLaVA-7B [24]	42.00	35.00	57.61	33.71	9.76	32.97	14.46	33.33	52.63	28.77	17.39	32.00	12.87	33.00
LLaVA-Video-7B [52]	67.00	2.00	70.65	17.98	17.07	43.96	25.30	5.56	57.90	20.55	21.74	40.00	11.88	34.39
LLaVA-NeXT-Video-7B [51]	65.00	41.00	68.48	33.71	2.44	47.25	18.07	0.00	54.74	38.36	13.04	63.00	27.72	42.94
LLaVA-NeXT-Video-34B [51]	53.00	47.00	55.44	39.33	7.32	48.35	21.69	0.00	56.84	39.73	13.04	39.00	19.80	39.36
VILA-8B [25]	59.00	56.00	60.87	37.08	4.88	51.65	28.92	5.56	60.00	53.43	21.74	53.00	23.76	45.33
Proprietary MLLMs														
Gemini-2.0-Flash	67.00	69.00	76.09	52.81	34.15	64.84	49.40	11.11	64.21	60.27	52.17	85.00	34.65	60.24
Gemini-2.5-Flash	73.00	59.00	80.44	61.80	**63.42**	86.81	75.90	66.67	76.84	61.65	65.22	94.00	78.22	74.25
Gemini-2.5-Pro	**88.00**	75.00	83.70	75.28	60.98	**95.60**	**86.75**	83.33	**82.11**	83.56	91.30	94.00	82.18	**83.80**
Gemini-3-Flash	87.00	**77.00**	83.70	69.66	**63.42**	89.01	57.83	**94.44**	**82.11**	68.49	**95.65**	94.00	78.22	79.32
Gemini-3.1-Pro	82.00	74.00	85.87	**78.65**	48.78	91.21	84.34	88.89	80.00	**87.67**	**95.65**	**97.00**	**85.15**	83.40
GPT-4.1	78.00	63.00	**86.96**	68.54	46.34	82.42	74.70	66.67	77.90	73.97	73.91	89.00	68.32	74.85
GPT-4o	77.00	61.00	82.61	66.29	39.02	82.42	67.47	44.44	75.79	83.56	78.26	86.00	61.39	72.27
GPT-4o (20 Frames)	78.00	58.00	81.52	60.67	36.59	80.22	67.47	61.11	72.63	75.34	78.26	85.00	66.34	70.97

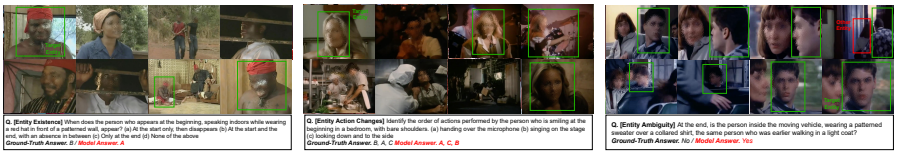
Table 2: Evaluation Results on NARRATIVETRACK. B, MC, and O denote binary, multiple-choice, and ordering questions, respectively. **Bold** and underline stands for the best and second. EE indicates entity existence; AC, OC, refer to entity action, outfit, scene changes, respectively; EA denotes entity ambiguity dimension in CRP. We evaluate all open-source MLLMs using 20 frames per video, while proprietary MLLMs are evaluated using 128 frames by default.

context capacity, while we further evaluate GPT-4o with 20 frames per video for fair comparison.

Evaluation Protocol. Our evaluation does not rely on LLM-based judging or open-ended response generation. All questions are formulated as binary-choice, multiple-choice, or option-ordering tasks. Model predictions are scored deterministically using exact-match accuracy against the ground-truth answers, eliminating ambiguity from natural-language phrasing, lexical variation, or subjective assessment.

4.2 Evaluation Results

Quantitative Results Table 2 presents the evaluation results on our benchmark, revealing a substantial performance gap across open-source and proprietary MLLMs. Among



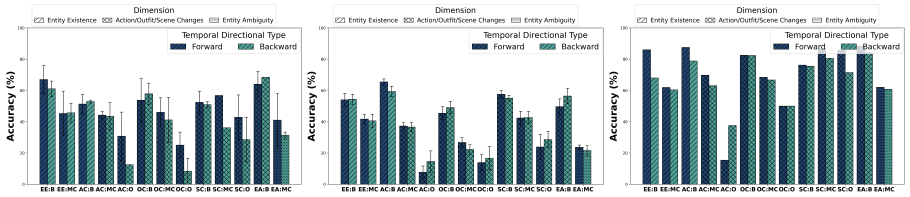
(a) Entity Existence **(b)** Entity Changes **(c)** Entity Ambiguity

Fig. 5: Examples of Model Failure in NARRATIVETRACK. Video sources are extracted from AVA [15] and the output is generated from Video-LLaMA2-7B [6].

OGP-MLLMs, Qwen-2.5-VL-32B achieves the highest accuracy (56.96%), surpassing the best OVS-MLLMs, Video-LLaMA2-72B (49.70%). Yet, both remain far behind proprietary MLLMs (e.g., Gemini-2.5-Pro) in maintaining consistent entity tracking. Across open-source models, performance drops sharply on tasks requiring dynamic attribute reasoning, such as action and outfit changes or entity disambiguation, while scenes change tasks show relatively better accuracy due to their lower visual variability. This pattern suggests that current open-source MLLMs struggle to integrate temporal coherence with fine-grained perceptual grounding. Scaling trends further reveal that larger models generally perform better (e.g., InternVL3-38B vs. 8B: +6.66%; Video-LLaMA2-72B vs. 7B: +5.66%), though gains are inconsistent, as seen in LLaVA-NeXT-Video-34B, which slightly underperforms its 7B version, indicating that size alone does not guarantee improved entity tracking.

In contrast, GPT-4o attains the highest overall performance (70.97%) under the same frame density, consistently outperforming all open-source baselines. Increasing the number of input frames to 128 further improves its average performance. Yet, even this strong proprietary model shows limitations in maintaining robust entity-level continuity over extended narratives. Further analysis confirms that NARRATIVETRACK requires genuine multimodal grounding rather than reliance on language priors: removing visual inputs reduces GPT-4o accuracy by 30.52%, approaching the random baseline, while reversing video frames drastically lowers performance on the entity change ordering task from 51.2% to 6.1% (Table 11), demonstrating the necessity of correct temporal reasoning. These findings confirm that narrative understanding is a core unsolved challenge, emphasizing the need for MLLMs that can jointly model fine-grained visual perception, temporal consistency, and entity-centric reasoning over extended video contexts.

Qualitative Results We conduct a qualitative analysis of representative failure cases to better understand model limitations across reasoning levels. At the entity existence level, models frequently overestimate continuity, predicting that an entity persists throughout the video even when it appears only briefly at the beginning (Fig. 5a). At the entity change level, errors compound: when entities undergo subtle attribute transitions, models often produce semantically plausible yet visually ungrounded responses (e.g., singing on the stage, handing over a microphone, and looking down), exposing deficiencies in fine-grained perception and temporal reasoning (Fig. 5b). The most severe failures occur at the entity ambiguity level, where multiple entities interact or reappear. Even strong proprietary models struggle here, often confusing identities or hallucinating entity presence



(a) OGP-MLLMs. **(b)** OVS-MLLMs. **(c)** Proprietary MLLMs.

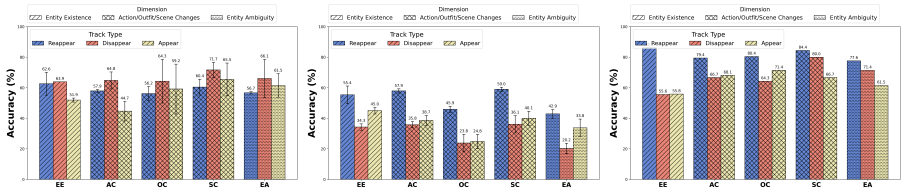
Fig. 6: Temporal Directional Bias of MLLMs. MLLMs encode temporal relations in a forward-only manner and fail to generalize to reversed or bidirectional temporal contexts.

(Fig. 5c). These error patterns reveal hierarchical failure cascades, where misperception at lower levels propagates upward, underscoring the persistent challenge of modeling entity continuity and temporal dependencies essential for coherent narrative understanding.

5 Analysis

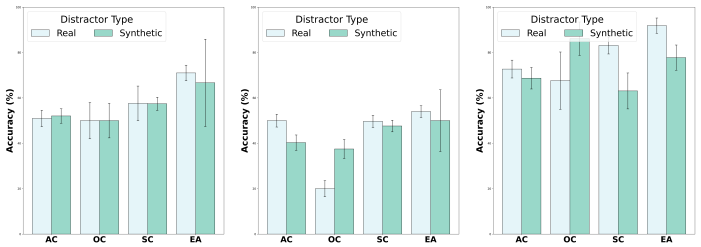
All analyses are conducted using 13 MLLMs: 12 open-source models from the Qwen2.5-VL, InternVL3, Video-LLaMA2, and LLaVA-NeXT-Video families, along with mPLUG-Owl3-7B, VideoChat2-7B, Video-LLaVA-7B and VILA-8B, plus GPT-4o. Following prior works [14, 39], we group them by training objective: OGP-MLLMs emphasize visual grounding through image-text alignment, whereas OVS-MLLMs model temporal dynamics through video-text alignment. We therefore analyze trends within each category rather than treating cross-category differences as direct measures of overall model quality.

Temporal Directional Bias. Our benchmark includes two reasoning types: *forward* and *backward* reasoning as described in §3.3. Across model types, we observe a consistent advantage in forward reasoning (Fig. 6), revealing a strong directional bias in temporal reasoning. The performance gap between forward and backward reasoning reaches 20.65%, 9.96%, and 17.55% for OGP-, OVS-, and proprietary MLLMs, respectively. This suggests that models can effectively propagate entity states along the video timeline but struggle to infer them in reverse. This phenomenon parallels the Reversal Curse in LLMs [3], where models trained on "A is B" fail to generalize to "B is A". Similarly, MLLMs encode temporal relations directionally, binding entities to sequentially observed events without learning an invertible mapping between earlier and later states. In essence, models can extend a narrative but cannot rewind it, rooted in the causal, left-to-right decoding paradigm inherited from their LLM backbones. We further introduce an *agnostic reasoning* in ordering questions, requiring bidirectional inference of entity states from a middle point. This condition yields the lowest accuracy (Fig. 11), underscoring that current MLLMs exhibit a forward bias and lack coherent temporal grounding mechanisms for reasoning across non-sequential or bidirectional contexts. Addressing this limitation may require bidirectional temporal modeling or contrastive reversal objectives that explicitly enforce symmetry between forward and backward temporal reasoning over entity states.



(a) OGP-MLLMs. **(b)** OVS-MLLMs. **(c)** Proprietary MLLMs.

Fig. 7: Performance Across Entity Continuity Types. OGP-MLLMs perform best on *disappear* cases, relying on static visual cues, whereas OVS- and proprietary MLLMs excel on *reappear* cases, reflecting stronger temporal integration but a higher tendency to hallucinate visual details.



(a) OGP-MLLMs. **(b)** OVS-MLLMs. **(c)** Proprietary MLLMs.

Fig. 8: Performance Across Distractor Types. OGP-MLLMs show stable performance across real and synthetic distractors, whereas OVS-MLLMs and proprietary MLLMs exhibit notable drops with synthetic distractors, revealing stronger hallucination tendencies.

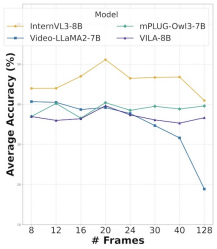


Fig. 9: Performance Across Frame Densities. Scaling temporal coverage does not guarantee performance gain.

Entity Continuity Types. We categorize entity continuity into three types: *appear*, *disappear*, and *reappear*. In appear and disappear cases, the target entity is visible in only one segment, entering or leaving the scene once. Conversely, reappear cases are the most temporally dynamic, requiring models to maintain identity across multiple disjoint segments. As shown in Fig. 7, OGP-MLLMs perform best on *disappear* cases, suggesting reliance on static visual evidence rather than reasoning over extended temporal sequences. In contrast, OVS-MLLMs achieve higher accuracy on *reappear* cases, reflecting stronger temporal integration but also a tendency to overpredict reappearances, often hallucinating entity reappearances or misattributing visual attributes. Proprietary MLLMs exhibit a similar behavior, suggesting that stronger temporal integration does not necessarily ensure precise entity grounding, particularly when identity must be maintained across temporally disjoint observations.

To verify that this pattern is not driven solely by basic entity-detection failures, we conduct a model-specific analysis conditioned on correct entity-existence predictions (Tab. 10). For each model, we evaluate entity-state changes and entity ambiguity only on videos in which the target entity is correctly identified as present or absent. Even under this controlled setting, OVS-MLLMs show larger performance drops on real (-11.5%) and synthetic (-14.9%) distractors, while remaining competitive on temporally driven tasks such as action changes. These results reveal a temporal–perceptual trade-off:

OVS-MLLMs better capture temporal continuity but are more susceptible to perceptual confusion, whereas OGP-MLLMs provide stronger visual grounding but weaker temporal integration.

Distractor Types. We compare model performance across two distractor types: *real* and *synthetic* as described in §3.3. Ideally, synthetic distractors should be easier since they are visually unrelated to the video. However, model behaviors diverge notably (Fig 8). OGP-MLLMs perform almost identically across both types, where they work slightly better on synthetic ones (0.44% higher in average), suggesting reliance on localized visual cues and effective rejection of irrelevant attributes. In contrast, OVS-MLLMs exhibit an accuracy drop on synthetic distractors (up to 9.69%), indicating stronger hallucination tendencies and weaker visual grounding. Proprietary MLLMs achieve the highest overall performance, yet demonstrate the largest real-synthetic gap (up to 20.03%), implying that even advanced models struggle to suppress contextually irrelevant generations. These results reveal a fundamental limitation of current MLLMs: while temporal modeling broadens contextual understanding, it also amplifies dependence on textual or global priors, undermining fine-grained visual discrimination required for robust multimodal reasoning.

Frame Density. Prior work shows that denser frame sampling improves long-video reasoning by enriching global context [39]. To examine whether greater temporal coverage similarly benefits entity-centric reasoning, we vary the number of input frames for open-source MLLMs, $k \in \{8, 12, 16, 20, 24, 28, 30, 40, 128\}$. The accuracy peaks $k = 20$ but degrades beyond that point, contrasting the monotonic gains seen in long-video reasoning tasks (Fig. 9). The drop is particularly pronounced for Video-LLaMA2, which is trained on only 8 frames and becomes unstable when supplied with 128. These results indicate that the bottleneck in entity tracking is not the amount of temporal coverage, but the model's ability to maintain temporal coherence and perceptual grounding as more frames are introduced. Simply increasing frame density adds redundant information rather than improving entity-centric performance.

6 Conclusion

We introduced NARRATIVETRACK, the first benchmark for evaluating narrative understanding from a bottom-up entity-centric perspective via fine-grained entity tracking. Grounded in a Compositional Reasoning Progression spanning entity existence, changes, and ambiguity with a fully automated pipeline, NARRATIVETRACK provides a scalable and diagnostic framework that systematically measures how MLLMs reason about entities and their temporal evolution. Our results reveal that current MLLMs excel at capturing static visual cues but fail to maintain coherent entity representations under temporal dynamics and visual ambiguity. This exposes a fundamental trade-off between temporal integration and perceptual precision: models can aggregate global context yet often lose fine-grained visual grounding, leading to hallucinated or inconsistent entity representations. Despite scaling and architectural advances, they remain limited by directional bias and weak cross-frame coherence, highlighting the need for entity-centric learning and bidirectional temporal modeling to move MLLMs beyond surface-level recognition toward fine-grained narrative reasoning.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Berglund, L., Tong, M., Kaufmann, M., Balesni, M., Stickland, A.C., Korbak, T., Evans, O.: The reversal curse: Llms trained on "a is b" fail to learn "b is a" (2024), https://arxiv.org/abs/2309.12288
4. Blume, A., Kim, J., Ha, H., Chatikyan, E., Jin, X., Nguyen, K.D., Peng, N., Chang, K.W., Hoiem, D., Ji, H.: Partonomy: Large multimodal models with part-level visual understanding (2025), https://arxiv.org/abs/2505.20759
5. Chen, Y., Wang, X., Peng, H., Ji, H.: Solo: A single transformer for scalable vision-language modeling. arXiv preprint arXiv:2407.06438 (2024)
6. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
7. Cho, J.H., Madotto, A., Mavroudi, E., Afouras, T., Nagarajan, T., Maaz, M., Song, Y., Ma, T., Hu, S., Jain, S., Martin, M., Wang, H., Rasheed, H., Sun, P., Huang, P.Y., Bolya, D., Ravi, N., Jain, S., Stark, T., Moon, S., Damavandi, B., Lee, V., Westbury, A., Khan, S., Krähenbühl, P., Dollár, P., Torresani, L., Grauman, K., Feichtenhofer, C.: Perceptionlm: Open-access data and models for detailed visual understanding (2025), https://arxiv.org/abs/2504.13180
8. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
9. Cutting, J.E.: Narrative theory and the dynamics of popular movies. Psychonomic bulletin & review **23**(6), 1713–1743 (2016)
10. Dalal, D., Vashishtha, G., Mishra, U., Kim, J., Kanda, M., Ha, H., Lazebnik, S., Ji, H., Jain, U.: Constructive distortion: Improving mllms with attention-guided image warping. arXiv preprint arXiv:2510.09741 (2025)
11. Feng, B., Lai, Z., Li, S., Wang, Z., Wang, S., Huang, P., Cao, M.: Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? arXiv preprint arXiv:2505.14321 (2025)
12. Feng, W., Li, J., Saxon, M., Fu, T.j., Chen, W., Wang, W.Y.: Tc-bench: Benchmarking temporal compositionality in text-to-video and image-to-video generation. arXiv preprint arXiv:2406.08656 (2024)
13. Feng, X., Yu, H., Wu, M., Hu, S., Chen, J., Zhu, C., Wu, J., Chu, X., Huang, K.: Narrlv: Towards a comprehensive narrative-centric evaluation for long video generation models. arXiv e-prints pp. arXiv–2507 (2025)
14. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: 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. pp. 24108–24118 (2025)
15. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6047–6056 (2018)

16. Ha, H., Zhan, Q., Kim, J., Bralios, D., Sanniboina, S., Peng, N., Chang, K.W., Kang, D., Ji, H.: Mm-poisonrag: Disrupting multimodal rag with local and global poisoning attacks. arXiv preprint arXiv:2502.17832 (2025)
17. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
18. Kim, J., Ji, H.: Finer: Investigating and enhancing fine-grained visual concept recognition in large vision language models. arXiv preprint arXiv:2402.16315 (2024)
19. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. arXiv preprint arXiv:1809.01696 (2018)
20. Li, B., Ge, Y., Ge, Y., Wang, G., Wang, R., Zhang, R., Shan, Y.: Seed-bench: Benchmarking multimodal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13299–13308 (2024)
21. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22195–22206 (2024)
22. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. arXiv preprint arXiv:2005.00200 (2020)
23. Li, Y., Wang, C., Jia, J.: Llama-vid: An image is worth 2 tokens in large language models. In: European Conference on Computer Vision. pp. 323–340. Springer (2024)
24. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122 (2023)
25. Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., Han, S.: Vila: On pre-training for visual language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26689–26699 (2024)
26. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)
27. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)
28. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
29. Minderer, M., Gritsenko, A., Houlsby, N.: Scaling open-vocabulary object detection. Advances in Neural Information Processing Systems **36**, 72983–73007 (2023)
30. Niu, K., Yu, H., Zhao, M., Fu, T., Yi, S., Lu, W., Li, B., Qian, X., Xue, X.: Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. arXiv preprint arXiv:2502.19958 (2025)
31. Patraucean, V., Smaira, L., Gupta, A., Recasens, A., Markeeva, L., Banarse, D., Koppula, S., Malinowski, M., Yang, Y., Doersch, C., et al.: Perception test: A diagnostic benchmark for multimodal video models. Advances in Neural Information Processing Systems **36**, 42748–42761 (2023)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
33. Saravanan, D., Gupta, V., Singh, D., Khan, Z., Gandhi, V., Tapaswi, M.: Velociti: Benchmarking video-language compositional reasoning with strict entailment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18914–18924 (2025)
34. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In:

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18221–18232 (2024)
35. Tapaswi, M., Zhu, Y., Stiefelhagen, R., Torralba, A., Urtasun, R., Fidler, S.: Movieqa: Understanding stories in movies through question-answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4631–4640 (2016)
36. Team, C.: Chameleon: Mixed-modal early-fusion foundation models, 2024. URL https://arxiv.org/abs/2405.09818 **9**(8) (2024)
37. Tsimpoukelli, M., Menick, J.L., Cabi, S., Eslami, S., Vinyals, O., Hill, F.: Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems **34**, 200–212 (2021)
38. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
39. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., et al.: Lvbench: An extreme long video understanding benchmark. arXiv preprint arXiv:2406.08035 (2024)
40. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
41. Wang, Z., Guo, X., Stoica, S., Xu, H., Wang, H., Ha, H., Chen, X., Chen, Y., Yan, M., Huang, F., Ji, H.: Perception-aware policy optimization for multimodal reasoning (2025), https://arxiv.org/abs/2507.06448
42. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
43. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019)
44. Xiao, J., Shang, X., Yao, A., Chua, T.S.: Next-qa: Next phase of question-answering to explaining temporal actions. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9777–9786 (2021)
45. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1645–1653 (2017)
46. Ye, M., Chen, S., Li, C., Zheng, W.S., Crandall, D., Du, B.: Transformer for object re-identification: A survey. International Journal of Computer Vision pp. 1–31 (2024)
47. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., et al.: mplug-owl: Modularization empowers large language models with multimodality. arXiv preprint arXiv:2304.14178 (2023)
48. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9127–9134. No. 01 (2019)
49. Zacks, J.M., Speer, N.K., Swallow, K.M., Braver, T.S., Reynolds, J.R.: Event perception: a mind-brain perspective. Psychological bulletin **133**(2), 273 (2007)
50. Zhang, J., Jiao, Y., Chen, S., Zhao, N., Tan, Z., Li, H., Chen, J.: Eventhallusion: Diagnosing event hallucinations in video llms. arXiv preprint arXiv:2409.16597 (2024)
51. Zhang, Y., Li, B., Liu, h., Lee, Y.j., Gui, L., Fu, D., Feng, J., Liu, Z., Li, C.: Llava-next: A strong zero-shot video understanding model (April 2024), https://llava-vl.github.io/blog/2024-04-30-llava-next-video/
52. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Llava-video: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)

53. Zheng, Z., Yang, M., Hong, J., Zhao, C., Xu, G., Yang, L., Shen, C., Yu, X.: Deepeyes: Incentivizing" thinking with images" via reinforcement learning. arXiv preprint arXiv:2505.14362 (2025)
54. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 3702–3712 (2019)
55. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)