

NexusQA: Spatio-Temporal Interlacement Grounded VideoQA

Xiaojun Ye*
yexiaojun@zju.edu.cn
Zhejiang University
Hangzhou, China

Sheng Zhou[✉]
zhousheng_zju@zju.edu.cn
Zhejiang University
Hangzhou, China

Bingxin Jia
bingxin.jbx@alibaba-inc.com
Alibaba Group
Hangzhou, China

Chun Wang*
zjuheadmaster@zju.edu.cn
Zhejiang University
Hangzhou, China

Rui Wang
wr246073@alibaba-inc.com
Alibaba Group
Hangzhou, China

Bin Li
yansheng@alibaba-inc.com
Alibaba Group
Hangzhou, China

Jiajun Bu
bjj@zju.edu.cn
Zhejiang University
Hangzhou, China

Junliang Chen*
211830066@smail.nju.edu.cn
Zhejiang University
Hangzhou, China

Zixu Yan
yan.yzx@alibaba-inc.com
Alibaba Group
Hangzhou, China

Yiren Song
songyiren@sjtu.edu.cn
National University of Singapore
Singapore, Singapore

Abstract

Video question answering (VideoQA) requires sophisticated interpretation of spatio-temporal dynamics. While grounded VideoQA has emerged to provide verifiable evidence, existing benchmarks are predominantly confined to Coarse-grained Spatio-Temporal (CST) evidence, such as event or object-level localization. Consequently, they fail to scrutinize the Fine-grained Spatio-Temporal (FST) evidence essential for clarifying subtle semantic states, allowing Multimodal Large Language Models (MLLMs) to exploit language priors instead of performing genuine visual grounding. To address this, we introduce NexusBench, a novel benchmark requiring joint reasoning across Multi-grained Spatio-Temporal evidence. Built upon a Hierarchical Video Scene Graph (HVSG), it explicitly models FST interactions to build logic-intensive queries. To diagnose current model limitations, we propose a metric termed NexusScore, assessing a model's capability to utilize evidence. Our analysis reveals that existing MLLMs exhibit inadequate FST grounding abilities, resulting in suboptimal performance, yet supplying them with the correct evidence drastically improves accuracy. Motivated by this, we present NexusQA, a training-free multi-agent baseline that accumulates evidence through collaborative reasoning. Extensive experiments demonstrate that NexusQA achieves state-of-the-art performance on both the proposed NexusBench and the public NExT-GQA dataset. Our code and benchmark will be released at <https://github.com/Daria8976/NexusQA>.

*These authors contributed equally to this work.



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

MM '26, Rio de Janeiro, Brazil

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11

<https://doi.org/10.1145/3767308.3835487>

CCS Concepts

• **Computing methodologies** → **Scene understanding**; *Multi-agent systems*; • **Information systems** → **Question answering**.

Keywords

Grounded VideoQA; Spatio-Temporal Clue; Fine-Grained

ACM Reference Format:

Xiaojun Ye, Chun Wang, Junliang Chen, Sheng Zhou, Rui Wang, Zixu Yan, Bingxin Jia, Bin Li, Yiren Song, and Jiajun Bu. 2026. NexusQA: Spatio-Temporal Interlacement Grounded VideoQA. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767308.3835487>

1 Introduction

Video question answering (VideoQA), which necessitates the interpretation of complex spatio-temporal dynamics to address open-ended queries, has recently emerged as a fundamental task [30, 36] for advancing Multimodal Large Language Models (MLLMs) [2, 49]. However, current MLLMs often rely on language priors rather than visual evidence, producing plausible but ungrounded answers [54]. Consequently, grounded VideoQA has emerged as an imperative paradigm, which shifts the focus of research toward verifiable visual evidence [4, 5, 25, 40]. As illustrated in Figure 1, grounded VideoQA requires models to move beyond global context to pinpoint fine-grained spatial regions—such as a “battery icon”—across specific temporal intervals to judge dynamic charging state changes.

The central challenge of grounded VideoQA is to identify the exact spatio-temporal evidence supporting an answer [6, 13]. However, existing video benchmarks [26, 40, 44] typically arrest at the Coarse-grained Spatio-Temporal (CST) level, evaluating answer accuracy and temporal localization without verifying fine-grained spatial evidence. Revisiting Figure 1, while existing models [6, 13]

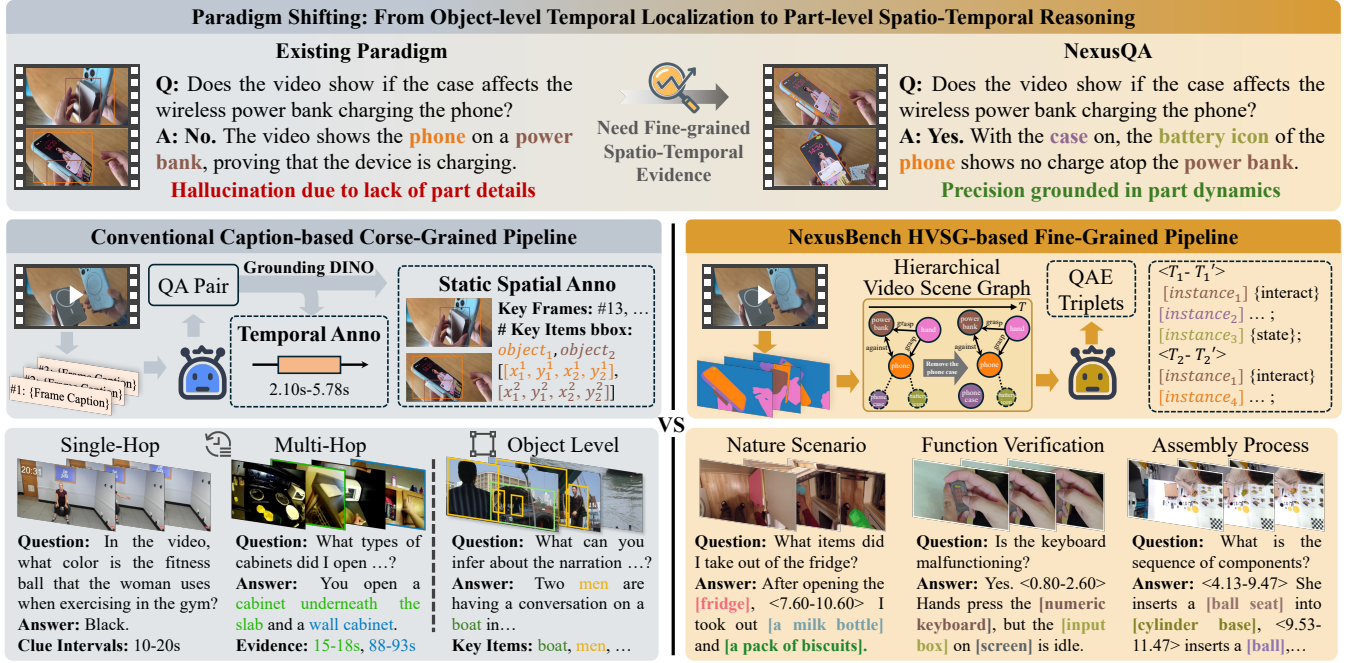


Figure 1: Overview of NexusBench and NexusQA. We introduce a novel grounded VideoQA benchmark, requiring models to pinpoint multi-grained spatio-temporal evidence rather than exploiting language priors.

may successfully localize the coarse interaction between a “phone” and a “power bank”, they often neglect the Fine-grained Spatio-Temporal (FST) evidence, such as the critical state of the “battery icon.” This lack of FST verification masks a significant unreliability, where MLLMs frequently generate logically plausible responses by exploiting linguistic priors while anchoring their reasoning on incorrect spatial evidence [50, 53].

To overcome the coarse-grained and static limitations inherent in prior paradigms, we introduce **NexusBench**, a novel benchmark dedicated to Multi-grained Spatio-Temporal Grounded VideoQA. NexusBench compels models to reason jointly across hierarchical spatio-temporal evidence, truly probing their evidence localization and reasoning capabilities. We construct this dataset by proposing the *Hierarchical Video Scene Graph (HVSG)*, which explicitly models the temporal interactions between parts and objects. HVSG provides structural multi-grained spatio-temporal priors that overcome the perception bottlenecks of standard QA generation pipelines, enabling the creation of logic-intensive queries grounded in fine-grained interactions. Furthermore, we introduce **NexusScore**, which compares direct and evidence-guided VideoQA performance, thereby assessing models’ intrinsic ability to use evidence.

Driven by the conclusion that evidence grounding is the primary bottleneck, we propose a competitive training-free baseline, **NexusQA**, which effectively localizes multi-grained evidence and reasons based on genuine visual evidence. Standalone MLLMs inherently struggle to simultaneously scan long temporal sequences while scrutinizing minute spatial details without suffering from attention dilution [15, 39]. NexusQA addresses this by implementing a “plan-act” collaborative reasoning cycle that decouples the

task into specialized roles to dynamically accumulate evidence within a shared *Nexus Context*. The Strategy Agent acts as the central controller, translating abstract user queries into concrete *Actual Question* and generating an initial *Tool Chain*. Execution is then transferred to the Grounding Agent, which collects evidence through three specific tool calls: the [Temporal Scout Tool] locates event intervals, the [Spatial Inspector Tool] tracks salient part-level masklets, and the [Reasoning Tool] synthesizes feedback for the Strategy Agent or derives the final answer.

The contributions of this work are summarized as follows:

- We develop a hybrid grounded VideoQA data construction pipeline. By combining domain-specific human validation with automated HVSG-guided QAE generation, our approach overcomes the limitation of coarse granularity inherent in conventional VideoQA dataset curation.
- We introduce **NexusBench**, a novel grounded VideoQA benchmark that mandates joint reasoning across multi-grained spatio-temporal evidence, featuring a dual-stage evaluation protocol, **NexusScore**, to independently assess models’ intrinsic capacity to utilize evidence for reasoning.
- We present **NexusQA**, a competitive training-free baseline that effectively localizes and reasons over evidence via collaboration between a Strategy Agent for task concretization and a Grounding Agent for evidence collection.

2 Related Work

Grounded VideoQA Benchmark. Video Question Answering has evolved from basic perception to complex reasoning [23, 31, 35, 48]. However, traditional multiple-choice benchmarks [11, 26, 55]

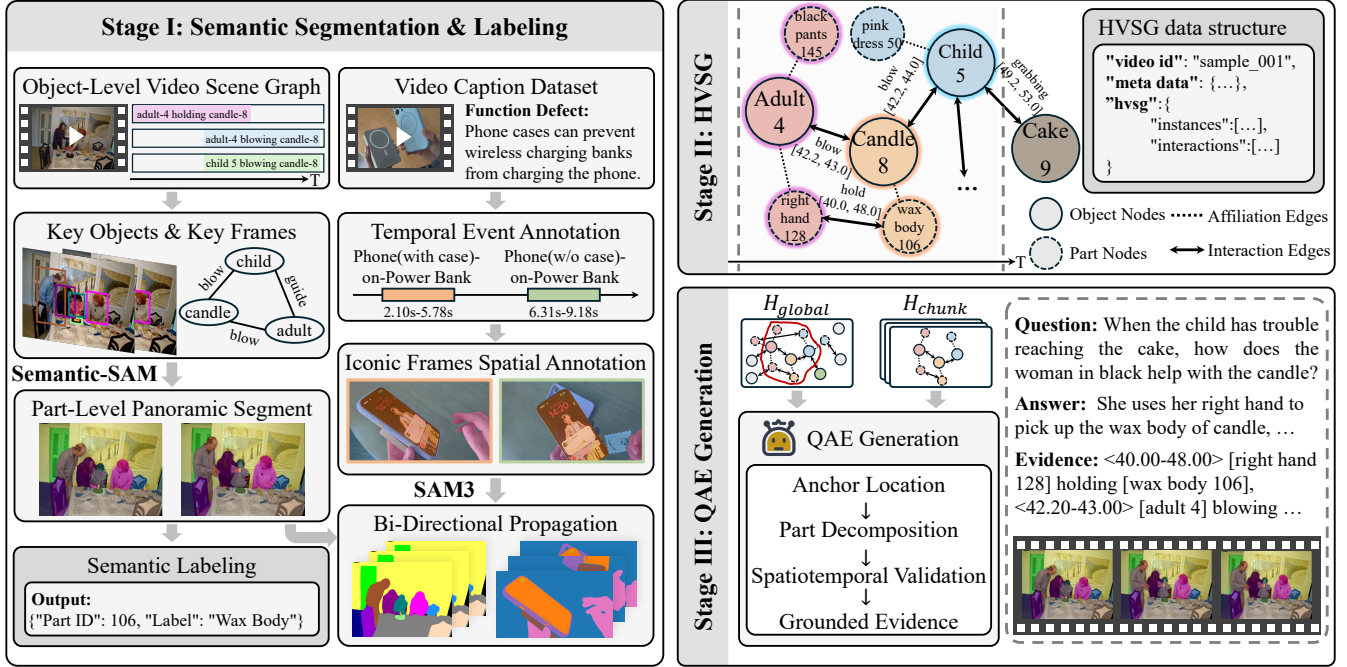


Figure 2: The graph-based generation pipeline of NexusBench. (i) Multi-Granularity Video Segmentation: We apply Semantic-SAM for object part segmentation and SAM3 for propagation, with an additional initial grounding step for newly collected videos. **(ii) HVSG Construction:** We expand the PVSG object-level scene graphs into subgraphs, refining interaction nodes from the object to the instance level. **(iii) QA & Evidence Chain Generation:** Using the video and HVSG, we prompt Gemini-3-flash to generate QA pairs and spatio-temporal evidence chains, followed by difficulty filtering and manual verification.

are susceptible to language priors, enabling models to eliminate distractors without adequate visual understanding [5]. Grounded VideoQA instead requires models to localize evidence supporting their answers [7, 8, 12, 18, 27]. NEX-T-GQA [40] and Multi-Hop EgoQA [6] primarily evaluate temporal localization, with limited spatial precision. VideoEspresso [13] introduces object-level boxes on keyframes but remains limited to static grounding. In contrast, NexusBench requires models to localize dynamic, multi-grained spatio-temporal evidence.

Grounded VideoQA Models. Recent Video MLLMs [41, 51, 57] adopt various temporal grounding strategies. TimeChat [33] and LITA [16] formulate timestamp prediction as next-token generation, whereas E.T. Chat [25] uses embedding matching to preserve numerical continuity. GroundVQA [9] and GeLM [6] additionally localize relevant clips during answer generation. Nevertheless, jointly capturing long temporal contexts and fine-grained spatial details remains challenging [10, 28, 34]. NexusQA addresses this trade-off by decoupling temporal and spatial localization and grounding answers with spatio-temporal mask tubes.

Multi-Agent Systems for Video Understanding. Monolithic MLLMs are prone to attention dilution and error propagation in complex video reasoning [21, 37]. Multi-agent systems mitigate these issues by coordinating specialized agents [14, 45]. VideoMultiAgents [17] and ReAgent-V [56] employ such agents for long-video reasoning

but rely largely on coarse narrative representations and object-level grounding. NexusQA instead separates high-level planning from fine-grained evidence localization, iteratively linking global temporal localization with part-level segmentation.

3 NexusBench

The construction of NexusBench follows a systematic pipeline designed to synthesize multi-granularity grounded VideoQA pairs, as shown in Figure 2. Unlike traditional pipelines that rely on coarse, unstructured raw video-to-text generation, we introduce a novel Hierarchical Video Scene Graph (HVSG) driven approach. This explicit graph representation provides crucial structural priors to guide both logic-intensive QA pair generation and precise spatio-temporal evidence chain construction. To fully evaluate multimodal reasoning capabilities, NexusBench investigates three core scenarios: **Natural Scenes**, **Function Verification**, and **Assembly Process**, encompassing 13 multi-grained sub-tasks.

3.1 Formalized Definition of HVSG for Evidence

To bridge the gap between coarse object-level actions and fine-grained part-level interactions, we formally define the HVSG, which maps a video stream $V \in \mathbb{R}^{T \times H \times W \times 3}$ into a structured representation consisting of: (i) **Hierarchical Nodes ($\mathcal{N}$)**: We define a two-level hierarchy comprising object-level CST nodes $\mathcal{N}_{CST}$ and part-level FST nodes $\mathcal{N}_{FST}$. (ii) **Temporal Interaction Edges ($\mathcal{E}$)**: Inspired by [29, 43], we define edges as dynamic interactions that

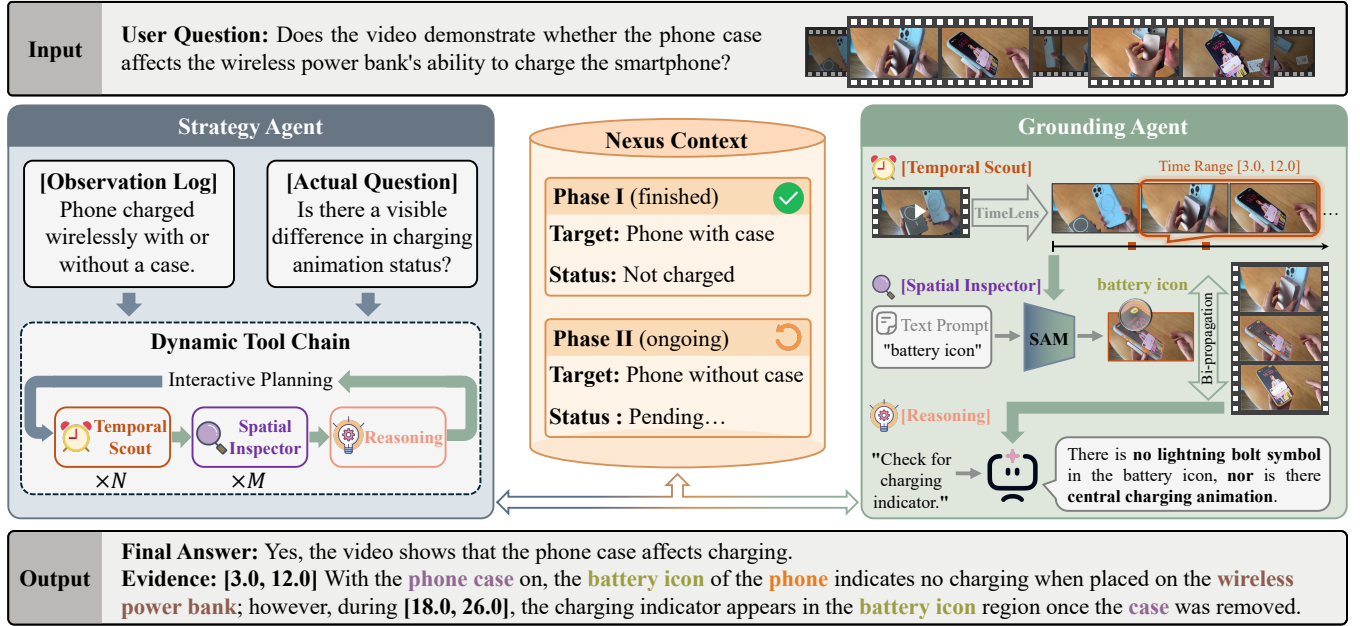


Figure 4: Architecture of NexusQA. The Strategy Agent (left) processes the initial user query to derive an *Actual Question* and formulating an iterative *Dynamic Tool Chain* plan. The Grounding Agent (right) executes this plan via modular tools.

strategy. Specifically, for each video V , we provide the LLM with the global HVSG (H_{global}) to offer a holistic understanding of the video, paired with a specific chunked sub-HVSG (H_{chunk}). The model is then instructed to generate a QAE list that specifically targets the interactions and part-level states described within the current H_{chunk} . Formally, for each video V and the selected chunked sub-HVSG H_{chunk} , the triplets comprising Question (Q), Answer (A), and Evidence (E_o) are generated with prompt ($\mathcal{P}$), denoted as:

$$\{(Q, A, E_o)\} = \text{LLM}(\mathcal{P}; H_{global}; H_{chunk}), \quad (1)$$

Adversarial Refinement & Manual Check. To mitigate language bias, we apply an adversarial refinement mechanism [32]. Specifically, we employ Gemini-3-flash as a “blind” LLM by strictly withholding visual inputs. This model attempts to answer questions based solely on text, and samples correctly answered by the blind model are discarded. Finally, four computer science graduate students serving as validators verify the logic, disambiguate object references, and correct spatio-temporal boundaries to ensure high fidelity. **More details about QAE generation and quality control are provided in Section A.6 and Section A.7.**

3.5 Benchmark Statistics and Analysis

As shown in table 1 and Figure 3, NexusBench stands out as the sole benchmark to offer simultaneous multi-granularity spatio-temporal credibility evaluation. NexusBench comprises 202 high-quality videos sourced from diverse scenarios, including 86 third-person and 116 egocentric videos. A total of 109K frames have been meticulously annotated with pixel-level instance masks, establishing a rich hierarchical structure with 1201 CST nodes (objects), 1113 FST nodes (parts), and 1531 interaction edges. NexusBench includes

1343 question-answer-evidence (QAE) triplets, and each triplet is anchored by a spatio-temporal mask tube ($\mathcal{M}_{CST}$ and $\mathcal{M}_{FST}$), serving as multi-granularity visual evidence that grounds reasoning in both object identity and part-level states. **Further details about sub-tasks are provided in Section A.1.**

4 NexusQA: A Baseline for NexusBench

Unlike conventional multi-agent systems that often rely on single-pass heuristic pipelines and static keyframe retrieval, NexusQA bridges the semantic gap between high-level reasoning queries and pixel-level spatio-temporal evidence through a dynamic “plan-act” paradigm. As illustrated in Figure 4, modeled after human cognitive process, our framework establishes a closed-loop interactive trajectory where the Strategy Agent (the cognitive planner) and the Grounding Agent (the low-level executor) iteratively collaborate via a shared *Nexus Context* to accumulate, verify, and synthesize multi-grained spatio-temporal evidence.

4.1 Strategy Agent

Serving as the system’s planner, the Strategy Agent synthesizes the video content, the user query, and preliminary observations to derive a concretized *Actual Question* and an initial *Tool Chain*, the latter of which is dynamically updated based on the feedback of preceding phases. By sparsely sampling the video, the agent extracts the *Observation Log* relevant to the question, which effectively guides the model to filter out redundant background noise and focus on critical frames. Subsequently, the agent performs question concretization, translating the abstract user query into a verifiable and objective *Actual Question* that focuses on specific interactions and state changes within the scene.

The Strategy Agent then determines the orchestration of the tools and their respective instructions to resolve the *Actual Question*. It constructs a phased *Dynamic Tool Chain* where each phase consists of multiple spatio-temporal localization stages followed by a terminal Reasoning Stage. Each stage directive within the plan is defined by a pair of [Tool Name] and its corresponding Instruction. At the conclusion of each phase, it dynamically adapts the subsequent plan based on the accumulated evidence and historical execution status stored in the *Nexus Context*.

4.2 Grounding Agent

The Grounding Agent acts as a low-level execution policy that instantiates each step of the plan into concrete actions a_t . Each action involves the invocation of specialized tools from a modular toolkit and subsequently collects the resulting output to feed back into the overarching reasoning cycle. We define three primary tools for the agent: the [Temporal Scout Tool], the [Spatial Inspector Tool], and the [Reasoning Tool].

Temporal Scout Tool. For each temporal hypothesis generated by the Strategy Agent, the executor invokes the Temporal Scout Tool, which utilizes TimeLens [47] to perform temporal pruning, thereby preventing attention dilution over long videos. By leveraging an interleaved textual-timestamp encoding mechanism, the tool localizes precise intervals $[t_{start}, t_{end}]$, restricting high-cost spatial analysis to critical short-duration events.

Spatial Inspector Tool. Within the pruned temporal interval, the Spatial Inspector Tool first employs a large language model (LLM) to parse and decompose the composite *Tool Chain* plan instruction into discrete target region descriptions. Subsequently, this tool utilizes a SAM 3 agent [3] to parse linguistic instructions, segment target regions across candidate keyframes, and bidirectionally propagate these masks. This outputs dynamic and multi-grained spatio-temporal masklets that capture fine-grained interactions while explicitly filtering out background noise.

Reasoning Tool. As the terminal step of an execution phase, the MLLM adheres to the strategy instruction to analyze the fine-grained visual states captured within the generated masklets (e.g., verifying the presence of a “lightning bolt” icon) and generates a textual analytical summary of the spatio-temporal evidence.

4.3 Interactive Loop and Termination Condition

Crucially, NexusQA avoids disjointed execution by tightly coupling the agents through the shared *Nexus Context*, which houses the Evidence Bank and monitors execution states. The analytical summary produced by the Reasoning Tool serves as an explicit interrupt signal transmitted back to the Strategy Agent, closing the current plan-act loop. Upon receiving this feedback, the Strategy Agent evaluates the sufficiency of the accumulated evidence against the *Actual Question*. If ambiguities persist, it formulates a new *Tool Chain* phase for further targeted grounding. However, once the Strategy Agent determines the evidence is sufficient, it triggers the termination condition, prompting the Grounding Agent to invoke the Reasoning Tool a final time to seamlessly synthesize the archived multi-phase evidence into a logically rigorous and visually grounded final answer.

5 Experiments

In this section, we first detail the metrics for our benchmark and assess the performance of existing methods. We then present the results of NexusQA, our training-free multi-agent system, establishing a robust baseline for the multi-granularity spatio-temporal grounded VideoQA task.

5.1 Experimental Setup

To comprehensively evaluate the capabilities of MLLMs on grounded VideoQA tasks, we select three categories of representative models: (i) closed-source large models, such as Gemini-3-flash and Qwen-VL-Max [1]; (ii) general-purpose MLLMs that claim strong video understanding capabilities on video benchmarks, such as InternVL3.5 series [38] and Qwen3-VL series [2]; and (iii) popular video MLLMs equipped with temporal grounding capabilities, such as GeLM [6] and VideoMind [24]. **More details about inference settings are presented in Section B.5 and Section B.6. More details on the implementation of NexusQA are presented in Section B.7.**

5.2 Evaluation Metrics.

NexusBench evaluates spatio-temporal grounding, direct QA accuracy, and scaffolded reasoning capability.

Spatio-Temporal Grounding. To robustly evaluate multi-segment temporal grounding, we propose a dynamically penalized mIoU with a bidirectional constraint. Given m predicted non-overlapping time spans $\hat{\mathcal{T}} = \{\hat{T}_1, \hat{T}_2, \dots, \hat{T}_m\}$ and n ground truth spans $\mathcal{T} = \{T_1, T_2, \dots, T_n\}$, the segment-matched **mIoU** is computed as:

$$\text{mIoU}(\mathcal{T}, \hat{\mathcal{T}}) = \frac{1}{\max(n, m)} \sum_{i=1}^m \max_{T_j \in \mathcal{T}} \text{IoU}(\hat{T}_i, T_j) \quad (2)$$

Additionally, to assess temporal precision, we compute **mIoP** [6] to evaluate the exact temporal overlap relative to the total predicted duration. To evaluate spatial evidence localization, we compute the spatial mean intersection over union (**SIoU**) across specific keyframes, first converting all predicted and ground-truth masks to minimal bounding boxes to ensure a fair comparison.

Open-Ended VideoQA. To prevent potential information leakage common in multiple-choice formats, we utilize Gemini-3-flash as an LLM judge to evaluate open-ended responses against ground-truth answers. We also report Acc@IoU [4] for comparison. **The human-LLM agreement analysis is detailed in Section B.3.**

Scaffolded Reasoning Evaluation. Given N evaluation instances, let $D_i \in \{0, 1\}$ and $S_i \in \{0, 1\}$ denote the correctness of the model’s prediction for the i -th instance under the direct QA setting and the scaffolded setting where spatio-temporal evidence is explicitly provided, respectively. We introduce **NexusScore** to measure a model’s ability to correct initial failures using external priors:

$$\text{NexusScore} = \frac{\sum_{i=1}^N \mathbb{I}(D_i = 0 \wedge S_i = 1)}{\sum_{i=1}^N \mathbb{I}(D_i = 0)} \quad (3)$$

where $\mathbb{I}(\cdot)$ is the indicator function. The denominator counts autonomous reasoning failures, while the numerator counts instances successfully corrected by the scaffolded prompt. **A rigorous statistical stability analysis is detailed in Section B.4.**

Table 2: Spatio-Temporal Grounding and Direct VideoQA Performance on NexusBench. Unlike existing MLLMs that struggle with fine-grained spatio-temporal evidence, NexusQA demonstrates robust grounding and reasoning capabilities.

Models	Natural Scenes					Function Verification					Assembly Process				
	mIoU	mIoP	SIoU	Acc	Acc@IoU	mIoU	mIoP	SIoU	Acc	Acc@IoU	mIoU	mIoP	SIoU	Acc	Acc@IoU
Closed-source MLLMs															
Gemini-3-flash	15.81	21.42	29.59	55.22	40.04	35.50	43.24	19.52	56.14	55.89	12.44	17.55	0.95	28.54	22.54
Qwen-VL-Max	16.01	21.89	18.10	<u>56.55</u>	<u>41.37</u>	31.08	38.12	13.27	33.33	32.83	10.03	13.27	0.92	20.86	16.55
Open-source MLLMs															
VideoMind-7B	5.56	6.58	-	35.67	28.08	26.59	32.98	-	25.81	25.81	8.92	9.72	-	15.83	11.51
VideoLLaMA3-7B	9.12	11.66	9.92	33.97	28.46	29.05	39.15	3.92	30.33	28.57	7.56	8.73	1.01	13.43	8.87
EgoQA-GeLM-7B	1.54	18.33	-	33.59	14.23	15.90	30.26	-	10.03	7.52	6.17	8.47	-	12.23	5.28
InternVL3.5-8B	4.22	7.09	10.14	36.81	20.68	8.45	9.84	3.72	29.82	25.56	0.64	0.84	1.26	15.59	8.39
InternVL3.5-38B	5.11	6.57	12.78	37.95	25.62	18.70	20.76	6.31	31.08	29.57	0.10	0.12	0.99	18.23	9.11
Qwen3-VL-8B	8.21	10.65	23.00	38.14	26.19	28.99	33.22	18.91	30.83	30.08	6.66	7.43	2.58	15.11	8.87
Qwen3-VL-235B	10.70	13.97	25.07	40.23	29.79	33.46	36.93	22.59	41.60	41.60	7.55	8.77	3.96	23.50	17.03
NexusQA-8B	<u>17.35</u>	<u>23.56</u>	28.79	45.92	30.74	33.60	<u>58.14</u>	<u>26.24</u>	31.58	31.33	<u>17.21</u>	<u>20.66</u>	<u>10.62</u>	23.02	19.90
NexusQA-235B	20.01	27.37	30.02	59.01	41.75	<u>34.60</u>	60.86	27.75	<u>48.62</u>	<u>47.37</u>	17.66	21.54	11.73	37.89	30.70

Table 3: Guided Reasoning Evaluation on NexusBench. Explicit spatio-temporal guidance substantially boosts QA accuracy across all models. This demonstrates that the primary bottleneck in complex video understanding lies in fine-grained evidence localization. T_{GT} and S_{GT} represent ground-truth temporal windows and spatial regions.

Models	Natural			Function			Assembly	
	Acc	NexusScore		Acc	NexusScore		Acc	NexusScore
VideoLLaMA3-7B	33.97			30.33			13.43	
+ T_{GT}	47.82 _{+13.85}	30.17		33.33 _{+3.00}	19.78		17.51 _{+4.08}	17.17
+ $T_{GT} + S_{GT}$	49.34 _{+15.37}			34.09 _{+3.76}			20.38 _{+6.95}	
InternVL3.5-8B	36.81			29.82			15.59	
+ T_{GT}	48.39 _{+11.58}	34.53		41.10 _{+11.28}	<u>29.29</u>		18.71 _{+3.12}	16.76
+ $T_{GT} + S_{GT}$	53.89 _{+17.08}			41.60 _{+11.78}			25.18 _{+9.59}	
InternVL3.5-38B	37.95			31.08			18.23	
+ T_{GT}	52.75 _{+14.80}	35.47		45.86 _{+14.78}	35.27		20.14 _{+1.91}	20.53
+ $T_{GT} + S_{GT}$	<u>55.79</u> _{+17.84}			<u>46.12</u> _{+15.04}			25.90 _{+7.67}	
Qwen3-VL-8B	38.14			30.83			15.11	
+ T_{GT}	51.23 _{+13.09}	<u>35.91</u>		43.36 _{+12.53}	25.83		25.18 _{+10.07}	<u>24.86</u>
+ $T_{GT} + S_{GT}$	55.60 _{+17.46}			44.11 _{+13.28}			31.18 _{+16.07}	
Qwen3-VL-235B	40.23			41.60			23.50	
+ T_{GT}	52.18 _{+11.95}	37.82		44.36 _{+2.76}	22.75		<u>31.41</u> _{+7.91}	25.08
+ $T_{GT} + S_{GT}$	58.25 _{+18.02}			47.12 _{+5.52}			35.49 _{+11.99}	

5.3 Results on NexusBench

Evidence Grounding Proficiency. As illustrated in Table 2, existing state-of-the-art MLLMs exhibit suboptimal performance in spatio-temporal evidence localization on NexusBench. Despite being the top-performing model among existing MLLMs, Gemini-3-flash achieves a limited weighted average mIoU of **20.61%**, mIoP of **26.70%**, and SIoU of **17.71%** across the evaluated scenarios, underscoring the challenges NexusBench poses by requiring fine-grained localization. Our proposed NexusQA-235B exhibits superior proficiency in evidence grounding, establishing a competitive baseline for the multi-grained grounded VideoQA task. Compared to its backbone model Qwen3-VL-235B, NexusQA-235B enhances mIoU, mIoP,

and SIoU by an weighted average of **7.13%**, **16.33%**, and **5.89%**, respectively. Notably, NexusQA-235B even surpasses Gemini-3-flash in the Natural Scenes and Assembly Process scenarios. We observed that existing MLLMs tend to terminate the evidence-seeking process once an initial, often coarse, visual cue is localized in the early segments of a video. This leads to failure in cases where the decisive evidence is situated in the latter half of the sequence. In contrast, NexusQA effectively mitigates this bias through a dynamic plan-and-act loop. The Strategy Agent evaluates gathered cues against the historical context, autonomously optimizing the Tool Chain to scrutinize subsequent windows until ambiguities are resolved.

Question Answering Accuracy. As illustrated in Table 2, based on precise spatio-temporal evidence localization, NexusQA-235B significantly improves direct QA performance compared to its base model, achieving absolute gains of **18.78%**, **7.02%**, and **14.39%** across the Natural, Function, and Assembly scenarios, respectively. Furthermore, NexusQA reaches a remarkable weighted average accuracy of **49.37%**, surpassing the leading closed-source model, Gemini-3-flash, by **2.16%**. These results confirm the effectiveness of NexusQA in bridging the semantic gap by decomposing abstract user queries into executable tool chain grounded in pixel-level evidence and mandating the localization of spatio-temporal evidence prior to formulating an answer. In addition, NexusQA achieves excellent results on Acc@IoU metric, proving that its responses are genuinely grounded in verifiable visual evidence rather than spurious correlations. To intuitively illustrate our proposed framework and further substantiate our quantitative findings, **we present comprehensive qualitative case studies in Section C.4.**

Scaffolded Reasoning Evaluation. We further analyze the guided reasoning capabilities of models as presented in Table 3. When explicitly provided with ground-truth temporal windows (T_{GT}) and spatial regions (S_{GT}), all models exhibit substantial performance leaps. This suggests that MLLMs’ reasoning potential is often hindered primarily by deficits in autonomous fine-grained localization rather than an inherent lack of logic. **NexusScore** is designed to quantify the error-correction rate. Comparing models with different parameters, we observe a positive correlation between parameter scale and evidence utilization efficiency. For example, Qwen3-VL-235B achieves a higher NexusScore than Qwen3-VL-8B in the Natural (37.82 vs. 35.91) and Assembly (25.08 vs. 24.86) scenarios, and InternVL3.5-38B similarly outperforms its 8B counterpart across the board. This demonstrates that models with larger parameter counts possess superior cognitive capacities to comprehend and actively leverage provided fine-grained evidence.

5.4 Ablation Studies

As shown in Table 4, we ablate the major components of the proposed MAS framework across three scenarios.

Ablation on Grounding Tools. To validate the relative importance of spatio-temporal evidence grounding, we independently ablate the Spatial Inspector and the Temporal Scout. A consistent trend emerges across both 8B and 235B models: removing the Spatial Inspector inflicts a more severe performance drop than removing the Temporal Scout. This highlights that for grounded VideoQA tasks, merely localizing the temporal range is insufficient for existing models. Spatial localization proves crucial, particularly in scenarios like Function that demand close attention to fine-grained state changes within specific regions.

Ablation on Dynamic Planning. We replace the MAS interactive communication with a single-hop baseline (*w/o Dynamic Planning*) that generates static plans. In Natural and Function scenarios, removing dynamic planning noticeably degrades performance, as conditional logic strictly requires multi-hop reasoning. In the Assembly scenario, removing multi-turn planning slightly improves the 235B model but degrades the lightweight 8B model, which lacks the capacity to robustly process entire sequences in a single pass.

Table 4: Ablation Study across Different Scenarios.

Settings	Natural	Function	Assembly
NexusQA-8B	45.92	31.58	23.02
w/o Dynamic Planning	37.19	17.54	20.38
w/o Spatial Inspector	30.17	10.03	8.39
w/o Temporal Scout	33.97	19.80	9.83
NexusQA-235B	59.01	48.62	37.89
w/o Dynamic Planning	54.46	42.11	38.85
w/o Spatial Inspector	51.61	23.31	35.01
w/o Temporal Scout	52.18	45.36	36.45

Table 5: Grounded VideoQA on NExT-GQA.

Method	IoU		IoP		Acc@GQA
	R@0.5	mIoU	R@0.5	mIoP	
LLoVi [46]	15.3	20.0	36.9	37.3	24.3
VideoChat-TPO [42]	23.4	27.7	32.8	35.6	25.5
VideoMind [24]	25.8	31.4	35.3	39.0	28.2
VideoChat-R1 [22]	15.83	20.46	24.29	27.32	19.68
VideoChat-R1.5 [41]	12.57	19.35	17.59	23.67	14.39
Temporal-R1 [20]	14.30	19.56	21.71	25.76	17.71
NexusQA-8B	34.25	38.71	46.79	50.12	36.40
NexusQA-235B	36.71	40.12	48.39	52.80	39.87

5.5 Results on NExT-GQA

To further validate the generalization and robustness of our proposed framework, we evaluate NexusQA on NExT-GQA, a widely recognized public benchmark for grounded VideoQA. As shown in Table 5, NexusQA-235B outperforms existing methods across all spatial and temporal localization metrics. Compared to the previous best model, it yields absolute gains of 8.72% in mIoU, 13.80% in mIoP, and 11.67% in overall Acc@GQA. These results demonstrate that our fine-grained spatio-temporal evidence grounding and dynamic plan-and-act loop are not limited to NexusBench, but are highly effective across other complex video understanding scenarios.

6 Conclusion

In this paper, we introduce **NexusBench**, a novel benchmark that shifts the Grounded VideoQA paradigm from coarse object-level recognition to multi-grained spatio-temporal grounding via a Hierarchical Video Scene Graph. To diagnose capability limitations, we propose a dual-stage evaluation metric termed **NexusScore**, assessing models’ intrinsic capacity to utilize visual evidence for reasoning. Experimental results reveal that explicitly providing multi-grained spatio-temporal evidence significantly improves grounded VideoQA performance. To address the evidence localization bottleneck exposed by our benchmark, we present **NexusQA**, a competitive training-free multi-agent baseline. Through a “plan-act” collaborative reasoning loop, NexusQA decouples complex video understanding into specialized tasks to accumulate verifiable visual evidence. Ultimately, we hope this work inspires the community to move beyond blind reasoning and develop more reliable MLLMs for complex video understanding.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No: 62372408), Ningbo Natural Science Foundation (Grant No: 2025J001).

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966* (2023).
- [2] 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, Sibao 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. 2025. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631* (2025).
- [3] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. 2025. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719* (2025).
- [4] Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. 2024. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075* (2024).
- [5] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Frank Wang. 2024. Rextime: A benchmark suite for reasoning-across-time in videos. *Advances in Neural Information Processing Systems* 37 (2024), 28662–28673.
- [6] Qirui Chen, Shangzhe Di, and Weidi Xie. 2025. Grounded multi-hop videoqa in long-form egocentric videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 2159–2167.
- [7] Weixing Chen, Yang Liu, Binglin Chen, Jiandong Su, Yongsun Zheng, and Liang Lin. 2025. Cross-modal causal relation alignment for video question grounding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 24087–24096.
- [8] Andong Deng, Tongjia Chen, Shoubin Yu, Taojiannan Yang, Lincoln Spencer, Yapeng Tian, Ajmal Saeed Mian, Mohit Bansal, and Chen Chen. 2025. Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 8625–8636.
- [9] Shangzhe Di and Weidi Xie. 2024. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12934–12943.
- [10] Yuchen Fang, Hao Miao, Yuxuan Liang, Liwei Deng, Yue Cui, Ximu Zeng, Yuyang Xia, Yan Zhao, Torben Bach Pedersen, Christian S Jensen, et al. 2026. Unraveling spatio-temporal foundation models via the pipeline lens: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering* (2026).
- [11] Chaoyou Fu, Yuhang Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 24108–24118.
- [12] Madeleine Grunze-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. 2021. AGQA: A Benchmark for Compositional Spatio-Temporal Reasoning. arXiv:2103.16002 [cs.CV] <https://arxiv.org/abs/2103.16002>
- [13] Songhao Han, Wei Huang, Hairong Shi, Le Zhuo, Xiu Su, Shifeng Zhang, Xu Zhou, Xiaojuan Qi, Yue Liao, and Si Liu. 2025. Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 26181–26191.
- [14] Junda He, Christoph Treude, and David Lo. 2025. Llm-based multi-agent systems for software engineering: Literature review, vision, and the road ahead. *ACM Transactions on Software Engineering and Methodology* 34, 5 (2025), 1–30.
- [15] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations*.
- [16] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. 2024. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision*. Springer, 202–218.
- [17] Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeep Khatua, Nidhi Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, et al. 2025. VideoMultiAgents: A Multi-Agent Framework for Video Question Answering. *arXiv preprint arXiv:2504.20091* (2025).
- [18] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020. TVQA+: Spatio-Temporal Grounding for Video Question Answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 8211–8225. doi:10.18653/v1/2020.acl-main.730
- [19] F Li, H Zhang, P Sun, X Zou, S Liu, J Yang, C Li, L Zhang, and J Gao. 2023. Semantic-sam: Segment and recognize anything at any granularity. *arXiv 2023. arXiv preprint arXiv:2307.04767* (2023).
- [20] Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. 2025. Reinforcement Learning Tuning for VideoLLMs: Reward Design and Data Efficiency. *arXiv preprint arXiv:2506.01908* (2025).
- [21] Haoran Li, Ziyi Su, Yun Xue, Zhiliang Tian, Yiping Song, and Minlie Huang. 2025. Advancing collaborative debates with role differentiation through multi-agent reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 22655–22666.
- [22] Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. 2025. VideoChat-R1: Enhancing Spatio-Temporal Perception via Reinforcement Fine-Tuning. *arXiv preprint arXiv:2504.06958* (2025).
- [23] Huabin Liu, Filip Ilievski, and Cees GM Snoek. 2025. Commonsense video question answering through video-grounded entailment tree reasoning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 3262–3271.
- [24] Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. 2026. VideoMind: A Chain-of-LoRA Agent for Temporal-Grounded Video Reasoning. In *International Conference on Learning Representations (ICLR)*.
- [25] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang W Chen. 2024. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems* 37 (2024), 32076–32110.
- [26] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems* 36 (2023), 46212–46244.
- [27] Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. 2025. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 19036–19046.
- [28] Muhammad Muzammul and Xi Li. 2025. Comprehensive review of deep learning-based tiny object detection: challenges, strategies, and future directions. *Knowledge and Information Systems* 67, 5 (2025), 3825–3913.
- [29] Trong-Thuan Nguyen, Pha Nguyen, and Khoa Luu. 2024. Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18384–18394.
- [30] Jean Park, Kuk Jin Jang, Basam Alasaly, Sriharsha Mopidevi, Andrew Zolensky, Eric Eaton, Insup Lee, and Kevin Johnson. 2025. Assessing modality bias in video question answering benchmarks with multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 19821–19829.
- [31] Haiyi Qiu, Minghe Gao, Long Qian, Kaihang Pan, Qifan Yu, Juncheng Li, Wenjie Wang, Siliang Tang, Yueting Zhuang, and Tat-Seng Chua. 2025. STEP: Enhancing Video-LLMs’ Compositional Reasoning by Spatio-Temporal Graph-guided Self-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3284–3294.
- [32] Ruchit Rawal, Khalid Saifullah, Miquel Farré, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. 2024. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813* (2024).
- [33] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14313–14323.
- [34] Yang Shi, Jiaheng Liu, Yushuo Guan, Zhenhua Wu, Yuanxing Zhang, Zihao Wang, Weihong Lin, Jingyun Hua, Zekun Wang, Xinlong Chen, et al. 2025. Mavors: Multi-granularity video representation for multimodal large language model. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 10994–11003.
- [35] Zihan Song, Xin Wang, Zi Qian, Hong Chen, Longtao Huang, Hui Xue, and Wenwu Zhu. 2025. Modularized self-reflected video reasoner for multimodal llm with application to video question answering. In *Forty-second International Conference on Machine Learning*.
- [36] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology* (2025).
- [37] Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms:

- A survey of llms. *arXiv preprint arXiv:2501.06322* (2025).
- [38] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265* (2025).
 - [39] Jianlong Wu, Wei Liu, Ye Liu, Meng Liu, Liqiang Nie, Zhouchen Lin, and Chang Wen Chen. 2025. A survey on video temporal grounding with multimodal large language model. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).
 - [40] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. 2024. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13204–13214.
 - [41] Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. 2025. VideoChat-R1.5: Visual Test-Time Scaling to Reinforce Multimodal Reasoning by Iterative Perception. *arXiv preprint arXiv:2509.21100* (2025).
 - [42] Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, et al. 2025. Task preference optimization: Improving multimodal large language models with vision task alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29880–29892.
 - [43] Jingkang Yang, Wenxuan Peng, Xiangtai Li, Zujin Guo, Liangyu Chen, Bo Li, Zheng Ma, Kaiyang Zhou, Wayne Zhang, Chen Change Loy, et al. 2023. Panoptic video scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18675–18685.
 - [44] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 9127–9134.
 - [45] Yanwei Yue, Guibin Zhang, Boyang Liu, Guancheng Wan, Kun Wang, Dawei Cheng, and Yiyan Qi. 2025. Masrouter: Learning to route llms for multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15549–15572.
 - [46] Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 21715–21737.
 - [47] Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, Ying Shan, and Limin Wang. 2025. TimeLens: Rethinking Video Temporal Grounding with Multimodal LLMs. *arXiv preprint arXiv:2512.14698* (2025).
 - [48] Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. 2025. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20626–20636.
 - [49] Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024. LLaVA-NeXT: A Strong Zero-shot Video Understanding Model. <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>
 - [50] Zefeng Zhang, Hengzhu Tang, Jiawei Sheng, Zhenyu Zhang, Yiming Ren, Zhenyang Li, Dawei Yin, Duohe Ma, and Tingwen Liu. 2025. Debiasing multimodal large language models via noise-aware preference optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9423–9433.
 - [51] Zijia Zhao, Yuqi Huo, Tongtian Yue, Longteng Guo, Haoyu Lu, Bingning Wang, Weipeng Chen, and Jing Liu. 2025. Efficient motion-aware video mllm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 24159–24168.
 - [52] Hao Zheng, Regina Lee, and Yuqian Lu. 2023. Ha-vid: A human assembly video dataset for comprehensive assembly knowledge understanding. *Advances in Neural Information Processing Systems* 36 (2023), 67069–67081.
 - [53] Xu Zheng, Chenfei Liao, Yuqian Fu, Kaiyu Lei, Yuanhuiyi Lyu, Lutaog Jiang, Bin Ren, Jiale Chen, Jiawen Wang, Chengxin Li, et al. 2025. Mllms are deeply affected by modality bias. *arXiv preprint arXiv:2505.18657* (2025).
 - [54] Chenyue Zhou, Mingxuan Wang, Yanbiao Ma, Chenxu Wu, Wanyi Chen, Zhe Qian, Xinyu Liu, Yiwei Zhang, Junhao Wang, Hengbo Xu, et al. 2025. From perception to cognition: A survey of vision-language interactive reasoning in multimodal large language models. *arXiv preprint arXiv:2509.25373* (2025).
 - [55] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. 2025. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 13691–13701.
 - [56] Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. 2025. ReAgent-V: A Reward-Driven Multi-Agent Framework for Video Understanding. *arXiv preprint arXiv:2506.01300* (2025).
 - [57] Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, et al. 2025. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 18891–18901.