Evidence-Backed Video Question Answering

Shijie Wang[1,2], Honglu Zhou[1], Ziyang Wang[1], Ran Xu[1], Caiming Xiong[1], Silvio Savarese[1], Chen Sun[2], and Juan Carlos Niebles[1]

[1] Salesforce, Palo Alto, CA, USA
[2] Brown University, Providence, RI, USA

Abstract. Current Video Large Language Models (Video LLMs) excel in question answering (QA) but largely operate as black boxes, providing textual answers without verifiable visual grounding. Existing explainability efforts rely on textual rationales or sparse bounding boxes, which struggle to capture complex video dynamics such as occlusions and non-rigid deformations. We propose **Evidence-Backed Video Question Answering (E-VQA)**, a novel task requiring models to jointly output a semantic answer and precise spatio-temporal evidence: temporal segments and dense, tracked object segmentation masklets. To support this, we introduce **ST-Evidence**, the first human-verified benchmark for both discriminative and generative pixel-level grounding. Evaluations of state-of-the-art models reveal a critical decoupling between QA accuracy and true visual perception that scaling alone fails to bridge. To address this, we develop scalable, automated generation pipelines to create **ST-Evidence-Instruct**, a 160k-scale dataset bridging high-level reasoning with fine-grained grounding. Fine-tuning grounded Video LLMs on this data yields substantial gains over the corresponding size-matched UniPixel baselines (e.g., +27.2 t-mean and +13.8 $\mathcal{J}\&\mathcal{F}$ on a 7B model), establishing a robust baseline for explainable, evidence-backed video understanding. Code and data are available at `https://github.com/SalesforceAIResearch/EVQA`.

1 Introduction

Recent Video Large Language Models (Video LLMs) have demonstrated impressive results on Video Question Answering (Video QA) benchmarks. However, most remain black-box systems that provide only textual answers, raising significant concerns regarding trust, explainability, and reliability. Without verifiable evidence, these models may rely on language priors or hallucinations rather than genuine visual perception; when errors occur, tracing their origin is nearly impossible. Although recent work explores video reasoning via language-based chain-of-thought (CoT) reasoning [12, 16, 22, 46, 47], these textual rationales often lack grounding in concrete spatio-temporal visual evidence. This lack of traceability is particularly critical in high-stakes domains such as autonomous driving, medical procedure analysis and human–robot collaboration, where decision-making requires rigorous visual justification.

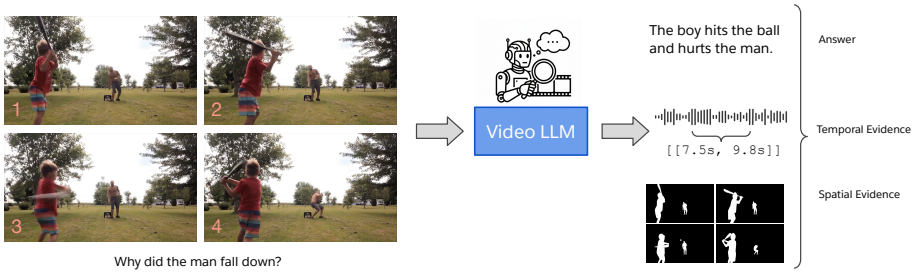

Fig. 1: E-VQA: Evidence-Backed Video Question Answering. Models provide textual answers to video questions while grounding their reasoning in spatio-temporal evidence, including relevant temporal video segments and densely tracked segmentation masks that highlight the spatio-temporal visual regions supporting the answer.

To bridge the gap between semantic reasoning and visual perception, we argue that effective video understanding requires evidence grounding: the ability to explicitly justify an answer with its corresponding spatio-temporal visual evidence. Existing grounded video reasoning work [8, 30, 52] predominantly focuses on answer grounding (identifying the object that constitutes the answer), often relying on spatially sparse bounding boxes over temporally limited keyframes. However, such sparse grounding lacks the precision to capture complex video dynamics, such as severe occlusions, continuous state changes, non-rigid deformations (e.g., liquids, wires), or identity preservation through crossovers (e.g., tracking a specific cup in a shell game). We argue that to unambiguously verify that a model has perceived the correct visual cues, it must perform dense, pixel-level spatio-temporal tracking as a formal evidence trail.

To formalize this task, we introduce **Evidence-Backed Video Question Answering** (**E-VQA**) (Fig. 1). Given a video and a question, a model must generate a triplet: (i) the semantic answer, (ii) the supporting temporal evidence (relevant video segments), and (iii) the supporting spatial evidence represented as dense, tracked spatio-temporal segmentation masks (masklets). By requiring these components, E-VQA explicitly couples high-level reasoning with low-level grounding. This formulation forces a model to not only solve the linguistic task but also to provide a verifiable, pixel-level justification of its perception.

As no existing datasets support this setting, we introduce **ST-Evidence**, the first benchmark specifically designed for E-VQA. ST-Evidence is constructed through a rigorous three-stage semi-automatic pipeline: *(1) Sample*: Select suitable video–question–answer pairs from existing Video QA datasets; *(2) Annotate*: Human annotators provide dense spatio-temporal evidence or reject low-quality samples; and *(3) Verify*: Independent human reviewers assess annotation quality and flag samples for re-annotation if necessary.

Since current General Video LLMs are typically not designed for dense, generative spatio-temporal grounding, we propose two benchmark variants:

ST-Evidence-Gen: A generative setting requiring the model to synthesize the complete triplet, coupling reasoning with dense pixel-level grounding.

ST-Evidence-MCQ: A multiple-choice formulation where the model must (1) select the correct answer, (2) identify the relevant temporal segment, and (3) choose the correct spatial mask from a candidate set. This variant focuses on discriminative and sparse grounding.

We evaluate leading General Video LLMs—including open-source models such as Qwen3-VL [37] and proprietary systems such as OpenAI-o3 [33] and Gemini-2.5-Pro [9]. Our results on ST-Evidence-MCQ show that high QA accuracy does not correlate with grounding proficiency; most open-source models perform near random-guess levels in spatial evidence selection. For ST-Evidence-Gen, we employ a two-step pipeline where General Video LLMs answer the question, generate temporal segments, and describe key evidence objects in text, which are then parsed by a multimodal grounding model (UniPixel [27]) to produce spatio-temporal masks. Under this regime, both temporal and spatial grounding scores remain low. Even specialized *Grounded* Video LLMs, such as UniPixel [27] and Sa2VA [55], struggle with the joint reasoning required for E-VQA. Our comprehensive evaluations suggest that scaling alone is insufficient; progress in E-VQA demands fundamental advances in architecture, data quality, and integrated training objectives.

To address these limitations, we introduce **ST-Evidence-Instruct**, a large-scale instruction-tuning dataset containing **160k** triplets of (QA pairs, temporal evidence, spatial evidence). While prior datasets are either purely *semantic* (QA) or *grounded* (localization), we bridge this gap by synthesizing aligned spatio-temporal evidence for existing QA pairs and by repurposing masklet-annotated videos for reasoning tasks. Our fully automatic pipelines enable generation of the required triplets at scale. Fine-tuning UniPixel [27] on **ST-Evidence-Instruct** yields improvements in QA accuracy and significant gains in dense grounding, validating our data-centric approach to enhancing E-VQA performance.

Our main contributions are summarized as follows: **(1) A new task:** We introduce Evidence-Backed Video Question Answering (**E-VQA**), a task requiring models to provide verifiable justification through the joint output of semantic answers, temporal segments, and dense spatial masklets. **(2) A comprehensive benchmark:** We construct ST-Evidence, the first E-VQA benchmark, featuring human-verified spatio-temporal annotations across generative (ST-Evidence-Gen) and discriminative (ST-Evidence-MCQ) evaluation variants. **(3) Large-scale instruction-tuning dataset:** We develop scalable, automated pipelines to bridge semantic and grounded video data, yielding ST-Evidence-Instruct with 160k triplets. Fine-tuning current grounded Video LLMs on this dataset achieves considerable gains in joint reasoning and pixel-level evidence grounding.

2 Related Work

Video Large Language Models. Video LLMs extend LLMs to reason over multimodal video inputs. While proprietary (e.g., GPT-4o [32] and o3 [33], Gemini 2.5 [9]) and open-source models (e.g., Video-LLaMA3 [57], InternVL-3.5 [43], Qwen3-VL [37]) show strong comprehension, they generate purely textual re-

sponses without visual justification. To improve logical depth, recent Reasoning Video LLMs apply reinforcement learning (RL) and test-time scaling to generate multi-step reasoning chains (e.g., Video-R1 [12], VideoChat-R1 [22], Video-RTS [46]). Some emerging models attempt to ground these chains: VITED [28] and Time-R1 [44] focus on temporal segments, while Open-O3 Video [30], Seg-R1 [54] and others [13] incorporate sparse bounding boxes or "think-before-segment" strategies. However, these works typically treat grounding as an intermediate "scratchpad" to improve textual accuracy. In contrast, E-VQA requires dense, spatio-temporal masklets as a formal component of the final answer triplet, demanding a pixel-level "proof" of the model's underlying reasoning.

Spatio-Temporal Grounded Video LLMs. Efforts to bridge language and perception have yielded Temporally Grounded Video LLMs (VTimeLLM [18], Momentor [36], VTG-LLM [15]) that predict segment boundaries but lack spatial grounding. To add spatial grounding, some models (VideoMolmo [1], NumPro [49], VGR [42]) utilize bounding boxes, while pixel-level models (Sa2VA [55], UniPixel [27], VideoLISA [5], VideoGLaMM [31], GLUS [25]) integrate segmentation decoders for dense masks. However, these methods often treat grounding and QA as separate tasks, rather than addressing reasoning and justification jointly. We address this via instruction tuning that compels models to synergize high-level reasoning with dense spatio-temporal tracking, moving beyond disjoint task execution.

Video QA and Grounding Datasets. Existing datasets do not fully support dense, evidence-backed reasoning. Semantic datasets, such as NeXT-QA [50] and STAR [48], focus on causal reasoning but lack grounding annotations. Conversely, recent grounded video datasets have emerged, such as NeXT-GQA [51], V-STaR [8], VideoEspresso [16], and CG-Bench [7], as well as referential comprehension video datasets (e.g., SAMA [40], VideoRefer [56], and Strefer [58]). Earlier grounded Video QA efforts expose supporting evidence in different forms: STAIR [45] audits intermediate spatio-temporal results, TranSTR [23] selects question-critical moments and objects as rationales, and TVQA+ [20] augments Video QA with temporal moments and sparse bounding boxes. In contrast, E-VQA requires temporal segments and dense tracked masklets as components of the final output rather than as intermediate or sparse rationales. Domain-specific efforts also include EgoMask [24] and InterRVOS [19]. However, E-VQA differs in two key ways: (a) Evidence vs. Answer Grounding: while prior work primarily focuses on *answer grounding* (locating the visual entity that *is* the answer), E-VQA requires *evidence grounding*—locating the visual cues that logically lead to the answer. (b) Dense vs. Sparse Representation: most existing benchmarks rely on sparse annotations (segments or bounding boxes on keyframes). As E-VQA demonstrates, these are insufficient for complex video dynamics such as severe occlusions, continuous state changes, and non-rigid object interactions. Our ST-Evidence benchmark unifies high-level reasoning with dense evidence generation, requiring tracked masklets at 6 FPS, and thus establishes a rigorous standard that binds linguistic logic to verifiable visual facts.

3 Evidence-Backed Video QA

We introduce **Evidence-Backed Video Question Answering (E-VQA)**. Unlike standard Video Question Answering (VQA), which only predicts an answer A, E-VQA requires a model F to justify its prediction via a unified triplet (A, E_t, E_s), given a video V and a question Q. The supporting spatio-temporal evidence is defined as:
Temporal Evidence (E_t): A set of non-overlapping time segments containing the critical time spans essential to answering Q:

$$E_t = \{[start_i, end_i]\}_{i=1}^{N}.$$

Spatial Evidence (E_s): The key objects or regions relevant to the reasoning process, represented as spatio-temporal masks (i.e., "masklets").

By requiring models to ground their reasoning in specific spatio-temporal regions, E-VQA mitigates reliance on static language priors (i.e., "language bias") and significantly enhances the explainability of video understanding models.

3.1 Benchmark Design: ST-Evidence

To rigorously evaluate models on the E-VQA task, we introduce **ST-Evidence**. The benchmark is constructed via a semi-automatic three-stage pipeline—*Sample*, *Annotate*, and *Verify*—ensuring high-quality, human-verified evidence.
(1) Sample. We sample high-quality video-question pairs from the validation and test splits of existing semantic Video QA benchmarks, including NeXT-QA [50], Perception Test [34], STAR [48], CLEVRER [53], and Ego4D [14]. As not all samples are suitable, we employ a vision-language model (VLM) as an initial filter. The filtering criteria require: (i) clear video quality and an appropriate duration (5–200 s), (ii) temporal evidence that spans neither the entire video nor a single frame, and (iii) a specific, visually determinable question.
(2) Annotate. Samples passing the VLM filter are assigned to human experts. Provided with the video, question, and ground-truth answer, annotators are tasked with: (i) identifying all critical temporal evidence segments, and (ii) annotating spatio-temporal tracked segmentation masks for each evidence object. Masks are densely annotated at 6 FPS, a standard sampling rate used in prominent video segmentation datasets (e.g., DAVIS [35]) to balance fine-grained temporal resolution with annotation feasibility. Annotators also filter out any remaining ambiguous or subjective samples.
(3) Verify. Finally, all annotations are reviewed by PhD-level researchers. Any suboptimal annotations are flagged for re-annotation, ensuring the benchmark maintains a rigorous standard for evaluating precise evidence grounding.

Given that dense mask annotation is exceptionally labor-intensive, we supplement our benchmark by sourcing data from the validation split of ViCaS [3], a dataset containing human-annotated dense masks and captions. We use a VLM to generate questions, answers, and distractor options conditioned on the video and dense captions. Human annotators then filter unsuitable QA pairs, annotate the temporal evidence, and map the correct spatial evidence to the existing

human-annotated ViCaS object masks. We use Qwen3-VL-235B-A22B [37] as the VLM.

Benchmark Variants. We propose ST-Evidence-Gen for generative, pixel-level grounding and ST-Evidence-MCQ for multiple-choice discriminative grounding. These distinct formats accommodate a broad spectrum of model architectures, evaluating E-VQA through both open-ended synthesis and structured selection.

ST-Evidence-Gen (Generative) requires a model to generate the complete triplet (A, E_t, E_s). This entails selecting the correct answer A from a list of options, predicting temporal segments E_t (start/end timestamps), and producing the spatial evidence E_s (as masklets). We evaluate A via accuracy; E_t using temporal Intersection over Union (tIoU) and Intersection over Prediction (IoP); and E_s via the standard $\mathcal{J}\&\mathcal{F}$ metric to jointly consider region similarity $\mathcal{J}$ and contour accuracy $\mathcal{F}$. See the Supplementary Material for details.

ST-Evidence-MCQ (Multiple-Choice Question) is designed for general Video LLMs not explicitly trained for generative localization. In this setting, all three subtasks are formulated as multiple-choice questions, requiring the model to: (1) select the correct answer A from a list of options, (2) select the correct temporal evidence E_t from a set of candidate segments, and (3) select the correct spatial evidence E_s from a set of candidate masks.

For temporal evidence, we use Qwen3-VL-235B-A22B [37] to generate plausible yet deliberately incorrect distractors, ensuring they are semantically relevant but have minimal overlap with the ground-truth segments.

For spatial evidence, we first sample the middle frame of a ground-truth evidence segment as the reference frame; the corresponding human-annotated mask serves as the answer. To ensure high quality and keep the task challenging, we then manually annotate three plausible but incorrect object masks as distractor options to formulate four-way multiple-choice questions.

All three subtasks are evaluated using standard accuracy.

4 Instruction Tuning Data Construction

Effective spatio-temporal reasoning requires large-scale training data aligning QA with dense grounded annotations. Existing datasets generally fall into two categories: *semantic* and *grounded*. *Semantic* datasets (e.g., video QA) focus on holistic understanding but lack grounding, while *grounded* datasets (e.g., temporal or spatial localization) target fine-grained perception and localization accuracy, often at the expense of complex reasoning.

The Evidence-Backed Video Question Answering (E-VQA) task unifies these two paradigms, demanding both high-level semantic reasoning and fine-grained spatio-temporal grounding. This dual requirement makes large-scale data annotation from scratch prohibitively challenging and costly. Therefore, to support future research on E-VQA, we propose scalable data-annotation pipelines designed to construct E-VQA training datasets by leveraging both existing semantic and grounded video understanding datasets.

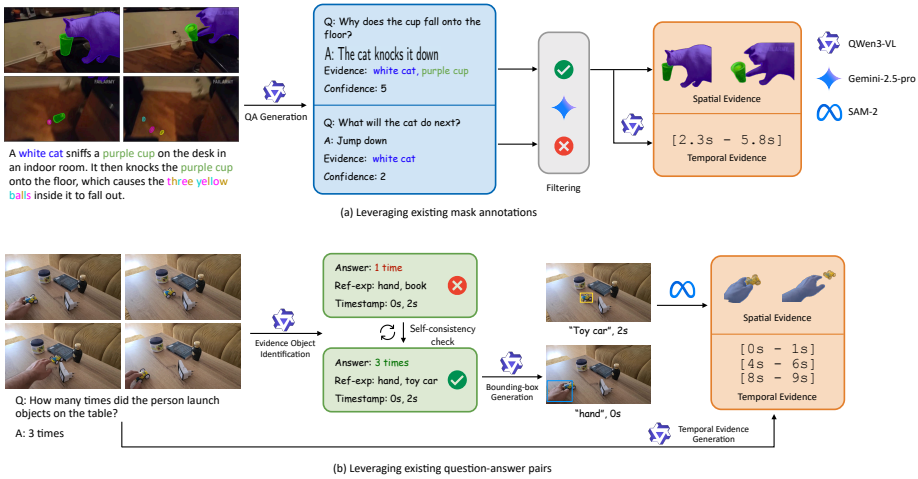

Fig. 2: Overview of our fully automated pipelines for constructing ST-Evidence-Instruct, a 160k-scale dataset for E-VQA instruction tuning.

Table 1: Statistics of our proposed datasets for the **E-VQA** task.

Name	# Questions	# Videos	Data sources	Temporal annotation	Mask annotation
ST-Evidence-MCQ	1298	348	NeXT-QA	Human	Human
ST-Evidence-Gen	2706	2548	Perception Test, STAR, CLEVRER, ViCaS, Ego4D	Human	Human
ST-Evidence-Instruct	160k	30k	Perception Test, STAR, CLEVRER, ViCaS	Model	Model + Human

Data Sources and Statistics. We construct our training dataset by aggregating samples from two benchmark categories: (1) Video Question Answering and (2) Video Segmentation. Specifically, as sources for our pipelines, we sample 20k video-question pairs from the training sets of Perception Test [34], STAR [48], and CLEVRER [53]. We also sample 20k videos with detailed captions and segmentation masks from the ViCaS [3] training set. Table 1 summarizes the data statistics.

4.1 Automatic Data Construction Pipeline

Each sample in the E-VQA dataset is formulated as a tuple (V, Q, A, C, E_t, E_s), comprising a video V, a question Q, an answer A, optional candidate choices C, temporal evidence E_t (time segments), and spatial evidence E_s (dense video masklets). To construct this at scale, we design a bidirectional, multi-step generation paradigm based on the source data type (illustrated in Figure 2).

Pathway 1: Grounding-to-Semantics (Leveraging Existing Masks). For datasets such as ViCaS [3], which already contain dense captions and pixel-level object masks paired with natural language object phrases (i.e., referring expressions), we employ a generation-verification pipeline to synthesize grounded QA pairs as shown in Figure 2(a).

(1) Generation. ViCaS provides human-written captions with phrase grounding linked to object masks. We utilize two VLMs (Qwen3-VL and Gemini) to independently generate tuples containing: a question Q, an answer A

with distractor options C, the associated evidence objects (specifically, object phrases), a confidence score, and a brief explanation. The models are conditioned on the video, the ground-truth captions, and the list of candidate objects. We employ rigorous prompting strategies (detailed in the Supplementary Material) to ensure the questions are answerable and grounded in the provided objects. For each video, each model generates three to four candidate questions, yielding typically six to eight candidates in total.

(2) Filtering. We employ a two-stage quality control process. First, we discard samples with formatting errors (e.g., missing answers, hallucinated evidence objects) or low confidence scores ($< \frac{3}{5}$). Second, we perform text-only validation using Gemini-2.5-Pro, which reviews the captions, object candidates, generated QA, evidence objects, and explanation to accept or reject each sample.

(3) Evidence Assignment. For each accepted video-question pair, we query Qwen3-VL again to predict the temporal evidence E_t. Notably, we process videos at a higher frame rate (4 FPS) during this stage than during the question-generation stage (2 FPS) to capture finer temporal dynamics. Finally, the spatial evidence E_s is directly derived from the pre-existing ViCaS mask annotations corresponding to the identified evidence objects.

Pathway 2: Semantics-to-Grounding (Leveraging Existing QA) For semantic datasets containing complex QA pairs but lacking visual grounding, the primary challenge lies in generating the supporting evidence (E_t and E_s). While temporal evidence E_t is generated using the method described in Pathway 1, generating accurate spatial masks E_s from complex questions remains non-trivial. A significant domain gap exists between *reasoning models* and *grounding models*. Reasoning models (typically trained on VQA) excel at processing complex queries but lack fine-grained spatial output capabilities (i.e., they usually generate pure text output). In contrast, grounding models (typically trained on localization tasks) produce precise pixel-level predictions but are limited to simple referring expressions or visual prompts, failing on complex reasoning tasks. To bridge this gap, we propose a three-step decomposition pipeline that leverages the strengths of both paradigms, as shown in Figure 2(b):

(1) Evidence Object Identification. We first employ Qwen3-VL-235B-A22B to identify the specific objects required to answer the question. To ensure quality, we adopt an intuitive principle: ideal evidence objects are those most likely to lead to the correct answer. We prompt the VLM to: (i) list referring expressions that uniquely describe the critical objects, (ii) provide a timestamp at which each object is clearly visible, and (iii) answer the original question based on these objects. If the generated answer matches the ground truth, the evidence is accepted. We allow up to five re-prompting attempts; if the model fails to predict the correct answer within five attempts, the sample is discarded.

(2) Bounding Box Generation. Given the verified referring expressions and their associated timestamps, we extract the corresponding frames and prompt Qwen3-VL to predict a spatial bounding box for each object. To ensure robust downstream tracking, we apply a filtering mechanism to remove degenerate predictions. Specifically, we discard bounding boxes exhibiting ex-

Table 2: Performance on ST-Evidence-Gen. † indicates closed-source proprietary models. For General Video LLMs, spatial evidence is generated via a proxy segmentation model (UniPixel-3B). Fine-tuning gains are shown relative to the size-matched UniPixel baseline (Ours-3B vs. UniPixel-3B; Ours-7B vs. UniPixel-7B).

Methods	QA	T-Evidence			S-Evidence		
	acc	mIoU	mIoP	t-mean	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
General Video LLM + Segmentation Model (UniPixel-3B)							
OpenAI-o3† [33]	82.6	30.6	41.6	36.1	39.4	44.4	41.9
Gemini-2.5-Flash† [9]	81.2	27.2	43.1	35.2	40.5	45.4	42.9
Gemini-2.5-Pro† [9]	**83.7**	**42.8**	**57.1**	**49.9**	41.8	46.3	44.0
Video-LLaMA3-7B [57]	51.0	13.7	18.9	16.3	12.3	14.2	13.2
LLaVA-OV-1.5-8B [2]	67.7	13.6	26.9	20.3	35.5	40.1	37.8
InternVL-3.5-8B [43]	72.0	17.5	27.8	22.6	33.7	39.7	36.7
Qwen2.5-VL-3B [4]	72.4	21.8	27.4	24.6	29.6	33.5	31.6
Qwen2.5-VL-7B [4]	73.4	18.1	28.5	23.3	27.7	30.9	29.3
Qwen2.5-VL-72B [4]	76.2	20.5	30.1	25.3	36.6	42.3	39.5
Qwen3-VL-4B [37]	76.9	33.0	44.3	38.6	37.6	42.3	39.9
Qwen3-VL-8B [37]	78.2	30.5	43.3	36.9	39.0	43.3	41.1
Qwen3-VL-235B-A22B [37]	83.6	38.4	50.6	44.5	39.6	44.1	41.9
Grounded Video LLM							
Sa2VA-Qwen2.5-VL-7B [55]	76.8	0.0	0.0	0.0	23.5	28.8	26.2
UniPixel-3B [27]	72.2	4.5	8.6	6.5	40.2	45.2	42.7
UniPixel-7B [27]	75.0	1.4	2.4	1.9	38.3	42.4	40.3
Ours-3B	72.4$_{(+0.2)}$	20.6$_{(+16.1)}$	33.2$_{(+24.6)}$	26.9$_{(+20.4)}$	51.2$_{(+11.0)}$	54.3$_{(+9.1)}$	52.7$_{(+10.0)}$
Ours-7B	76.0$_{(+1.0)}$	21.1$_{(+19.7)}$	37.1$_{(+34.7)}$	29.1$_{(+27.2)}$	**52.8**$_{(+14.5)}$	**55.4**$_{(+13.0)}$	**54.1**$_{(+13.8)}$

treme area ratios ($< 1\%$ or $> 90\%$), anomalous aspect ratios, or severe spatial inconsistencies compared to the object's bounding boxes in adjacent sampled frames.

(3) Dense Mask Propagation. Finally, we utilize SAM-3 [6] to propagate dense spatio-temporal segmentation masks across the entire video volume, using the verified multi-frame bounding boxes as prompt inputs. To eliminate redundancy and ensure a concise set of spatial evidence, we apply a post-processing filter that deduplicates highly overlapping masklets (IoU > 0.9).

This pipeline yields **ST-Evidence-Instruct**, a 160k-scale dataset containing aligned (Q, A, E_t, E_s) tuples, providing the necessary data foundation to train models capable of explainable, evidence-backed video reasoning.

5 Experiments

We evaluate models on the E-VQA task using the ST-Evidence benchmark, from two categories: *General Video LLMs* and *Grounded Video LLMs*.

General Video LLMs. These models are designed for broad multimodal understanding and primarily output text. For closed-source models, we select three state-of-the-art multimodal LLMs: OpenAI-o3, Gemini-2.5-Flash, and Gemini-2.5-Pro as representatives. We also evaluate multiple recent open source MLLMs [2, 4, 37, 43, 57], with parameter counts ranging from 3B to 235B. Notably, while some general models (e.g., Gemini and Qwen3-VL) support image-level grounding by predicting bounding boxes or masks as strings, they

Table 3: Performance on ST-Evidence-MCQ. † indicates closed-source models.

Methods	QA-acc	T-Evidence-acc	S-Evidence-acc
Random Guess	20.00	25.00	25.00
General Video LLM			
OpenAI-o3† [33]	81.21	50.67	<u>84.32</u>
Gemini-2.5-Flash† [9]	80.43	**78.81**	36.29
Gemini-2.5-Pro† [9]	**82.82**	70.88	81.28
Video-LLaMA3-7B [57]	67.64	48.84	27.12
LLaVA-OV-1.5-8B [2]	69.03	66.02	23.42
InternVL-3.5-8B [43]	79.04	58.17	27.43
Qwen2.5-VL-3B [4]	71.42	45.69	26.35
Qwen2.5-VL-7B [4]	74.65	27.12	46.15
Qwen2.5-VL-72B [4]	79.04	45.38	71.80
Qwen3-VL-4B [37]	76.44	63.41	77.04
Qwen3-VL-8B [37]	77.81	70.34	76.43
Qwen3-VL-30B-A3B [37]	81.14	<u>71.54</u>	67.04
Qwen3-VL-235B-A22B [37]	<u>81.59</u>	70.96	**86.90**
Grounded Video LLM			
Sa2VA-Qwen2.5-VL-7B [55]	62.40	34.21	23.04
UniPixel-3B [27]	68.57	45.45	24.88
UniPixel-7B [27]	74.04	35.29	22.03

cannot directly produce tracked video masks due to length and efficiency constraints.

Grounded Video LLMs. We evaluate two recent methods specialized for segmentation: Sa2VA [55] and UniPixel [27]. These models equip the Multimodal LLM with SAM-2.1 [38] as the segmentation head, enabling simultaneous text reasoning and pixel-level video mask prediction.

5.1 Evaluation on ST-Evidence

We conduct a comprehensive evaluation of baseline models on both variants of our benchmark. **ST-Evidence-Gen** requires models to predict temporal segments and object masks directly rather than selecting from options. Temporal segments are generated as text strings following the same format above. For spatial evidence, grounded Video LLMs can produce masks directly. However, since general Video LLMs lack native dense mask generation heads, we employ a two-step proxy evaluation for spatial evidence: the Video LLM first generates textual referring expressions for the evidence objects, which are then fed into a frozen video segmentation model (UniPixel-3B) to produce the final masks. To assess models' ability to reason jointly about answers and evidence, general Video LLMs are evaluated in a single inference pass and prompted to simultaneously output the answer, temporal segments, and spatial referring expressions. However, as current grounded LLMs often struggle with long, multi-task instructions and strict output formatting, we evaluate them using a sequential, multi-turn dialogue strategy in which the instructions and outputs from previous steps serve as context for subsequent subtasks.

Table 4: Performance on video segmentation and general understanding benchmarks. Gains or losses are shown relative to the size-matched UniPixel baseline.

Methods	Size	Segmentation ($\mathcal{J}\&\mathcal{F}$)			Understanding (acc)
		MeViS_u	Ref-DAVIS17	ReVOS	MVBench
VideoLISA	3.8B	51.7	68.8	–	–
VISA	7B	–	70.4	47.1	–
Sa2VA	4B	52.1	73.8	53.2	–
UniPixel	3B	59.7	74.2	62.1	62.5
UniPixel	7B	61.7	76.4	63.9	64.3
Ours	3B	61.8(+2.1)	74.0(-0.2)	62.7(+0.6)	62.3(-0.2)
Ours	7B	**62.5**(+0.8)	**77.8**(+1.4)	**64.2**(+0.3)	**65.6**(+1.3)

On **ST-Evidence-MCQ**, we treat the three subtasks as independent multiple-choice questions. First, for answer prediction (A), the model is presented with the video, the question, and candidate answers, and must select the correct option. Second, for temporal evidence (E_t), we provide the video, the original question, and four candidate temporal segments formatted as text strings (e.g., "A. $[[s_1, e_1], \ldots]$"), querying the model to identify the time intervals that best support the answer. Finally, for spatial evidence (E_s), we construct the options as four distinct images, each visualizing a candidate mask overlaid with a red boundary on the target frame. The model receives the video, question, and these four visual options, and must select the one highlighting the correct objects that serve as the evidence.

5.2 Experimental Results on ST-Evidence

Evaluation results on ST-Evidence-Gen and ST-Evidence-MCQ are shown in Table 2 and Table 3, respectively.

Performance on ST-Evidence-Gen. Table 2 details model performance across the three subtasks. Gemini-2.5-Pro leads in QA and Temporal Evidence (E_t), although its t-mean of 49.9 leaves substantial room for improvement. Among open-source General Video LLMs, the Qwen3-VL series exhibits substantial gains over Qwen2.5-VL. This improvement is likely attributable to its novel text-timestamp alignment mechanism, which adopts an interleaved "timestamp-frame" input format to enable fine-grained alignment between temporal information and visual content, thereby directly benefiting temporal evidence prediction. Consequently, Qwen3-VL-235B-A22B achieves the second-best performance on E_t.

Regarding Spatial Evidence (E_s), even the strongest General Video LLM obtains only 44.0 $\mathcal{J}\&\mathcal{F}$: Gemini-2.5-Pro ranks first at 44.0, followed by Gemini-2.5-Flash at 42.9, while Qwen3-VL-235B-A22B is the strongest open-source General Video LLM at 41.9. This underscores an inherent limitation in existing models: they often rely on language priors or shortcuts rather than grounding answers in specific visual evidence. Turning to Grounded Video LLMs, Sa2VA-Qwen2.5-VL-7B achieves a QA accuracy of 76.8% but fails entirely on temporal segments

(0.0 mIoU). Its low S-Evidence score further indicates a weakness in handling segmentation tasks that require complex reasoning rather than simple phrase grounding. Similarly, while UniPixel-3B produces temporal outputs, its poor performance reflects the lack of explicit temporal supervision during training. On S-Evidence, UniPixel-3B reaches 42.7 $\mathcal{J}\&\mathcal{F}$, outperforming all open-source General Video LLMs and OpenAI-o3, but remaining below Gemini-2.5-Pro and Gemini-2.5-Flash. Overall, these results highlight the significant challenges current models face in performing reliable spatio-temporal evidence-backed video question answering.

Performance on ST-Evidence-MCQ. In the multiple-choice setting (Table 3), most General Video LLMs achieve competitive QA accuracies (approximately 68%–83%). While all three closed-source models consistently perform near 80%, Gemini-2.5-Pro attains the highest QA accuracy (82.82%), while the open-source Qwen3-VL-235B-A22B is very competitive (81.59%). However, a critical observation is that comparable QA performance does not imply similar grounding capabilities. For instance, while Video-LLaMA3-7B and LLaVA-OV-1.5-8B have similar QA scores (67.64% and 69.03%), they exhibit a substantial 17.18-point gap in temporal evidence accuracy. Regarding spatial evidence, OpenAI-o3 demonstrates significantly stronger visual grounding capabilities than Gemini-2.5-Flash, while Gemini-2.5-Pro is also competitive. Among open-source baselines, the Qwen-VL family—particularly the latest Qwen3-VL models—displays consistently strong performance in identifying evidence objects. In contrast, most other models perform near the random-guess baseline ($\sim$25%), highlighting a widespread lack of fine-grained discriminative capability to distinguish between correct and incorrect spatial regions. Notably, Qwen3-VL-235B-A22B delivers the strongest overall open-source performance, achieving the best open-source QA and S-Evidence scores, while Qwen3-VL-30B-A3B attains the best open-source T-Evidence score (71.54). Gemini-2.5-Flash achieves the strongest overall T-Evidence performance (78.81).

Influence of Model Scaling. We analyze the impact of model scaling using the Qwen2.5-VL and Qwen3-VL families on **ST-Evidence-Gen**. While increasing model size within a single family consistently boosts QA accuracy, improvements in Temporal and Spatial Evidence are marginal. Conversely, comparing across generations reveals significant gains. When comparing models of similar scale (e.g., Qwen2.5-VL-3B vs. Qwen3-VL-4B), the newer generation demonstrates clear improvements across all metrics. Most notably, the compact Qwen3-VL-4B matches or outperforms the massive Qwen2.5-VL-72B on every metric, with the most dramatic improvement observed in T-Evidence. This indicates that simple parameter scaling is insufficient to solve the E-VQA task. Instead, fundamental advances in model architecture, data quality, and training objectives are the primary drivers of performance improvement.

5.3 A Baseline Model for the E-VQA Task

To validate the effectiveness of our automated annotation pipeline and the quality of the proposed ST-Evidence-Instruct dataset, we train a baseline model

building upon the UniPixel architecture. Specifically, we fine-tune the 3B and 7B variants of UniPixel, which integrate Qwen2.5-VL (3B/7B) as the base Multimodal LLM and SAM-2.1-Base+ as the mask decoder.

To prevent overfitting to the specific E-VQA task and preserve general capabilities, we incorporate existing video segmentation [3, 10, 11, 35, 39, 41, 52, 55] and general video understanding datasets [26,29] into the training mix, following the UniPixel training recipe. We employ LoRA [17] to efficiently fine-tune the visual encoder and LLM, while the mask decoder is fully trained. The model is optimized to jointly predict the answer, temporal segments, and evidence object masks by minimizing a combination of next-token prediction loss (for QA and temporal segments) and mask decoding losses [38] (for spatial evidence).

The quantitative results are presented in the final rows of Table 2. Compared to the original UniPixel baseline, our fine-tuned model yields slight gains in QA accuracy but achieves substantial improvements in both Temporal and Spatial Evidence. Notably, on T-Evidence, our 3B model achieves a competitive score of 26.9, outperforming the significantly larger QWen2.5-VL-72B. On S-Evidence, both our 3B and 7B models significantly outperform all other baselines. These results validate the effectiveness of our annotation pipeline and demonstrate the potential of our ST-Evidence-Instruct dataset to advance future research in spatio-temporal evidence grounding.

We further evaluate our model on several video segmentation and general video understanding benchmarks [11, 21, 35, 52], comparing it with other grounded Video LLMs in Table 4. Our 7B model consistently surpasses all baseline models on both segmentation and general understanding tasks. These results demonstrate that our model not only excels at the proposed E-VQA task but also retains robust pixel-level and semantic-level capabilities.

5.4 Ablations and Deeper Analyses

Is the Annotated Evidence Truly Evidential? We examine whether the annotated evidence object masks truly contain answer-supporting evidence, rather than merely salient but non-essential regions. We conduct experiments on a 200-sample subset of ST-Evidence-Gen using two representative general Video LLMs (Gemini-2.5-Flash and Qwen3-VL-8B). For each model, we evaluate three settings: (1) **Original**, with the unmodified video; (2) **Evidence Mask Corruption**, where annotated evidence regions are pixelated to make them visually unrecognizable; and (3) **Non-evidence Mask Corruption**, where randomly selected non-evidence regions are pixelated using the same procedure. If the annotations indeed capture evidential objects, QA performance should degrade substantially more under evidence corruption than under non-evidence corruption. As shown in Table 5, corrupting evidence objects causes a large QA drop, while corrupting non-evidence objects yields only a minor drop for both models. This result supports the validity of our task formulation and the reliability of our evidence annotations.

Is the Proxy Segmenter an Evaluation Bottleneck? A potential concern when evaluating General Video LLMs via a proxy segmenter is whether

Table 5: QA accuracy with corrupted regions on a subset of ST-Evidence-Gen.

Model	Original	Evidence Corr.	Non-evidence Corr.
Gemini-2.5-Flash	**81.32**	61.12(-20.20)	76.43(-4.89)
Qwen3-VL-8B	**79.31**	58.32(-20.99)	73.66(-5.65)

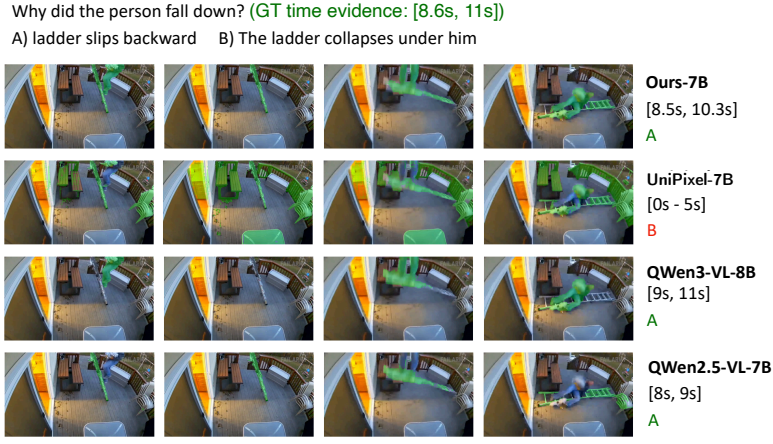

Fig. 3: Comparison of models on ST-Evidence-Gen. E-VQA requires models to jointly perform complex reasoning and provide dense spatial and temporal evidence.

low spatial scores ($\mathcal{J}\&\mathcal{F}$) reflect the LLM's poor reasoning or simply the segmenter's inability to follow text descriptions. To isolate this, we choose 100 samples with manually annotated referring expressions and feed them into our proxy segmenter (UniPixel-3B), achieving a $\mathcal{J}\&\mathcal{F}$ score of 44.7, significantly outperforming the expressions generated by top models (e.g., Gemini-2.5-Pro at 25.6). This massive performance gap confirms that the primary bottleneck in the E-VQA task mainly lies in the LLM's inability to ground to the correct evidence objects.

Validation of the Automated Annotation Pipeline. To empirically validate the quality of the evidence masks generated by our automated pipeline, we evaluated the pipeline's outputs directly against the human-verified ST-Evidence-Gen benchmark. The pipeline achieved a $\mathcal{J}\&\mathcal{F}$ score of 62.8, comfortably outperforming both our fine-tuned 7B model (54.1) and the 235B baseline (41.9). This verifies the effectiveness of our automatic annotation pipeline.

5.5 Visualization

Figure 3 compares our model against baselines on ST-Evidence-Gen. To answer "Why did the person fall down?", a model needs to causally link the fall to the ladder slipping backward and localize the temporal interval in which both events occur. Furthermore, valid spatial evidence requires simultaneous, dense tracking of both entities. As illustrated, existing baseline models struggle to satisfy these strict evidence requirements: they either fail to capture the complete interaction

between the person and the ladder (Qwen3-VL and Qwen2.5-VL) or ground unrelated background objects and output a highly inaccurate temporal window (UniPixel). In contrast, our model successfully segments both critical entities throughout the correct causal time span, demonstrating a clear synergy between high-level semantic reasoning and dense visual grounding. Additional qualitative results and failure-mode analyses are provided in the Supplementary Material.

6 Conclusion

We introduce Evidence-Backed Video Question Answering (E-VQA), a framework unifying question answering with dense spatio-temporal evidence grounding. We provide ST-Evidence, a human-verified benchmark, and ST-Evidence-Instruct, an instruction-tuning dataset that aligns answers with visual evidence. Fine-tuning on ST-Evidence-Instruct enhances both accuracy and grounding. Our work establishes a foundation for explainable and evidence-grounded Video LLMs that reason transparently over time and space. By requiring dense, video-aligned evidence alongside the answer, E-VQA bridges reasoning and perception, improves explainability and debuggability, and enables reliable use in applications that demand precision and transparency.

References

1. Ahmad, G.S., Heakl, A., Gani, H., Shaker, A., Shen, Z., Khan, F.S., Khan, S.: VideoMolmo: Spatio-temporal grounding meets pointing. arXiv preprint arXiv:2506.05336 (2025)
2. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Wu, C., et al.: LLaVA-OneVision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
3. Athar, A., Deng, X., Chen, L.C.: ViCaS: A dataset for combining holistic and pixel-level video understanding using captions with grounded segmentation. CVPR (2025)
4. Bai, S., et al.: Qwen2.5-VL technical report (2025), `https://arxiv.org/abs/2502.13923`, accessed: 2026-06-26
5. Bai, Z., He, T., Mei, H., Wang, P., Gao, Z., Chen, J., Zhang, Z., Shou, M.Z.: One token to seg them all: Language instructed reasoning segmentation in videos. Advances in Neural Information Processing Systems **37**, 6833–6859 (2024)
6. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., Lei, J., Ma, T., Guo, B., Kalla, A., Marks, M., Greer, J., Wang, M., Sun, P., Rädle, R., Afouras, T., Mavroudi, E., Xu, K., Wu, T.H., Zhou, Y., Momeni, L., Hazra, R., Ding, S., Vaze, S., Porcher, F., Li, F., Li, S., Kamath, A., Cheng, H.K., Dollár, P., Ravi, N., Saenko, K., Zhang, P., Feichtenhofer, C.: SAM 3: Segment anything with concepts (2025), `https://arxiv.org/abs/2511.16719`, accessed: 2026-06-26
7. Chen, G., Liu, Y., Huang, Y., He, Y., Pei, B., Xu, J., Wang, Y., Lu, T., Wang, L.: CG-Bench: Clue-grounded question answering benchmark for long video understanding. arXiv preprint arXiv:2412.12075 (2024)

8. Cheng, Z., Hu, J., Liu, Z., Si, C., Li, W., Gong, S.: V-STaR: Benchmarking Video-LLMs on video spatio-temporal reasoning (2025), https://arxiv.org/abs/2503.11495, accessed: 2026-06-26
9. Comanici, G., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities (2025), https://arxiv.org/abs/2507.06261, accessed: 2026-06-26
10. Deng, A., Chen, T., Yu, S., Yang, T., Spencer, L., Tian, Y., Mian, A.S., Bansal, M., Chen, C.: Motion-grounded video reasoning: Understanding and perceiving motion at pixel level. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8625–8636 (2025)
11. Ding, H., Liu, C., He, S., Jiang, X., Loy, C.C.: MeViS: A large-scale benchmark for video segmentation with motion expressions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 2694–2703 (2023)
12. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-R1: Reinforcing video reasoning in MLLMs. arXiv preprint arXiv:2503.21776 (2025)
13. Gong, S., Zhang, L., Zhuge, Y., Jia, X., Zhang, P., Lu, H.: Reinforcing video reasoning segmentation to think before it segments. arXiv preprint arXiv:2508.11538 (2025)
14. Grauman, K., et al.: Ego4D: Around the world in 3,000 hours of egocentric video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18995–19012 (2022)
15. Guo, Y., Liu, J., Li, M., Tang, X., Chen, X., Zhao, B.: VTG-LLM: Integrating timestamp knowledge into video LLMs for enhanced video temporal grounding. arXiv preprint arXiv:2405.13382 (2024)
16. Han, S., Huang, W., Shi, H., Zhuo, L., Su, X., Zhang, S., Zhou, X., Qi, X., Liao, Y., Liu, S.: 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. pp. 26181–26191 (2025)
17. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
18. Huang, B., Wang, X., Chen, H., Song, Z., Zhu, W.: VTimeLLM: Empower LLM to grasp video moments. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14271–14280 (2024)
19. Jin, W., Kim, S., Lee, J., Kim, S.: InterRVOS: Interaction-aware referring video object segmentation. arXiv preprint arXiv:2506.02356 (2025)
20. Lei, J., Yu, L., Berg, T., Bansal, M.: TVQA+: Spatio-temporal grounding for video question answering. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8211–8225 (2020). https://doi.org/10.18653/v1/2020.acl-main.730
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, X., Yan, Z., Meng, D., Dong, L., Zeng, X., He, Y., Wang, Y., Qiao, Y., Wang, Y., Wang, L.: VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. arXiv preprint arXiv:2504.06958 (2025)
23. Li, Y., Xiao, J., Feng, C., Wang, X., Chua, T.S.: Discovering spatio-temporal rationales for video question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13869–13878 (2023)

24. Liang, S., Zhong, Y., Hu, Z.Y., Tao, Y., Wang, L.: Fine-grained spatiotemporal grounding on egocentric videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9385–9395 (2025)
25. Lin, L., Yu, X., Pang, Z., Wang, Y.X.: GLUS: Global-local reasoning unified into a single large language model for video segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8658–8667 (2025)
26. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
27. Liu, Y., Ma, Z., Pu, J., Qi, Z., Wu, Y., Ying, S., Chen, C.W.: UniPixel: Unified object referring and segmentation for pixel-level visual reasoning. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
28. Lu, Y., Song, Y., Wang, W., Torresani, L., Nagarajan, T.: VITED: Video temporal evidence distillation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8501–8511 (2025)
29. Maaz, M., Rasheed, H., Khan, S., Khan, F.: VideoGPT+: Integrating image and video encoders for enhanced video understanding. arXiv preprint arXiv:2406.09418 (2024)
30. Meng, J., Li, X., Wang, H., Tan, Y., Zhang, T., Kong, L., Tong, Y., Wang, A., Teng, Z., Wang, Y., et al.: Open-o3 Video: Grounded video reasoning with explicit spatio-temporal evidence. arXiv preprint arXiv:2510.20579 (2025)
31. Munasinghe, S., Gani, H., Zhu, W., Cao, J., Xing, E., Khan, F., Khan, S.: VideoGLaMM: A large multimodal model for pixel-level visual grounding in videos. ArXiv (2024), https://arxiv.org/abs/2411.04923, accessed: 2026-06-26
32. OpenAI: GPT-4o system card (2024), https://arxiv.org/abs/2410.21276, accessed: 2026-06-26
33. OpenAI: OpenAI o3 model. https://platform.openai.com/docs/models/o3 (2024), accessed: 2025-11-14
34. 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)
35. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 DAVIS challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
36. Qian, L., Li, J., Wu, Y., Ye, Y., Fei, H., Chua, T.S., Zhuang, Y., Tang, S.: Momentor: Advancing video large language model with fine-grained temporal reasoning. arXiv preprint arXiv:2402.11435 (2024)
37. Qwen Team: Qwen3-VL: Sharper vision, deeper thought, broader action (sep 2025), https://qwen.ai/blog?id=99f0335c4ad9ff6153e517418d48535ab6d8afef, accessed: 2025-11-13
38. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
39. Seo, S., Lee, J.Y., Han, B.: URVOS: Unified referring video object segmentation network with a large-scale benchmark. In: European conference on computer vision. pp. 208–223. Springer (2020)
40. Sun, Y., Zhang, H., Ding, H., Zhang, T., Ma, X., Jiang, Y.G.: SAMA: Towards multi-turn referential grounded video chat with large language models. arXiv preprint arXiv:2505.18812 (2025)

41. Wang, H., Yan, C., Wang, S., Jiang, X., Tang, X., Hu, Y., Xie, W., Gavves, E.: Towards open-vocabulary video instance segmentation. In: proceedings of the IEEE/CVF international conference on computer vision. pp. 4057–4066 (2023)
42. Wang, J., Kang, Z., Wang, H., Jiang, H., Li, J., Wu, B., Wang, Y., Ran, J., Liang, X., Feng, C., et al.: VGR: Visual grounded reasoning. arXiv preprint arXiv:2506.11991 (2025)
43. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
44. Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., et al.: Time-R1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377 (2025)
45. Wang, Y., Wang, Y., Chen, K., Zhao, D.: STAIR: Spatial-temporal reasoning with auditable intermediate results for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38(17), pp. 19215–19223 (2024). https://doi.org/10.1609/aaai.v38i17.29890
46. Wang, Z., Yoon, J., Yu, S., Islam, M.M., Bertasius, G., Bansal, M.: Video-RTS: Rethinking reinforcement learning and test-time scaling for efficient and enhanced video reasoning. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 28114–28128 (2025)
47. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
48. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: Star: A benchmark for situated reasoning in real-world videos. In: NeurIPS (2021)
49. Wu, Y., Hu, X., Sun, Y., Zhou, Y., Zhu, W., Rao, F., Schiele, B., Yang, X.: Number it: Temporal grounding videos like flipping manga. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 13754–13765 (2025)
50. 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)
51. Xiao, J., Yao, A., Li, Y., Chua, T.S.: Can I trust your answer? visually grounded video question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13204–13214 (2024)
52. Yan, C., Wang, H., Yan, S., Jiang, X., Hu, Y., Kang, G., Xie, W., Gavves, E.: VISA: Reasoning video object segmentation via large language models. In: European Conference on Computer Vision. pp. 98–115. Springer (2024)
53. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: CLEVRER: collision events for video representation and reasoning. In: ICLR (2020)
54. You, Z., Wu, Z.: Seg-R1: Segmentation can be surprisingly simple with reinforcement learning. arXiv preprint arXiv:2506.22624 (2025)
55. Yuan, H., Li, X., Zhang, T., Huang, Z., Xu, S., Ji, S., Tong, Y., Qi, L., Feng, J., Yang, M.H.: Sa2VA: Marrying SAM 2 with LLaVA for dense grounded understanding of images and videos. arXiv preprint arXiv:2501.04001 (2025)
56. Yuan, Y., Zhang, H., Li, W., Cheng, Z., Zhang, B., Li, L., Li, X., Zhao, D., Zhang, W., Zhuang, Y., et al.: VideoRefer suite: Advancing spatial-temporal object understanding with video LLM. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18970–18980 (2025)
57. Zhang, B., et al.: VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)

58. Zhou, H., Peng, X., Kendre, S., Ryoo, M.S., Savarese, S., Xiong, C., Niebles, J.C.: STRefer: Empowering video LLMs with space-time referring and reasoning via synthetic instruction data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4289–4300 (2025)