Clue Matters: Empower Video Reasoning with Brain-Inspired Latent Clue Learning

Kaixin Zhang[1,2,3,4*], Xiaohe Li[1,2*, †], Jiahao Li[1,2,3,4], Haohua Wu[1,2,3,4], Xinyu Zhao[1,2,3,4†], Zide Fan[1,2,3,4], and Lei Wang[1,2,3,4]

[1] Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] Key Laboratory of Target Cognition and Application Technology (TCAT), Beijing 100094, China
[3] University of Chinese Academy of Sciences, Beijing 101408, China
[4] School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China

Abstract. Multi-modal Large Language Models (MLLMs) have significantly advanced video reasoning, yet Video Question Answering (VideoQA) remains challenging due to its demand for temporal causal reasoning and evidence-grounded answer generation. Prevailing end-to-end MLLM frameworks lack explicit structured reasoning between visual perception and answer derivation, causing severe hallucinations and poor interpretability. Existing methods also fail to address three core gaps: faithful visual clue extraction, utility-aware clue filtering, and end-to-end clue-answer alignment. Inspired by hierarchical human visual cognition, we propose ClueNet, a clue-aware video reasoning framework with a two-stage supervised fine-tuning paradigm without extensive base model modifications. Decoupled supervision aligns clue extraction and chain-based reasoning, while inference supervision with an adaptive clue filter refines high-order reasoning, alongside lightweight modules for efficient inference. Experiments on NExT-QA, STAR, and MVBench show that ClueNet outperforms state-of-the-art methods by $\geq 1.1\%$, with strong benchmark generalization, hallucination mitigation, and improved inference efficiency. This work bridges the perception-to-generation gap in MLLM video understanding, providing an interpretable, faithful reasoning paradigm for real-world VideoQA applications.

Keywords: Video Question Answering · Multi-modal Large Language Models · Visual Reasoning · Video Understanding

1 Introduction

Recent advances in Multi-modal Large Language Models (MLLMs) have driven remarkable breakthroughs in video reasoning for a wide range of real-world applications, including embodied robotics, intelligent surveillance, and human-computer interaction [3, 46, 54]. A core and uniquely challenging task is Video

* Equal contribution. †Corresponding author

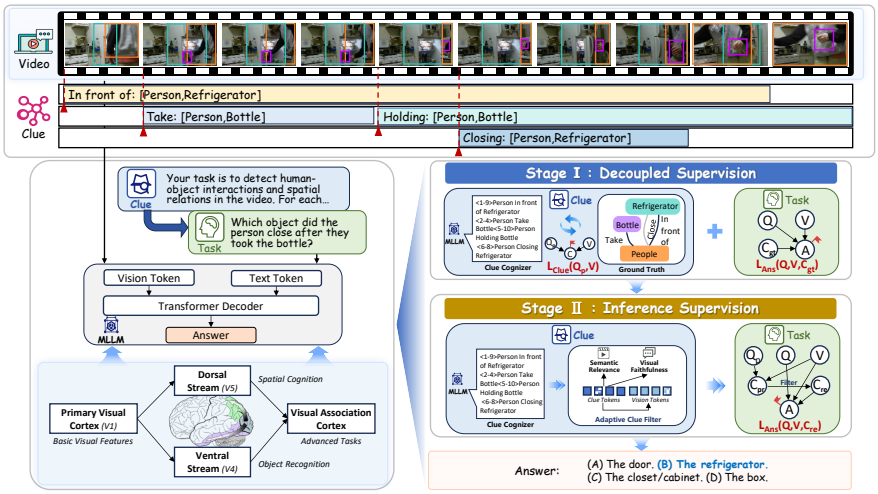

Fig. 1: Overall framework of the proposed ClueNet, which mimics the hierarchical human visual cognitive process to enable explicit clue-aware video reasoning.

Question Answering (VideoQA), which requires processing temporally continuous content, comprehending evolving entities and causal interrelations across frames, and generating accurate, evidence-grounded answers [26, 28, 49, 50].

Despite impressive performance on short-form benchmarks, prevailing MLLM-based frameworks follow a monolithic end-to-end paradigm: aligning visual tokens with textual representations via feature projectors [6, 30, 55] or cross-modal interaction modules [30, 45], and generating answers in a one-shot "intuitive" manner. This omits an explicit, structured reasoning step bridging low-level perception and high-level answer derivation, leading to two fundamental limitations: (1) dense temporal visual content exacerbates catastrophic hallucinations [12, 29, 47]; and (2) the black-box pipeline suffers from severely limited interpretability. For example, when answering "Which object did the person pick up after setting down the mug?", state-of-the-art (SOTA) models frequently hallucinate by ignoring the temporal order clue while exposing no reasoning process for validation. Recent works attempt to enhance VideoQA reasoning: CCoT [38] constructs frame-level scene graphs via CoT [48] prompting; VoT [9] introduces a scene graph module for step-by-step reasoning; VideoEspresso [13] provides a 200K-sample CoT fine-tuning dataset. Yet none fundamentally addresses faithful visual clue extraction and utility-aware clue judgment. Existing approaches overlook three critical gaps: (1) lack of faithful, accurate temporal visual clue mining causes missed critical event dynamics and spurious temporal reasoning; (2) missing explicit supervision for clue relevance and utility assessment leads to reasoning path error propagation; and (3) disjoint clue extraction and answer generation lack end-to-end alignment between intermediate clues and final outputs. We conduct a detailed analysis of the causes of reasoning failures in existing MLLMs in Section 3.

Inspired by the hierarchical mechanism of human visual cognition [8, 10], where V1 handles low-level parsing, V2–V5 extract perceptual clues [37], and visual association areas coordinate with the prefrontal cortex to filter and integrate clues for high-order reasoning [4,35], we propose **ClueNet**: a reliable video reasoning framework that discovers, filters, and leverages latent visual clues via a two-stage supervised fine-tuning paradigm, with no extensive base model modifications.

In the first stage, *Decoupled Supervision* jointly optimizes semantic clue extraction and clue chain-based reasoning. Via instruction tuning [31], ClueNet builds a full-video dynamic clue cognizer to track cross-temporal entity states, ensuring end-to-end alignment between clue extraction and downstream reasoning. In the second stage, *Inference Supervision* fine-tunes high-order reasoning to ground answers strictly in structured clue chains. We iteratively refine reasoning with model-mined raw video clues, and propose an Adaptive Clue Filter to enhance critical clues via a differentiable gating mechanism guided by Semantic Relevance and Visual Faithfulness. The refined clue chain is iteratively input to the model, enabling explicit step-by-step reasoning without the computational overhead of RL-based selection.

At inference time, Keyframe Selection and Visual Compression modules implement preliminary keyframe extraction and clue-aligned token selection respectively, ensuring inference efficiency and strengthened clue-visual consistency.

Extensive experiments on NExT-QA [50], STAR [49], and MVBench [28] demonstrate that ClueNet outperforms SOTA video reasoning methods by at least 1.1% across all benchmarks, with strong test-time generalization on MVBench, robust hallucination mitigation, and improved inference efficiency. These results indicate that explicit clue mining is a general reasoning principle for improving evidence-grounded VideoQA, rather than a benchmark-specific heuristic. Our core contributions are summarized as:

- We propose ClueNet, an end-to-end framework that jointly models hierarchical clue extraction and utility-aware clue filtering for reliable MLLM-based video reasoning.
- We design a progressive two-stage fine-tuning paradigm with Decoupled Supervision for clue-reasoning alignment and Inference Supervision for high-order reasoning enhancement, at minimal training cost.
- We develop an Adaptive Clue Filter (ACF) to suppress spurious clues, alongside training-free Keyframe Selection and clue-aligned Visual Compression for efficient inference.
- Extensive experiments on NExT-QA, STAR, and MVBench show ClueNet outperforms SOTA by $\geq 1.1\%$, with strong benchmark generalization, hallucination mitigation, and accelerated inference.

2 Related Work

MLLMs for Video Reasoning. VideoQA demands spatio-temporal reasoning to align dynamic visual sequences with linguistic queries. Early works use

recurrent neural networks or 3D CNNs [20, 26], advancing to spatio-temporal attention and graph-based memory networks [25, 50]. The field has since shifted to MLLMs [31, 60], with video-specific models (Video-ChatGPT [32], Video-LLaVA [30], VideoLLaMA series [6, 54, 55]) mapping sampled frames to LLM embedding space via visual encoders and alignment modules. Despite strong generalization, these monolithic architectures treat spatio-temporal reasoning as a black box, causing severe temporal hallucinations [12, 27, 29, 47] driven by language priors rather than grounded visual evidence.

Efficient Video Context Reduction. Unconstrained video processing incurs prohibitive computational cost due to extreme spatio-temporal redundancy, addressed mainly via frame sampling or token compression. Frame-level methods [42,51,59] leverage motion priors, text-conditioned querying or temporal grounding to outperform naive uniform sampling [32], but rely on disjointed multi-stage pipelines or rigid external components, adding overhead and breaking end-to-end differentiability. Token-level pruning strategies [5, 18, 52, 57] dynamically remove redundant visual features during encoding, but their visual-centric design lacks structured scene abstraction, risking irreversible loss of prompt-critical fine-grained details. Existing paradigms lack a unified end-to-end mechanism to preserve high-value visual evidence, leaving critical semantics vulnerable to premature, unrecoverable elimination.

Intermediate Reasoning and Symbolic Representations. Structured intermediate representations bridge low-level perception and cognitive reasoning. Entity relationship or scene graph, a formalism for encoding objects and their relations, is established as a core video understanding task via Action Genome [21]. Subsequent works advance relational modeling with hierarchical graph architectures [39], object-centric relational encoders [15], and multi-step graph reasoning for VideoQA [22,34], but these methods decouple graph construction from downstream reasoning, limiting joint optimization. For structured reasoning without explicit graph supervision, Multimodal Chain-of-Thought approaches use scene-graph-guided prompting [38], sequential reasoning decomposition [9], and multi-step visual rationale generation [58], yet their reasoning chains lack robust anchoring to verifiable spatio-temporal locations, making them error-prone on spatially demanding tasks.

3 Question Analysis

This section systematically dissects the failure causes of canonical VideoQA models and the core limitations of existing MLLMs. All diagnostic visualizations and case studies are validated on mainstream open-source VideoLLaMA3 [54] and Qwen2.5-VL [3] for generality. We categorize failures into three types of cognitive biases: visual perception bias, clue cognition bias, and inductive reasoning bias.

Visual Perception Bias. Human primary visual cortex uses task-driven selective attention to rapidly identify query-relevant high-salience visual clues and adaptively allocate attention to critical regions, even for sparse, transient cross-

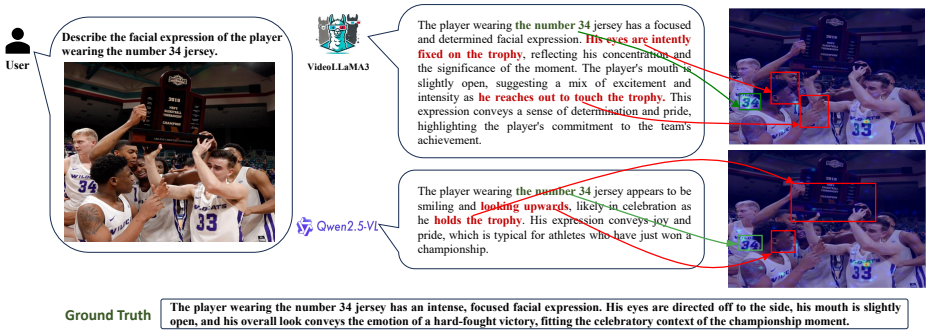

Fig. 2: MLLM attention heatmaps and paired VideoQA results: misaligned attention strongly correlates with incorrect answers, confirming visual perception bias.

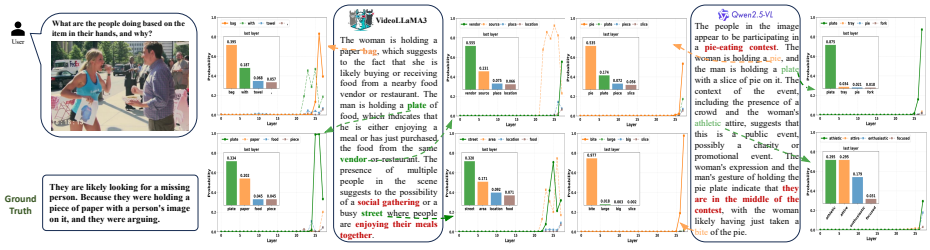

Fig. 3: Intermediate layers top-probability output token visualization: erroneous visual clue semantic interpretation in intermediate layers propagates to hallucinated reasoning and incorrect final answers, confirming clue cognition bias.

frame clues [7]. Existing video understanding methods fail to systematically emulate this query-guided, temporally-aware attention paradigm. We validate this via attention heatmap visualization (Fig. 2), plotting layer-averaged cross-attention weights of VideoLLaMA3 and Qwen2.5-VL. In the visualized case, the model misassociates a critical query clue (jersey number) with the wrong subject, producing erroneous evidence that triggers cascaded reasoning errors. This limitation motivates us to reorient the model's input visual tokens toward valid, query-critical regions and strengthen coupling between fine-grained visual features and high-level logical reasoning clues.

Clue Cognition Bias. Human higher-order visual cortex transforms raw perceptual inputs into linguistically structured, semantically coherent interpretations, and uses working memory to maintain cross-frame temporal consistency of visual episodes for evidence-aligned descriptions [36]. Existing MLLMs lack systematic modeling of this clue-to-semantic mapping, resulting in a "knowing *what* but not *why*" phenomenon: superficially plausible answers without grounded, interpretable logical support. As shown in Fig. 3, VideoLLaMA3 and Qwen2.5-VL misclassify a woman's held portrait poster as a bag or pie, triggering fully erroneous reasoning. Without explicit clue semantic grounding, these misinter-

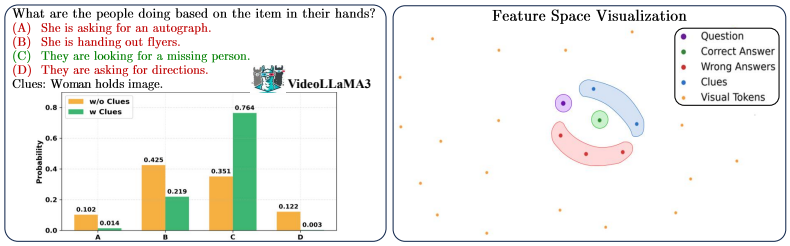

Fig. 4: t-SNE visualization of question, correct/incorrect candidate answer, reasoning clues, and sampled visual token features: clue-answer feature proximity, causal reasoning utility, and ground-truth clue performance boost confirm inductive reasoning bias.

pretations propagate through the generation pipeline to cause final VideoQA errors, motivating our proposed clue cognizer module that maps accurately perceived visual clues to verifiable semantic interpretations.

Inductive Reasoning Bias. The human prefrontal cortex and visual association regions integrate cross-frame visual evidence with reasoning chains, generate evidence-aligned answers, and self-verify consistency across perception, reasoning, and conclusions [35]. Existing MLLMs rarely emulate this hierarchical cross-modal fusion paradigm, leading to weak consistency constraints between clue representations, reasoning chains, and predictions. This decouples perception, logic, and final outputs, drastically elevating hallucination risk. We validate this via feature space analysis (Fig. 4) and comparative experiments: query-relevant valid clues have significantly higher similarity to ground-truth answers than spurious clues, and guiding models with these valid clues consistently improves reasoning faithfulness and accuracy. These findings confirm that reliable video reasoning demands faithful evidence grounding, motivating our approach to leverage query-aligned clues.

4 Method

4.1 Framework Overview

The ClueNet architecture overview is shown in Fig. 1, with two core components: (1) a Clue Cognizer, which uses instruction prompting to guide an MLLM to generate structured, temporally grounded latent visual clue descriptions; (2) a principled two-stage training scheme that maximizes extracted clues' utility for video understanding, while enforcing cognitive consistency between visual perception and high-level reasoning.

4.2 Clue Cognizer

The Clue Cognizer transforms raw visual inputs into structured logical evidence via a dynamic $\mathcal{G}$ for the full video, where each visual clue $C_i \in \mathcal{G}$ is a structured

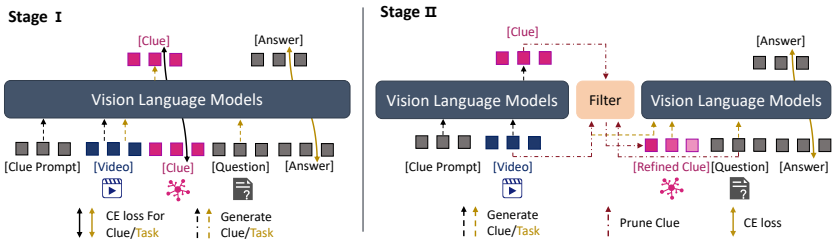

Fig. 5: Overview of our two-stage training scheme.

quintuplet:

$$C_i = \langle t_{\text{start}},\, t_{\text{end}},\, s_i,\, r_i,\, o_i \rangle, \tag{1}$$

with s_i, o_i as subject and object entities, r_i the interaction predicate, and $[t_{\text{start}}, t_{\text{end}}]$ the temporal frame lifespan. Unlike static image scene graphs [15,24], this representation explicitly tracks the temporal evolution of entity interactions, optimized for temporal VideoQA reasoning [21,39]. For instance, a generated clue is instantiated as: `"<2-4> person take bottle"`. A complete clue example is provided in Fig. 1. To ensure temporal consistency, the collection is updated incrementally across video segments to maintain alignment with the evolving scene state.

During training, we supervise clue generation with a standard autoregressive language modeling objective using timestamp-aligned scene annotation tuples. Let $c_i = f_\theta(Q_P, V)$ denote the i-th token of the generated clue sequence, where V is the input video and Q_P the full-video clue extraction instruction prompt. The visual clue generation loss for an L-frame video is:

$$\mathcal{L}_{\text{clue}} = -\frac{1}{L}\sum_{i=1}^{L} \ell_{CE}\left(C_{gt,i}, f_\theta(Q_P, V)\right), \tag{2}$$

where ℓ_{CE} is cross-entropy loss. We minimize $\mathcal{L}_{\text{clue}}$ to produce video- and question-grounded intermediate reasoning descriptions, with the resulting multi-frame clue sequence fed into the downstream answer generation module.

4.3 Stage 1: Decoupled Supervision of Clue Discovery and Reasoning

To enable reliable endogenous clue discovery and utilization, we decouple video reasoning into two complementary sub-tasks: (1) video-based clue discovery, and (2) clue-grounded answer generation, optimized with separate dedicated supervision signals.

To enforce exclusive clue-grounded reasoning, the model is trained to generate answers conditioned on the question, ground-truth clues, and video content. The Stage 1 training objective is:

$$\mathcal{L}_{\text{Stage1}} = \mathcal{L}_{\text{clue}} + \ell_{CE}\left(A, f_\theta(C_{gt}^{full}, Q, V)\right), \tag{3}$$

where C_{gt}^{full} is the full ground-truth clue token sequence. The two loss terms correspond to the decoupled sub-tasks: the first supervises accurate video-based clue generation, the second supervises ground-truth clue-guided answer generation. To stabilize training and avoid conflicting gradients, we optimize the two terms alternately per batch.

4.4 Stage 2: Adaptive Clue Filter and Inference Supervision

Stage 1 decoupled supervision trains effective ground-truth clue generation and utilization, but does not resolve the critical training-inference gap: during end-to-end inference, the model relies on self-generated clues, which may contain noise, irrelevant content, or hallucinations from imperfect visual grounding and MLLM backbone knowledge gaps. To close this gap and enforce end-to-end perception-reasoning consistency, we introduce the Adaptive Clue Filter (ACF), which retains only high-fidelity, question-relevant clues.

After obtaining candidate scene clues C_{pr} via the Clue Cognizer, ACF evaluates each candidate C_i along two dimensions to ensure question relevance and visual grounding. Since each clue C_i is grounded to a specific temporal segment, we use V_i to denote the sequence of frames encompassed by C_i.

- **Semantic Relevance**: Question informativeness, measured by alignment between the MLLM final-layer average token representations of the clue $f(C_i)$ and question $f(Q)$.
- **Visual Faithfulness**: Visual evidence support, measured by alignment between $f(C_i)$ and $f(V_i)$, where $f(V_i)$ is the average visual token representation of frames in V_i from the visual encoder.

The final clue gating weight g_i is computed via a two-branch network (Fig. 6):

$$g_i = \sigma\left(\mathcal{F}_{\text{sem}}(f(Q), f(C_i)) + \mathcal{F}_{\text{vis}}(f(V_i), f(C_i))\right), \tag{4}$$

where $\mathcal{F}_{\text{sem}}$ and $\mathcal{F}_{\text{vis}}$ each consist of linear layer, LayerNorm, and ReLU activation, and $\sigma(\cdot)$ is the sigmoid function. Learned gating weights scale the input embeddings of all tokens within each variable-length clue, yielding refined representations $C_{\text{re}} = \{g_i \cdot C_i\}$. We add ℓ_1 sparsity regularization on gating weights to encourage focus on critical, non-redundant evidence. The Stage 2 training objective is:

$$\mathcal{L}_{\text{Stage2}} = \ell_{CE}\left(A, f_\theta(C_{re}^{full}, Q, V)\right) + \frac{\lambda}{N}\sum_{j=1}^{N}|g_j|, \tag{5}$$

where λ is the loss balance hyperparameter, and C_{re}^{full} is the full refined clue token sequence. This end-to-end supervision makes g_i a differentiable filter enforcing logical consistency from clue generation to answer induction, closing the training-inference gap and maintaining clue-driven reasoning with self-generated inference-time clues.

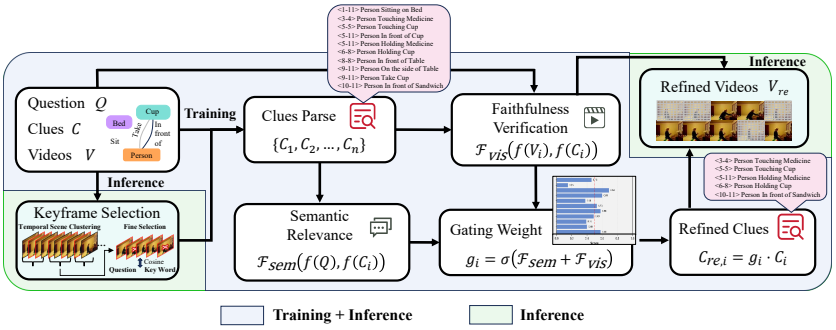

Fig. 6: Refining clues and visual information during the training and inference stages.

We note that although $\mathcal{F}_{\text{sem}}$ and $\mathcal{F}_{\text{vis}}$ both read from the model's own representations, ACF does not suffer from circular self-reinforcement: the gating weights g_i are supervised end-to-end by the answer cross-entropy loss, an external signal independent of ACF's own scores that anchors learning against degenerate solutions; the ℓ_1 sparsity term further suppresses non-discriminative gating.

4.5 Inference Process

The complete inference pipeline sequentially performs Keyframe Selection, Clue Generation, ACF, and Visual Compression before final answer generation, as illustrated in Fig. 6.

Keyframe Selection (KS). To reduce spatio-temporal redundancy in long videos and sample critical keyframes, we propose a two-step Keyframe Selection module:

- **Redundancy elimination**: We adopt LVNet's Temporal Scene Clustering (TSC) algorithm [40], extracting frame-level features via pre-trained ResNet-18 [14], then grouping visually similar frames via TSC's dynamic thresholding to produce a condensed candidate set V_c.
- **Semantic-driven Fine Selection**: For each frame in V_c, we extract patch-level visual embeddings $E_v \in \mathbb{R}^{N\times P\times d_v}$ (N = candidate frames, P = patches per frame, d_v = embedding dimension). We use spaCy [16] to select nouns/verbs as semantic keywords from the question and obtain their initial text token set as E_k (assuming $|E_k| = J$), plus the global question embedding E_q from the MLLM's textual encoder. The relevance score S_i for frame $i \in [1, N]$ is computed via patch-level max-pooling:

$$S_i = \alpha \max_{1\le p\le P} \text{sim}(E_q, E_v^{(i,p)}) + \beta \max_{1\le j\le J, 1\le p\le P} \text{sim}(E_k^{(j)}, E_v^{(i,p)}). \quad (6)$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, and α, β balance global question and local keyword alignment. We select the top 16 frames by S_i to form the Keyframe V'.

Visual Compression (VC). Given a cropped video V' and a question Q, the MLLM generates initial clues C_{pr}, which are refined by ACF to get refined clues C_{re}. Formally, we reuse the Visual Faithfulness score $s_v^{(i,t)}$ from $\mathcal{F}_{\text{vis}}$, which measures the alignment between clue C_i and frame $V_t \in V'$. Since multiple clues may overlap temporally, the retention score for V_t is defined by averaging the products of gating weights and Visual Faithfulness scores across all clues encompassing it:

$$S_{\text{ret}}(V_t) = \frac{1}{|\mathcal{M}_t|} \sum_{i \in \mathcal{M}_t} g_i \cdot s_v^{(i,t)}, \tag{7}$$

where $\mathcal{M}_t = \{i \mid C_i \in C_{pr},\ t \in [t_{i,\text{start}}, t_{i,\text{end}}]\}$.

Based on S_{ret}, each frame in V' is processed according to a retention threshold τ_{vc}: frames with $S_{\text{ret}}(V_t) \geq \tau_{vc}$ are retained in full; frames below τ_{vc} undergo adaptive average pooling to reduce redundancy while preserving background context; if all frames score below τ_{vc}, the frame with the highest S_{ret} is retained without compression. The refined visual input V_{re} and filtered clues C_{re} are interleaved as input to the LLM for final answer generation. Full details are provided in Appendix A.

4.6 Theoretical Analysis

We provide an information-theoretic justification for ClueNet via the Information Bottleneck (IB) principle [43]. Let V, Q, C, and A denote the input video, question, extracted clues, and target answer, respectively. We formulate C as a task-relevant bottleneck variable that mediates the reasoning process. The IB principle seeks a minimal sufficient representation by maximizing predictive power $I(C; A)$ while minimizing input redundancy $I(C; V, Q)$, yielding the objective:

$$\mathcal{L}_{\text{IB}}(C) = I(C; V, Q) - \gamma I(C; A). \tag{8}$$

where $\gamma > 0$ balances input compression and predictive power.

Fundamentally, all core components of ClueNet are designed to optimize this trade-off: our decoupled Stage 1 training maximizes the sufficiency term $I(C; A)$, while the ACF and evidence-driven visual compression aggressively minimize the redundancy term $I(C; V, Q)$. A comprehensive analysis of this theoretical alignment, along with further proofs, is provided in Appendix B.

5 Experiments

5.1 Experimental Setup

Datasets and Benchmarks We evaluate ClueNet on three challenging video reasoning benchmarks:

- **NExT-QA** [50]: VideoQA benchmark with Multi-Choice (MC) and Open-Ended (OE) tasks across Causal, Temporal, and Descriptive question types.

- **STAR** [49]: Real-world situated reasoning benchmark with four splits (*Interaction*, *Sequence*, *Prediction*, *Feasibility*), requiring evidence-grounded logical reasoning.
- **MVBench** [28]: Multi-modal evaluation suite with 4,000 QA pairs across 20 spatio-temporal tasks, spanning low-level visual perception to high-level cognitive reasoning.

For both NExT-QA and STAR, scene graph clue annotations are provided and converted into our structured quintuplet format (Eq. 1).

Metrics. For multiple-choice tasks (NExT-QA MC, STAR, MVBench), we report standard **Accuracy (%)**. For the open-ended NExT-OE task, we adopt Wu-Palmer Similarity (WUPS) [33], which measures semantic similarity between predictions and ground truths based on taxonomic depth. We report both the lenient WUPS@0 (retaining raw scores for any semantic overlap) and the strict WUPS@0.9 (0.9 similarity threshold, penalizing imprecise predictions with $0.1\times$ scaling).

Baselines. We compare against state-of-the-art models, spanning both closed-source models (GPT-4o [19], Qwen-VL-Max [1]) and open-source video-language models, Results for Flipped-VQA [23], ViLA [44], and VideoChat2 [28] are from their original papers. All other open-source baselines, Video-LLaVA [30], VideoLLaMA3 [54], LongVA [56], Qwen2.5-VL [3], Qwen3-VL [2], Cos [17], InternVideo2.5 [46] are re-implemented and evaluated with a maximum of 16 input frames to ensure fair comparison.

Implementation Details. We adopt VideoLLaMA3 [54] as our base model for fine-tuning. The visual encoder is a frozen pre-trained SigLIP-SO400M [53]; the LLM is initialized with Qwen2.5-7B-Instruct [41]. For fair comparison, VideoLLaMA3 is also fine-tuned for one round of standard supervised fine-tuning (SFT) (input: V, Q; output: A) under identical training settings. Full hyperparameter details are provided in Appendix C.

5.2 Main Results

Quantitative Evaluation on STAR and NExT-QA. As shown in Table 1, ClueNet achieves state-of-the-art (SOTA) performance on both benchmarks: 77.6% on STAR and 84.3% on NExT-QA, outperforming all open-source and closed-source competitors. It gains +5.5% over VideoLLaMA3 (STAR) and +2.8% over Qwen2.5-VL (NExT-QA), validating the superiority of explicit clue-based reasoning over pure model scaling.

The gains are most pronounced where temporal and causal understanding is required. On *Sequence* and *Temporal* splits, ClueNet outperforms VideoLLaMA3 by +5.1% and Qwen2.5-VL by +4.7%, confirming that two-stage training directs the model to reason over temporal clue trajectories rather than static visual shortcuts. On *Interaction* and *Causal* splits, the dynamic clue collection in the Clue Cognizer disambiguates complex entity-action relationships that confound end-to-end baselines (+5.8% over ViLA, +1.8% over VideoLLaMA3). The sole exception is the *Feasibility* split: ClueNet trails InternVideo2.5 by 1.3%, as its

Table 1: SOTA method comparison on STAR and NExT-QA: multiple-choice task accuracy (%). Best and second-best results are **bolded** and underlined. Total (Tot) denotes the overall accuracy over all test samples.

Model	STAR					NExT-QA			
	Int	Seq	Pred	Feas	Tot	Cau	Tem	Des	Tot
Closed-source MLLMs									
GPT-4o [19]	66.5	75.3	70.2	67.1	71.1	80.1	73.8	84.7	78.9
Qwen-VL-Max [1]	67.6	74.5	67.3	65.8	70.7	79.7	74.1	84.3	78.7
Open-source MLLMs									
LongVA [56]	50.2	54.5	57.2	54.3	53.3	56.9	62.8	65.8	60.2
Video-LLaVA [30]	40.0	41.1	45.7	42.9	41.3	55.2	53.6	69.4	57.0
Flipped-VQA [23]	66.2	67.9	57.2	52.7	65.4	72.7	69.2	75.8	72.0
ViLA [44]	70.0	70.4	65.9	62.2	69.2	75.3	71.8	82.1	75.3
VideoChat2 [28]	58.4	60.9	55.3	53.1	59.0	57.4	61.9	69.9	60.8
Cos [17]	50.7	57.3	62.0	54.6	55.2	67.3	66.1	69.0	67.2
InternVideo2.5 [46]	63.8	73.6	67.2	**68.0**	69.1	81.0	76.5	87.0	80.6
Qwen2.5-VL [3]	63.5	73.3	65.9	64.2	68.4	81.5	79.0	86.5	81.5
Qwen3-VL [2]	54.4	59.1	60.3	58.2	57.5	77.1	71.5	83.1	76.3
VideoLLaMA3 [54]	68.6	75.6	71.0	67.7	72.1	81.7	77.2	87.8	81.3
ClueNet (Ours)	**75.8**	**80.7**	**77.6**	66.7	**77.6**	**83.5**	**83.7**	**88.1**	**84.3**

Table 2: Evaluation on the NExT-OE generative task using WUPS. Best and second-best results are **bolded** and underlined.

Model	NExT-OE							
	All		Causal		Temporal		Descriptive	
	WUPS@0	WUPS@0.9	WUPS@0	WUPS@0.9	WUPS@0	WUPS@0.9	WUPS@0	WUPS@0.9
LongVA	27.7	21.3	20.8	14.2	22.3	16.1	50.3	44.1
Cos	27.8	21.4	21.0	14.3	22.3	16.3	50.4	44.3
Video-LLaVA	25.6	19.5	17.9	12.0	21.7	15.5	47.9	41.5
InternVideo2.5	28.4	22.0	25.1	18.1	27.3	20.1	37.4	33.1
Qwen2.5-VL	30.6	24.6	23.2	17.1	25.0	18.6	54.6	49.6
Qwen3-VL	29.1	23.1	22.0	15.8	23.7	17.3	52.2	46.7
VideoLLaMA3	32.3	25.7	25.6	18.4	27.3	20.2	53.9	48.7
ClueNet(Ours)	**35.5**	**29.2**	**26.2**	**19.3**	**28.9**	**22.2**	**64.9**	**60.5**

questions demand unobservable commonsense knowledge, creating a minor, acceptable trade-off from the ACF's evidence-first filtering given gains elsewhere.

Open-Ended Generation Performance. Table 2 reports results on the NExT-OE generative task. Unlike multi-choice tasks, open-ended generation demands precise free-form visual grounding. To comprehensively evaluate semantic accuracy, we report both WUPS@0 and WUPS@0.9. ClueNet consistently achieves state-of-the-art performance across all categories under both metrics. The most striking gain is in *Descriptive*: ClueNet reaches WUPS@0/0.9 of 64.9/60.5, outperforming Qwen2.5-VL by a significant margin (+10.3/+10.9). This improvement stems from ClueNet's ability to guide the LLM to focus on fine-grained object attributes and state transitions. ClueNet also demonstrates remarkable

Table 3: Evaluation on the MVBench benchmark. Best and second-best results are **bolded** and underlined, respectively. ClueNet achieves the highest average score, excelling particularly in fine-grained temporal and object interaction tasks.

Model	Avg	AS	AP	AA	FA	UA	OE	OI	OS	MD	AL	ST	AC	MC	MA	SC	FP	CO	EN	ER	CI
Closed-source MLLMs																					
GPT-4o	60.8	65.5	75.5	82.5	48.5	83.5	69.5	72.0	37.5	46.0	41.5	90.0	49.0	46.0	70.5	57.5	47.5	72.0	41.0	59.5	61.0
Qwen-VL-Max	61.4	64.5	77.0	82.0	50.0	**85.5**	74.0	75.5	36.0	52.0	33.5	91.0	47.0	54.5	64.5	61.0	46.5	74.0	**45.0**	62.5	51.5
Open-source MLLMs																					
Cos	49.6	57.5	58.0	57.0	44.0	69.0	50.5	61.0	32.0	37.5	35.0	86.5	48.5	27.5	58.0	47.5	44.5	58.5	39.5	57.0	42.0
LongVA	49.4	53.5	59.0	67.0	43.5	67.5	50.0	59.0	33.0	35.0	33.0	85.5	50.5	30.0	58.5	49.5	43.5	59.0	39.0	51.0	41.0
VideoChat2	51.1	66.0	47.5	83.5	49.5	60.0	58.0	71.5	42.5	23.0	23.0	88.5	39.0	42.0	58.5	44.0	49.0	36.5	35.0	40.5	65.5
Video-LLaVA	42.3	44.0	52.0	54.5	39.0	53.5	45.0	45.5	38.5	27.0	24.5	84.5	33.0	27.0	51.0	41.0	42.0	51.5	31.5	42.5	37.5
InternVideo2.5	68.7	**83.5**	70.0	**91.0**	48.0	75.5	89.5	77.5	41.0	64.0	41.0	90.0	57.0	77.0	**95.0**	67.0	51.0	72.0	41.5	62.5	**80.0**
Qwen2.5-VL	65.2	76.5	70.5	84.5	48.0	79.0	**94.0**	71.0	34.5	53.0	42.0	91.5	41.5	**81.0**	91.5	55.0	50.0	77.5	39.0	53.0	71.5
Qwen3-VL	64.5	64.5	58.0	78.5	48.5	78.0	82.0	60.5	42.0	**66.5**	35.5	82.5	37.5	67.5	87.0	59.0	48.0	71.0	36.0	55.0	64.0
VideoLLaMA3	68.3	73.5	77.5	83.0	50.0	83.5	92.5	79.0	42.0	56.0	48.0	92.0	56.5	78.5	91.5	**67.5**	51.0	**79.5**	32.0	57.5	74.5
ClueNet(Ours)	**69.8**	76.5	**79.5**	90.0	**51.5**	80.5	93.5	**83.0**	**43.0**	58.5	**48.5**	**92.5**	**57.5**	78.0	92.0	65.5	**51.5**	79.0	35.0	**64.0**	77.0

resilience to the strict $\alpha = 0.9$ penalty on *Temporal* and *Causal* tasks, confirming that structured clue chains provide a reliable symbolic anchor that mitigates hallucination in long-form event reasoning.

Generalization Evaluation on MVBench. To verify the generalization performance of ClueNet in the out-of-distribution scenario, we conduct experiments on the MVBench benchmark (Table 3), which comprises 20 fine-grained video understanding tasks. ClueNet achieves state-of-the-art performance with an average score of 69.8%, outperforming strong open-source baselines such as InternVideo2.5 (68.7%) and VideoLLaMA3 (68.3%), while also surpassing closed-source models like GPT-4o.

ClueNet leads on dynamic tasks where frame-averaging baselines typically struggle: Action Prediction (AP: 79.5%) and Object Interaction (OI: 83.0%) are both top-ranked, confirming that temporally-grounded clues faithfully track entity state transitions across frames. Strong performance on counting-centric tasks (AC: 57.5%, MC: 78.0%) further demonstrates that the instance-level dynamic visual clues preserve discriminative temporal structure in complex scenes. In static attribute tasks such as Object Existence, where temporal reasoning offers little advantage, ClueNet is competitive but does not lead, consistent with our design focus on temporal dynamics.

We further evaluate ClueNet on VideoMME [11] and VideoHallucer [47], confirming competitive performance against both open-source and proprietary models (full results in Appendix C).

5.3 Ablation Study

Table 4 reports ablations on STAR and NExT-QA using VideoLLaMA3 [54] as base model (Row 1).

Two-Stage Training Paradigm. The performance drop in Stage 1 alone (-1.3% on STAR) empirically illustrates the Inductive Reasoning Bias identi-

Table 4: ClueNet ablation on STAR and NExT-QA; deltas relative to Row 1 (VideoLLaMA3 [54]) baseline. Row 1 inference time omitted (different pipeline).

Components					Benchmarks		Inference	
Stage 1	Stage 2	KS	ACF	VC	STAR	NExT-QA	Time (s)	GFLOPs
-	-	-	-	-	72.1	81.3	-	-
✓	-	-	-	-	70.8 (-1.3)	80.6 (-0.7)	6.5060	54,197.96
✓	✓	-	-	-	75.1 (+3.0)	83.3 (+2.0)	5.9963	54,026.85
✓	✓	✓	-	-	75.4 (+3.3)	83.4 (+2.1)	6.0443	54,075.02
✓	✓	✓	✓	-	**77.6** (+5.5)	**84.3** (+3.0)	4.9963	51,355.89
✓	✓	✓	✓	✓	77.3 (+5.2)	84.1 (+2.8)	**3.9298**	**40,622.84**

fied in Sec. 3: the model overfits to perfect semantic anchors and its reasoning chain collapses when inference-time clues introduce inevitable noise. Stage 2 inference supervision acts as a critical regularization mechanism, forcing the model to deduce answers robustly from imperfect, self-generated evidence, yielding a significant +3.0% recovery and gain on STAR. Additional ablations (Appendix C) further confirm Stage 1's necessity: skipping Stage 1 drops STAR accuracy to 33.1%, as the model cannot produce well-formed quintuplets for ACF.

Keyframe Selection (KS) & Adaptive Clue Filter (ACF). As a heuristic pre-processing step, KS yields a modest +0.3% gain, suggesting its benefit is primarily realized in conjunction with downstream clue filtering rather than as a standalone component. Integrating the trainable ACF yields a substantial +2.5% surge on STAR by rigorously gating clues based on Visual Faithfulness and Semantic Relevance, directly mitigating the Clue Cognition Bias and preventing hallucinations from propagating into the final reasoning process.

Visual Compression (VC). The ∼21% reduction in GFLOPs (51.4K to 40.6K) and acceleration to 3.93s with only a negligible −0.3% accuracy trade-off empirically validates our Information Bottleneck hypothesis: once reasoning is anchored to ACF-verified clues, unanchored visual tokens become safely removable without loss of task-critical semantics.

5.4 In-Depth Analysis of Leveraging Clues

To locate the primary bottleneck (high-level reasoning vs. low-level visual perception) of current VideoQA systems, we conduct an oracle clue experiment on STAR. Oracle experiments supply perfect ground-truth intermediate information to eliminate upstream errors and isolate downstream component performance; our setup feeds models ground-truth dynamic clue collections with video and question inputs, bypassing visual clue extraction to set a perfect perception upper bound.

As shown in Fig. 7, all models exhibit dramatic accuracy improvements with oracle clues: Qwen2.5-VL and InternVideo2.5 improve by +22.9% and +21.1%, respectively. The near-uniform gain magnitude (∼20%) across architecturally diverse backbones suggests that these failures are systemic properties of the end-to-end paradigm itself, rather than artifacts of any specific architecture. These

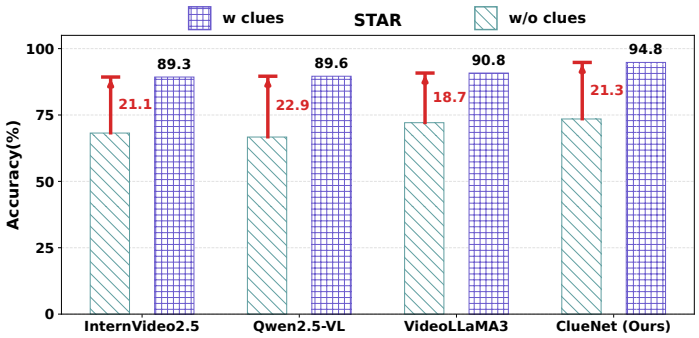

Fig. 7: Oracle experiment results on STAR. Striped and gridded bars denote standard and oracle inference.

results validate ClueNet's core design philosophy: explicitly mining structured logical evidence helps mitigate the cognitive biases observed across representative video MLLMs, and the clue-aware pipeline provides a practical reasoning paradigm for addressing the perception-cognition gap in current video understanding systems.

Notably, ClueNet achieves the highest oracle accuracy of 94.8%, outperforming all baselines by at least +4.0% despite a comparable delta gain (+21.3%), confirming that Stage 2 inference supervision optimizes the model's ability to exploit structured clues beyond clue quality alone. The residual ∼5.2% error gap under oracle conditions corresponds to the Feasibility split underperformance in Table 1, where unobservable commonsense knowledge poses an irreducible reasoning limit orthogonal to visual clue quality.

Other experimental results can be found in Appendix C.

6 Conclusion

We present ClueNet, a clue-aware video reasoning framework bridging visual perception and logical deduction. Decomposing QA into explicit visual clue localization and logical clue cognition, it mitigates hallucinations prevalent in standard end-to-end models. Extensive experiments on STAR, NExT-QA, and MVBench validate its state-of-the-art performance, with strong benchmark generalization, hallucination mitigation, and accelerated inference.

Acknowledgements

This work has been supported by the Strategic Priority Research Program of Chinese Academy of Sciences (Grant No. XDA0310502), Future Star of Aerospace Information Research Institute, Chinese Academy of Sciences, China (Grant No. E3Z10701), National Key Laboratory of Space Target Awareness, China (Grant No. STA2025ZCA0102).

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), `https://arxiv.org/abs/2502.13923`
4. Buschman, T.J., Miller, E.K.: Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices. science **315**(5820), 1860–1862 (2007)
5. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision. pp. 19–35. Springer (2024)
6. Cheng, Z., Leng, S., Zhang, H., Xin, Y., Li, X., Chen, G., Zhu, Y., Zhang, W., Luo, Z., Zhao, D., et al.: Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. arXiv preprint arXiv:2406.07476 (2024)
7. Desimone, R., Duncan, J., et al.: Neural mechanisms of selective visual attention. Annual review of neuroscience **18**(1), 193–222 (1995)
8. DiCarlo, J.J., Zoccolan, D., Rust, N.C.: How does the brain solve visual object recognition? Neuron **73**(3), 415–434 (2012)
9. Fei, H., Wu, S., Ji, W., Zhang, H., Zhang, M., Lee, M.L., Hsu, W.: Video-of-thought: Step-by-step video reasoning from perception to cognition. arXiv preprint arXiv:2501.03230 (2024)
10. Felleman, D.J., Van Essen, D.C.: Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex (New York, NY: 1991) **1**(1), 1–47 (1991)
11. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24108–24118 (2025)
12. Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al.: Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14375–14385 (2024)
13. 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)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Herzig, R., Raboh, M., Chechik, G., Berant, J., Globerson, A.: Mapping images to scene graphs with permutation-invariant structured prediction. Advances in Neural Information Processing Systems **31** (2018)

16. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A., et al.: spacy: Industrial-strength natural language processing in python (2020)
17. Hu, J., Cheng, Z., Si, C., Li, W., Gong, S.: Cos: Chain-of-shot prompting for long video understanding. arXiv preprint arXiv:2502.06428 (2025)
18. Huang, X., Zhou, H., Han, K.: Prunevid: Visual token pruning for efficient video large language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 19959–19973 (2025)
19. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
20. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2758–2766 (2017)
21. Ji, J., Krishna, R., Fei-Fei, L., Niebles, J.C.: Action genome: Actions as compositions of spatio-temporal scene graphs. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10236–10247 (2020)
22. Jiang, P., Han, Y.: Reasoning with heterogeneous graph alignment for video question answering. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 11109–11116 (2020)
23. Ko, D., Lee, J., Kang, W.Y., Roh, B., Kim, H.: Large language models are temporal and causal reasoners for video question answering. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 4300–4316 (2023)
24. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017)
25. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9972–9981 (2020)
26. Lei, J., Yu, L., Bansal, M., Berg, T.: Tvqa: Localized, compositional video question answering. In: Proceedings of the 2018 conference on empirical methods in natural language processing. pp. 1369–1379 (2018)
27. Li, C., Im, E.W., Fazli, P.: Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13723–13733 (2025)
28. 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)
29. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 conference on empirical methods in natural language processing. pp. 292–305 (2023)
30. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
31. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023)

32. Maaz, M., Rasheed, H., Khan, S., Khan, F.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12585–12602 (2024)
33. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. Advances in neural information processing systems **27** (2014)
34. Mao, J., Jiang, W., Wang, X., Feng, Z., Lyu, Y., Liu, H., Zhu, Y.: Dynamic multistep reasoning based on video scene graph for video question answering. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 3894–3904 (2022)
35. Miller, E.K., Cohen, J.D.: An integrative theory of prefrontal cortex function. Annual review of neuroscience **24**(1), 167–202 (2001)
36. Milner, A.D., Goodale, M.A.: Two visual systems re-viewed. Neuropsychologia **46**(3), 774–785 (2008)
37. Mishkin, M., Ungerleider, L.G., Macko, K.A.: Object vision and spatial vision: two cortical pathways. Trends in neurosciences **6**, 414–417 (1983)
38. Mitra, C., Huang, B., Darrell, T., Herzig, R.: Compositional chain-of-thought prompting for large multimodal models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14420–14431 (2024)
39. Nguyen, T.T., Nguyen, P., Luu, K.: 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. pp. 18384–18394 (2024)
40. Park, J., Ranasinghe, K., Kahatapitiya, K., Ryu, W., Kim, D., Ryoo, M.S.: Too many frames, not all useful: Efficient strategies for long-form video qa. arXiv preprint arXiv:2406.09396 (2024)
41. Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., Qiu, Z.: Qwen2.5 technical report (2025)
42. Ren, S., Yao, L., Li, S., Sun, X., Hou, L.: 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. pp. 14313–14323 (2024)
43. Tishby, N., Pereira, F.C., Bialek, W.: The information bottleneck method. arXiv preprint physics/0004057 (2000)
44. Wang, X., Liang, J., Wang, C.K., Deng, K., Lou, Y., Lin, M.C., Yang, S.: Vila: Efficient video-language alignment for video question answering. In: European Conference on Computer Vision. pp. 186–204. Springer (2024)
45. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: European conference on computer vision. pp. 396–416. Springer (2024)
46. Wang, Y., Li, X., Yan, Z., He, Y., Yu, J., Zeng, X., Wang, C., Ma, C., Huang, H., Gao, J., et al.: Internvideo2. 5: Empowering video mllms with long and rich context modeling. arXiv preprint arXiv:2501.12386 (2025)

47. Wang, Y., Wang, Y., Zhao, D., Xie, C., Zheng, Z.: Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. arXiv preprint arXiv:2406.16338 (2024)
48. 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)
49. Wu, B., Yu, S., Chen, Z., Tenenbaum, J.B., Gan, C.: Star: A benchmark for situated reasoning in real-world videos. arXiv preprint arXiv:2405.09711 (2024)
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. Yu, S., Jin, C., Wang, H., Chen, Z., Jin, S., Zuo, Z., Xu, X., Sun, Z., Zhang, B., Wu, J., et al.: Frame-voyager: Learning to query frames for video large language models. arXiv preprint arXiv:2410.03226 (2024)
52. Yuan, B., You, S., Bao, B.K.: Dtoma: Training-free dynamic token manipulation for long video understanding. In: Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence. pp. 2314–2322 (2025)
53. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 11975–11986 (2023)
54. Zhang, B., Li, K., Cheng, Z., Hu, Z., Yuan, Y., Chen, G., Leng, S., Jiang, Y., Zhang, H., Li, X., et al.: Videollama 3: Frontier multimodal foundation models for image and video understanding. arXiv preprint arXiv:2501.13106 (2025)
55. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations. pp. 543–553 (2023)
56. Zhang, P., Zhang, K., Li, B., Zeng, G., Yang, J., Zhang, Y., Wang, Z., Tan, H., Li, C., Liu, Z.: Long context transfer from language to vision. arXiv preprint arXiv:2406.16852 (2024)
57. Zhang, Q., Liu, M., Li, L., Lu, M., Zhang, Y., Pan, J., She, Q., Zhang, S.: Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. arXiv preprint arXiv:2506.10967 (2025)
58. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
59. Zhao, Z., Huo, Y., Yue, T., Guo, L., Lu, H., Wang, B., Chen, W., Liu, J.: Efficient motion-aware video mllm. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24159–24168 (2025)
60. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. arXiv preprint arXiv:2304.10592 (2023)