Video-HolmesV2: Can MLLMs Reason with Spatio-Temporal Audio-Visual Evidence in Long Videos?

Zhaoyang Wei[1,2], Zipeng Wang[1], Yushe Cao[3], Chenhui Qiang[2], Shuaibing Cheng[1], Xuesong Yang[1], Sen Nie[1], Bowen Jiang[1], Wenchao Ding[2], Yanchao Hao[2⋆], Zheng Wei[2], Xuehui Yu[2], and Zhenjun Han[1⋆⋆]

[1] University of Chinese Academy of Sciences, China
[2] Tencent, China
[3] Tsinghua University, China
weizhaoyang23@mails.ucas.ac.cn

Abstract. Multimodal Large Language Models have demonstrated impressive video understanding, yet their ability to reason over long-form narratives is often masked by visual-centric evaluations and inefficient context processing. Existing benchmarks over-rely on visual heuristics while marginalizing auditory cues, effectively reducing models to "silent observers" that bypass genuine cross-modal reasoning. Moreover, standard dense sampling creates an **evidence–context trade-off**: increasing frames to capture evidence inevitably leads to attention distraction and token explosion. To bridge these gaps, we present **Video-HolmesV2**, a novel benchmark designed for Deep Audio-Visual Coupling. Unlike previous works, it enforces an Evidence-Based Evaluation, requiring models to justify answers with precise **spatio-temporal audio-visual evidence**, thereby reducing confounding effects of guessing and hallucinated evidence. To support this, we introduce: (1) a Multi-Model Cross-Verification pipeline to ensure task rigor; (2) a Spatio-temporal Evidence-Aware Metric for fine-grained calibration. Furthermore, we propose an Audio-Text Guided Token Compression framework. By fusing task intent with auditory anchors, our method distills high-value reasoning cues to mitigate long-context noise. In our evaluation, even strong proprietary models achieve below 60% accuracy, while our approach outperforms over comparable open-source omni-models.

Keywords: Video Understanding · Reasoning Evidence · Audio-visual Combination

1 Introduction

The landscape of video understanding has been fundamentally transformed by the rapid evolution of Multimodal Large Language Models (MLLMs) [1,14,17,29,

⋆ Co Corresponding author.
⋆⋆ First Corresponding author.

Benchmarks	Modality	# Videos	Avg. Len	# QA Pairs	A-V Corr.	S-T Evid.	Evid. Chain
Video-Only Benchmarks							
ActivityNet-QA [40]	V	5,800	180s	58,000	✗	✗	✗
MMBench-Video [5]	V	609	165s	1,998	✗	✗	✗
EgoSchema [18]	V	5,063	180s	5,063	✗	✗	✗
Video-MME [6]	V	900	1018s	2,700	✗	✗	✗
Audio-Visual Benchmarks							
WorldSense [10]	A+V	1,662	141s	3,172	✓	✗	✗
OmniVideoBench [12]	A+V	628	384s	1000	✓	✗	✗
Our Series							
Video-Holmes (V1) [2]	A+V	270	<300s	1,837	✗	✗	✓
Video-HolmesV2 (Ours)	**A+V**	**784**	**600-7500s**	**4000**	✓	✓	✓

Table 1: Comparison with existing benchmarks. Video-HolmesV2 stands out by offering large-scale QA pairs, long-form video understanding, and unique spatio-temporal audio-visual evidence, surpassing both video-only and recent audio-visual benchmarks.

41]. While these models exhibit remarkable proficiency in short-clip perception, comprehending long-form narratives remains a formidable challenge [18, 25, 30]. To address this, recent "omni-modal" architectures—such as GPT-4o [19], Gemini 2.5 Pro [9], and Qwen2.5-Omni [38]—natively integrate auditory streams to facilitate holistic understanding of extended temporal contexts. However, evaluation paradigms have lagged significantly behind these architectural advances. Existing benchmarks—from short-clip assessments [2, 4, 10, 15] to hour-long evaluations like Video-MME [11] and LongVideoBench [36]—remain predominantly *visual-centric*, inherently relegating the auditory channel to a secondary, often neglected signal. This systematic bias reduces MLLMs to "silent observers" that overlook dynamic acoustic semantics while over-relying on visual patterns. Such visual fixation creates a critical bottleneck for long video understanding, where visual and auditory cues are deeply intertwined rather than merely additive.

Beyond the necessity of cross-modal coupling, traditional metrics that rely solely on answer accuracy fail to capture long-range cross-modal causality, as they cannot distinguish genuine reasoning from visually-biased hallucinations [3]. To bridge this critical gap, we present Video-HolmesV2, a large-scale benchmark comprising 784 long-form videos (averaging 43 minutes) specifically designed to enforce deep audio-visual coupling (A-V Corr.). As detailed in Tab. 1, unlike existing datasets [11, 42] that prioritize visual dominance or lack intermediate supervision, Video-HolmesV2 pioneers an Evidence-Based Evaluation. This rigorous paradigm mandates that models go beyond merely predicting the correct outcome; they must actively construct a logical deduction chain (Evid. Chain) substantiated by precise, spatio-temporal audio-visual evidence (S-T Evid.), thereby proving that their answers are grounded in concrete multimodal facts.

To ensure uncompromised data quality, Video-HolmesV2 is curated from recent high-fidelity cinematography and meticulously annotated via a multi-model cross-verification pipeline to eliminate unimodal biases. Consequently, our task taxonomy transcends simple perception to challenge high-order cognition, including causal and counterfactual deduction (Fig. 2a). A representative example of this intricate coupling is illustrated in Fig. 1: visually identifying a "blue

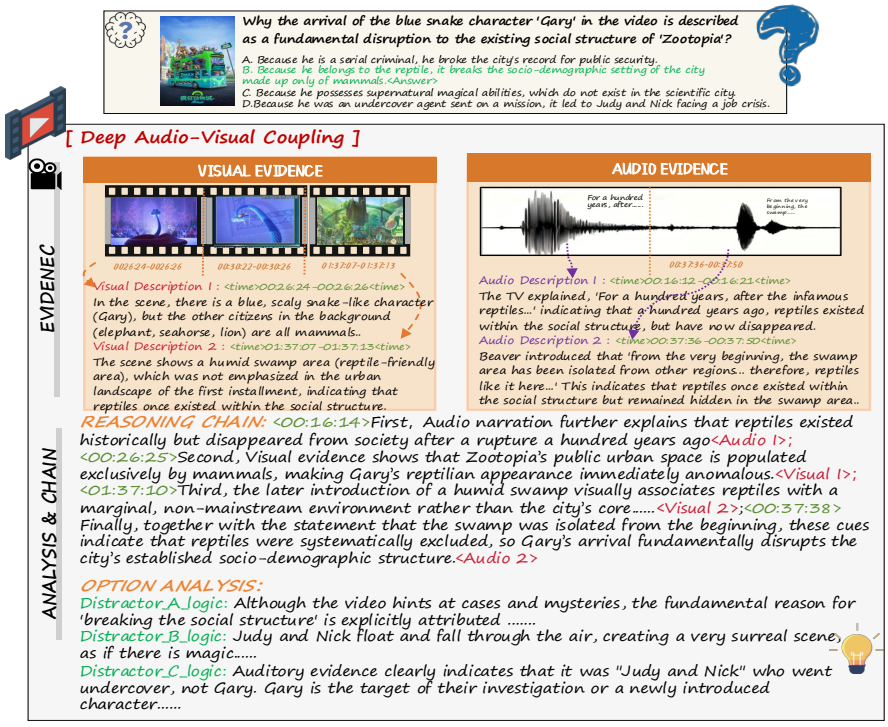


Fig. 1: Deep Audio-Visual Coupling in Video-HolmesV2. Using *Zootopia*, we demonstrate the necessity of cross-modal synthesis. Visual cues alone (identifying a blue snake) fail to imply "social disruption" without the auditory historical context (reptiles disappeared centuries ago). Uniquely, our benchmark provides an Evidence-Based Evaluation, requiring models to ground answers in precise spatio-temporal timestamps from both visual and auditory channels.

snake" only implies a social threat when synthesized with the auditory context that reptiles vanished centuries ago. Furthermore, to robustly assess these complex reasoning chains, we introduce a Spatio-Temporal Evidence-Aware Evaluation. This metric replaces rigid thresholds with Gaussian-based soft scoring and strictly conditions the final assessment on valid evidence extraction, effectively disentangling genuine multimodal reasoning from spurious guessing.

To effectively navigate the intricate causal evidence in ultra-long videos, high-density frame sampling is a prerequisite to capture fleeting evidence. However, this precipitates a severe token explosion that overwhelms the context windows of current MLLMs. While token compression offers a remedy, prevailing strategies often rely on naive top-K truncation or unimodal guidance, inadvertently severing vital cross-modal links or discarding macroscopic context. To resolve this efficiency-accuracy dilemma, we propose a **Audio-Text Guided Token Compression** framework. By fusing text-guided subjective intent with audio-

guided objective events, our approach comprehensively evaluates visual token relevance. Instead of rigid pruning, it performs fine-grained deduplication on redundant high-scoring patches and dynamically reallocates the saved budget to retain global context, ensuring both microscopic evidence and narrative atmosphere are preserved. Extensive experiments demonstrate that this paradigm not only excels on Video-HolmesV2 but also achieves state-of-the-art performance across multiple established video understanding benchmarks.

Our main contributions are summarized as follows:

- We introduce Video-HolmesV2, a benchmark of 784 long-form videos featuring *Deep Audio-Visual Coupling*. It challenges models to resolve complex causal dependencies relying on inextricably linked visual and auditory clues.
- We curate precise Spatio-Temporal Audio-Visual Evidence via multi-model cross-verification, enabling both high-fidelity bimodal annotations and a novel evaluation metric that strictly conditions reasoning accuracy on valid temporal grounding, effectively penalizing ungrounded speculation.
- We propose an Audio-Text Guided Token Compression framework. By combining bimodal scoring with differentiated deduplication, it preserves critical evidence and achieves superior reasoning accuracy compared to existing token compression methods on long-video tasks.

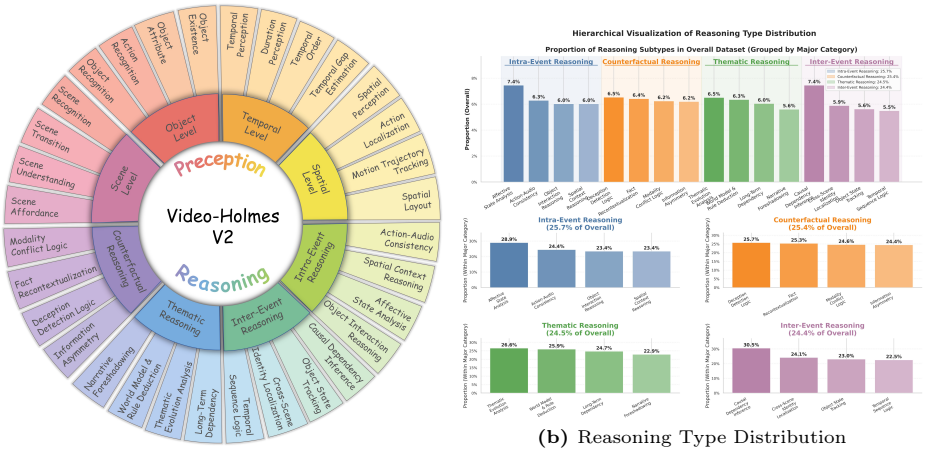
2 Related Works

2.1 Multimodal Models for Video Understanding

The field of video understanding has rapidly evolved with Multimodal Large Language Models (MLLMs). Initially, models treated videos as sequences of images [14, 17, 41], leveraging image-based foundations like CLIP [23] and modular encoders to perform unified inference [1, 29, 44]. Recently, the frontier has shifted towards "Omni-modal" architectures capable of natively processing interwoven data streams. Proprietary models like GPT-4o [19] and Gemini 2.5 Pro [9], alongside open-source efforts such as Qwen2.5-Omni [38], VITA [7, 8], Video-SALMONN [26], and InteractiveOmni [28], have explored end-to-end capabilities for simultaneous audio-video understanding. Despite these architectural advancements, evaluating and processing thousands of frames introduces severe computational burdens. While recent methods have explored token pruning [20, 24], keyframe selection [31], and memory-augmented frameworks [16] to alleviate this token explosion, efficiently bridging the gap between multimodal inputs and complex reasoning in long contexts remains a significant challenge.

2.2 General and Long Video Understanding Benchmarks

Benchmarks act as the compass for model development. Early evaluations primarily assessed fundamental perception capabilities on short clips [33, 35, 37, 40, 43]. As models progressed, the community introduced benchmarks targeting sophisticated reasoning, comprehensive short-clip abilities, Chain-of-Thought



(a) Hierarchical Task Taxonomy

(b) Reasoning Type Distribution

Fig. 2: The Taxonomy and Data Distribution of Video-HolmesV2. (a) Our **Hierarchical Taxonomy** bridges foundational Perception (inner rings) with high-order Reasoning (outer rings), uniquely emphasizing *Causal*, *Counterfactual*, and *Thematic* logic. (b) The **Statistical Distribution** validates this design, revealing a balanced composition across four major reasoning pillars ($\approx 25\%$ each), ensuring the benchmark evaluates diverse cognitive capabilities without category bias.

(CoT), and academic domain knowledge [2,4,10,15,21,32,42]. To assess temporal coherence and memory retention, datasets like MovieChat-1K [25], EgoSchema [18], LVBench [30], Video-MME [11], and LongVideoBench [36] have extended evaluations to hour-long sequences. Concurrently, in the realm of multi-modal perception, benchmarks like AVQA [39] and Music-AVQA [13] have integrated audio cues. However, existing long-video benchmarks often treat audio as a secondary modality, while audio-visual datasets are largely confined to short durations. More importantly, they do not strictly require models to substantiate their answers with precise spatio-temporal evidence, a gap highlighted by recent calls for process-level annotation [3,22,34,35]. Video-HolmesV2 directly addresses this by introducing a benchmark centered on omni-modal, event-based reasoning over truly long durations.

3 Benchmark Construction.

3.1 Initial Video Collection and Technical Curation

Our data construction begins by collecting 3704 high-quality, post-2024 cinematic videos (ranging from 10 to 150 minutes) to evaluate genuine zero-shot capabilities and prevent data contamination. We prioritize visually and acoustically rich narratives (e.g., award-winning shorts, episodic series, movies) while systematically excluding unimodal-dominant or causally sparse genres, such as

news reports and sports matches. Detailed duration stratification and engagement thresholds are provided in Appendix A.

Multi-Dimensional Analyzability Assessment. To rigorously quantify the multimodal reasoning potential of the filtered candidates, we employ a consortium of state-of-the-art native MLLMs (e.g., Gemini 3 Pro, GPT-5.2, Qwen3-VL-235B-A22B-thinking) to evaluate each video. Specifically, videos are scored on a scale from 1 to 10 across five core dimensions:
1. Audio-Visual Complementarity: The strict necessity of synthesizing both visual and auditory channels, ensuring neither modality is sufficient in isolation.
2. Cross-Modal Causal Complexity: The depth of multi-hop cause-and-effect chains that inextricably span across modalities and time.
3. Dynamic Character Arc: The meaningful evolution of key entities over time, evidenced concurrently by visual changes and auditory shifts.
4. Narrative Coherence: The structural integrity and logical consistency of the storyline, filtering out videos with disjointed editing or fallacies.
5. Audience Critical Reception: A meta-metric gauging narrative depth based on whether the video provokes high-quality intellectual discussion. A comprehensive elaboration of these criteria and the hierarchical processing strategy for ultra-long sequences can be found in Appendix A.

Quantitative Curation and Stratification. To synthesize the multi-dimensional assessments into a robust decision, we implement a rigorous three-stage quantitative pipeline. This process filters low-quality content, enforces strict multimodal dependencies, and identifies high-difficulty samples based on model uncertainty.
1. Bias Mitigation via Z-Score Normalization. Different VLMs exhibit intrinsic scoring biases (e.g., varying scales of leniency). To mitigate this, we standardize the raw score $x_{d,i}$ of model i on dimension d to a Z-score via $z_{d,i} = (x_{d,i} - \mu_i)/\sigma_i$. This step aligns the evaluator distributions, ensuring the selection is driven by relative ranking rather than absolute score artifacts.
2. Priority-Weighted Aggregation with Modality Veto. We derive a unified Analyzability Score (S_{final}) via a weighted aggregation of the normalized dimensions. We assign higher importance weights (w_d) to Audio-Visual Complementarity and Cross-Modal Causal Complexity to emphasize reasoning:

$$S_{\text{final}} = \sum_{d \in \mathcal{D}} w_d \cdot \left(\frac{1}{N} \sum_{i=1}^{N} z_{d,i} \right). \tag{1}$$

Crucially, to prevent the inclusion of "high-quality but single-modality" videos, we enforce a **Modality Veto**. Regardless of the total S_{final}, any video scoring below a specific threshold ($z < -0.5$) in the *Audio-Visual Complementarity* dimension is automatically discarded.
3. Uncertainty-Driven Data Stratification. To quantify evaluator consensus and highlight narrative ambiguity, we define a **Divergence Score (D_{video})** as the arithmetic mean of the standard deviations across all dimensions:

$$D_{\text{video}} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \sqrt{\frac{\sum_{i=1}^{N} (z_{d,i} - \bar{z}_d)^2}{N}}. \tag{2}$$

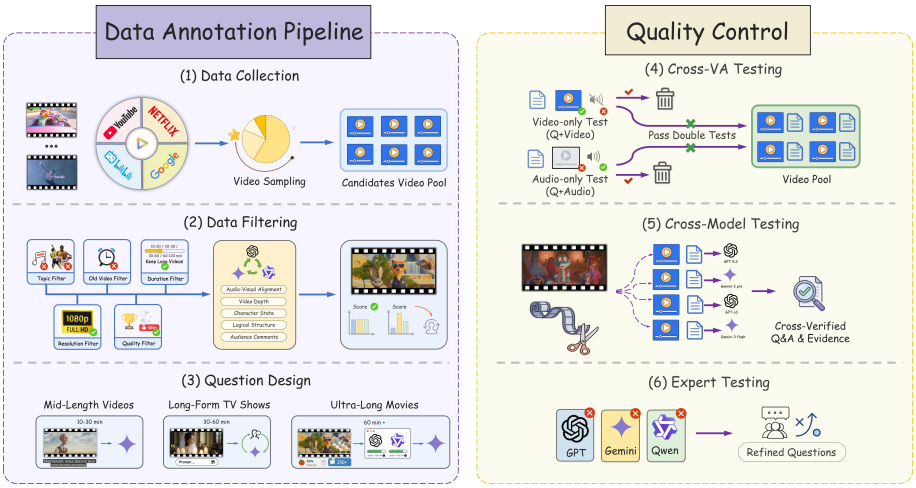


Fig. 3: The Scalable Annotation and Quality Control Pipeline. Our construction process integrates native long-context MLLM generation with rigorous human-in-the-loop verification. The pipeline moves from (1-3) diverse sourcing and stratified question generation to a strict (4-6) multi-stage validation protocol. A critical innovation is step (4) *Cross-VA Testing*, which acts as a "Unimodal Trap" to automatically discard questions solvable by audio or vision alone, ensuring that every sample in Video-HolmesV2 strictly enforces multimodal interdependence. Finally, expert review (6) adjudicates hard negatives to guarantee logical determinism.

A high D_{video} indicates significant SOTA model disagreement. Instead of a fixed arbitrary threshold, we define an adaptive high-divergence cutoff $\tau_{div} = \mu_D + \sigma_D$ to statistically flag the top $\approx 16\%$ most controversial samples. Consequently, videos are strictly categorized into: (1) **Core Set** ($S_{\text{final}} \geq 0, D_{\text{video}} < \tau_{div}$); (2) **Rejected** ($S_{\text{final}} < 0$ or Vetoed); and (3) **Hard Candidates** ($D_{\text{video}} \geq \tau_{div}$). These high-variance candidates are directly routed to expert human review for further adjudication.

3.2 Scalable Annotation via Native Long-Context Understanding

As illustrated in Fig. 3, we propose a unified annotation pipeline leveraging native long-context MLLMs. Unlike traditional clip-based methods, we process complete narrative streams to capture long-range dependencies. The workflow comprises four rigorous phases, with detailed process provided in Appendix A.1. **Phase 1: Stratified Question Generation** (Fig. 3, Step 3)**.** Recognizing varying narrative complexities, we employ a length-differentiated strategy: direct *Constraint-Based Prompting* for short films, *Contextual Priming* via injected lore for episodic series, and *Comment-Driven Inverse Reasoning* for feature films. Crucially, the models are constrained to output not only the complex QA pairs but also explicit grounded *spatio-temporal audio-visual evidence*.

Phase 2: The "Unimodal Trap" Blind Filtration (Fig. 3, Step 4)**.** To strictly enforce multimodal reasoning, we introduce an adversarial filter. Generated QA pairs are evaluated against ablated *Visual-Only* (silent frames) and *Audio-Only* (subtitles) inputs. Any question solvable by a single modality with high confidence is automatically discarded as a "pseudo-multimodal" instance.
Phase 3: Segment-Level Cross-Verification (Fig. 3, Step 5)**.** We isolate the specific narrative segments containing the proposed evidence and deploy a diverse consortium of reasoning-heavy SOTA models as validators. A sample is retained only if a majority consensus is reached on both the correct answer and the precise evidence timestamps (Intersection over Union > 0.75).
Phase 4: Expert Adjudication for Hard Negatives (Fig. 3, Step 6)**.** Samples where the automated ensemble disagrees or fails are routed to a Double-Blind Expert Review. Specialists independently assess whether the failure stems from model hallucination or genuine, deep logical complexity. Validated samples from this queue are salvaged to form the *Hard Subset* of our benchmark, representing the upper bound of current AI reasoning capabilities.

4 Evaluation Metrics

We introduce a comprehensive evaluation protocol that moves beyond rigid thresholding to jointly assess multiple-choice reasoning accuracy and spatio-temporal grounding quality in long videos.
4.1 Answer Accuracy ($Acc_{\mathbf{MCQ}}$). The correctness of the reasoning conclusion is binary. For a chosen option $\hat{y}$ and ground truth y:

$$Acc_{\mathrm{MCQ}} = \mathbb{I}(\hat{y} = y) \tag{3}$$

4.2 Evidence Quality ($S_{\mathbf{evidence}}$). Measuring evidence quality in long videos is challenging due to variable event durations. We propose a robust metric that combines length-adaptive temporal alignment with semantic matching.

1. Temporal Soft Alignment (S_{temp}). To accommodate systematic temporal drift ($\tau_{drift} = avg\ duration * 0.05$) while severely penalizing duration mismatch, we formulate a soft score that multiplies a length-similarity ratio by a Gaussian decay. For a predicted interval p and ground-truth g with lengths $L_{(\cdot)}$ and centers $C(\cdot)$:

$$S_{\mathrm{raw}}(p, g) = \left(\frac{\min(L_p, L_g)}{\max(L_p, L_g)} \right) \cdot \exp\left(- \frac{\max(0, \|C(p) - C(g)\| - \tau_{drift})^2}{2 \cdot (\alpha_{gauss} \cdot L_g)^2} \right) \tag{4}$$

where $\alpha_{gauss} = 0.5$. The Gaussian factor elegantly absorbs minor localization shifts within τ_{drift}, while the leading ratio strictly enforces structural consistency. To handle fragmented outputs, we coalesce proximal predictions (gap $\leq \tau$) into unified spans $\mathcal{P}_c$. Following bipartite matching with ground-truths, the temporal score is strictly prediction-centric: $S_{\mathrm{temp}} = \frac{1}{|\mathcal{P}_c|} \sum_{p \in \mathcal{P}_c} S_{\mathrm{raw}}(p, g^*)$. Unmatched predictions score zero to penalize hallucinations. Crucially, since logical deduction often requires only partial clues, unretrieved ground-truths are not penalized here; this recall penalty is inherently deferred to S_{sem}.

2. Semantic Matching (S_{sem}). Beyond mere temporal overlap, models must accurately identify causal events. Thus, we utilize Qwen3-235B-A22B-Thinking

to evaluate semantic equivalence exclusively for temporally intersecting predictions ($S_{\text{temp}} > 0$). Since complex reasoning often requires only a subset of clues, we apply the $F_{0.5}$ score (weighting precision twice over recall) to heavily penalize hallucinations while tolerating partial, sufficient evidence:

$$S_{\text{sem}} = (1 + 0.5^2) \cdot \frac{P_{\text{sem}} \cdot R_{\text{sem}}}{0.5^2 \cdot P_{\text{sem}} + R_{\text{sem}}} \tag{5}$$

4.3 Overall Performance (S_{total}). The final score enforces a strict dependency between the answer and evidence. We first combine the temporal and semantic scores into a unified evidence metric:

$$S_{\text{evidence}} = \alpha_{evi} \cdot S_{\text{temp}} + (1 - \alpha_{evi}) \cdot S_{\text{sem}} \tag{6}$$

where $\alpha_{evi} = 0.5$ balances temporal precision and semantic correctness. The total performance score is then defined as:

$$S_{\text{total}} = \lambda_1 \cdot Acc_{\text{MCQ}} + \lambda_2 \cdot (Acc_{\text{MCQ}} \cdot S_{\text{evidence}}) \tag{7}$$

We strictly set weights $\lambda_1 = 0.6$ and $\lambda_2 = 0.4$. The conditional term ($Acc_{\text{MCQ}} \cdot S_{\text{evidence}}$) ensures evidence quality is only rewarded when the final answer is correct, effectively penalizing ungrounded guessing.

5 Method

5.1 Lightweight Text-Guided Visual Scoring. Long-form video understanding entails processing an exorbitant number of visual tokens, rendering the direct application of massive MLLM computationally prohibitive. To efficiently identify query-relevant visual signals before heavy autoregressive decoding, we introduce a lightweight, rank-supervised token selector (Qwen2.5 0.5B) distilled from a large-scale teacher model (Qwen2.5-VL 72B).

Given a textual query q and a sequence of visual tokens $\mathbf{V} = \{v_1, v_2, \ldots, v_N\}$, our goal is to compute a text-guided relevance score $S_{text} \in \mathbb{R}^N$. Prior studies indicate that intermediate transformer layers in VideoLLMs exhibit the highest semantic alignment between text queries and visual features. Therefore, we extract the cross-modal attention from a designated reference layer L_{ref} of the teacher model. For the i-th visual token, the teacher's semantic relevance score r_i^{ref} is defined by averaging the attention weights across all H heads:

$$r_i^{ref} = \frac{1}{H} \sum_{h=1}^{H} \mathbf{A}_{q \to v_i}^{(L_{ref}, h)} \tag{8}$$

where $\mathbf{A}_{q \to v_i}^{(L_{ref}, h)}$ denotes the attention weight from the query tokens to v_i. To circumvent the prohibitive computational cost of the teacher model during inference, we train a shallow, lightweight selector to predict these relevance scores, denoted as $\hat{\mathbf{r}} = [\hat{r}_1, \ldots, \hat{r}_N]$. Instead of minimizing the absolute mean squared error, we optimize the ranking consistency between the student and the teacher using a Spearman Rank Correlation Loss. The training objective is:

$$\mathcal{L}_{rank} = 1 - \frac{\sum_{i=1}^{N} (R(r_i^{ref}) - \bar{R}^{ref})(R(\hat{r}_i) - \bar{\hat{R}})}{\sqrt{\sum_{i=1}^{N} (R(r_i^{ref}) - \bar{R}^{ref})^2 \sum_{i=1}^{N} (R(\hat{r}_i) - \bar{\hat{R}})^2}} \tag{9}$$

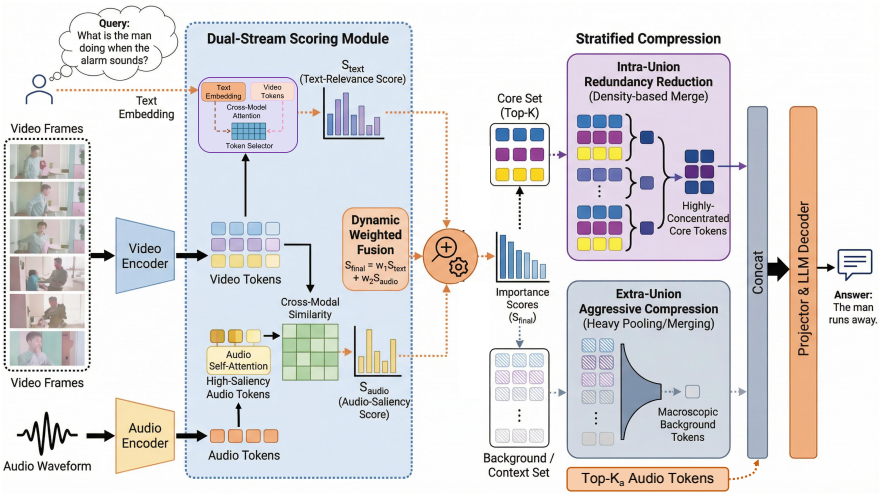


Fig. 4: The Audio-Text Guided Token Compression Framework. We fuse text-guided subjective intent (S_{text}) with audio-guided objective cues (S_{audio}) to dynamically assess visual importance. A Stratified Compression strategy then preserves fine-grained evidence in the Core Set while aggressively pooling the Background Set to maximize context efficiency.

where $R(\cdot)$ denotes the continuous approximation of the ranking function via differentiable sorting, and $\bar{R}$ is the mean rank. Once trained, this lightweight selector efficiently outputs the text-guided visual score $S_{text} = \hat{\mathbf{r}}$, serving as the top-down semantic prior for our subsequent dual-stream compression. More details in Appendix C.1.

5.2 Training-Free Audio-Guided Visual Scoring. To capture crucial out-of-view or transient auditory events often omitted by text queries, we introduce a complementary, training-free audio-guided scoring mechanism that anchors visual tokens to salient acoustic cues.

The process consists of two stages: intra-modal audio filtering and cross-modal saliency propagation. Given the raw audio tokens $\mathbf{A} \in \mathbb{R}^{N_a \times d}$ extracted by the audio encoder, we first identify the most informative audio anchors. Since continuous background noise provides marginal semantic value, we leverage the intrinsic self-attention matrix $\mathbf{M} \in \mathbb{R}^{N_a \times N_a}$ from the final layer of the audio encoder. The saliency of the j-th audio token is quantified by the average attention it receives from all other audio tokens:

$$s_a^{(j)} = \frac{1}{N_a} \sum_{k=1}^{N_a} \mathbf{M}_{k,j} \tag{10}$$

We select the top-K_a (e.g., $K_a = 2048$) tokens with the highest s_a to form a compact set of salient audio anchors $\hat{\mathbf{A}} \in \mathbb{R}^{K_a \times d}$.

Subsequently, we project this acoustic saliency into the visual domain. For each visual token $v_i \in \mathbf{V}$, we compute its cross-modal semantic alignment with the salient audio anchors using normalized cosine similarity. The audio-guided visual score $S_{audio}^{(i)}$ is determined by its maximum similarity to any of the audio

anchors:

$$S_{audio}^{(i)} = \max_{1 \le j \le K_a} \left(\frac{v_i \cdot \hat{a}_j}{\|v_i\|_2 \|\hat{a}_j\|_2} \right) \tag{11}$$

where $\hat{a}_j \in \hat{\mathbf{A}}$. By employing the max-pooling operation over the audio dimension, $S_{audio} \in \mathbb{R}^{N_v}$ effectively highlights specific visual regions that synchronously align with transient auditory peaks, serving as an indispensable counterpart to the text-guided scores. More details in Appendix C.2.

5.3 Dual-Stream Fusion and Differentiated Compression. The text-guided score S_{text} and audio-guided score S_{audio} capture subjective intents and objective transient events, respectively. To construct a holistic importance metric, we perform normalization on both scores, denoted as $\tilde{S}_{text}$ and $\tilde{S}_{audio}$, and compute the final fused score:

$$S_{final} = \alpha_{fuse} \cdot \tilde{S}_{text} + (1 - \alpha_{fuse}) \cdot \tilde{S}_{audio} \tag{12}$$

where $\alpha_{fuse} \in [0, 1]$ balances the bimodal influences (empirically set to 0.7).

Given a strict sequence length budget B_{vis} (e.g., $B_{vis} = 8192$) for the LLM, a naive Top-B_{vis} truncation based on S_{final} suffers from high semantic redundancy and a complete loss of macro-context. To address this, we propose a Masked Differentiated Compression strategy. We first generate a binary spatial-temporal mask $\mathcal{M} \in \{0, 1\}^{T \times H \times W}$ by thresholding the Top-B_{vis} tokens of S_{final}. Tokens with $\mathcal{M} = 1$ constitute the **Core Set** $\mathcal{C}$, while those with $\mathcal{M} = 0$ form the **Background Set** $\mathcal{B}$.

Intra-Set Fine-Grained Compression. The Core Set $\mathcal{C}$ contains semantically dense but visually redundant tokens (e.g., homogeneous sky patches scoring high). We apply a masked spatio-temporal merging operation exclusively within $\mathcal{C}$. For spatially or temporally adjacent tokens $v_i, v_j \in \mathcal{C}$, if their visual feature cosine similarity exceeds a predefined threshold τ_{intra}, they are merged into a single token. The fused visual feature is computed via average pooling, while its importance score inherits the maximum value: $S_{final}^{(new)} = \max(S_{final}^{(i)}, S_{final}^{(j)})$. This ensures that visual fidelity is preserved without diluting the semantic saliency. This fine-grained deduplication reduces the size of $\mathcal{C}$ from B_{vis} to n ($n \le B_{vis}$), dynamically freeing up a token quota $R = B_{vis} - n$.

Extra-Set Aggressive Scavenging. The liberated quota R is utilized to rescue discarded tokens from the Background Set $\mathcal{B}$, reinstating macro-context. If $R > 0$, we retrieve the top $M \times R$ tokens from $\mathcal{B}$ based on S_{final}, where M is the aggressive compression ratio (e.g., $M = 4$). Since background elements primarily provide global atmosphere rather than fine-grained details, we perform aggressive M:1 pooling on these tokens to strictly fit them into the R budget.

Sequence Reassembly. Finally, the n highly-concentrated core tokens and the R macroscopic background tokens are concatenated. By strictly preserving their original spatio-temporal positional embeddings during merging, the reconstructed B_{vis}-token and top-K_a audio token sequence retains integral causal and spatial structures for the LLM's autoregressive decoding. More details in Appendix C.3.

Model	Perception					Reasoning					Overall
	Obj.	*Scn.*	*Spa.*	*Tem.*	**Ave.**	*Intra*	*Inter*	*Count.*	*Them.*	**Ave.**	
Proprietary Models											
GPT-4o	52.1	60.2	65.4	46.7	56.1	43.5	40.2	50.1	51.8	46.4	46.9
Gemini-2.5-Flash	63.5	68.2	71.4	64.9	67.0	63.5	62.8	66.4	68.1	65.2	65.3
Gemini-2.5-Pro	**74.3**	**77.3**	**81.8**	**70.6**	**75.8**	**71.5**	**70.8**	**74.6**	**75.5**	**73.1**	**73.2**
Open Source Models											
Qwen3-VL-4B	38.7	43.6	72.3	33.3	47.3	25.0	20.7	28.3	34.2	27.0	28.0
Qwen2.5-VL-7B	30.7	56.4	46.8	38.9	41.9	37.3	31.2	40.7	42.7	37.9	38.1
Qwen3-VL-8B	33.8	56.4	75.1	35.2	49.5	37.3	31.3	41.6	46.2	39.0	39.5
Qwen2.5-VL-32B	43.6	53.9	70.2	44.4	52.7	41.0	37.1	46.7	47.9	43.2	43.6
MiniCPM-V-4.5	38.7	23.1	51.1	38.9	38.6	43.8	43.0	48.7	54.3	47.4	47.0
InternVL3-8B	48.2	52.5	60.4	39.5	50.6	44.3	43.5	51.1	53.8	48.0	48.1
Qwen3-VL-30B-A3B	49.5	58.2	62.8	45.2	53.7	42.2	41.8	52.2	56.1	48.1	48.4
Qwen3-VL-32B	43.1	66.5	73.2	37.8	54.4	44.3	42.1	55.3	55.8	49.4	49.6
Qwen2.5-VL-72B	43.6	51.3	72.3	41.7	52.2	44.8	42.7	56.5	57.3	50.1	50.2
Qwen3-VL-235B-A22B	**54.8**	**67.2**	**75.9**	**48.4**	**58.5**	**48.5**	**48.5**	**60.4**	**63.7**	**55.1**	**55.3**
Qwen2.5-Omni-7B	33.9	59.0	53.2	27.8	42.9	40.7	35.6	43.0	43.1	40.6	40.7
Ours	37.1	66.7	59.6	25.0	46.7	42.7	38.4	47.7	48.8	44.4	44.5 (↑ 3.8)

Table 2: Comprehensive Benchmark Results for Perception and Reasoning. The final **Overall** score is strictly weighted by the sample size of each module. Models are sorted by their final weighted Overall score within their categories. The best scores among proprietary and open-source models are highlighted in **bold**, respectively.

6 Experiment

6.1 Experimental Settings

Evaluation Setup. To establish a comprehensive baseline, we evaluate both closed-source frontier models (GPT-4o, Gemini) and recent open-source models (Qwen series). The open-source evaluations are deployed using the MS-SWIFT framework on 2 × H20 GPUs. For generation hyperparameters, we set the temperature to 0.01 to ensure deterministic reasoning, and crucially expand max tokens to 8192. This extended context allows models sufficient generation capacity to output detailed visual and acoustic evidence descriptions without truncation.
Training Implementation. For our token compression method, we evaluate on standard long-video benchmarks (e.g., VideoMME, WorldScene), strictly aligning our evaluation protocol with OmniZip [27] for fair comparison. Our token selector is trained on 4 × H20 GPUs with a total batch size of 8 and a learning rate of 1×10^{-5}. The training data comprises 50K long-duration samples, randomly selected from the longest videos within the LLaVA-Video-178k dataset.

6.2 Main Results on Video-HolmesV2

Tab. 2 presents the comprehensive evaluation of state-of-the-art MLLMs on our Video-HolmesV2 benchmark.

Model	Reasoning (MCQ Accuracy %)					Evidence (Score)					Overall
	Intra	*Inter*	*Count.*	*Them.*	**Ave.**	*Intra*	*Inter*	*Count.*	*Them.*	**Ave.**	
Open Source Models											
Qwen3-VL-235B-A22B	48.5	48.5	60.4	63.7	55.1	7.2	6.5	9.8	10.5	8.5	34.9
Proprietary Models											
Gemini-2.5-Flash	63.5	62.8	66.4	68.1	65.2	26.5	25.8	30.2	32.1	28.6	46.6
Gemini-2.5-Pro	**71.5**	**70.8**	**74.6**	**75.5**	**73.1**	**33.2**	**31.5**	**36.8**	**39.4**	**35.2**	**54.2**

Table 3: Comparison between Multiple-Choice Reasoning Accuracy and Evidence Score. The **Overall** score is calculated as $S_{\text{total}} = 0.6 \cdot Acc_{\text{MCQ}} + 0.4 \cdot (Acc_{\text{MCQ}} \cdot S_{\text{evidence}})$, strictly penalizing models that fail to provide correct evidence for their choices. The best scores in each column are highlighted in **bold**.

1. The Capability Gap and Proprietary Dominance. There remains a substantial performance gap between proprietary and open-source models. Gemini-2.5-Pro emerges as the state-of-the-art, achieving the highest overall accuracy (73.2) and demonstrating robust consistency across all Perception and Reasoning sub-categories. Among the open-source community, the massive Qwen3-VL-235B-A22B establishes the frontier baseline with an overall score of 55.3, yet it still lags significantly behind the leading closed-source counterpart. Interestingly, GPT-4o struggles on our benchmark (46.9). We attribute this primarily to its restrictive input constraints (capped at 50 sampled frames). In long-form narratives requiring dense, multi-timestamp audio-visual integration, such extreme sparse sampling inevitably misses fleeting critical evidence and severs long-range causal chains, rendering it ineffective for video reasoning.

2. The Perception-to-Reasoning Degradation. The sub-module breakdown exposes a sharp cognitive cliff in smaller open-source models. For instance, Qwen3-VL-4B achieves a reasonable 47.3 in foundational Perception but collapses to 27.0 in higher-order Reasoning. This indicates that while basic spatio-temporal recognition is improving, bridging these observations into multi-hop causal chains over long sequences remains a profound bottleneck.

3. Superior Efficiency of Our Approach. Notably, our framework (**Ours**, 44.5) effectively outperforms parameter-heavy models (e.g., Qwen2.5-VL-32B, 43.6) and omni-models (Qwen2.5-Omni, 40.7). Despite severe token compression, its strong performance—particularly in *Scene* and *Thematic* reasoning—proves that our Audio-Text Guided strategy successfully preserves critical macro-context and micro-evidence without token explosion.

4. Severe Degradation over Extreme Contexts. Tab. 10 reveals a universal vulnerability: reasoning accuracy degrades significantly as video length increases. The leading Qwen3-VL-235B-A22B drops sharply from ~ 70 on Short videos to below 45 on Ultra-long movies, where it is even outperformed by smaller models like Qwen2.5-VL-72B. This proves that simply scaling up parameters cannot resolve "attention dilution" in massive contexts, highlighting the critical necessity of our efficient token compression method (detailed analysis in Appendix D.2).

Method	Retained Ratio	AVUTBench							VideoMME	ShortVid-Bench	Avg. Ratio
		EL	OR	OM	IE	CC	CM	Avg.	wo	Avg. Score	
Qwen2.5-Omni-7B											
Full Tokens	100%	38.2	67.8	59.6	85.6	44.1	66.7	64.5	66.0	70.5	100%
Random	40%	31.7	58.5	53.3	74.9	43.2	59.0	56.9	65.0	67.7	94.3%
Random	55%	38.2	64.9	55.6	80.1	34.7	65.0	61.0	65.4	68.3	96.9%
FastV	35%	24.1	60.7	54.3	81.6	40.7	58.3	57.8	-	67.9	93.8%
FastV	50%	34.1	64.3	57.1	77.6	36.4	56.4	58.4	-	68.0	94.3%
DyCoke (V&A)	35%	32.9	62.1	54.9	74.5	39.0	58.3	57.4	65.2	68.0	94.7%
DyCoke (V&A)	50%	**38.8**	67.2	58.2	81.9	39.0	62.4	62.0	65.5	68.5	97.5%
OmniZip	35%	34.1	67.5	54.6	83.7	42.4	61.2	61.0	66.1	69.0	97.6%
OmniZip	45%	38.4	67.2	56.9	**85.3**	42.4	66.0	63.0	66.3	69.9	99.1%
Ours	-	37.8	**68.5**	**61.2**	85.0	**46.3**	**68.2**	**65.5**	**67.8**	**70.8**	**101.5%**

Table 4: Performance on AVUTBench, VideoMME, and ShortVid-Bench.

Method	Retained Ratio	Tech & Science	Culture & Politics	Daily Life	Film & TV	Perform.	Games	Sports	Music	Avg.
Qwen2.5-Omni-7B										
Full Tokens	100%	52.4	50.1	48.5	44.6	43.8	41.6	41.6	47.3	46.8
Random	55%	47.1	47.0	44.4	41.2	40.0	40.1	40.1	46.3	43.6
FastV	50%	48.8	47.4	44.2	44.1	41.2	38.3	40.0	46.6	44.3
DyCoke (V&A)	50%	48.4	49.9	46.7	41.4	39.9	40.8	40.2	46.5	44.6
OmniZip	35%	48.3	49.5	47.6	42.5	40.1	40.2	42.3	46.3	45.3
Ours	-	**51.5**	**50.5**	**48.8**	**45.1**	**42.5**	**41.9**	**42.8**	**47.6**	**47.2**

Table 5: Detailed performance across different categories on WorldScene.

6.3 The Illusion of Accuracy: Why Evidence Matters.

Tab. 3 shows Multiple-Choice (MCQ) accuracy with spatio-temporal evidence-aware capabilities (smaller models are excluded due to near-zero evidence scores). A striking paradox emerges with **Qwen3-VL-235B-A22B**: despite a respectable 55.1 MCQ accuracy, its Evidence score collapses to 8.5. This exposes a critical flaw in traditional evaluations: the model's performance is largely an illusion. It relies heavily on parametric memory to guess answers but hallucinates the spatio-temporal proof. Conversely, **Gemini-2.5-Pro** shows healthier reasoning-grounding synchronization (Evidence 35.2). By strictly computing S_{total} to penalize ungrounded guessing, our benchmark evaluates *epistemic honesty*—ensuring models genuinely deduce from video facts rather than exploiting statistical priors (More experiments in Appendix D).

6.4 Results on Other Benchmarks.

As shown in Tab. 4, our method using Qwen2.5-Omni-7B not only outperforms recent compression baselines but remarkably surpasses the uncompressed *Full Tokens* baseline (101.5% relative performance). This proves our bimodal selection acts as a powerful noise filter to prevent attention dilution. Crucially, unlike ratio-based methods that still suffer from token explosion on long videos,

TGC	AGC	DSFDC	VideoMME	WorldScene
✗	✗	✗	66.0	46.8
✓	✗	✗	67.0	46.5
✗	✓	✗	63.3	45.1
✓	✓	✗	67.4	46.9
✓	✓	✓	**67.8**	**47.2**

Table 6: Ablation study of components (Text-Guided, Audio-Guided, Dual-Stream Compress).

Settings	Token Combinations					
	Comb 1	Comb 2	Comb 3	**Ours (Best)**	Comb 5	Full
Visual Tokens	4096	4096	8192	**8192**	12288	-
Audio Tokens	1024	2048	1024	**2048**	4096	-
VideoMME	66.5	66.7	67.4	**67.8**	67.6	66.0
WorldScene	45.7	46.2	46.9	**47.2**	47.3	46.8

Table 7: Ablation on Token combinations. The 8192/2048 setting achieves the optimal balance.

our framework operates on a **fixed token budget**. While matching OmniZip's $\sim$ 40% retention under a standard 128-frame setting, our fixed-budget design uniquely scales to massive frame counts without memory bottlenecks. This robust superiority generalizes across diverse genres (Tab. 5). Our approach achieves SOTA scores across all eight WorldScene categories, yielding a 47.2 average that again beats the Full Tokens baseline (46.8). This confirms our dual-stream scoring seamlessly adapts to varying narrative paces, enhancing global understanding despite extreme token reduction.

6.5 Ablation Study

Effectiveness of Key Components. Tab. 6 isolates module contributions. Using Audio-Guided Compression (AGC) alone degrades performance by discarding silent yet critical visual frames. However, fusing it with Text-Guided Compression (TGC) boosts VideoMME from 67.0 to 67.4. Finally, adding Dual-Stream Fusion and Differentiated Compression (DSFDC) achieves the optimal 67.8, proving that deduplicating core tokens while preserving background context massively outshines naive hard-pruning.

Optimal Token Allocation. Tab. 7 examines token budgets. Accuracy steadily improves as budgets scale up to our default 8192 visual and 2048 audio tokens. Crucially, expanding further (12288 visual / 4096 audio) actually causes a slight performance drop. This empirically validates the "attention dilution" issue: excessive tokens reintroduce background noise and distract reasoning. Thus, the 8192/2048 setting strikes the balance between reasoning accuracy and computational efficiency.

7 Conclusion

We introduce a novel audio-visual benchmark for extreme-length videos, alongside a Spatio-Temporal Evidence-Aware Evaluation protocol to rigorously assess fine-grained multimodal reasoning. We observe that massive token counts in long videos severely overwhelm MLLM context limits, causing attention dilution and the loss of critical details. To address this, we propose a Audio-Text Guided Token Compression framework. By prioritizing the retention of crucial audio-visual tokens within a fixed budget, our method effectively filters redundancy, allowing models to remarkably surpass uncompressed full-token baselines.

References

1. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024)
2. Cheng, J., Ge, Y., Wang, T., Ge, Y., Liao, J., Shan, Y.: Video-holmes: Can mllm think like holmes for complex video reasoning? arXiv preprint arXiv:2505.21374 (2025)
3. Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al.: Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 (2021)
4. Fang, X., Mao, K., Duan, H., Zhao, X., Li, Y., Lin, D., Chen, K.: Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. ArXiv **abs/2406.14515** (2024)
5. 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. arXiv preprint arXiv:2405.21075 (2024)
6. 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)
7. Fu, C., Lin, H., Long, Z., Shen, Y., Zhao, M., Zhang, Y., Wang, X., Yin, D., Ma, L., Zheng, X., et al.: Vita: Towards open-source interactive omni multimodal llm. arXiv preprint arXiv:2408.05211 (2024)
8. Fu, C., Lin, H., Wang, X., Zhang, Y.F., Shen, Y., Liu, X., Cao, H., Long, Z., Gao, H., Li, K., et al.: Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. arXiv preprint arXiv:2501.01957 (2025)
9. Google: Gemini-2.5-pro (2025)
10. Hong, J., Yan, S., Cai, J., Jiang, X., Hu, Y., Xie, W.: Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. arXiv preprint arXiv:2502.04326 (2025)
11. Hu, K., Wu, P., Pu, F., Xiao, W., Zhang, Y., Yue, X., Li, B., Liu, Z.: Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. arXiv preprint arXiv:2501.13826 (2025)
12. Li, C., Chen, Y., Ji, Y., Xu, J., Cui, Z., Li, S., Zhang, Y., Wang, W., Song, Z., Zhang, D., et al.: Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. arXiv preprint arXiv:2510.10689 (2025)
13. Li, G., Wei, Y., Tian, Y., Xu, C., Wen, J.R., Hu, D.: Learning to answer questions in dynamic audio-visual scenarios. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
14. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
15. 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)
16. Liu, Y., Lin, K.Q., Chen, C.W., Shou, M.Z.: Videomind: A chain-of-lora agent for long video reasoning (2025), `https://arxiv.org/abs/2503.13444`

17. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv preprint arXiv:2306.05424 (2023)
18. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. Advances in Neural Information Processing Systems **36**, 46212–46244 (2023)
19. OpenAI: Hello gpt-4o (2024), https://openai.com/index/hello-gpt-4o/
20. Peng, D., Yang, X., Guo, Z., Zhang, Y., Chen, C., Zhang, Y., Yao, Y., Wan, F., Ke, W., Sun, M.: Flexivideo: Variation-aware temporal dynamics modeling for efficient video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9804–9814 (2026)
21. Qi, Y., Zhao, Y., Zeng, Y., Bao, X., Huang, W., Chen, L., Chen, Z., Zhao, J., Qi, Z., Zhao, F.: Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. arXiv preprint arXiv:2504.07956 (2025)
22. Qiang, C., Wei, Z., Han, X., Wang, Z., Li, S., Lan, X., Jiao, J., Han, Z.: Ver-bench: Evaluating mllms on reasoning with fine-grained visual evidence. In: Gurrin, C., Schoeffmann, K., Zhang, M., Rossetto, L., Rudinac, S., Dang-Nguyen, D., Cheng, W., Chen, P., Benois-Pineau, J. (eds.) Proceedings of the 33rd ACM International Conference on Multimedia, MM 2025, Dublin, Ireland, October 27-31, 2025. pp. 12698–12705. ACM (2025). https://doi.org/10.1145/3746027.3758208, https://doi.org/10.1145/3746027.3758208
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020
24. Shao, K., Tao, K., Zhang, K., Feng, S., Cai, M., Shang, Y., You, H., Qin, C., Sui, Y., Wang, H.: When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios (2025), https://arxiv.org/abs/2507.20198
25. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Chi, H., Guo, X., Ye, T., Zhang, Y., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18221–18232 (2024)
26. Sun, G., Yu, W., Tang, C., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Wang, Y., Zhang, C.: video-salmonn: Speech-enhanced audio-visual large language models. arXiv preprint arXiv:2406.15704 (2024)
27. Tao, K., Shao, K., Yu, B., Wang, W., Wang, H., et al.: Omnizip: Audio-guided dynamic token compression for fast omnimodal large language models. arXiv preprint arXiv:2511.14582 (2025)
28. Tong, W., Guo, H., Ran, D., Chen, J., Lu, J., Wang, K., Li, K., Zhu, X., Li, J., Li, K., et al.: Interactiveomni: A unified omni-modal model for audio-visual multi-turn dialogue. arXiv preprint arXiv:2510.13747 (2025)
29. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
30. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Gu, X., Huang, S., Xu, B., Dong, Y., et al.: Lvbench: An extreme long video understanding benchmark. arXiv preprint arXiv:2406.08035 (2024)
31. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos (2025), https://arxiv.org/abs/2405.19209

32. Wei, Z., Chen, P., Yu, X., Li, G., Jiao, J., Han, Z.: Semantic-aware SAM for point-prompted instance segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024. pp. 3585–3594. IEEE (2024). https://doi.org/10.1109/CVPR52733.2024.00344, https://doi.org/10.1109/CVPR52733.2024.00344
33. Wei, Z., Ding, W., Hao, Y., Chen, X.: Guiding the inner eye: A framework for hierarchical and flexible visual grounded reasoning. CoRR **abs/2511.22172** (2025). https://doi.org/10.48550/ARXIV.2511.22172, https://doi.org/10.48550/arXiv.2511.22172
34. Wei, Z., Qiang, C., Jiang, B., Han, X., Yu, X., Han, Z.: Adˆ2-bench: A hierarchical cot benchmark for MLLM in autonomous driving under adverse conditions. CoRR **abs/2506.09557** (2025). https://doi.org/10.48550/ARXIV.2506.09557, https://doi.org/10.48550/arXiv.2506.09557
35. Wei, Z., Wang, G., Gao, G., Hao, Y., Li, M., Ding, W., Chen, X., He, S., Yu, X.: Seeing is believing: Grounding long-video understanding in spatio-temporal visual evidence. In: Koenig, S., Jenkins, C., Taylor, M.E. (eds.) Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026. pp. 10584–10592. AAAI Press (2026). https://doi.org/10.1609/AAAI.V40I13.38031, https://doi.org/10.1609/aaai.v40i13.38031
36. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. Advances in Neural Information Processing Systems **37**, 28828–28857 (2024)
37. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: ACM Multimedia (2017)
38. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., et al.: Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215 (2025)
39. Yang, P., Wang, X., Duan, X., Chen, H., Hou, R., Jin, C., Zhu, W.: Avqa: A dataset for audio-visual question answering on videos. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3480–3491 (2022)
40. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9127–9134 (2019)
41. Zhang, Y., Wu, J., Li, W., Li, B., Ma, Z., Liu, Z., Li, C.: Video instruction tuning with synthetic data. arXiv preprint arXiv:2410.02713 (2024)
42. Zhao, Y., Xie, L., Zhang, H., Gan, G., Long, Y., Hu, Z., Hu, T., Chen, W., Li, C., Song, J., et al.: Mmvu: Measuring expert-level multi-discipline video understanding. arXiv preprint arXiv:2501.12380 (2025)
43. Zhong, Y., Xiao, J., Ji, W., Li, Y., Deng, W., Chua, T.S.: Video question answering: Datasets, algorithms and challenges. arXiv preprint arXiv:2203.01225 (2022)
44. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Duan, Y., Tian, H., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)