Putting Captions to the Test: Evaluating Video Caption Quality through Multiple-Choice Question Answering

Anonymous ACL submission

Abstract

Evaluating video captioning remains a critical challenge for Visual Large Language Models (VLLMs). Existing metrics primarily rely on matching generated text against ground-truth references. This paradigm suffers from the "one-to-many" nature of video description, where high-quality captions are often penalized for lexical mismatches or valid shifts in visual focus. Furthermore, such assessments are typically one-dimensional, failing to provide a fine-grained analysis of caption quality. To address this, we redefine caption quality via information fidelity: *A caption must maximize the coverage of salient visual information while ensuring strict factuality.* We introduce CapQuiz, a novel reference-free benchmark that assesses captions based on their utility in answering human-verified, fine-grained, multiple-choice questions derived from the video. CapQuiz features a hierarchical taxonomy of 10 question types (spanning *Descriptive* and *Inferential* categories) across 24 diverse video domains. We further formulate CapF1, a composite metric that synthesizes CapP (measuring factuality) and CapR (measuring coverage). Extensive experiments demonstrate that CapQuiz correlates significantly better with human judgments than existing metrics and offers interpretable insights into model performance. We will release the benchmark to facilitate reproducible research.

1 Introduction

Recent advancements in Visual Large Language Models (VLLMs) have significantly improved the performance of vision-language tasks. Among these, video captioning remains a fundamental challenge, requiring models to perceive temporal visual dynamics and synthesize them into coherent natural language. This capability is essential for various downstream applications, such as video retrieval, content indexing, and accessibility tools for the visually impaired. As VLLMs become more capable of generating detailed and lengthy descriptions, the need for a robust benchmark to assess their quality has become increasingly critical.

The dominant evaluation paradigm remains text-based comparison with ground-truth references, relying on n-gram overlap metrics like BLEU (Papineni et al., 2002) and CIDEr (Vedantam et al., 2015), or semantic judges like BERTScore (Zhang et al., 2019) and LLM-based evaluators (Liu et al., 2023; Fu et al., 2024). However, these methods fundamentally suffer from the "one-to-many" nature of video description, often penalizing valid captions that diverge lexically from the ground truth. To address this, cross-modal metrics have emerged to incorporate visual grounding into the evaluation loop. Approaches range from leveraging matching models and scene graphs (Jiang et al., 2019; Wang et al., 2021) to utilizing pre-trained vision-language embeddings (Hessel et al., 2021; Hu et al., 2023). Parallel to this, a distinct research strand has shifted toward question-based evaluation, such as VDC (Chai et al., 2024) and VCapsBench (Zhang et al., 2025), which probe caption fidelity through question answering. Crucially, however, neither paradigm successfully provides a fine-grained, diagnostic analysis capable of effectively decoupling factual trustworthiness from information coverage.

To address these limitations, we propose evaluating captions based on information fidelity. Fundamentally, a high-quality caption should act as an effective textual surrogate for the video, preserving the key visual details accurately. The core criterion thus becomes: *can a user correctly answer fine-grained questions about the video solely by reading the caption?* This reference-free approach encourages the model to maximize the coverage of salient information while ensuring factuality regarding the source video.

Guided by this philosophy, we introduce CapQuiz, a benchmark that quantifies caption quality through human-verified, fine-grained, multiple-

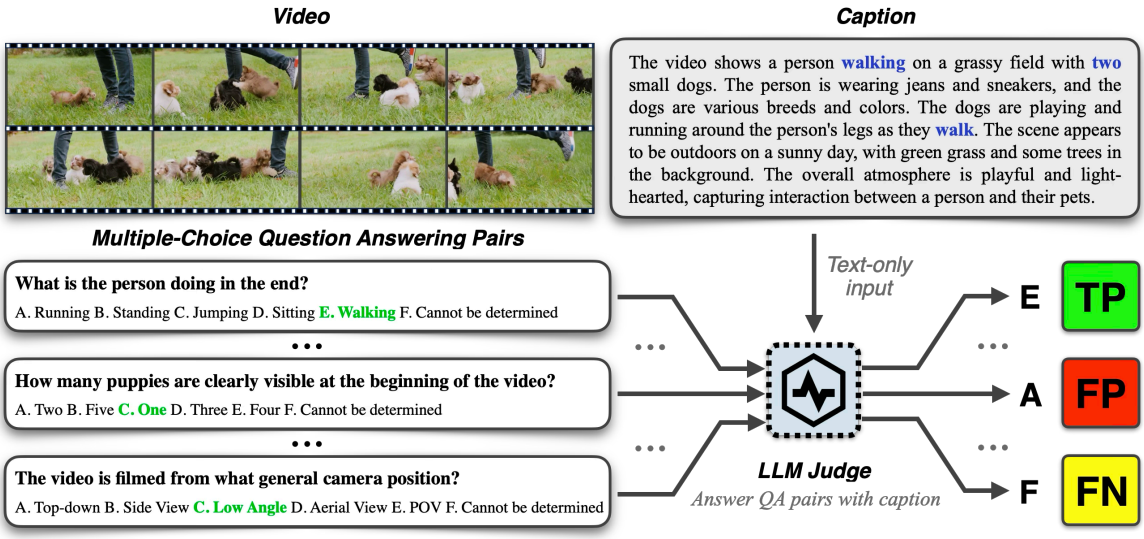


Figure 1: Overview of the CapQuiz evaluation pipeline. We assess caption quality by leveraging an LLM to answer human-verified fine-grained multiple-choice questions based solely on the generated caption. The green options denote the ground-truth answers, while blue text highlights the key evidence within the caption.

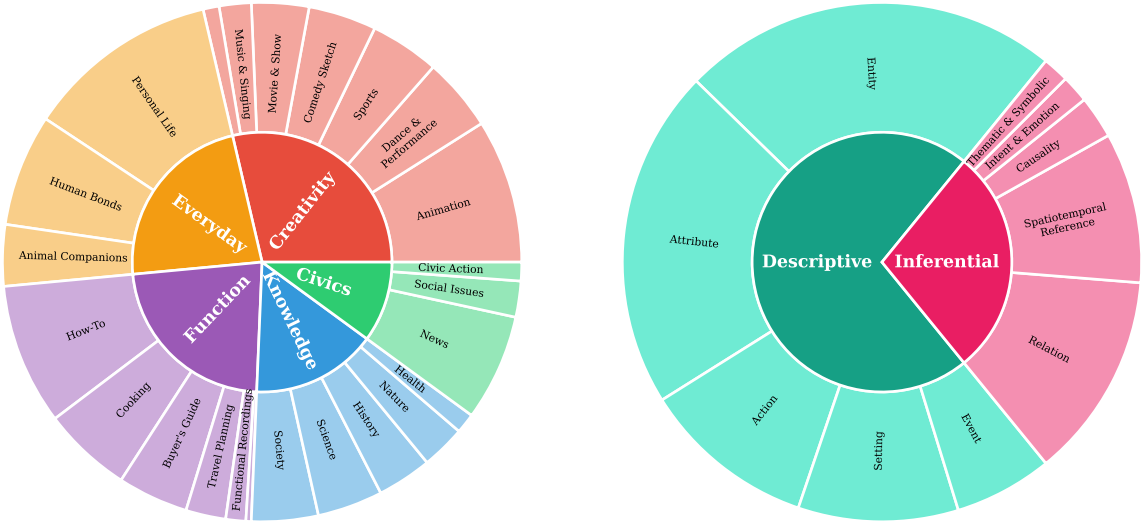
choice question answering, as shown in Figure 1. Unlike binary verification, the multiple-choice format forces the model to discriminate between correct details and plausible distractors, offering a more rigorous test of fine-grained understanding. CapQuiz is built upon a hierarchical taxonomy comprising 10 specific question types, categorized under *Descriptive* (e.g., entity, attribute) and *Inferential* (e.g., relation, causality). This structure enables a multi-faceted analysis of model capabilities across 24 diverse video domains. We categorize the QA results into True Positives (TP), False Positives (FP), and False Negatives (FN). Building upon this taxonomy, we introduce CapP and CapR to quantify factuality and coverage, respectively, and further aggregate them into a unified metric, CapF1.

Our main contributions are summarized as follows:

- We propose CapQuiz, a novel reference-free benchmark grounded in the principle of information fidelity. By utilizing human-verified, fine-grained multiple-choice questions with plausible distractors, it robustly assesses caption quality in terms of factuality (CapP), coverage (CapR), and the unified metric (CapF1).
- We release a comprehensive benchmark comprising 1,204 videos spanning 24 diverse domains. It features 23,632 human-verified multiple-choice question-answer pairs, organized into a rigorous taxonomy of 10 specific types under broader *Descriptive* and *Inferential* categories.
- Extensive experiments demonstrate that CapQuiz achieves superior alignment with human preferences. Our fine-grained analysis further reveals VLLM disparities and diagnostic insights missed by traditional one-dimensional metrics.

2 Related Works

Text-based Evaluation relies on comparing candidates against human-authored reference captions. Traditional n-gram metrics (e.g. BLEU (Papineni et al., 2002), ROUGE (Lin, 2004)) focus on surface-level lexical overlap, while METEOR (Banerjee and Lavie, 2005) incorporates synonymy via WordNet (Miller, 1995). CIDEr (Vedantam et al., 2015), specifically designed for image captioning, computes cosine similarity using TF-IDF weighting to emphasize distinctive terms. To capture structural semantics, SPICE (Anderson et al., 2016) parses captions into scene graphs composed of objects, attributes, and relationships. However, these rigid matching paradigms suffer from the "one-to-many" nature of video description, penalizing valid captions that deviate lexically from the ground truth. Recent semantic metrics like BERTScore (Zhang et al., 2019) and LLM-based judges (Liu et al., 2023; Fu et al., 2024) move beyond exact matches



(a) Video Categories Distribution

(b) Question Categories Distribution

Figure 2: Hierarchical statistics of the proposed CapQuiz.

but remain inherently reference-dependent. They are constrained by the limited coverage of ground-truth annotations and often suffer from reference bias, where high-quality captions are undervalued simply for describing valid visual details absent in the specific reference texts.

Cross-modal Grounded Evaluation mitigates reference reliance by directly incorporating visual information into the evaluation loop. Early approaches leveraged image-text matching models (Lee et al., 2018; Jiang et al., 2019) or scene graphs (Wang et al., 2020, 2021) to score fidelity. With the advent of large-scale pre-training, CLIPScore (Hessel et al., 2021) has become a de facto standard, computing the cosine similarity in a shared semantic space provided by models like CLIP (Radford et al., 2021). Similarly, InfoMetIC (Hu et al., 2023) builds on VLMs to provide both coarse-grained and token-level quality scores. Despite their popularity, these metrics typically yield a global similarity score that treats the caption as a "bag of words", often failing to distinguish fine-grained semantic nuances such as object relations or action directionality. Crucially, they lack explicit modeling of temporal dynamics and logical reasoning, making them less effective in diagnosing whether a model truly understands the complex events unfolding in a video.

Question-based Evaluation assesses caption quality via information fidelity. While early works like QACE (Lee et al., 2021) and VQAScore (Lin et al., 2024) utilize visual question answering to verify consistency, they are primarily tailored for static images. In the video domain, Dream1K (Wang et al., 2024) compares answers derived from candidates against those from references; however, this reintroduces the "one-to-many" limitation where valid but non-overlapping information is penalized. To enable reference-free evaluation, VDC (Chai et al., 2024) and VCapsBench (Zhang et al., 2025) introduce QA sets to bypass this issue. Nevertheless, such binary or open-ended formats are susceptible to random guessing or instability, lacking the plausible distractors necessary for fine-grained discrimination. Critically, most existing QA metrics predominantly verify factual correctness, largely overlooking whether the caption provides comprehensive coverage of salient events. In contrast, our work explicitly decomposes caption quality into factual trustworthiness and information coverage, treating question answerability not merely as a verification signal but as the primary evaluation objective.

3 The CapQuiz Benchmark

In this section, we introduce the methodology behind CapQuiz, a reference-free benchmark designed to evaluate video caption quality through human-verified, fine-grained multiple-choice question answering.

3.1 Hierarchical Taxonomy Design

To ensure a holistic assessment of VLLM caption capabilities, we ground CapQuiz in a rigorous two-

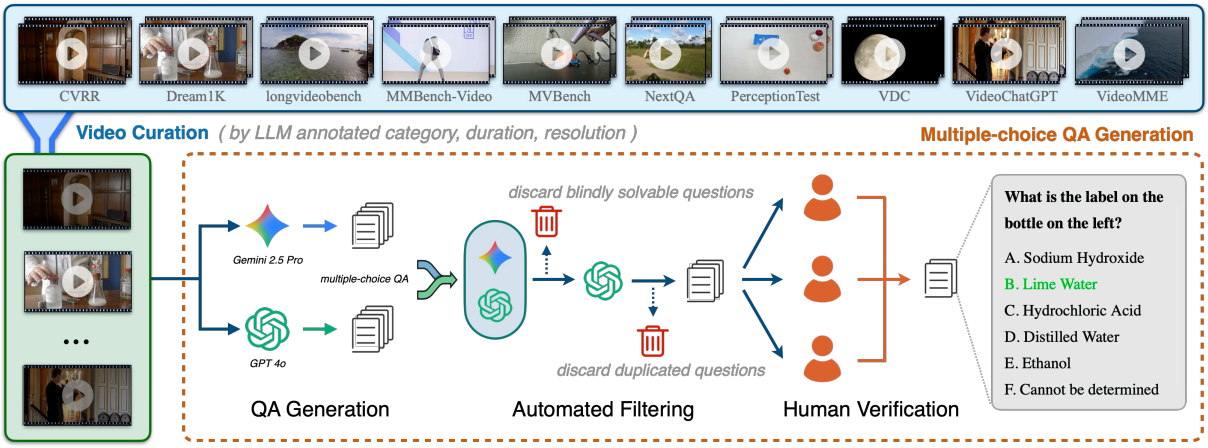


Figure 3: The construction pipeline of CapQuiz.

level taxonomy governing both visual domains and the probing question types. As shown in Figure 2(a), we curate videos across 5 super-categories (e.g., *Knowledge*, *Everyday*) branching into 24 fine-grained sub-categories. This stratification maximizes semantic diversity, ranging from dynamic events in *Sports* to information-dense scenes in *News*, ensuring models are tested against distinct visual distributions and temporal dynamics. Complementing this visual breadth, our question taxonomy (Figure 2(b)) probes information fidelity across two cognitive dimensions: *Descriptive* that focusing on visual grounding tasks like *Entity* and *Action*, and *Inferential* that targeting higher-order logic such as *Relation* and *Causality*. By organizing 10 specific question types under these categories, we effectively decouple basic recognition from complex interpretation, enabling fine-grained diagnostic analysis. Detailed definitions of both taxonomies are provided in Appendix A.

3.2 Benchmark Construction

Figure 3 shows the construction pipeline of the benchmark.

3.2.1 Video Curation

Preventing data contamination was a paramount priority in our construction process. To minimize the risk of training set leakage, we strictly sourced candidate videos from the test or held-out validation splits of 10 public benchmarks, including VideoMME (Fu et al., 2025), VideoChatGPT (Maaz et al., 2024), NextQA (Xiao et al., 2021), MVBench (Li et al., 2024), MMBench-Video (Fang et al., 2024), CVRR (Khattak et al., 2025), PerceptionTest (Patraucean et al., 2023), longvideobench (Wu et al., 2024), VDC (Chai et al., 2024) and Dream1K (Wang et al., 2024).

We extracted essential metadata (i.e., resolution, duration) and employed `Gemini-2.5-Pro` (Comanici et al., 2025) to annotate video categories, which guided the subsequent selection process. We adopted a duration ratio of approximately *6:3:1* for short (0–30s), medium (30–60s), and long (60–120s) videos. While ensuring ample coverage of short-form clips, this distribution also incorporates narratively rich content through longer videos.

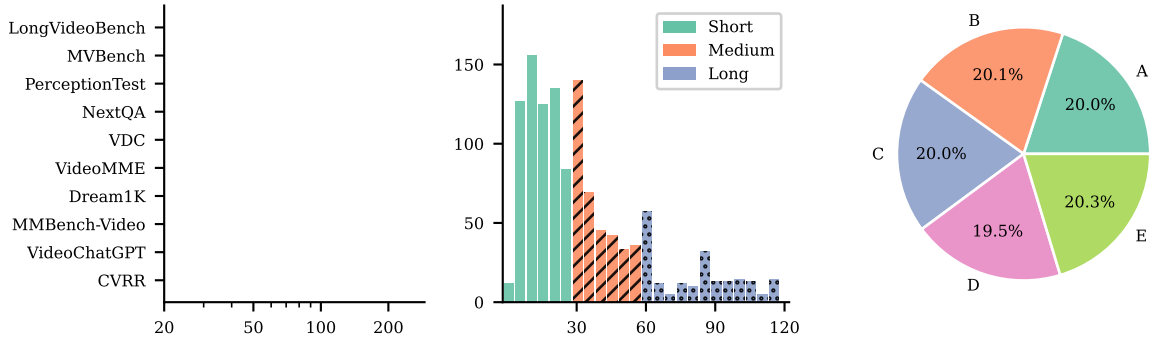
3.2.2 Multiple-Choice QA Generation

We implemented a rigorous *Over-generate then Filter* pipeline to construct the multiple-choice question-answer set. In the generation phase, utilizing `Gemini-2.5-Pro` and `GPT-4o` (Hurst et al., 2024), we produced diverse question-answer pairs rooted in our hierarchical taxonomy. Specifically, we prompted the models with the raw video, the specific definition of the target question type, and at least five few-shot reference examples. The models were instructed to generate a list of candidate QA pairs, where each pair comprises a distinct question body and five answer options.

Subsequently, these candidates underwent a strictly controlled filtration phase, beginning with an automated stage designed to ensure validity and information density. We first addressed language bias via a *Blind Solvability Check*. In this step, QA pairs were fed to LLMs (i.e., `Gemini-2.5-Pro` and `GPT-4.1`) in a video-blind setting, with option orders shuffled across three independent trials for each LLM. A question was deemed *Blindly Solvable* if both models answered it correctly in at least two out of the three trials, suggesting the answer could be inferred solely from textual patterns without visual context. Following this, we performed semantic de-duplication to eliminate redundancy. We utilized `GPT-4.1` to analyze semantic similarity

Benchmark	Reference-free	Human Verified	QA Format	# Videos	Avg. Duration (s)	Avg. Q/V
MSVD (2011)	✗	-	-	1,970	9.65	-
MSR-VTT (2016)	✗	-	-	10,000	15.01	-
ActivityNet Captions (2017)	✗	-	-	9,802	118.21	-
VATEX (2019)	✗	-	-	4,478	144.78	-
Dream1K (2024)	✗	-	-	1,000	8.87	-
VDC (2024)	✓	✗	Open-Ended	1,027	28.18	94.35
VCapsBench (2025)	✓	✓	Yes/No	5,677	9.79	18.38
CapQuiz	✓	✓	Multiple-Choice	1,204	34.69	19.63

Table 1: **Comparison of CapQuiz with existing video captioning benchmarks.** Our benchmark provides human-verified fine-grained multiple-choice question answering pairs designed for caption information fidelity evaluation. Avg. Q/V indicates Average number of questions per video.



(a) Video Sources (b) Video Durations (c) Golden Options

Figure 4: Detailed statistics of the proposed CapQuiz.

and cluster related questions. Within each cluster, we retained only the single most challenging instance, quantified as the one yielding the lowest accuracy during the blind pass, thereby maximizing the discriminative power of the dataset.

Finally, the surviving candidates advanced to the human stage for rigorous verification. Each QA pair was evaluated by at least three annotators based on three strict criteria:

- **Visual Relevance:** A question is considered valid only if it is strictly grounded in the video content.
- **Factual Correctness:** Each question must have exactly one correct answer that is objectively verifiable through visual evidence in the video.
- **Question difficulty:** To ensure the benchmark evaluates fine-grained visual understanding, we require that at least one incorrect option be a "hard negative." A hard negative is an option that appears plausible but is factually incorrect, requiring careful inspection of the video details to rule out.

To validate the reliability of our human review process, we calculated the inter-annotator agreement on a subset of the data, achieving a Gwet's AC1 (Gwet, 2001) of 0.92, which indicates the high consistency and quality of our ground-truth annotations.

The related prompts are listed in Appendix B.

3.3 Dataset Statistics

As shown in Table 1, CapQuiz contains 1,204 videos, with an average of 19.63 QA pairs per video. Unlike reference-based benchmarks (e.g., MSR-VTT), CapQuiz adopts a reference-free paradigm to circumvent the "one-to-many" constraints of text matching, thereby directly assessing information fidelity. Its human-verified multiple-choice format ensures deterministic and reliable evaluation, superior to the reference-free attempts limited by unverified generation (VDC) or binary tasks (VCapsBench). Furthermore, featuring a hierarchical taxonomy of 24 video domains and 10 question types, CapQuiz enables a more nuanced diagnosis of model capabilities than prior coarse-grained datasets.

As visually detailed in Figure 2, our benchmark is structured around a sophisticated hierarchical video and question taxonomy. This design meticulously balances visual richness with linguistic complexity, ensuring that the benchmark covers a wide spectrum of semantic granularity, from coarse-grained object recognition to fine-grained reason-

ing. Furthermore, the source distribution presented in Figure 4(a) demonstrates that we aggregate video data from a highly heterogeneous array of sources, while minimizing the risk of training set leakage. This strategy is intended to maximize visual diversity and domain coverage, thereby testing the generalization ability of models across different visual styles. In terms of video duration (Figure 4(b)), while the benchmark is primarily anchored in short-form videos (< 30s) to capture atomic events, we deliberately maintain a substantial proportion of medium- and long-form content. This diverse duration distribution serves as a rigorous test for models' temporal reasoning capabilities and their ability to model long-range dependencies. Finally, to ensure fair evaluation, the answer options are strictly uniformly distributed as shown in Figure 4(c). This balance is critical for mitigating potential position bias and prevents models from bypassing genuine understanding by exploiting statistical shortcuts or spurious correlations.

3.4 Evaluation Methodology

To evaluate the quality of video caption C, we define a question set $\mathcal{Q}$ derived from the video. Each question comprises a stem q_i and 5 options $\mathcal{O}_i = \{o_{i,1}, o_{i,2}, \ldots, o_{i,5}\}$, where $o_{i,gt\in[1,\ldots,5]}$ is the correct option. We introduce an extra *"Cannot be determined"* option o_{unk} to form the evaluation space $\mathcal{O}'_i = \mathcal{O}_i \cup \{o_{unk}\}$, allowing the `judge` to explicitly signal uncertainty due to missing information. Given the caption C and question q_i, the `judge` selects the most likely option $\hat{o}_i$ as the answer from options $\mathcal{O}'_i$. The outcomes are categorized as:

- **True Positive (TP)** ($\hat{o}_i = o_{i,gt}$): The caption contains the correct visual information, enabling the `judge` to select the ground-truth option.
- **False Negative (FN)** ($\hat{o}_i = o_{unk}$): The caption lacks the necessary information to answer the question. This results in an *omission*, forcing the `judge` to choose "Cannot be determined."
- **False Positive (FP)** ($\hat{o}_i \neq o_{i,gt} \wedge \hat{o}_i \neq o_{unk}$): The caption contains incorrect or misleading details consistent with a wrong option. This reflects *hallucination*, leading the `judge` to a specific incorrect answer.

Method	Spearman (ρ)			Kendall (τ)		
	Factuality	Coverage	Overall	Factuality	Coverage	Overall
VLLM-as-a-Judge	0.556	0.579	0.522	0.500	0.515	0.463
Ours	**0.690**	**0.663**	**0.643**	**0.581**	**0.551**	**0.533**

Table 2: **Correlation analysis between automated metrics and human judgments.** Ours CapQuiz achieves the highest correlation across all dimensions. All correlations are significant ($p < 0.01$).

Based on the categorization outcomes, we propose three metrics to quantify caption quality:

- **Factuality (CapP).** This metric measures the precision of determinate answers, penalizing hallucinations (FP) while disregarding uncertainty (FN).

$$CapP = \frac{TP}{TP + FP}$$

- **Coverage (CapR).** This metric evaluates the completeness of the caption by measuring the proportion of correctly retrieved details against the total number of questions N.

$$CapR = \frac{TP}{N}$$

- **Overall (CapF1).** To provide a holistic assessment, we compute the harmonic mean of factuality and coverage:

$$CapF1 = \frac{2 \cdot CapP \cdot CapR}{CapP + CapR}$$

To ensure reproducibility and minimize prior knowledge bias, we employ `GPT-4.1` as the judge in our experiments, instructing it to strictly ground its answers in the provided caption C.

4 Experiments

4.1 Alignment with Human Judgments

To validate the effectiveness of CapQuiz, we assess the alignment between automated metrics and human judgments using Spearman (ρ) (Spearman, 1961) and Kendall (τ) (Kendall, 1948) rank coefficients. We randomly sampled 200 videos from the benchmark, where the captions are generated by three different models (i.e. `GPT-4o`, `InternVL3.5-8B` and `QWen3VL-8B`), yielding a total of 600 video-caption pairs. Human experts and the VLLM-as-a-Judge rated the generated captions on a Likert scale of 1-5 across *Factuality*, *Coverage*, and *Overall Quality*. For the VLLM-as-a-Judge,

Model	Overall			Descriptive			Inferential		
	F1	P	R	F1	P	R	F1	P	R
Proprietary Models									
Gemini-2.5-flash (2025)	76.79	83.49	71.08	78.68	85.92	72.57	72.01	<u>77.45</u>	67.28
Gemini-2.5-pro (2025)	<u>78.42</u>	<u>84.56</u>	73.10	<u>80.79</u>	<u>87.50</u>	75.04	<u>72.44</u>	77.28	<u>68.17</u>
GPT-4o-2024-11-20 (2024)	72.07	78.46	66.65	74.98	82.21	68.92	64.85	69.38	60.88
GPT-4.1-2025-04-14 (2025)	76.54	81.53	72.12	79.34	84.95	74.42	69.53	73.13	66.28
GPT-5.2-2025-12-11 (2025)	**83.47**	**86.45**	**80.68**	**85.77**	**88.98**	**82.79**	**77.78**	**80.24**	**75.46**
Open-source Models									
AuroraCap-7B (2024)	39.86	63.00	29.16	43.86	68.40	32.28	29.48	48.26	21.22
LLaVA-Video-7B (2024)	63.16	73.61	55.30	66.83	78.02	58.45	53.85	62.51	47.30
Tarsier2-7b (2025)	42.37	55.80	34.16	43.55	57.91	34.90	39.44	50.70	32.27
InternVL3.5-8B (2025)	45.69	67.57	34.51	47.67	72.46	35.52	40.89	56.75	31.95
InternVL3.5-30B-A3B (2025)	54.46	70.03	44.55	57.92	75.03	47.16	45.80	57.83	37.92
Qwen3-VL Series (2025)									
Qwen3-VL-2B	67.20	74.81	60.99	70.81	79.28	63.97	58.16	63.85	53.40
Qwen3-VL-4B	71.47	77.75	66.13	74.53	81.43	68.70	63.80	68.66	59.58
Qwen3-VL-8B	72.39	78.41	67.22	75.44	82.27	69.65	64.79	69.01	61.06
Qwen3-VL-32B	77.08	81.01	<u>73.51</u>	79.90	84.22	<u>76.00</u>	69.97	73.01	67.18
Qwen3-VL-30B-A3B	73.31	79.25	68.20	76.24	82.89	70.58	65.99	70.32	62.17
Qwen3-VL-235B-A22B	75.60	80.68	71.12	78.33	84.03	73.36	68.75	72.44	65.41

Table 3: **Main results on CapQuiz**. We report performance across the Overall metric and its two question categories: Descriptive and Inferential. P, R, and F1 correspond to CapP, CapR, and CapF1, respectively. The best performance is marked in **bold** and the second best is <u>underlined</u>.

we prompted `Gemini-2.5-pro` to judge based on the video content (prompt detailed in Appendix B). For CapQuiz, we map the metric components to the evaluation dimensions: CapP evaluates *Factuality*, CapR measures *Coverage*, and CapF1 represents *Overall Quality*. As presented in Table 2, our method demonstrates a stronger correlation with human judgments compared to the VLLM-as-a-Judge. These results underscore that decomposing evaluation into fine-grained QA tasks yields a more reliable and interpretable assessment than direct scoring.

4.2 Evaluation on SOTA Models

We evaluated several popular proprietary and open-source models. To ensure a fair comparison, all models were queried with a standardized prompt: "`Describe this video in detail`", , with visual inputs uniformly sampled at 32 frames per video. The quantitative results are summarized in Table 3, revealing three key observations.

Model Capabilities and Scaling Trend. CapQuiz effectively differentiates model tiers. Generally, proprietary models outperform open-source counterparts, with GPT-5.2 achieving the highest Overall CapF1 score of 83.47. Within the open-source landscape, we observe a distinct scaling trend in the Qwen3-VL series. As activated model size increases from 2B to 32B, performance improves generally (e.g., Overall CapF1 rises from 67.20 to 77.08), underscoring that increased parameter count correlates strongly with video caption quality.

The Trade-off between Factuality and Coverage. A pervasive trend across all models is that CapP (*Factuality*) consistently exceeds CapR (*Coverage*). For instance, GPT-4o achieves a high factuality of 78.46 but a significantly lower coverage of 66.65. This suggests that current VLLMs exhibit a conservative generation strategy: *they tend to generate trustworthy descriptions but often fail to exhaustively cover the salient visual details*. This indicates room for improvement in increasing caption density without introducing hallucinations.

The Reasoning Gap. Performance on Inferential questions consistently lags behind Descriptive ones, validating the hierarchical difficulty of our taxonomy. Crucially, this performance gap widens significantly for smaller models. While top-tier models like GPT-5.2 see a moderate degradation ($\sim$10%) when transitioning from Descriptive to Inferential tasks, smaller models like AuroraCap-7B suffer a significant drop of $\sim$30%. This indicates that while visual recognition is becoming a baseline capability, complex visual reasoning remains the primary differentiator for superior VLLMs.

4.3 Prompt Sensitivity Analysis

It is widely recognized that VLLM performance can be heavily influenced by prompt design. To

Model	Overall			Descriptive			Inferential		
	F1	P	R	F1	P	R	F1	P	R
Proprietary Models									
Gemini-2.5-flash (2025)	76.69 −0.13%	83.28 −0.25%	71.06 −0.03%	78.65 −0.04%	85.78 −0.16%	72.61 +0.06%	71.73 −0.39%	77.06 −0.50%	67.09 −0.28%
Gemini-2.5-pro (2025)	78.59 +0.22%	84.69 +0.15%	73.31 +0.29%	80.89 +0.12%	87.47 −0.03%	75.23 +0.25%	72.80 +0.50%	77.77 +0.63%	68.43 +0.38%
GPT-4o-2024-11-20 (2024)	74.72 +3.68%	78.20 −0.33%	71.54 +7.34%	78.65 +4.89%	82.47 +0.32%	75.18 +9.08%	64.79 −0.09%	67.50 −2.71%	62.29 +2.32%
GPT-4.1-2025-04-14 (2025)	79.16 +3.42%	80.96 −0.70%	77.43 +7.36%	83.23 +4.90%	85.13 +0.21%	81.40 +9.38%	68.82 −1.02%	70.37 −3.77%	67.34 +1.60%
GPT-5.2-2025-12-11 (2025)	85.26 +2.14%	87.15 +0.81%	83.45 +3.43%	87.57 +2.10%	89.47 +0.55%	85.75 +3.58%	79.36 +2.03%	81.23 +1.23%	77.58 +2.81%
Open-source Models									
AuroraCap-7B (2024)	27.00 −32.26%	57.24 −9.14%	17.67 −39.40%	29.31 −33.17%	62.35 −8.85%	19.16 −40.64%	21.15 −28.26%	44.46 −7.87%	13.88 −34.59%
LLaVA-Video-7B (2024)	60.63 −4.01%	69.56 −5.50%	53.73 −2.84%	64.66 −3.25%	74.65 −4.32%	57.03 −2.43%	50.54 −6.15%	57.11 −8.64%	45.32 −4.19%
Tarsier2-7b (2025)	9.95 −76.52%	11.92 −78.64%	8.54 −75.00%	10.20 −76.58%	12.21 −78.92%	8.75 −74.93%	9.32 −76.37%	11.16 −77.99%	8.00 −75.21%
InternVL3.5-8B (2025)	64.03 +40.14%	71.89 +6.39%	57.72 +67.26%	67.53 +41.66%	76.05 +4.95%	60.73 +70.97%	55.20 +35.00%	61.51 +8.39%	50.06 +56.68%
InternVL3.5-30B-A3B (2025)	64.97 +19.30%	73.61 +5.11%	58.14 +30.51%	68.37 +18.04%	77.94 +3.88%	60.88 +29.09%	56.47 +23.30%	63.02 +8.97%	51.15 +34.89%
Qwen3-VL Series (2025)									
Qwen3-VL-2B	68.75 +2.31%	73.40 −1.88%	64.66 +6.02%	73.06 +3.18%	78.08 −1.51%	68.65 +7.32%	57.84 −0.55%	61.58 −3.56%	54.53 +2.12%
Qwen3-VL-4B	76.50 +7.04%	80.13 +3.06%	73.18 +10.66%	80.12 +7.50%	83.91 +3.05%	76.65 +11.57%	67.15 +5.25%	70.37 +2.49%	64.21 +7.77%
Qwen3-VL-8B	76.68 +5.93%	80.60 +2.79%	73.12 +8.78%	79.93 +5.95%	84.09 +2.21%	76.17 +9.36%	68.43 +5.62%	71.79 +4.03%	65.38 +7.08%
Qwen3-VL-32B	78.74 +2.15%	81.77 +0.94%	75.92 +3.28%	81.73 +2.29%	84.83 +0.72%	78.84 +3.74%	71.11 +1.63%	73.96 +1.30%	68.48 +1.94%
Qwen3-VL-30B-A3B	78.65 +7.28%	82.03 +3.51%	75.53 +10.75%	81.71 +7.17%	85.20 +2.79%	78.49 +11.21%	70.86 +7.38%	73.96 +5.18%	68.01 +9.39%
Qwen3-VL-235B-A22B	79.01 +4.51%	82.38 +2.11%	75.89 +6.71%	82.16 +4.89%	85.67 +1.95%	78.93 +7.59%	70.97 +3.23%	74.02 +2.18%	68.17 +4.22%

Table 4: **Evaluation results under the detailed caption prompt setting.** The colored subscripts indicate the relative performance change compared to the baseline standardized prompt results reported in Table 3. Green and red denote performance improvement and decline, respectively.

validate our findings, we experimented with a more complex and detailed caption prompt (see Appendix B) in this section, comparing it against the concise standardized prompt from our main experiments. As shown in Table 4, these results reveals divergent sensitivities across models. Proprietary models demonstrate remarkable stability against prompt variations. For instance, `GPT-4o` exhibits a maximum change in Overall CapF1 of only $+3.68\%$. Similarly, the `Qwen3-VL` series shows consistent performance gains with the detailed prompt. In contrast, other open-source models display high sensitivity to prompt changes. For example, `InternVL3.5-8B` shows a substantial improvement of 40.14% in Overall CapF1, whereas `Tarsier2-7B` suffers a drastic decline of 76.52%.

Crucially, however, the main conclusions of our benchmark remain robust against these variations. The key trends from the previous section (i.e., *Model Capabilities and Scaling Trend*, *The Trade-off between Factuality and Coverage*, and *The Reasoning Gap*) persist regardless of the prompt complexity, demonstrating that CapQuiz effectively captures fundamental model capabilities independent of prompting strategies.

5 Conclusion

In this paper, we introduce CapQuiz, a novel reference-free benchmark designed to assess video captioning quality via information fidelity. By leveraging human-verified, fine-grained multiple-choice questions, it decouples and quantifies caption factuality and coverage. Experiments demonstrate that CapQuiz achieves robust alignment with human judgments compared to existing evaluators. Moreover, our analysis exposes a critical limitation in current SOTA VLLMs: *While exhibiting high factual precision, they often struggle with comprehensive coverage and show significant degradation on inferential tasks compared to descriptive ones.* We envision CapQuiz as a vital testbed to guide future research toward more robust and grounded video understanding models.

Limitations

First, the question set of CapQuiz may not be exhaustive. Although we generate an average of 19.63 QA pairs per video to capture a wide range of visual information, guaranteeing the complete coverage of every visual detail in a complex video remains practically challenging. Consequently, our coverage metric (CapR) serves as a proxy based on identified salient information rather than an absolute measure of total visual content. In cases where videos contain extremely dense or subtle background details not captured by our QA generation pipeline, the reported coverage scores might be slightly overestimated. Moreover, the current QA pairs are exclusively in English, which limits the evaluation of VLLMs in multilingual contexts.

Furthermore, to maintain a scalable and reference-free evaluation, we utilize `GPT-4.1` as the judge. While exhibiting high alignment with human experts, the evaluation is inherently constrained by the judge model's upper bound and susceptible to the closed-source nature of the API, where model updates may affect reproducibility. Additionally, potential biases in the judge model regarding ambiguous visual descriptions could introduce noise.

Ethical Considerations

We prioritize ethical concerns associated with video content in benchmark construction, particularly regarding privacy and safety. To mitigate these risks, we curated video samples exclusively from established open-source datasets distributed under Creative Commons or compatible licenses, ensuring compliance with their usage policies. Regarding the text annotation, while commercial APIs (e.g., GPT and Gemini) implement built-in safety guardrails, we acknowledge the residual risk of introducing model-inherent biases or toxicity. Furthermore, we conducted a rigorous manual inspection to filter out any content containing potential Not Safe For Work (NSFW) elements or sensitive Personally Identifiable Information (PII). We ensured that all human annotators involved in this verification process were compensated at a rate exceeding the local minimum wage, adhering to fair labor practices.

Acknowledgments

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. 2024. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.

Kilem Gwet. 2001. Handbook of inter-rater reliability: How to estimate the level of agreement between two or multiple raters. *Gaithersburg, MD: STATAXIS Publishing Company*.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

Anwen Hu, Shizhe Chen, Liang Zhang, and Qin Jin. 2023. Infometic: An informative metric for reference-free image caption evaluation. *arXiv preprint arXiv:2305.06002*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ming Jiang, Qiuyuan Huang, Lei Zhang, Xin Wang, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. 2019. Tiger: Text-to-image grounding for image caption evaluation. *arXiv preprint arXiv:1909.02050*.

Maurice George Kendall. 1948. Rank correlation methods.

Muhammad Uzair Khattak, Muhammad Ferjad Naeem, Jameel Hassan, Muzammal Naseer, Federico Tombari, Fahad Shahbaz Khan, and Salman Khan. 2025. How good is my video-lmm? complex video reasoning and robustness evaluation suite for video-lmms. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3642–3651.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Hwanhee Lee, Thomas Scialom, Seunghyun Yoon, Franck Dernoncourt, and Kyomin Jung. 2021. Qace: Asking questions to evaluate an image caption. *arXiv preprint arXiv:2108.12560*.

Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2024. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. 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)*, pages 12585–12602.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

OpenAI. 2025. Gpt-5.1: A smarter, more conversational chatgpt. https://openai.com/index/gpt-5-1/.

OpenAI. 2025. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2025-04-14.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Charles Spearman. 1961. " general intelligence" objectively determined and measured.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. 2024. Tarsier: Recipes for training and evaluating large video description models. *arXiv preprint arXiv:2407.00634*.

Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1508–1517.

Sijin Wang, Ziwei Yao, Ruiping Wang, Zhongqin Wu, and Xilin Chen. 2021. Faier: Fidelity and adequacy ensured image caption evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14050–14059.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4581–4591.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. 2025. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding. *Preprint*, arXiv:2501.07888.

Shi-Xue Zhang, Hongfa Wang, Duojun Huang, Xin Li, Xiaobin Zhu, and Xu-Cheng Yin. 2025. Vcaps-bench: A large-scale fine-grained benchmark for video caption quality evaluation. *arXiv preprint arXiv:2505.23484*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024. Video instruction tuning with synthetic data. *Preprint*, arXiv:2410.02713.

A Taxonomy

A.1 Video Taxonomy

Knowledge Content that systematically presents knowledge about the natural world, human societies, historical developments, and scientific or technological principles. The primary purpose is to inform, explain, and deepen understanding through structured, evidence-based narratives.

- **Nature**: Content about the natural world on Earth, including ecosystems, wildlife, environmental processes, and conservation efforts. Focuses on non-human-driven phenomena and the interdependence of living organisms and their habitats.

- **Science**: Systematic knowledge of the physical and technological world, including fundamental principles in physics, chemistry, biology, astronomy, and engineering. Covers how things work, from subatomic particles to space exploration, and the development of technologies such as AI and robotics.

- **Health**: Knowledge about the human body, medical science, disease prevention, mental well-being, and public health. Emphasizes evidence-based understanding of health conditions, treatments, and lifestyle impacts on physical and psychological wellness.

- **History**: Documented understanding of past human events, civilizations, conflicts, discoveries, and cultural developments. Based on historical records, archaeological findings, and scholarly analysis of how societies have evolved over time.

- **Society**: Insights into human social structures, behaviors, institutions, and collective thought. Includes economics, psychology, education, ethics, philosophy, and the study of how individuals and groups interact within cultural and organizational contexts.

Everyday Authentic recordings of ordinary life that capture personal experiences, relationships with people and animals, and moments of solitude. Focuses on unscripted, non-performance-based content that reflects how individuals live, connect, and exist in their daily environments—whether alone, with others, or alongside companion animals.

- **Human Bonds**: Authentic moments of connection and coexistence with family, friends, partners, or acquaintances, emphasizing emotional intimacy, shared experiences, and the warmth of human relationships.

- **Animal Companions**: Daily life and emotional bonding between humans and their animal companions, highlighting care, spontaneity, and the unique non-verbal intimacy shared across species.

- **Personal Life**: Recordings of an individual's daily existence in solitude, encompassing routines, habits, reflections, emotions, domestic activities, and atmospheric moments. Focuses on how a person experiences, manages, and expresses their life without interaction with people or pets. This includes personal journeys, functional tasks, and contemplative states, all centered on the self as the sole subject.

Creativity Content that expresses imagination, emotion, or aesthetic vision through artistic performance, storytelling, or physical excellence. Includes movies, music, dance, animation, comedy, and sports events. The primary intent is to be seen, heard, or experienced as a form of personal or collaborative expression—not for instruction, commerce, or information alone.

- **Movie & Show**: Fictional or dramatic videos that tell a story, including movies, TV series, web dramas, and short films. Typically feature actors, scripts, and narrative structure.

- **Dance & Performance**: Choreographed or expressive performances centered on movement, including original dance routines, dance covers, stage shows, spoken word poetry, and artistic recitations. Emphasizes physical expression, rhythm, and emotional delivery.

- **Music & Singing**: Original or performed musical works, including official music videos, song releases, vocal covers, instrumental performances, and creative audio-visual compositions. Focuses on auditory artistry and musical expression.

- **Animation**: Animated works created through 2D, 3D, stop-motion, or digital techniques, including short films, creative explainers, and experimental visual stories. Emphasizes visual imagination and motion design.

- **Comedy Sketch**: Short, scripted humorous videos designed to entertain, including parodies, satirical scenes, original comedy skits, and creative spoofs. Often feature exaggerated characters and comedic timing.

- **Game**: Creative content made within or about games, such as custom maps, mods, character designs, in-game art projects, or narrative-driven gameplay. Emphasizes originality, design, and virtual world-building.

- **Sports**: Content centered on athletic competitions and physical performance, including live events, highlights, athlete stories, news, and expert analysis. Emphasizes the drama, skill, and emotional intensity of sports as a form of visual and emotional entertainment.

Civics Coverage of real-world public events, societal issues, political developments, and collective experiences that impact communities or nations. Focuses on factual reporting, public discourse, and awareness of civic life.

- **News**: Reporting on recent, impactful public events such as natural disasters, accidents, conflicts, or major societal incidents. Focuses on what happened, where, and when, with emphasis on timeliness and factual accuracy.

- **Social Issues**: Coverage of ongoing societal challenges and public debates, such as education inequality, mental health awareness, housing affordability, gender rights, racial justice, and environmental policy. Focuses on current events, stakeholder perspectives, and civic discourse.

- **Civic Action**: Recordings of collective efforts to address social or environmental issues, such as protests, volunteer work, humanitarian aid, and community organizing. Highlights public participation and social change.

Function Content designed to help users accomplish a practical goal, such as learning how to cook a meal, perform a task, make a purchase decision, plan a trip, organize daily life, or review a recording for reference. The primary intent is utility—providing actionable guidance, decision support, or functional documentation—rather than entertainment, knowledge explanation, or personal expression.

- **How-To**: Step-by-step instructions for completing practical tasks in daily life, work, or learning—excluding cooking—such as repairing a device, using software, organizing space, crafting, or performing a physical skill. Covers both short-term actions and repeatable routines, with a focus on actionable guidance and immediate application.
- **Cooking**: Step-by-step instructions for preparing meals, dishes, or beverages, including recipe demonstrations, cooking techniques, meal prep, and kitchen tips. Focuses on food creation, flavor development, and practical kitchen skills.
- **Buyer's Guide**: Content that helps viewers decide what to buy, including product reviews, comparisons, recommendations, unboxing, and live commerce. Emphasizes real-world usage, value assessment, and decision support.
- **Travel Planning**: Guides for designing a trip, including itinerary creation, budgeting, transportation, accommodation, and visa planning. Helps viewers prepare for travel with practical, organized advice.
- **Life Guide**: Guides that help viewers design sustainable, personalized life systems, such as minimalism, daily routines, or personal workflows. Focuses on the philosophy, structure, and long-term optimization of everyday living—beyond step-by-step instructions.
- **Functional Recordings**: Videos recorded for practical purposes, such as screen recordings, surveillance, dashcams, meeting logs, or training replays. Not intended for entertainment or artistic expression.

A.2 Question Taxonomy

Descriptive Focus on factual information that is directly observable in the video content.

- **Entity**: Pertains to the identification of specific people, animals, objects, texts, or symbols present in the scene.
- **Attribute**: Involves visual properties (e.g., color, shape), quantity counts, or the physical states (e.g., open/closed) of the entities.
- **Action**: Captures physical movements, simple behaviors performed by a single entity, or direct physical interactions between entities.
- **Event**: Summarizes the composite activity, situation, or overarching happening depicted throughout the video clip.
- **Setting**: Describes the environmental context, background scenery, or visual cues indicating location and time of day.

Inferential Focus on abstract information derived from visual cues through logical reasoning or contextual understanding.

- **Relation**: Captures the spatial configurations, temporal ordering, or comparative relationships between entities or actions.
- **Causality**: Explains the logic behind events, including causes (why?), immediate effects (results), or potential counterfactuals.
- **Intent & Emotion**: Implies the agents' underlying goals, motivations, emotional states, or the psychological purpose behind their actions.
- **Thematic & Symbolic**: Involves high-level understanding of the plot, abstract themes, or symbolic meanings embedded in the content.
- **Spatiotemporal Reference**: Includes specific deictic cues or references (e.g., "on the left", "at the beginning") that ground the text to precise spatial regions or temporal segments.

B Prompts

- Video Taxonomy Prompt, see Figure 5
- QA Generation Prompt, see Figure 6
- Blind Solvability Check Prompt, see Figure 7
- Question Deduplication Prompt, see Figure 7
- Multiple-Choice QA Prompt, see Figure 8
- VLLM-as-a-Judge Prompt, see Figure 8
- Detailed Caption Prompt, see Figure 9

C Annotation Details

We recruited annotators through our internal data annotation platform. Our annotation workforce exhibits a diverse international background, roughly evenly split into four groups. India, Singapore, and China each contribute approximately 25% of the personnel, while the final cohort represents a mix of

European and North American countries, including the United States, Spain, and Ireland. To ensure high-quality text generation and comprehension, we enforced a strict prerequisite of proficiency in written English. The annotation interface used by the workers is illustrated in Figure 10.

Video Taxonomy Prompt

You are a video understanding expert. Analyze the following sequence of video frames in chronological order and classify the video into up to 3 most likely subcategories from the predefined taxonomy. For each predicted subcategory, provide a confidence probability (0.0–1.0). If the video cannot be confidently classified into any defined category, include "Others" with an appropriate probability. Output only a JSON dictionary in the format: {"subcategory1": probability1, "subcategory2": probability2, ...}. Probabilities do not need to sum to 1.0. Do not include any explanation or formatting.

Taxonomy Structure:
Knowledge
- **Nature**: Content about the natural world on Earth, including ecosystems, wildlife, environmental processes, and conservation efforts. Focuses on non-human-driven phenomena and the interdependence of living organisms and their habitats.
- **Science**: Systematic knowledge of the physical and technological world, including fundamental principles in physics, chemistry, biology, astronomy, and engineering. Covers how things work, from subatomic particles to space exploration, and the development of technologies such as AI and robotics.
- **Health**: Knowledge about the human body, medical science, disease prevention, mental well-being, and public health. Emphasizes evidence-based understanding of health conditions, treatments, and lifestyle impacts on physical and psychological wellness.
- **History**: Documented understanding of past human events, civilizations, conflicts, discoveries, and cultural developments. Based on historical records, archaeological findings, and scholarly analysis of how societies have evolved over time.
- **Society**: Insights into human social structures, behaviors, institutions, and collective thought. Includes economics, psychology, education, ethics, philosophy, and the study of how individuals and groups interact within cultural and organizational contexts.
Everyday
- **Human Bonds**: Authentic moments of connection and coexistence with family, friends, partners, or acquaintances, emphasizing emotional intimacy, shared experiences, and the warmth of human relationships.
- **Animal Companions**: Daily life and emotional bonding between humans and their animal companions, highlighting care, spontaneity, and the unique non-verbal intimacy shared across species.
- **Personal Life**: Recordings of an individual's daily existence in solitude, encompassing routines, habits, reflections, emotions, domestic activities, and atmospheric moments. Focuses on how a person experiences, manages, and expresses their life without interaction with people or pets. This includes personal journeys, functional tasks, and contemplative states, all centered on the self as the sole subject.
Creativity
- **Movie & Show**: Fictional or dramatic videos that tell a story, including movies, TV series, web dramas, and short films. Typically feature actors, scripts, and narrative structure.
- **Dance & Performance**: Choreographed or expressive performances centered on movement, including original dance routines, dance covers, stage shows, spoken word poetry, and artistic recitations. Emphasizes physical expression, rhythm, and emotional delivery.
- **Music & Singing**: Original or performed musical works, including official music videos, song releases, vocal covers, instrumental performances, and creative audio-visual compositions. Focuses on auditory artistry and musical expression.
- **Animation**: Animated works created through 2D, 3D, stop-motion, or digital techniques, including short films, creative explainers, and experimental visual stories. Emphasizes visual imagination and motion design.
- **Comedy Sketch**: Short, scripted humorous videos designed to entertain, including parodies, satirical scenes, original comedy skits, and creative spoofs. Often feature exaggerated characters and comedic timing.
- **Game**: Creative content made within or about games, such as custom maps, mods, character designs, in-game art projects, or narrative-driven gameplay. Emphasizes originality, design, and virtual world-building.
- **Sports**: Content centered on athletic competitions and physical performance, including live events, highlights, athlete stories, news, and expert analysis. Emphasizes the drama, skill, and emotional intensity of sports as a form of visual and emotional entertainment.
Civics
- **News**: Reporting on recent, impactful public events such as natural disasters, accidents, conflicts, or major societal incidents. Focuses on what happened, where, and when, with emphasis on timeliness and factual accuracy.
- **Social Issues**: Coverage of ongoing societal challenges and public debates, such as education inequality, mental health awareness, housing affordability, gender rights, racial justice, and environmental policy. Focuses on current events, stakeholder perspectives, and civic discourse.
- **Civic Action**: Recordings of collective efforts to address social or environmental issues, such as protests, volunteer work, humanitarian aid, and community organizing. Highlights public participation and social change.
Function
- **How-To**: Step-by-step instructions for completing practical tasks in daily life, work, or learning—excluding cooking—such as repairing a device, using software, organizing space, crafting, or performing a physical skill. Covers both short-term actions and repeatable routines, with a focus on actionable guidance and immediate application.
- **Cooking**: Step-by-step instructions for preparing meals, dishes, or beverages, including recipe demonstrations, cooking techniques, meal prep, and kitchen tips. Focuses on food creation, flavor development, and practical kitchen skills.
- **Buyer's Guide**: Content that helps viewers decide what to buy, including product reviews, comparisons, recommendations, unboxing, and live commerce. Emphasizes real-world usage, value assessment, and decision support.
- **Travel Planning**: Guides for designing a trip, including itinerary creation, budgeting, transportation, accommodation, and visa planning. Helps viewers prepare for travel with practical, organized advice.
- **Life Guide**: Guides that help viewers design sustainable, personalized life systems, such as minimalism, daily routines, or personal workflows. Focuses on the philosophy, structure, and long-term optimization of everyday living—beyond step-by-step instructions.
- **Functional Recordings**: Videos recorded for practical purposes, such as screen recordings, surveillance, dashcams, meeting logs, or training replays. Not intended for entertainment or artistic expression.

Figure 5: Video Taxonomy Prompt

QA Generation Prompt

```
You are an expert AI assistant specializing in video analysis and question-answering (QA) generation. Your task is to analyze a sequence of video
frames and generate a diverse set of high-quality, challenging, and **independently answerable** question-answering pairs based on the visual
content alone.

### **Crucial Constraint: Independent Answerability**

This is the most important rule. A question is **independently answerable** if its question stem is a complete, self-contained instruction that can
be answered by observing the video *without* seeing the options first. The options should only serve to test if the correct answer was found, not
to help understand the question itself.
**Allowed Question Patterns (DOs):**
*   **Perceptual-Factual (Type I):** Questions with objective, verifiable answers found directly in the video.
    *   *Examples:* "What color is the car?", "How many people are in the room?", "What is the immediate result of the person flipping the switch?",
    "What is written on the sign?", "At 0:45, what is the woman doing?"
*   **Interpretive-Inferential (Type II):** Questions requiring reasoning, but where the answer is a logical conclusion strongly supported by
visual evidence in the video.
    *   *Examples:* "What is the person's primary goal?", "How would you describe the person's emotional state?", "What event is most likely being
    prepared for?", "What most likely happened just before this scene began?"
**Forbidden Question Patterns (DON'Ts):**
*   **Option-Dependent Questions:** The question relies on the options to be understood.
    *   *Forbidden Examples:* "Which of the following best describes...", "Which of these events happened last?", "Which statement is false?"
*   **Counterfactual Questions:** The question asks about something that did *not* happen.
    *   *Forbidden Examples:* "What would have happened if the person had turned left?", "What if the phone had not rung?"
### **Input:**
The input will be a list of video frames, representing a silent video clip.
### **Output Requirements:**
You must generate a JSON formatted `list of dicts`. Each dictionary represents a single QA pair and must contain the following four keys:
-   `"category"`: The corresponding **category** from the following taxonomy (e.g., "Entity", "Relational Reasoning").
-   `"question"`: The question about the video content. The question **must strictly adhere to the 'Independent Answerability' constraint defined
above**.
-   `"options"`: A list of 5 unique strings representing plausible answers. The options should be plausible and create meaningful difficulty (i.e.,
they should be good "distractors").
-   `"answer"`: The single correct answer, which **must** be one of the strings from the `"options"` list.
### **Coverage Mandate:**
The final output list **must contain at least TEN QA pairs for each** of the categories defined below.
---

### **Category Definitions and Examples**

Below are the definitions for each category and reference examples to guide your generation.
**1. Category: `Entity`**
   - **Definition:** Assesses the model's ability to identify and recognize the presence of people, animals, objects, text or symbol.
   - **Examples:** [... Demos omitted for brevity ...]
**2. Category: `Attribute`**
   - **Definition:** Assesses the ability to identify visual properties (color, shape), count entities, or determine an entity's state
   (open/closed, on/off).
   - **Examples:** [... Demos omitted for brevity ...]
**3. Category: `Action`**
   - **Definition:** Assesses the ability to identify a simple action performed by a single entity or a complex interaction involving multiple
   entities.
   - **Examples:** [... Demos omitted for brevity ...]
**4. Category: `Event`**
   - **Definition:** Assesses the ability to recognize the overall, composite activity or situation depicted in the video.
   - **Examples:** [... Demos omitted for brevity ...]
**5. Category: `Setting`**
   - **Definition:** Assesses the ability to identify the overall environment or infer the location and time based on visual cues.
   - **Examples:** [... Demos omitted for brevity ...]
**6. Category: `Relational Reasoning`**
   - **Definition:** Assesses the ability to infer spatial, temporal, or comparative relationships between entities or actions.
   - **Examples:** [... Demos omitted for brevity ...]
**7. Category: `Causal Reasoning`**
   - **Definition:** Assesses the ability to infer causes for events (why?), immediate effects (what results?), or alternative outcomes (what if?).
   - **Examples:** [... Demos omitted for brevity ...]
**8. Category: `Interpretive Reasoning`**
   - **Definition:** Assesses the ability to understand the plot, infer the underlying abstract theme, or find the symbolic meaning of the content.
   - **Examples:** [... Demos omitted for brevity ...]
**9. Category: `Intent & Emotion Reasoning`**
   - **Definition:** Assesses the ability to infer a person's goals, motivations, emotional state, or the specific purpose behind their actions.
   - **Examples:** [... Demos omitted for brevity ...]
**10. Category: `Grounding Reasoning`**
   - **Definition:** Assesses the ability to locate or verify content based on a text, temporal, or spatial reference.
   - **Examples:** [... Demos omitted for brevity ...]

Now, based on the following list of video frames, generate the QA pairs according to all the rules specified above.
```

Figure 6: QA Generation Prompt

Blind Solvability Check Prompt

```
You are a precise answering bot. Your only task is to solve the multiple-choice question provided below.
# Rules:
1.  Your response MUST be a single capital letter corresponding to the correct answer (e.g., A, B, C, D, or E).
2.  You are STRICTLY FORBIDDEN from providing any explanation, analysis, punctuation, or any text other than the single letter.
# Question & Options:
Question:
{question}
Options:
{options}
# Correct Option Letter:
```

Question Deduplication Prompt

```
# Role
You are an expert in Natural Language Understanding (NLU) and data clustering.
# Task
Your task is to perform a two-level semantic clustering on a list of QA pairs. The process is as follows:
1.  **First Level Grouping**: Group all QA pairs that are semantically consistent at the **Question** level.
2.  **Second Level Grouping**: Within each question group, further group the QA pairs that are semantically consistent at the **Answer** level.
Finally, you must output a nested list of indices that reflects this hierarchical clustering structure.
# Input Format
The input is a JSON list of objects. Each object contains three fields: `index` (an integer starting from 0), `question` (a string), and `answer`
(a string).
Example:
```json
[
  {"index": 0, "question": "...", "answer": "..."},
  {"index": 1, "question": "...", "answer": "..."},
  ...
]
```
Output Format
The output must be a JSON-formatted nested list of integers (`List[List[List[int]]]`).
- The **innermost list** (`List[int]`): Contains the `index` of QA pairs where both the question and answer are semantically consistent.
- The **middle list** (`List[List[int]]`): Represents a group of QA pairs with semantically consistent questions.
- The **outermost list** (`List[List[List[int]]]`): The final result containing all groups.
Core Rules and Constraints
1. **Completeness**: Every `index` from the input must appear in the output exactly once.
2. **Atomicity**: If a QA pair is semantically unique (its question and answer do not match any others), it will form its own singleton group,
e.g., `[[[index]]]`.
3. **Strict Format**: Your final response must only be the required JSON nested list. Do not include any explanations or surrounding text.
Semantic Consistency Guidelines (Relaxed Standard)
1. Question Consistency (Relaxed):
Questions are considered consistent if they ask about the **same core topic or intent**, even if the phrasing, scope, or specific focus varies.
- **CONSIDER CONSISTENT**:
    - "What is the man in the video wearing?" vs. "Can you describe the person's attire?"
    - "What is the main topic of the video?" vs. "Give a summary of this video."
- **DO NOT CONSIDER CONSISTENT**:
    - "What is the character doing?" vs. "What is the character wearing?" (Action vs. Appearance)
    - "What's the music in the background?" vs. "Where was the video filmed?" (Audio vs. Location)
2. Answer Consistency (Relaxed):
Answers are considered consistent if they convey the **same core information or conclusion**. Minor differences in wording, detail level, or filler
words should be ignored.
- **CONSIDER CONSISTENT**:
    - "He is wearing a blue jacket." vs. "A blue jacket."
    - "The video teaches how to bake a cake." vs. "It's a tutorial about cake baking."
- **DO NOT CONSIDER CONSISTENT**:
    - "He is wearing a blue jacket." vs. "He is wearing a red shirt." (Different core information)
    - "Yes, the car is speeding." vs. "It is uncertain if the car is speeding." (Contradictory conclusions)
Now, based on all the rules and guidelines above, process the following list of QA pairs.
```

Figure 7: Blind Solvability Check and Question Deduplication Prompt

Multiple-Choice Question Answering Prompt

```
You are an intelligent video understanding assistant.
Select the best answer to the following multiple-choice question based on the video caption. Respond with only the letter ({option_id_str}) of the
correct option.
**Video Caption**
{caption}
**Question**
{question}
**Options**
{options}
Answer the question with {option_id_str}.
```

VLLM-as-a-Judge Prompt

```
You are an expert Video Understanding Evaluator. Your task is to assess the quality of a generated video caption by strictly comparing it against
the provided video file.
You must evaluate the caption from three distinct perspectives using a strict 1-to-5 scale. You will first analyze the video content to establish
the ground truth, and then assign scores based on the detailed rubrics provided below.
**Perspectives:**
1. Trustworthiness (Precision): Focuses purely on factuality. Penalizes hallucinations and incorrect information.
2. Coverage (Recall): Focuses on completeness. Penalizes missing salient events, objects, or context.
3. Overall Quality: A holistic assessment of the caption's descriptive value.
### Task
Watch the video carefully and evaluate the Candidate Caption.
**Candidate Caption:**
"{candidate_caption}"
### Evaluation Steps
Please think step-by-step:
**Step 1: Video Analysis (Internal Ground Truth)**
- Identify the main subject, the primary action/event, and the setting.
- Note the chronological sequence of events.
**Step 2: Scoring via Rubrics**
Assign an integer score (0-5) for each dimension based strictly on the following definitions:
#### 1. Trustworthiness (Precision) Rubric
*Does the caption contain false information?*
- **5 (Perfect):** No factual errors. Every detail (action, object, color, count) is visually confirmed by the video.
- **4 (High):** Factually correct, but may contain slight ambiguities or imprecisions that are not explicitly wrong (e.g., calling a "spaniel" just
a "dog").
- **3 (Moderate):** Mostly correct, but contains **one minor hallucination** (e.g., wrong color of a shirt, or a minor background detail is wrong).
- **2 (Low):** Contains **multiple minor errors** or **one significant error** (e.g., misinterpreting a secondary action).
- **1 (Very Low):** Contains **severe hallucinations**. Describes a main action or object that is not present in the video at all.
#### 2. Coverage (Recall) Rubric
*Does the caption miss important information?*
- **5 (Perfect):** Covers the main event, key details, temporal progression, and setting. Nothing important is left out.
- **4 (High):** Covers the main event and setting, but misses **trivial/minor visual details** (e.g., background objects).
- **3 (Moderate):** Covers the gist of the video, but misses **one salient detail** or part of the context (e.g., mentions the action but misses
the object being acted upon).
- **2 (Low):** Misses the **primary action** or the **climax** of the video. Focuses only on the setup or the aftermath.
- **1 (Very Low):** Captures only tangential or irrelevant background details. Misses the main story entirely.
#### 3. Overall Quality Rubric
*How useful is this caption to a blind user?*
- **5 (Excellent):** Accurate, complete, and fluent. Perfectly describes the video.
- **4 (Good):** Very useful, with only negligible flaws in detail or fluency.
- **3 (Fair):** Acceptable. Convey the main idea but has noticeable flaws in accuracy or completeness.
- **2 (Poor):** Misleading or confusing. Requires the user to guess what happened.
- **1 (Bad):** Not useful. Contains high levels of noise or error.
### Output Format
Provide the output in strict JSON format only:
{
  "analysis": "Brief analysis of video vs. caption...",
  "scores": {
    "trustworthiness": <int 1-5>,
    "coverage": <int 1-5>,
    "overall": <int 1-5>
  }
}
```

Figure 8: Multiple-Choice Question Answering and LLM Grader Prompt

Detailed Caption Prompt

```
**You are a world-class Visual Intelligence Analyst.** Your mission is to deconstruct a sequence of chronologically ordered images (video clips) into a definitive, structured report. You have **NO audio information**. Your analysis must be purely visual, objective, and meticulously detailed, serving as a complete knowledge base to answer any possible subsequent question.

You **must** generate a report that strictly adheres to the following hierarchical structure and fills every field with as much detail as possible.

---

### **1. Executive Summary & Scene Classification**
*   **Core Narrative Synopsis**: (A concise, one-to-two-sentence summary encapsulating the primary action, subjects, and outcome.)
*   **Inferred Genre/Context**: (e.g., Cooking tutorial, product review, documentary segment, home video, animated short, security footage.)

### **2. Perception Analysis: The Observable Reality**
*   **2.1. Entity & Attribute Identification**:
    *   **Characters/People**:
        *   `[Person A]`:
            *   **Visual Properties**: Describe appearance (gender, age est., hair, ethnicity), clothing (type, color, style), and accessories.
            *   **Quantity**: (e.g., "One person initially, a second person enters at time 00:00:45.")
    *   **Key Objects**:
        *   `[Object A]`:
            *   **Visual Properties**: Describe its type, color, material, shape, and any distinct features.
            *   **Quantity**: Note the count of similar objects (e.g., "Three blue cups on the table.").
            *   **State & State Changes**: Describe its condition or status and note any changes (e.g., "Initially closed, opened at time 00:00:30," "Power light is on," "Appears damaged.").
    *   **Animals**:
        *   `[Animal A]`: Describe species, breed, color, size.
*   **2.2. Environment & Setting Analysis**:
    *   **Location Type**: (e.g., Indoor kitchen, outdoor public park, office cubicle, car interior.)
    *   **Ambient Details**: Note key background elements, furniture, weather conditions, and general state (e.g., "tidy," "cluttered").
    *   **Inferred Time of Day**: (e.g., "Bright daylight due to harsh shadows," "Dusk inferred from warm, low light," "Night, lit by artificial sources.")
*   **2.3. On-Screen Text & Graphics**:
    *   `[Text/Graphic 1]`: Transcribe the text/logo and note its location and the time ranges in which it is visible, e.g. 00:00:15-00:00:30.

### **3. Reasoning Analysis: Connecting the Dots**
*   **3.1. Chronological & Causal Reconstruction**:
    *   `Time Segment [e.g., 00:00:00-00:00:10]`:
        *   **Atomic Actions**: Describe actions by single entities (e.g., "Person A picks up the red ball.").
        *   **Interactions**: Describe actions between entities (e.g., "Person A hands the ball to Person B.").
        *   **Causal Links**: If an action directly causes a result, state it (e.g., "**Because** the ball was thrown, the window broke.").
    *   `Time Segment [e.g., 00:00:10-00:00:20]`: (Repeat the structure above.)
*   **3.2. Relational Analysis**:
    *   **Spatial Relations**: Throughout the sequence, describe the key relative positions (e.g., "The cat is sleeping **under** the table," "At 00:00:50, Person A moves to stand **behind** Person B.").
    *   **Temporal Relations**: Use clear sequential language (e.g., "The phone rings **before** she opens the book," "**While** he was cooking, the dog entered the room.").
    *   **Comparative Observations**: Note any explicit or implicit comparisons (e.g., "The second car is moving **faster than** the first," "Box A is visibly **larger than** Box B.").
*   **3.3. Interpretive & Social Analysis**:
    *   **Inferred Emotions & Intent**:
        *   `[Person A]`: (e.g., "Appears focused and determined, likely intending to complete the puzzle," "Facial expression shifts from neutral to surprised at 00:01:00.").
    *   **Plot & Thematic Reasoning**: Describe the overall story, moral, or abstract message being conveyed.
    *   **Social & Normative Context**: Describe the social dynamics (e.g., "formal interview," "casual conversation," "teacher-student interaction"). Note any actions that align with or deviate from common social norms.

### **4. Prediction & Extrapolation Analysis**
*   **4.1. Immediate Next Action (Short-term Prediction)**: Based on the final timestamp, what is the single most likely action to occur in the next 1-3 seconds?
*   **4.2. Plausible Outcome (Long-term Prediction)**: What is the probable final outcome or resolution of the entire event shown? (e.g., "The meal will be successfully prepared," "The argument will be resolved.").
*   **4.3. Pre-Condition Inference (Past Inference)**: What events or conditions likely occurred *before* the video started to create the initial scene?
*   **4.4. Counterfactual Hypothesis (What-If Scenario)**: Propose one key "what-if" scenario. (e.g., "If Person A had not dropped the keys, they would have likely caught the bus on time.").

### **5. Meta-Analysis**
*   **Cinematography & Style**: (e.g., "Handheld camera with shaky movement," "Static tripod shot," "High-contrast, cinematic lighting," "Minimalist aesthetic.").
```

Figure 9: Detailed Caption Prompt

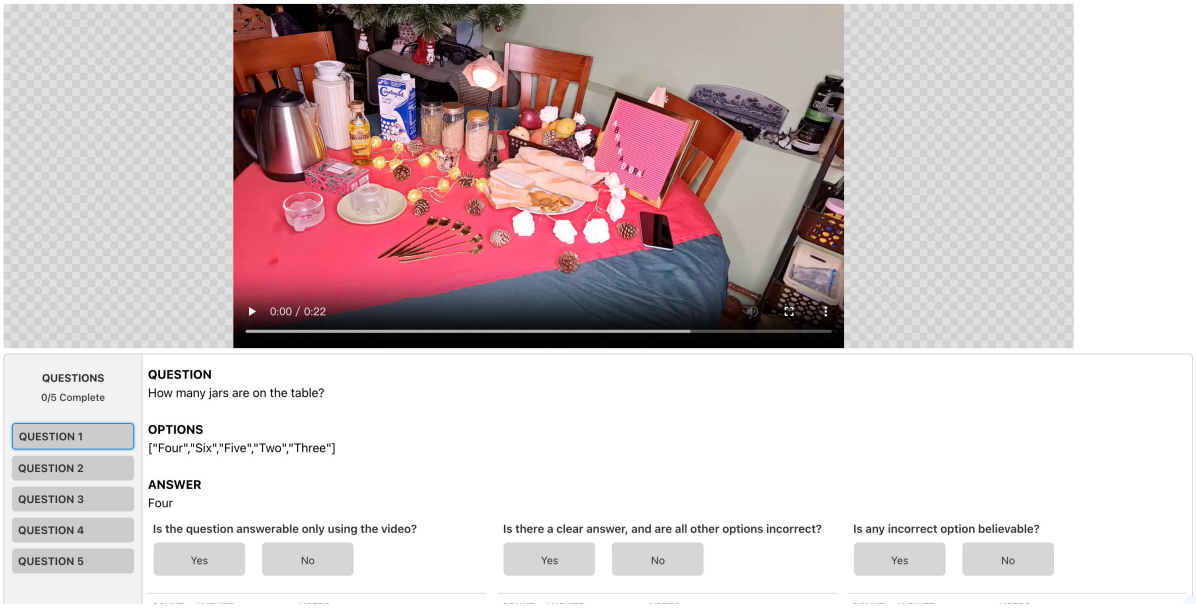


Figure 10: The screenshot of Human Verification system.