ViSIL: Unified Evaluation of Information Loss in Multimodal Video Captioning

Anonymous authors
Paper under double-blind review

Abstract

Multimodal video captioning condenses dense footage into a structured format of keyframes and natural language. By creating a cohesive multimodal summary, this approach anchors generative AI in rich semantic evidence and serves as a lightweight proxy for high-efficiency retrieval. However, traditional metrics like BLEU or ROUGE fail to quantify information coverage across disparate modalities, such as comparing a paragraph of text to a sequence of keyframes. To address this, we propose the **Video Summary Information Loss** (**ViSIL**) score, an information-theoretic framework that quantifies the video information not captured by a summary via vision-language model (VLM) inference. By measuring the information loss, ViSIL is a unified metric that enables comparison across multimodal summary formats despite their structural discrepancies. Our results demonstrate that ViSIL scores show a statistically significant correlation with both human and VLM performance on Video Question Answering (VQA) tasks. ViSIL also enables summary selection to optimize the trade-off between information loss and processing speed, establishing a Pareto-optimal frontier that outperforms text summaries by 7% in VQA accuracy without increasing processing load.

1 Introduction

The surge in high-resolution video has rendered multimodal summarization essential. By unifying visual keyframes with linguistic descriptors, multimodal summaries effectively bridge the gap between massive raw datasets and meaningful understanding. Unlike unimodal descriptions, these multimodal summaries provide the rich semantic grounding required to evaluate text-to-video generation models and the dense indexing necessary for precise retrieval-augmented generation (RAG). This cross-modal synergy is also critical for human-in-the-loop applications like security surveillance, where combined visual and textual cues enable rapid analysis without reviewing full-length footage. While keyframes capture instantaneous context, text is vital for synthesizing temporal dynamics and providing high-level reasoning that images alone may obscure. The synergy creates a spectrum of multimodal video summaries, from text-only to hybrid formats with varying keyframe densities, as shown in Figure 1.

However, this diversity renders traditional evaluation metrics, *e.g.*, BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), or METEOR (Banerjee & Lavie, 2005), insufficient for capturing the holistic information contribution across heterogeneous modalities. These metrics are restricted to unimodal text-to-text comparison and cannot capture information distribution across disparate modalities. Also, it remains unclear which format optimally balances information richness with processing efficiency, such as human response time or the input tokens for a vision-language model (VLM). It is not yet established whether increasing the number of images in a summary necessarily leads to better video understanding and faster processing.

To unify the evaluation of these heterogeneous formats of modalities, we propose the **Video Summary Information Loss** (ViSIL) score, an information-theoretic framework that quantifies the semantic information loss when compressing a video V into a summary $\tilde{V}$. As illustrated in Figure 1, we first generate a detailed caption C, either through a VLM or human annotation, to act as a comprehensive textual proxy for the source video V. ViSIL then measures the information loss by evaluating a VLM's ability to recover caption C using the multimodal summary $\tilde{V}$ relative to the original video V. Defined as the conditional pointwise mutual information $I(C;V|\tilde{V}) = \log \frac{P(C|V)}{P(C|\tilde{V})}$, the metric

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

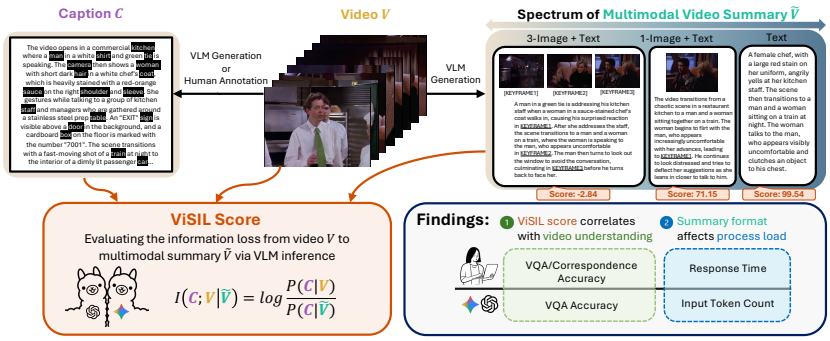


Figure 1: **A Unified Evaluation for Multimodal Video Captions.** Given a video V, VLM-generated detailed caption C, and several multimodal video summaries $\tilde{V}$, the ViSIL score quantifies the information loss within the summaries relative to the original video content. Our results show that ViSIL correlates with video understanding (VQA accuracy) for both humans and VLMs, while the summary format dominates the process load (response time and token count).

captures visual details that remain "unaccounted for" by the summary. By measuring information loss, where lower scores signify better coverage, ViSIL offers a unified metric that aligns with both human and VLM comprehension across the multimodal summary spectrum.

Contributions. We introduce ViSIL, an information-theoretically motivated framework that evaluates text-only, visual, and multimodal video summaries through a common caption-recovery objective. We investigate its relationship with human and VLM question-answering performance and use it to characterize trade-offs between summary information loss and processing load. Our work demonstrates that summary format primarily dictates process load, such as response time and token consumption, rather than inherent video understanding. By leveraging ViSIL for summary selection, we establish a Pareto-optimal frontier that outperforms pure text summaries by 7% in VQA accuracy without increasing processing overhead.

2 Related Works

Video Captioning (Qasim et al., 2025; Abdar et al., 2024) is the task of using VLMs to generate natural language descriptions that semantically summarize the visual and auditory content of a video. High-quality, precise captions are critical for modern generative AI and data retrieval systems; they are indispensable for semantic grounding in text-to-video generation models (Chen et al., 2024; OpenAI, 2025b) and crucial for efficient indexing in RAG-based storage and retrieval systems (Zhu et al., 2023; han Li et al., 2024). While Kudo et al. (2023) also explores multimodal (keyframes + text) captioning, they rely on existing unimodal evaluation metrics. In contrast, ViSIL is for multimodal captions, and we verify it using VLM-based and human-based video understanding tests.

Video Caption and Keyframe Evaluation. Robust evaluation metrics for high-quality video captions fall into two categories: reference-based methods requiring ground truth captions (Kudo et al., 2023; Chai et al., 2025), and reference-free methods. Reference-free methods typically rely on multimodal embedding similarity (Lee et al., 2020; Hessel et al., 2021; han Li et al., 2025) or mutual information between text and video (Chen et al., 2025). However, embedding-based metrics struggle with cross-format summaries; disparate architectures and incompatible latent spaces are often limited to unimodal text representations (OpenAI, 2025a; Google, 2025) and preclude a unified metric for heterogeneous data. In contrast, ViSIL sidesteps embedding limitations by focusing on cross-modal signal preservation. This work also departs from prior work, which evaluates modalities in isolation, such as standalone (Liang et al., 2024) or VQA-based (Ye et al., 2025) keyframe selection.

Human-Centric Evaluation. Although automatic metrics enable scalable evaluation, human-centric assessment remains the gold standard for measuring the practical utility of video summaries. Prior human evaluations have primarily focused on fluency and informativeness (Belz & Reiter, 2006; Graham et al., 2017), but these methods are difficult to scale and do not always correlate with practical utility. Recent work demonstrates that aligning model representations with human perception,

including attention (Linsley et al., 2018), temporal visual dynamics (Parthasarathy et al., 2023), and conceptual structures (Muttenthaler et al., 2022; 2025), improves robustness, interpretability, and generalization. Extending this philosophy to video summarization, we argue that evaluation should reflect human task performance. We therefore adopt an extrinsic evaluation paradigm (Nenkova et al., 2011; Pu et al., 2023), measuring how summaries affect human response time and accuracy on multiple-choice questions grounded in video content.

3 Video Summary Information Loss (ViSIL)

To ground ViSIL, we define Mutual Information (MI) and its pointwise form.

Mutual Information (MI) (Kreer, 1957) quantifies the mutual dependence between two random variables, $\mathbf{X}$ and $\mathbf{Y}$. It measures how much information is obtained about one random variable through observing the other, thus other literature calls it "information gain," which is defined as:

$$\mathbf{I}(\mathbf{X};\mathbf{Y}) = \mathbb{E}_{X,Y}\left[\log \frac{P(x,y)}{P(x)P(y)}\right] \geq 0. \tag{1}$$

Pointwise Mutual Information (PMI) (Bouma, 2009), in contrast, provides a measure of association between individual events or outcomes X and Y, rather than random variables. It is defined as:

$$\mathcal{I}(X;Y) = \log \frac{P(X,Y)}{P(X)P(Y)} = \log \frac{P(X|Y)}{P(X)} \in (-\infty, \infty). \tag{2}$$

MI averages over a distribution and is non-negative; PMI evaluates a single pair and remains unbounded. Since we only use PMI, the notation $\mathcal{I}$ denotes PMI rather than MI throughout this paper.

3.1 Problem Formulation

For the evaluated setting, let $V = I$ denote the visual content of a video with N frames $I = \{I_i\}_{i=1}^{N}$. Now, suppose we have a video summary $\tilde{V} = \tilde{I} \cup T$ that consists of a subset of keyframes $\tilde{I} \subseteq I$ and a textual summary T. This formulation provides a flexible definition covering a broad **spectrum of video summaries**. While $\tilde{V}$ is inherently multimodal, containing both visual and textual components, it seamlessly accommodates unimodal scenarios when either $\tilde{I} = \emptyset$ or $T = \emptyset$.

We measure the quality of a summary by calculating the PMI as in Equation (2) between the original video and the summary. A higher PMI score indicates a better summary, as it signifies that the summary contains a greater amount of shared information derived from the video:

$$\mathcal{I}(V;\tilde{V}) = \underbrace{\log \frac{P(V|\tilde{V})}{P(V)}}_{\text{(3.a)}} = \underbrace{\log \frac{P(\tilde{V}|V)}{P(\tilde{V})}}_{\text{(3.b)}} \text{ (PMI of video and summary; higher is better).} \tag{3}$$

However, direct calculation of the mutual information in Equation (3) is doubly intractable. First, we can only approximate these terms using machine learning models. In term (3.a), the numerator is the conditional probability to generate the original video given the summary, $P(V|\tilde{V})$. It is inaccessible by any diffusion models, which are the state-of-the-art video generation models. While the denominator, $P(V)$, is the notoriously intractable existential likelihood of a data point. The same difficulty holds for the term (3.b) in Equation (3) as well. Therefore, we must approximate the video conditional generation probability and the existential likelihood in Equation (3).

3.2 Approximation with Autoregressive Models

To approximate Equation (3) into a solvable form, we draw inspiration from generative models operating in other modalities. Our key insight is that while video generation models struggle to estimate the necessary conditional probabilities, **autoregressive VLMs** can leverage their inherent next-token prediction mechanism to estimate conditional probabilities effectively, even when relating different modalities. The mathematical formulation of VLMs is fundamentally suited, as they

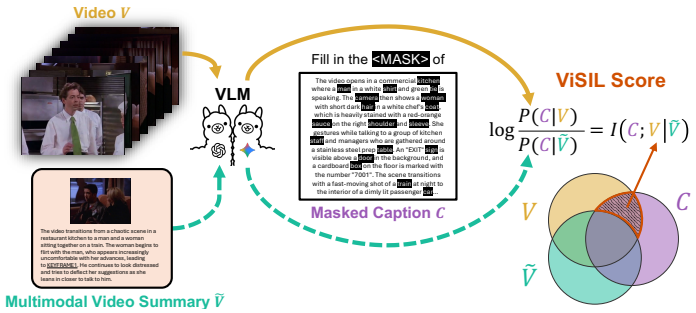


Figure 2: **ViSIL Implementation via VLM Inference.** ViSIL assesses information loss by comparing a VLM's ability to recover masked tokens in caption C from video V versus summary $\tilde{V}$. ViSIL is defined as the pointwise mutual information between the video and caption conditioned on the summary, representing the information in the video that remains unaccounted for by the summary. A lower ViSIL score indicates better information preservation.

model the conditional probability $P_{\text{VLM}}(Y|X)$, which represents the token probabilities of the output sentence Y given the multimodal input X.

Based on this insight about the solvability of next-token mechanisms, we introduce two necessary reformulations: (1) We require a **language form of the video** to serve as a textual proxy for the raw video V since VLMs cannot output videos with token probabilities. The need for textual reference is common in the evaluation of captions (Lee et al., 2020; Chai et al., 2025), despite variations in the underlying motivations. (2) To handle the intractable marginal likelihood term in the denominator, we replace it with a conditional probability of text (*e.g.*, by conditioning it on a prompt or context). This critical transformation converts the difficult marginal $P(\tilde{V})$ into a conditional probability, $P(\tilde{V}|\text{Auxiliary Text})$, which fits perfectly with the next-token prediction mechanism of autoregressive models.

3.3 ViSIL–A Unified Framework to Evaluate Video Summaries

To approximate Equation (3), we first use a VLM to caption the raw video V into a long and detailed caption C, which can also be annotated by humans. Caption-generator sensitivity is examined in Appendix F. Then, we evaluate the information loss between video V and summary $\tilde{V}$, the amount of information contained in the video but missed in the summary, the lower the better. Using the video caption C as the proxy, we define ViSIL score as:

$$\mathcal{I}(C; V \mid \tilde{V}) = \log \frac{P(C \mid V, \tilde{V})}{P(C \mid \tilde{V})} = \log \frac{P(C \mid V)}{P(C \mid \tilde{V})} \quad \text{(ViSIL Score; lower is better)}. \tag{4}$$

The derivation follows from the assumption that $\tilde{V}$ is contained within V (*i.e.*, $\tilde{V} \subseteq V$). $\mathcal{I}(C; V \mid \tilde{V})$ quantifies how much new information the caption C adds about the raw video V, given the multimodal summary $\tilde{V}$. If the multimodal summary already captures all video information, this value should be minimal. For an illustrative example, see Figure 2.

Equation (4) defines a unified metric that measures information loss in any multimodal summary $\tilde{V}$ relative to video V, enabling direct comparison and selection by minimizing the information loss. In our experiments, we demonstrate that the ViSIL correlates strongly with video understanding tasks assessed by both advanced VLMs and humans.

Venn Diagram Interpretation. The red shaded area in the Venn diagram (Figure 2, right) visually represents the information loss $(V \cap C) \setminus \tilde{V}$ that ViSIL aims to minimize, quantifying the information contained in the video-caption overlap that the summary fails to capture. By utilizing a comprehensive caption C such that $V \cap C \approx V$ (i.e., C effectively "covers" V), ViSIL measures how much of the video's core content is missing from the summary $\tilde{V}$. As the summary becomes more comprehensive, this shaded area shrinks, yielding a lower (better) score.

To link Equation (3) to Equation (4), note that maximizing the overlap $V \cap \tilde{V}$ (as in Equation (3), shown by the yellow-green intersection) is equivalent to minimizing the set difference $V \setminus \tilde{V}$. When the caption is sufficiently descriptive ($V \cap C$ is large enough), the shaded region measured by Equation (4) essentially measures $V \setminus \tilde{V}$.

Sensitivity to Hallucinated Captions. In a controlled Gemini experiment with 100 captions, injected factual errors significantly increase ViSIL scores (Wilcoxon and t-test $p < 0.001$). This sensitivity is evaluator-dependent. Appendix G has the protocol, results, and cross-family comparison.

3.4 Information Loss-Process Efficiency Trade-off

We propose an ViSIL-based summary selection strategy to balance information loss and processing load by minimizing the joint objective:

$$\min_{\tilde{V}} \quad \mathcal{I}(C; V|\tilde{V}) + \alpha \cdot \tau(\tilde{V}), \tag{5}$$

where $\tau(\tilde{V})$ denotes the token count (processing load), and α is the Lagrange multiplier controlling the trade-off. Varying α traverses the Pareto-optimal frontier, selecting summaries that balance semantic completeness with the processing speed of VLMs or humans. As shown in Figure 3, ViSIL dominates fixed baselines like *Random* (randomly selected summaries), *1-Image*, and *3-Image* summaries. The setup is detailed later in Section 4. Notably, at the highest processing speeds, our selected summaries achieve 63% accuracy, significantly outperforming the 56% accuracy of pure text summaries at equivalent efficiency. Figure 3 confirms that our optimization effectively preserves understanding utility while minimizing overhead. The evidence scope is discussed in Appendix I.

3.5 ViSIL Computation via Inference

The information-theoretic formulation in Equation (4) motivates our practical score, which approximates caption likelihoods through masked-keyword prediction. This computation provides a model-based proxy for information loss; it does not directly estimate the underlying video–summary mutual information. Its coverage depends on the content captured in C, and its values depend on the evaluator's conditional predictions.

To compute ViSIL, we only require next-token probabilities from VLMs, which all major APIs and local models provide. Since most online APIs produce non-deterministic probabilities (He & Lab, 2025), we estimate stable values by repeated sampling and taking the geometric mean. To accelerate evaluation, we follow prior work (Chen et al., 2025; Jung et al., 2024; Bhatt et al., 2025) and approximate sentence probability using keyword prediction. We mask key semantic tokens and compute

$$P(C|V) \simeq \prod_{i=1}^{n} P(k_i|V), \quad P(C|\tilde{V}) \simeq \prod_{i=1}^{n} P(k_i|\tilde{V}),$$

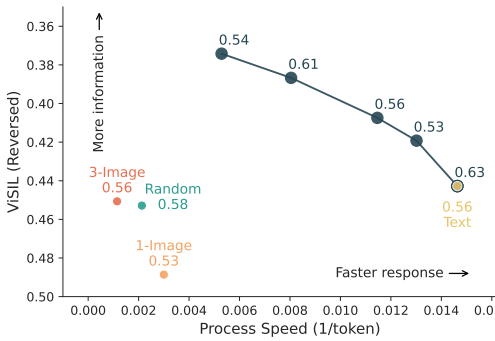


Figure 3: **Pareto Frontier of Process Speed and ViSIL Score** showing that static formats are sub-optimal for the process speed–information trade-off. *The annotated VQA accuracy* confirms ViSIL identifies high-utility summaries that outperform pure text and fixed-image formats while preserving processing speed.

where k_i denotes keywords in C. Keyword-based estimation reduces VLM inference cost, improves stability, and ignores low-information tokens such as 'a', 'the', etc.

4 Experiment and User Study

Let $\mathcal{D} = \{(V_i, Q_i, A_i)\}_{i=1}^{M}$ denote a visual question-answering (VQA) dataset where each sample contains a video V, a question Q, and an answer A. Our evaluation focuses on two specialized subsets: Episodic Reasoning (EpR) from MVBench (Li et al., 2024), which tests long-term temporal understanding, and the SSS (Sequence of Scenes) subset of LongVideoBench (Wu et al., 2024). For each video V in the dataset, we generate a detailed caption C using `Gemini 2.5 Pro`, as it is the most powerful model when we conducted the experiments. Recall that the multimodal summary $\tilde{V} = \tilde{I} \cup T$ consists of keyframes $\tilde{I}$ and text T. We constrain the number of keyframes in $\tilde{I} = \{f_1, \ldots, f_k\}$ to $k \leq 3$. See dataset selection details and all prompts used in Appendices A and J.

Captioning & Keyword Masking C. We employ a two-stage pipeline. First, `Gemini 2.5 Pro` generates comprehensive captions capturing events and legible text. Second, `GPT-5` extracts up to 20 fine-grained keywords, preserving their original morphology and sequential order.

Summary Construction $\tilde{V}$. We use `Gemini 2.5 Pro` to both select representative keyframes $\tilde{I}$ and generate the textual component T. Crucially, T is conditioned on $\tilde{I}$ to ensure the summary is visually grounded and contextually faithful to the underlying content. The final summary $\tilde{V}$ is thus a composite of the generated text description and the retrieved keyframes.

4.1 Research Questions

We investigate the role of multimodal video summaries in supporting both VLM and human understanding of video content. We address the following research questions (RQ):

RQ1. ViSIL as a predictive metric. To what extent does the ViSIL score correlate with downstream VLM and human video understanding, measured by performance on VQA tasks?

RQ2. Impact of summary format. How do different summary formats (*Text-Only*, *1-Image*, *3-Image*, and *Full Video*) affect comprehension performance for both VLMs and human users?

4.2 VLM Evaluation

We compute the ViSIL score with `Gemini 2.0 Flash`, the latest model revealing token probabilities in its API. To assess evaluator dependence, we also score all summary formats with GPT-4o on matched observations (Appendix E.1). Then, we evaluate VLM performance on the VQA task, where we employ `Gemini 2.5 Pro` as the answering model, following the evaluation protocol detailed in Appendix J.7 and J.8.

For each video and each multimodal summary, we sample 3 independent runs and take the geometric average of token probabilities to account for generation variability. For each VQA question, the model is provided only with the corresponding summary and is tasked with answering the associated question. We then compute (1) the VQA accuracy achieved under each summary format and (2) the ViSIL score of the same summary, enabling a paired analysis at the instance level.

Our evaluation reports two metrics: (1) VQA accuracy per summary format and (2) correlation between ViSIL scores and VQA accuracy across formats and videos. This analysis allows us to quantify how well ViSIL reflects downstream task performance and to assess its effectiveness as a proxy for summary informativeness. We benchmark ViSIL against a direct VLM judge on matched observations (Appendix H) and quantify uncertainty with pooled confidence intervals (Table 1).

4.3 User Study

We conducted two controlled user studies to evaluate human understanding of video summaries. Both utilized a within-subjects balanced Latin square design (Keedwell & Dénes, 2015) to mitigate ordering effects. Detailed demographics and interfaces are provided in Appendices C and D. To isolate human recall from active information retrieval, such as re-watching or scrubbing through footage, we restrict

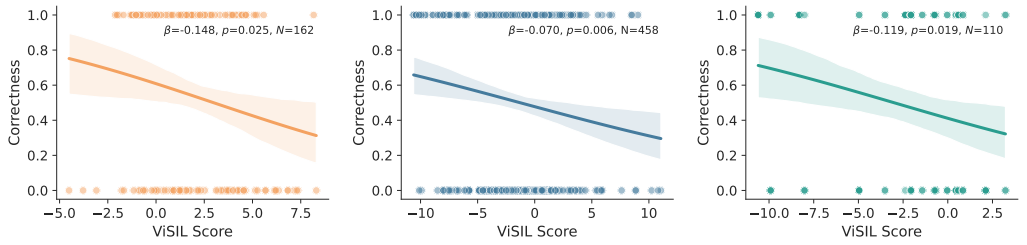
users to a single-pass viewing. It ensures that response times measure immediate comprehension rather than the latency involved in searching for specific details. This single-viewing approach better captures the essence of video understanding by measuring human information acquisition rather than the navigational efficiency of video.

We measured *accuracy*, *response time*, and *confidence* (5-point Likert (Likert, 1932)) across 4 summary conditions: *Text-Only*, *1-Image*, *3-Image*, and *Full Video*. Participants answered questions for 4 LongVideoBench videos, seeing each video in only one format to prevent learning effects.

5 Results and Analysis

Our results show that ViSIL predicts human and VLM video understanding, while summary format governs efficiency and human sensitivity to inconsistency.

5.1 RQ1: ViSIL as a Predictor of Video Understanding



(a) MVBench (EpR) (b) LongVideoBench (SSS) (c) Human VQA

Figure 4: **Logistic regression analysis** between individual sample correctness and corresponding ViSIL scores. The downward trend in general indicates that higher ViSIL scores (representing more information loss) correlate with decreased VLM accuracy.

To address RQ1, we investigate the extent to which the ViSIL score correlates with downstream video understanding in both Vision-Language Models (VLMs) and human participants, as measured by performance on Visual Question Answering (VQA) tasks. Since VQA correctness is binary, we first examine this relationship using logistic regression. Across both VLM datasets and the human study, lower ViSIL scores correspond to higher VQA accuracy, with significant associations in the original pooled analysis (Figure 4).

On MVBench, higher information loss (higher ViSIL score) effectively predicts lower VQA accuracy ($\beta = -0.148, p = 0.025, N = 162$). This predictive power extends to LongVideoBench as well, where we observe a significant negative correlation ($\beta = -0.070, p = 0.006, N = 458$), indicating that information density is a critical factor even for long-form video understanding. This trend is further validated by human study, which closely mirrors the VLM results: human VQA correctness exhibits a significant negative correlation with ViSIL scores ($\beta = -0.119, p = 0.019, N = 110$).

Permutation tests (Moore, 1999; Fisher, 1971; Pitman, 1937) support this negative relationship in all three pooled analyses (Table 1). The effects are modest, and neither VLM association remains significant after video-level clustering.

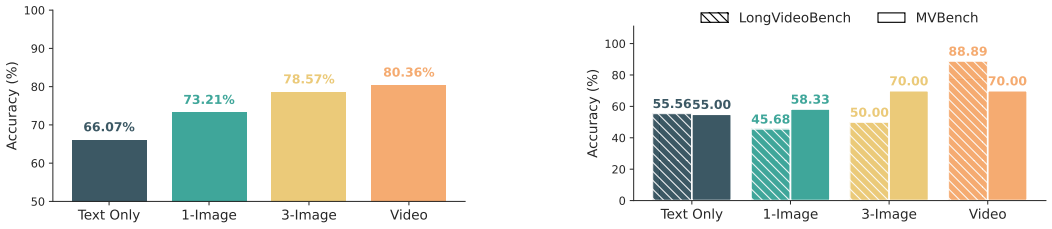
Table 1: Pooled Pearson correlations, 95% confidence intervals, and permutation p-values (10,000 shuffles). N counts observations; Videos counts source videos.

Dataset	N	**Videos**	r	**95% CI**	p**-value**
MVBench	162	18	-0.178	$[-0.324, -0.025]$	0.021^{*}
LongVideoBench	458	35	-0.129	$[-0.218, -0.038]$	0.006^{**}
Human VQA	110	4	-0.228	$[-0.398, -0.042]$	0.017^{*}

$^{*}\ p < 0.05$, $^{**}\ p < 0.01$.

All three pooled 95% confidence intervals in Table 1 lie below zero, supporting the association between lower ViSIL and higher VQA accuracy. The observed association remains negative when we replace Gemini with GPT-4o as evaluator (Appendix E.1), and both caption generators yield the same summary-format ordering (Appendix F). Against a direct VLM judge on matched observations, ViSIL achieves a larger correlation magnitude on MVBench ($|r| = 0.183$ vs. 0.034); the judge leads on LongVideoBench (0.170 vs. 0.090). Appendix H reports confidence intervals for these comparisons.

5.2 RQ2: Impact of Summary Format



(a) Human (b) VLM

Figure 5: **VQA Accuracy across Summary Formats.** (a) Human accuracy improves as visual context increases, with the 3-Image format approaching Video-level performance. (b) VLM accuracy similarly improves as the number of visual keyframes increases. * denotes statistical significance at $p < 0.05$; ** denotes $p < 0.01$.

On Human Performance. We analyze how summary formats impact human performance and subjective certainty. As shown in Figure 5a, the full Video format achieves the highest VQA accuracy (80.36%). However, the 3-Image summary is remarkably competitive, achieving 78.57% accuracy–within 2% of the full video baseline–while reducing human response time by nearly 20 seconds on average (see 65.94s vs. 85.23s in Table 2). It identifies that the 3-Image format preserves semantic information without the temporal redundancy of video. We observe a similar trend in VLM performance (Figure 5b), where increasing visual density enhances grounding.

While performance is similar, user perception differs. As shown in Figure 6, participants reported significantly higher confidence when viewing full videos (4.04 ± 1.17) compared to the 3-Image format (3.43 ± 1.29). It suggests that while static summaries are sufficient for *correct* reasoning, the dynamic context of video provides a psychological layer of reassurance that static keyframes lack.

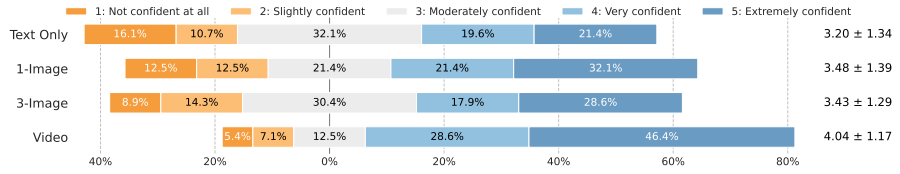


Figure 6: **Distribution of Human Confidence Ratings.** A diverging stacked bar chart of Likert responses (1–5) shows that dynamic visual context (Video) shifts the distribution toward higher confidence compared to static summaries.

Table 2: **Comparison of VLM and Human Process Load across Summary Formats.** Video incurs the highest process load among all formats, while the text-only format yields the lowest.

Metric	Text Only	1-Image	3-Image	Video
MVBench (EpR) (token)	59.05 ± 8.58	326.34±10.84	866.88±11.04	7148.02 ± 2348.49
LongVideoBench (SSS) (token)	77.86±11.68	336.98±10.76	873.27±11.57	54,587.73 ± 54,131.49
Human Response Time (sec)	62.60±35.98	64.78±41.61	65.94±35.68	85.23 ± 78.57

On Process Load. Table 2 highlights the variance in computational and cognitive costs. While full video inputs provide maximum context, they incur a computational overhead, consuming up to $700\times$ more VLM tokens than text-only summaries (54.5k vs. 77 tokens on LongVideoBench) and approximately $62\times$ more than the 3-Image summary (873 tokens). Human cognitive load follows a similar trend; participants required significantly more time to process full video (85.23s) compared to static summaries (≈ 64s). Notably, the 3-Image format incurs only a marginal increase in response time over text ($+3.34$s) while maintaining a relatively low token footprint, suggesting it offers an efficient middle ground.

Takeaways. ViSIL quantifies cross-modal information retention rather than semantic overlap, enabling effective multimodal evaluation beyond the limits of unimodal BLEU-style evaluation. Appendix E shows consistent behavior across evaluation VLMs, with stable relative rankings despite scale differences, allowing small VLMs to serve as reliable evaluators. ViSIL also provides strong robustness to hallucinations by filtering ungrounded captions, as shown in Table 8.

5.3 Limitations

The framework remains largely insensitive to hallucinations in the caption C, as explained earlier in Section 3.3. To reduce correlated failure modes, we recommend using different models for caption generation and evaluation, which avoids cases where both C and $\tilde{V}$ hallucinate in the same way. Such aligned hallucinations remain difficult to detect, since the evaluator VLM may lack the capability to recognize the error. Built on existing VLMs, ViSIL does not extend model capabilities but focuses on evaluating the quality of multimodal summaries. ViSIL targets evaluation only and does not address better keyframe and event selection, which requires solving NP-hard combinatorial subset selection problems over temporal segments (Sun et al., 2025).

6 Conclusion

We introduce ViSIL, an information-theoretic framework for unified evaluation of multimodal video summaries. Unlike modality-specific metrics, ViSIL quantifies information loss from the source video across any format. It measures information richness, while the summary format determines processing load (*e.g.*, human response time, VLM tokens), making ViSIL a proxy for information density that separates content from efficiency.

Future work can expand ViSIL to handle interactive summarization where the "information need" shifts based on user queries. It can include audio in the spectrum of video summaries while still leveraging the ViSIL framework. We aim to explore the integration of ViSIL into the training loop of video summarization models to optimize for information preservation rather than just evaluation.

AI use statement

Generative AI models are components of the research pipeline described in Section 4 and Appendix J, including caption and summary generation, keyword extraction, and model-based evaluation. We also used generative AI tools to assist with coding and rephrasing. The authors are responsible for the final content of this work, including its text, claims, and artifacts produced.

Ethics statement

The human-subject experiments were conducted under institutional review board approval, with risks limited to those encountered in daily life. Participant recruitment and demographics are described in Appendix C; study instructions and interfaces are provided in Appendix D.

Reproducibility statement

The method and inference procedure are described in Section 3.3 and Section 3.2. Section 4 specifies the models and evaluation protocol. The appendix provides dataset selection (Appendix A), user-study materials (Appendices C and D), and prompts (Appendix J).

REFERENCES

Moloud Abdar, Meenakshi Kollati, Swaraja Kuraparthi, Farhad Pourpanah, Daniel McDuff, Mohammad Ghavamzadeh, Shuicheng Yan, Abduallah Mohamed, Abbas Khosravi, Erik Cambria, and Fatih Porikli. A review of deep learning for video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–20, 2024. doi: 10.1109/TPAMI.2024.3522295.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss (eds.), *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Anja Belz and Ehud Reiter. Comparing automatic and human evaluation of nlg systems. In *11th conference of the european chapter of the association for computational linguistics*, pp. 313–320, 2006.

Neel P. Bhatt, Po han Li, Kushagra Gupta, Rohan Siva, Daniel Milan, Alexander T. Hogue, Sandeep P. Chinchali, David Fridovich-Keil, Zhangyang Wang, and Ufuk Topcu. Uncap: Uncertainty-guided planning using natural language communication for cooperative autonomous vehicles, 2025. URL https://arxiv.org/abs/2510.12992.

Gerlof Bouma. Normalized (pointwise) mutual information in collocation extraction. *From Form to Meaning: Processing Texts Automatically. Proceedings of the Biennial GSCL Conference 2009*, 2009.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark, 2025. URL https://arxiv.org/abs/2410.03051.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. ShareGPT4video: Improving video understanding and generation with better captions. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=EiH6WWLzlu.

Shenghui Chen, Po han Li, Sandeep P. Chinchali, and Ufuk Topcu. Vibe: Annotation-free video-to-text information bottleneck evaluation for TL;DR. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=C35FCYZBXp.

Ronald Aylmer Fisher. *The design of experiments*. Springer, 1971.

Google. Embeddings | gemini api. https://ai.google.dev/gemini-api/docs/embeddings, December 2025. Accessed 18 Dec 2025.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30, 2017.

Po han Li, Yunhao Yang, Mohammad Omama, Sandeep Chinchali, and Ufuk Topcu. Any2any: Incomplete multimodal retrieval with conformal prediction, 2024. URL https://arxiv.org/abs/2411.10513.

Po han Li, Sandeep P. Chinchali, and Ufuk Topcu. CSA: Data-efficient mapping of unimodal features to multimodal features. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=6Mg7pjG7Sw.

Horace He and Thinking Machines Lab. Defeating nondeterminism in llm inference. *Thinking Machines Lab: Connectionism*, 2025. doi: 10.64434/tml.20250910. https://thinkingmachines.ai/blog/defeating-nondeterminism-in-llm-inference/.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Jaehun Jung, Ximing Lu, Liwei Jiang, Faeze Brahman, Peter West, Pang Wei Koh, and Yejin Choi. Information-theoretic distillation for reference-less summarization. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=JXcXnJJSuL.

A Donald Keedwell and József Dénes. *Latin Squares and Their Applications: Latin Squares and Their Applications*. Elsevier, 2015.

J. Kreer. A question of terminology. *IRE Transactions on Information Theory*, 3(3):208–208, 1957. doi: 10.1109/TIT.1957.1057418.

Keito Kudo, Haruki Nagasawa, Jun Suzuki, and Nobuyuki Shimizu. A challenging multimodal video summary: Simultaneously extracting and generating keyframe-caption pairs from video. *arXiv preprint arXiv:2312.01575*, 2023.

Hwanhee Lee, Seunghyun Yoon, Franck Dernoncourt, Doo Soon Kim, Trung Bui, and Kyomin Jung. Vilbertscore: Evaluating image caption using vision-and-language bert. In *Proceedings of the first workshop on evaluation and comparison of NLP systems*, pp. 34–39, 2020.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, 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.

Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. Keyvideollm: Towards large-scale video keyframe selection, 2024. URL https://arxiv.org/abs/2407.03104.

Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Drew Linsley, Dan Shiebler, Sven Eberhardt, and Thomas Serre. Learning what and where to attend. *arXiv preprint arXiv:1805.08819*, 2018.

Jason H Moore. Bootstrapping, permutation testing and the method of surrogatedata. *Physics in Medicine & Biology*, 44(6):L11, 1999.

Lukas Muttenthaler, Jonas Dippel, Lorenz Linhardt, Robert A Vandermeulen, and Simon Kornblith. Human alignment of neural network representations. *arXiv preprint arXiv:2211.01201*, 2022.

Lukas Muttenthaler, Klaus Greff, Frieda Born, Bernhard Spitzer, Simon Kornblith, Michael C Mozer, Klaus-Robert Müller, Thomas Unterthiner, and Andrew K Lampinen. Aligning machine and human visual representations across abstraction levels. *Nature*, 647(8089):349–355, 2025.

Ani Nenkova, Kathleen McKeown, et al. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233, 2011.

OpenAI. Embeddings | openai api. https://platform.openai.com/docs/guides/embeddings, December 2025a. Accessed 18 Dec 2025.

OpenAI. Sora 2 is here. https://openai.com/index/sora-2/, December 2025b. Accessed 18 Dec 2025.

Stefan Palan and Christian Schitter. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Nikhil Parthasarathy, SM Eslami, Joao Carreira, and Olivier Henaff. Self-supervised video pretraining yields robust and more human-aligned visual representations. *Advances in Neural Information Processing Systems*, 36:65743–65765, 2023.

Edwin JG Pitman. Significance tests which may be applied to samples from any populations. *Supplement to the Journal of the Royal Statistical Society*, 4(1):119–130, 1937.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. Is summary useful or not? an extrinsic human evaluation of text summaries on downstream tasks. *arXiv preprint arXiv:2305.15044*, 2023.

Iqra Qasim, Alexander Horsch, and Dilip Prasad. Dense video captioning: A survey of techniques, datasets and evaluation protocols. *ACM Computing Surveys*, 57(6):1–36, 2025.

Hui Sun, Shiyin Lu, Huanyu Wang, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Ming Li. Mdp3: A training-free approach for list-wise frame selection in video-llms. *arXiv preprint arXiv:2501.02885*, 2025.

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

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8579–8591, 6 2025.

Cunjuan Zhu, Qi Jia, Wei Chen, Yanming Guo, and Yu Liu. Deep learning for video-text retrieval: a review. *International Journal of Multimedia Information Retrieval*, 12(1):3, 2023.

APPENDIX

A DATASET SELECTION

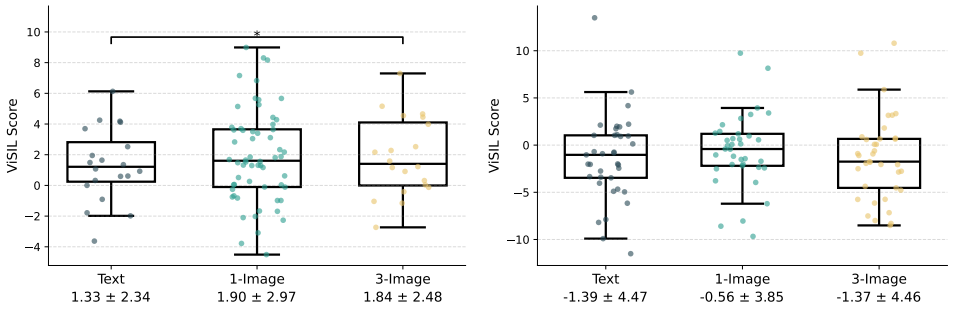
To evaluate the capabilities of our framework in long-context video understanding and reasoning, we utilize two challenging benchmarks: LongVideoBench (Wu et al., 2024) and MVBench (Li et al., 2024). We specifically select tasks that necessitate sustained temporal attention and high-level semantic synthesis rather than simple object recognition.

From LongVideoBench, we focus on the `Sequence of Scenes (SSS)` task, which requires identifying the correct chronological order or relationship between disparate events. From MVBench, we utilize the `Episodic Reasoning (EpR)` task, which tests a model's ability to infer causal links and overarching narratives across extended durations. The details of these subsets are summarized in Table 3.

Table 3: Statistics of Selected Dataset Subsets

Task Subset	Videos	QA Pairs	Multimodal Summaries	Avg. Duration (sec)
LongVideoBench (`SSS`)	37	54	333	207.52 ± 203.16
MVBench (`EpR`)	20	20	180	39.81 ± 13.66

B VISIL DISTRIBUTION SCATTERS



(a) MVBench (b) LongVideoBench

Figure 7: **Distribution of ViSIL Scores.** We visualize the scatter of ViSIL scores on (a) MVBench and (b) LongVideoBench.

As shown in Figure 7, we analyze the distribution of ViSIL scores to verify the metric's discriminative power across different datasets. On both MVBench and LongVideoBench, the scores exhibit consistent distributions and overlapping quartiles across Text, 1-Image, and 3-Image formats. The comparable mean scores suggest that ViSIL is modality-agnostic, evaluating the intrinsic informativeness of a summary rather than its specific format.

C PARTICIPANT RECRUITMENT AND DEMOGRAPHICS

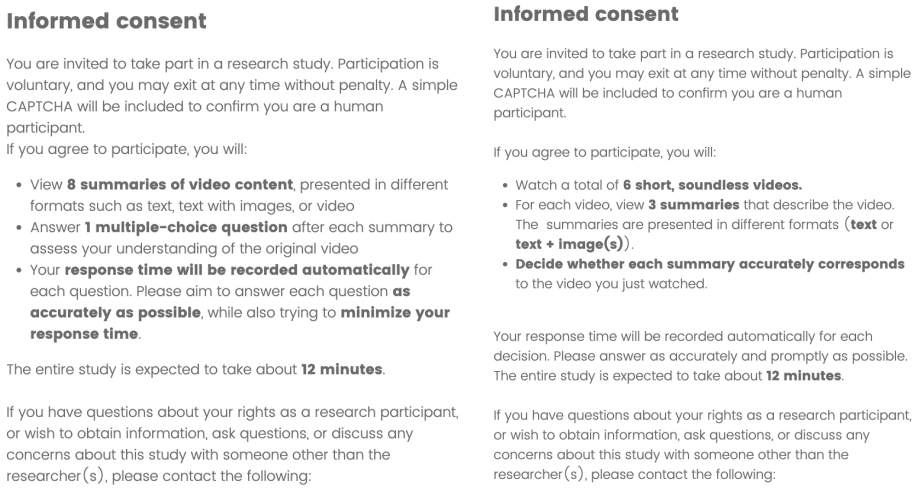
We conducted a Video Question Answering (VQA) test and a correspondence test. Participants were recruited via Prolific (Palan & Schitter, 2018). All were adults (age > 18) with normal or corrected-to-normal vision and English proficiency. Demographic details are summarized in Table 4.

Table 4: Participant Demographics

Study	N	**Mean Age** (SD)	**Gender Distribution**
VQA Test	37	34.32 (11.76)	64.9% Male, 35.1% Female
Correspondence Test	29	34.21 (11.97)	79.3% Male, 17.2% Female, 3.4% Non-binary

D User Study Instruction and Interfaces

All participants provide informed consent before participation and are compensated at a rate consistent with Prolific and institutional standards. The user instructions of the VQA and the correspondence test are presented in Figure 8.



(a) Instructions for VQA test.

(b) Instructions for correspondence test.

Figure 8: User Instructions for the user study.

E Sensitivity of ViSIL to Evaluation Model Choice

We compare ViSIL scores produced by `GPT-4o` and `Gemini-2.0-Flash` to evaluate robustness to evaluator choice on $n = 125$ summaries. Table 5 reports the correlation between scores from the two models. Both Pearson ($r = 0.1867$) and Spearman ($\rho = 0.1863$) correlations are positive and statistically significant ($p < 0.05$), indicating consistent ranking behavior across models despite differences in scale. These results suggest that ViSIL generalizes across evaluation models and does not depend strongly on a specific model choice.

Table 5: Correlation between ViSIL scores computed using GPT-4o and Gemini 2.0 Flash ($n = 125$).

Correlation Metric	Coefficient	p-value
Pearson (r)	0.1867	0.037059
Spearman (ρ)	0.1863	0.037500

E.1 Evaluator comparison across summary formats

We extend the text-only score-agreement check in Appendix E to downstream VQA evaluation across Text-Only, 1-Image, and 3-Image summary formats. We hold the caption source fixed and replace the Gemini 2.0 Flash evaluator with GPT-4o, repeating the logistic analysis on matched observations.

Table 6 reports regression coefficients β, rather than Pearson correlations. The comparison covers 100 summary–question pairs from 20 MVBench videos and 270 pairs from 37 LongVideoBench videos.

Table 6: Downstream VQA association under different evaluators, using matched observations across summary formats.

Dataset	Evaluator	N	Videos	β	p
MVBench	GPT-4o	100	20	-0.154	0.250
MVBench	Gemini 2.0 Flash	100	20	-0.138	0.072
LongVideoBench	GPT-4o	270	37	-0.093	0.088
LongVideoBench	Gemini 2.0 Flash	270	37	-0.043	0.140

All coefficients retain the expected negative sign: lower ViSIL corresponds to better downstream performance. However, none of these four estimates reaches $p < 0.05$. The direction is not specific to the Gemini evaluator, but this experiment does not establish evaluator independence or robust predictive validity.

F Caption-generator ablation

We replace Gemini 2.5 Pro captions with GPT-4o captions, re-extract keywords using the same GPT-5 prompt, and rescore the available summaries with a fixed GPT-4o evaluator (540 scoring calls, no failures). This compares a cross-family captioner–evaluator pairing with a same-family pairing. GPT-4o captions are approximately 29% shorter, so the swap also changes caption detail; it does not isolate caption length as a separate causal factor.

Table 7: Captioner sensitivity with a fixed GPT-4o evaluator. Word and keyword counts are means; downstream associations are Pearson correlations.

Captioner	Words	Keywords	r	95% CI
Gemini 2.5 Pro	292.7	19.1	-0.105	$[-0.221, 0.015]$
GPT-4o	208.4	17.1	-0.096	$[-0.215, 0.027]$

Both captioners yield the same format ordering by ViSIL score: Text-Only $<$ 3-Image $<$ 1-Image. The downstream associations have similar signs and magnitudes, with both confidence intervals including zero. The same-family condition does not exhibit a larger observed association; this check limits, but does not eliminate, concern about captioner–evaluator dependence.

G Hallucination sensitivity

G.1 Controlled Gemini experiment

Hallucinations in the caption proxy C can hinder reliable ViSIL evaluation, as it may contain content not grounded in the source video V. To assess how sensitive ViSIL is to caption hallucination, we conduct an experiment comparing two caption sets of 100 video captions scored by the same `Gemini` model: (i) original captions generated by our pipeline, and (ii) hallucinated variants constructed by injecting factual errors while preserving sentence structure. For example, an original caption describing a *woman* in a *studio* with a *burger* and *kitchen countertop*" is altered to a *man* in a *garage* with a *sandwich* and *bathroom tabletop*." As shown in Table 8, injected factual errors significantly increase ViSIL scores in this controlled Gemini experiment (Wilcoxon $p < 0.001$, t-test $p < 0.001$).

Table 8: ViSIL scores (lower is better) for original vs. hallucinated captions. Original captions receive lower scores, whereas hallucinated ones receive higher scores, indicating that ViSIL penalizes ungrounded content and prefers video-aligned content.

Condition	ViSIL Score (↓) =	$\log P(C\|V)$ −	$\log P(C\|\tilde{V})$
Original Captions	**-3.189**	-11.214	-8.025
Hallucinated Captions	-1.597	-18.469	-16.872
Δ (Hallucinated − Original)	+1.592	-7.255	-8.847

Although we cannot formally prove that hallucinated captions consistently increase the ViSIL score, we observe a clear empirical trend. We hypothesize that multimodal summaries $\tilde{V}$, which contain both text and images, provide a more structured and semantically explicit representation than raw video, making inconsistencies more detectable to VLMs. When hallucinated details appear in C, the conditional likelihood $P(C \mid \tilde{V})$ decreases more sharply ($\Delta = -8.847$) than the corresponding decrease in $P(C \mid V)$ ($\Delta = -7.255$). This asymmetric drop suggests that $\tilde{V}$ penalizes ungrounded content more strongly than raw video. This imbalance explains the observed increase in score for hallucinated captions.

G.2 Cross-family evaluator check

Table 8 is a controlled sanity check under the original Gemini evaluator. To test whether its penalty generalizes across evaluator families, we keep the caption source fixed and replace the evaluator with GPT-4o. The penalty does not robustly persist under this change. Thus, the original result should not be interpreted as a general hallucination-detection guarantee: ViSIL is defined relative to a caption proxy and can inherit errors in that proxy. This additional test was reported qualitatively; no cross-family effect size or significance value is available to quantify the change.

H Direct VLM-as-judge baseline

We evaluate GPT-4o as a direct judge of information preservation. The judge receives 12 uniformly sampled video frames, the candidate summary, and its selected keyframes, and returns a score from 0 to 100. We obtain scores for 285 summaries with no failures and evaluate the resulting scores on the same 370 ViSIL–VQA pairs (270 LongVideoBench and 100 MVBench observations). The number of unique summaries differs from the number of summary–question pairs.

Table 9: Pearson association with VQA correctness for ViSIL and the direct judge. Lower ViSIL and higher judge scores indicate better summaries.

Dataset	Metric	r	95% CI
LongVideoBench	ViSIL	-0.090	$[-0.208,\ 0.029]$
($N = 270$)	VLM judge	$+0.170$	$[0.052,\ 0.284]$
MVBench	ViSIL	-0.183	$[-0.366,\ 0.014]$
($N = 100$)	VLM judge	$+0.034$	$[-0.164,\ 0.229]$

The LongVideoBench judge association is significant under the reported permutation ($p = 0.005$) and within-video ($p = 0.016$) tests. The other three confidence intervals include zero. Agreement between the two metrics is weak overall (Pearson $r = -0.069$, $p = 0.19$) and stronger on MVBench ($r = -0.223$, $p = 0.025$); the negative sign reflects their opposite score polarities. These results do not establish that ViSIL predicts downstream performance better than a direct judge. Its contribution is a common caption-recovery-based score across summary formats, while the judge supplies a prompt-defined rating.

I SCOPE OF SUMMARY-SELECTION EVIDENCE

The summary-generation procedure is fixed, and ViSIL evaluates the resulting candidates; it does not itself generate keyframes. The original Pareto analysis in Figure 3 reports 63% versus 56% VQA accuracy at the compared operating point, a seven-percentage-point difference. No additional held-out-video or independent-dataset validation of this gain, or new matched-budget candidate-generation ablation, is reported here. Computing the score without VQA labels does not by itself establish out-of-sample selection performance. The existing result is an exploratory within-evaluation comparison, and prospective utility remains to be validated. Token counts characterize the evaluated inference setting, not measured wall-clock latency or system-level cost.

J PROMPTS USED

Prompt J.1: Captioning Prompt

You are a video content analyst. Your task is to generate a single, detailed, and objective descriptive paragraph for the provided video file.
Please ensure your description faithfully includes, in chronological order:

- **Setting(s)**: Describe the environment(s) where the video takes place (e.g., indoor/outdoor, specific locations if identifiable).
- **Subjects and Objects**: Highlight the key people, animals, or significant objects. Include details on appearance, clothing, expressions, and notable features.
- **Sequence of Events**: Provide a clear, step-by-step account of the actions and events as they unfold from beginning to end.
- **Key Visual Details**: Note important visual information such as lighting, weather, or significant on-screen elements.
- **OCR (Optical Character Recognition)**: Transcribe any clearly visible and legible text seen in the video (e.g., signs, labels, graphics).

Constraints:

- Output only the descriptive paragraph—no introductions, explanations, or bullet points.
- Do NOT include interpretation, speculation, or information that cannot be directly observed in the video.
- Maintain an objective and neutral tone throughout.

Prompt J.2: Keyword masking prompt

Extract up to 20 keywords from the provided paragraph by selecting the most distinctive descriptors—specifically objects, motions, or events—that are directly relevant to the video content.
Exclude the word “video" as a keyword.
Include each remaining keyword only once, using the original word form as they appear in the paragraph (do not apply stemming or lemmatization), and ensure all keywords are presented as a single lowercase word.
Arrange the keywords sequentially as they appear in the paragraph.
If fewer than 20 suitable keywords are identifiable, return only those present. Avoid duplicates.

Output Format
Output a list of words as a JSON array, for example: [“dog", “jump", “frisbee", “park"]

Prompt J.3: Keyframe selection

You are a video analysis API. Your task is to process the provided video and extract key moments.
Analyze the video and identify the three most significant keyframes that summarize the core action or story. For each keyframe, output a `"timestamp"` that is as precise as possible, using the SMPTE timecode format `"HH:MM:SS:FF"`, where `"FF"` represents the exact frame number within the second (not just to the nearest second). This enables frame-accurate referencing.

Return your response as a valid JSON array of objects. Each object must contain two keys:

- `"timestamp"`: A string of the timestamp in SMPTE timecode `"HH:MM:SS:FF"` format.

- `"description"`: A string containing a brief, neutral description of the scene and its importance.

Do not include any text or explanation outside of the JSON array.

Example Response

```
[
    {
        "timestamp": "00:00:11:15",
        "description": "An intense explosion rocks an industrial
            structure, establishing the scene's chaotic and
            dangerous stakes."
    },
    {
        "timestamp": "00:00:16:03",
        "description": "A character in a tactical vest braces for
            impact inside an elevator, grounding the action with a
            human perspective."
    },
    {
        "timestamp": "00:00:20:15",
        "description": "A man in a white parka approaches a massive
            , high-tech vault, revealing the objective of the
            sequence."
    }
]
```

Prompt J.4: Summary generation prompt

You are a visual narrative analyst. You will be provided with:

- A Video File: The complete motion clip.
- Keyframes (Optional): A sequence of $N \leq 3$ static images extracted from the video..

Your task is to write a concise, narrative summary. Use the video file as your primary source to understand the motion, transitions, and actions that happen between the static keyframes. The keyframes serve as the fixed anchor points for your narrative.

Guidelines:

If **keyframes are provided** ($N \geq 0$):

- Refer to each keyframe chronologically using placeholders: [KEYFRAME1], [KEYFRAME2], ..., [KEYFRAMEn].
- Your text must form the narrative "glue." Describe only the essential actions leading up to, between, and following the keyframes.
- Do not describe the visual content of the keyframes themselves. The placeholders represent those visual moments; your role is to explain the transitions connecting them.
- Example structure: "The clip opens with [action] leading to [KEYFRAME1]... [action between frames]... resulting in [KEYFRAMEn]..."

If **no keyframes are provided** ($N = 0$):

- Describe the essential actions, transitions, or scene changes in chronological order to form a coherent story from start to finish.

Constraints:

- The output must be a short paragraph of 2-3 sentences.
- Your output should be the summary text ONLY.

Prompt J.5: Distractor generation prompt

I will give you a correct summary of a video. Do not modify the first sentence. Extract the key facts (actor, action, object, location, event order). Then create 3 plausible distractor summaries that differ from the correct summary by exactly one or two factual changes. Use only these modification types: Attribute change (color, number, size, timing), Actor or location change, Event order change. Do not change facts related to purpose or causal information. Keep each distractor fluent, realistic, and similar in length and style to the original. Do not introduce impossible or absurd details. Do not repeat

the original summary. Output only the {total_distractor_num} distractor summaries, formatted as a {format}. Do not include labels, explanations, or numbering—just the array. Do not ask any questions.

Prompt J.6: VLM correspondence test prompt

Please watch the video carefully. After the video, you will see several summaries. Each summary may include text, images, or both.
For each summary, decide whether it correctly corresponds to the video you just watched. Consider the summary as a whole (text + images together). The alignment between the text and the images within a summary is not important; judge whether the overall summary matches what happened in the video.
If a summary does not correspond to the video, this may be because the keyframe image does not match the video, or because the text contradicts the video—for example, differences in order of events, colors of objects, or locations.
Please answer as accurately as possible. Also, give a confidence score between 1 and 5 where 1 is the lowest confidence and 5 is the highest confidence.

Prompt J.7: ViSIL score computation

You are asked to recover masked words that describe the content of a video. The input varies depending on the summary modality.

- **For video:** Given the video frames
- **For 3-image summary:** Given the three keyframe images [KEYFRAME1], [KEYFRAME2], and [KEYFRAME3] extracted from a video in the correct sequential order and a textual TLDR describing a video: {summary}.
- **For 1-image summary:** Given the single keyframe image [KEYFRAME1] extracted from a video and a textual TLDR describing a video: {summary}.
- **For text-only summary:** Given a textual TLDR describing a video: {summary}.

Additionally, you are given the masked caption of the video: {masked_caption}.

Task Guess all [MASK] words originally representing any words describing the video, e.g., first_guess second_guess. Return only the answers, without any explanation. Do not use quotes or commas; separate tokens with a single space.

Prompt J.8: VQA evaluation

Given {input_format}, answer concisely using only the provided information.
Textual description of the video: {summary}
Respond only with the letter corresponding to the correct option (A, B, C, or D). Do not include any other symbol or text.
Question: {question}
Options: {options}