VIBE: Annotation-Free Video-to-Text Information Bottleneck Evaluation for TL;DR

Shenghui Chen,* **Po-han Li**,* **Sandeep Chinchali**, **Ufuk Topcu**
The University of Texas at Austin
{shenghui.chen, pohanli, sandeepc, utopcu}@utexas.edu

Abstract

Many decision-making tasks, where both accuracy and efficiency matter, still require human supervision. For example, tasks like traffic officers reviewing hour-long dashcam footage or researchers screening conference videos can benefit from concise summaries that reduce cognitive load and save time. Yet current vision-language models (VLMs) often produce verbose, redundant outputs that hinder task performance. Existing video caption evaluation depends on costly human annotations and overlooks the summaries' utility in downstream tasks. We address these gaps with **V**ideo-to-text **I**nformation **B**ottleneck **E**valuation (VIBE), an annotation-free method that scores VLM outputs using two metrics: *grounding* (how well the summary aligns with visual content) and *utility* (how informative it is for the task). VIBE selects from randomly sampled VLM outputs by ranking them according to the two scores to support effective human decision-making. Human studies on `LearningPaper24`, `SUTD-TrafficQA`, and `LongVideoBench` show that summaries selected by VIBE consistently improve performance—boosting task accuracy by up to 61.23% and reducing response time by 75.77% compared to naive VLM summaries or raw video. [2]

1 Introduction

Efficiently extracting relevant information from extensive video is a major bottleneck for human decision-making, where both accuracy and efficiency matter [1–4]. Tasks demanding human supervision, such as a traffic officer analyzing hours of dashcam footage to determine fault or a researcher distilling key insights from a lengthy oral presentation, are often limited by the time and cognitive load required to process raw video streams. In this work, we aim to improve the *quality* and *brevity* of video summaries to boost human task performance compared to existing vision-language model (VLM) outputs and raw video, especially for longer clips where summarization offers greater utility.

Existing video caption evaluation metrics, however, rely heavily on reference-based comparisons to human-annotated summaries [5–8]. These metrics face two main issues. First, these works require human annotators to watch video clips and write gold-standard captions, which contradicts the goal of reducing human response time and limits generalization to unseen video clips. Second, they are oblivious to downstream tasks and fail to measure how well captions support the tasks.

We propose **V**ideo-to-text **I**nformation **B**ottleneck **E**valuation (VIBE), an annotation-free method for selecting task-relevant video summaries without model retraining. As shown in Figure 1, VIBE defines two metrics—grounding and utility scores—based on the information bottleneck principle [9]. It uses pointwise mutual information to quantify how well a summary reflects video evidence and supports the downstream task. We leverage next-token prediction in VLMs to access the probability

*Equal contribution (Order determined by coin toss).
[2] Project Website, Code, and LearningPaper24 Dataset.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

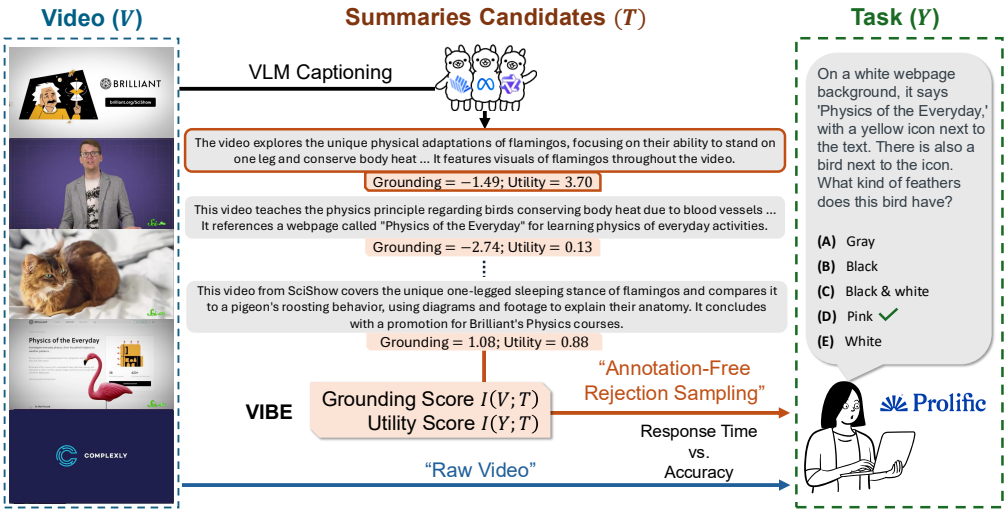


Figure 1: **VIBE for Video-to-Text Summary Selection.** Given a video, a task, and VLM-generated summaries, VIBE ranks the summaries using the proposed grounding and utility scores, which assess video alignment and task relevance. It selects the summary most conducive to helping human users achieve higher task accuracy and lower response time compared to watching the full video.

of generating summaries or task answers. By comparing these probabilities with and without key information, we measure how well one modality (text or video) compensates for missing information in the other to assess grounding and task relevance, as shown later in Figure 2. Using these scores, VIBE selects the most decision-supportive summary for humans from randomly sampled VLM outputs via annotation-free rejection sampling, which filters candidates without human labels for faster task completion than watching full video clips.

To evaluate VIBE, we conduct between-subjects user studies with 243 participants across three datasets—`LearningPaper24` (self-curated), `LongVideoBench` [10], and `SUTD-TrafficQA` [11]—measuring human performance in terms of accuracy, response time, and inverse efficiency score, the ratio of response time to accuracy [12]. Results show that summaries selected with maximal utility score improve task accuracy by up to 40%, while those chosen by maximizing grounding score yield up to 27.6% gains, both on the `LongVideoBench` dataset. VIBE-selected summaries also significantly reduce response time compared to raw video. We also observe a strong positive correlation between utility score and human accuracy, and a strong negative correlation between summary length and response time per word. These patterns highlight the value of concise, relevant information for efficient human decision-making.

Contributions. Our contributions are threefold: (a) We identify the need and propose the problem of annotation-free, task-aware evaluation for video-to-text summarization, improving human response time and accuracy without relying on gold-standard captions. (b) We propose VIBE, an annotation-free evaluation framework that combines grounding and utility scores of video clips to rank and select high-quality summaries from VLM outputs without requiring retraining. (c) We demonstrate through user studies that VIBE summaries significantly boost humans' task accuracy by up to 61.23% and reduce response time by up to 75.77% compared to standard VLM outputs.

Critique and Open Problems. This work reframes the video-to-text evaluation problem through the lens of human decision support, offering an annotation-free, scalable alternative to costly reference-based comparisons. VIBE's ability to score and select summaries without training makes it a practical plug-in for both closed- and open-source VLMs. Looking ahead, VIBE opens several promising directions. One is exploring the joint optimization of summary generation and selection through self-supervised fine-tuning of VLMs. Another is extending beyond human supervision tasks to investigate whether task-aware captions can improve VLM performance on downstream reasoning tasks. These directions point to the broader applicability of VIBE beyond evaluating video summarization.

2 Related Work

Reference-based Video Caption Evaluation. Evaluating video caption quality is crucial for tasks like video question answering [13, 14], text-to-video retrieval [5, 15, 16], and multimodal language model training. The quality of captions greatly impacts the downstream performance of tasks and models. While existing benchmarks compare VLM-generated captions to gold-standard references using metrics like ROUGE [17], BLEU [18], and CIDEr [19], these are costly to curate and focus solely on captions rather than video context. In contrast, we propose VIBE, a new evaluation metric that incorporates video context, requires no gold standard, and can be applied to unseen video-caption pairs. VIBE extends the information-theoretic approach of [20], which uses pointwise mutual information to assess news summaries for language model tuning. We adapt it for video captions and examine their effect on human task accuracy and response time.

Human-Centric Evaluation via Response Time and Accuracy. In the context of human-agent interaction involving natural language [21], traditional human evaluations have focused on assessing summaries based on fluency and informativeness [22, 23], but these methods are often subjective and hard to scale. To address this, we adopt an extrinsic evaluation approach that measures how summaries affect human performance on downstream tasks [24, 25]. Specifically, we evaluate captions based on human response time and accuracy on multiple-choice questions grounded in video content. To account for the speed-accuracy tradeoff [26], we also report the inverse efficiency score [12], which normalizes response time by accuracy.

3 Preliminaries

Mutual Information. Our work heavily relies on the concept of mutual information. For the ease of readers, we briefly introduce it here. Mathematically, for two random variables, X, Z, we can calculate their mutual information:

$$\mathbf{I}(X;Z) = \mathbf{E}\left[\log \frac{\mathbf{P}(X,Z)}{\mathbf{P}(X)\mathbf{P}(Z)}\right] = \mathbf{E}\left[\log \frac{\mathbf{P}(X|Z)}{\mathbf{P}(X)}\right], \quad \text{(Mutual Information)} \tag{1}$$

where the expectation operator is over the joint probability distribution of X and Z.

Intuitively, mutual information measures the reduction in uncertainty about X after observing Z. It is zero if and only if X and Z are independent, meaning knowledge of Z provides no information about X. Higher mutual information indicates stronger dependency between the two variables. In this work, we leverage mutual information to quantify how much the summaries retain relevant information from the raw video clips and how well they support the target downstream task.

Information Bottleneck. The information bottleneck (IB) framework extracts relevant information from input data while compressing irrelevant details. Tishby et al. [9] first introduced it to formalize the trade-off between accuracy and complexity in learning systems. Later, researchers applied it to domains such as neuroscience [27], natural language processing [28, 29], and computer vision [30, 31]. To ensure clarity, we modify the notation from [9] to align with our later sections, where V is the raw video input, T is the summary of the video, and Y is the downstream task of our interest.

We now introduce the IB framework, which seeks to learn a representation T that discards sensitive and irrelevant information from the input V while preserving information useful for predicting the target task Y. IB relies on minimizing the mutual information between input V and a compressed representation T, while preserving as much information as possible about a target variable Y:

$$\min_{\mathbf{P}(t|v)} \underbrace{\mathbf{I}(V;T)}_{\text{compression}} - \beta \underbrace{\mathbf{I}(T;Y)}_{\text{informativeness}}, \quad \text{(Information Bottleneck)} \tag{2}$$

where $\mathbf{I}(\cdot\,;\cdot)$ denotes mutual information, and $\beta \geq 0$ controls the trade-off between compression and prediction. The optimization variable is the conditional distribution $\mathbf{P}(t|v)$, which defines a stochastic encoder that maps $v \in V$ to $t \in T$. The IB principle in eq. (2) formalizes the trade-off between compression and prediction. Minimizing $\mathbf{I}(V;T)$ encourages stronger compression, while maximizing $\mathbf{I}(T;Y)$ promotes the preservation of task-relevant information. The parameter β balances these two competing objectives: a higher β prioritizes retaining more information about Y, while a lower β encourages stronger compression of V.

Researchers use the IB principle to analyze generalization, regularization, and the role of hidden representations of neural networks [32, 33], where V is the input of neural networks, T is the output of intermediate layers, and Y is the classification label. It also drives recent advances in representation learning, where models aim for compact, informative encodings [34, 35].

4 Video-to-Text Information Bottleneck Evaluation (VIBE)

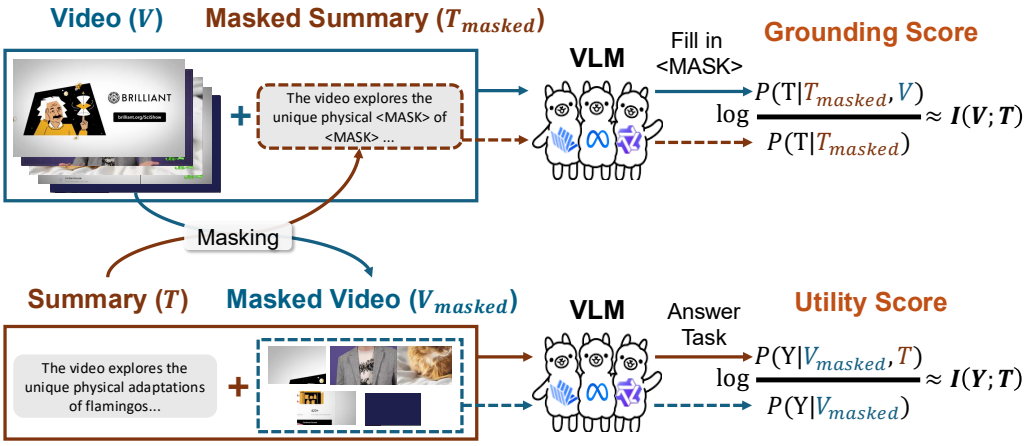


Figure 2: **Computing VIBE Scores via Masked Inference.** VIBE estimates grounding and utility scores using the next-token prediction mechanism of VLMs. The grounding score measures how well the video helps reconstruct a masked summary, while the utility score captures how much the summary improves task prediction given a masked video.

We now formally describe our video-to-text information bottleneck evaluation (VIBE) method. Inspired by the IB formulation in eq. (2), we adapt its two terms for the purpose of evaluating video-to-text summary quality. We denote the raw video clips that are fed into the VLMs as V, the summaries from the VLM response as T, and the task target we care about as Y. For instance, let V be a video of a paper presentation, T be the summary of the presentation, and the task Y be to determine the primary area to see if it is of one's interest.

Grounding Score. In eq. (2), the mutual information $\mathbf{I}(V;T)$ between raw data and representation is the compression term. In our setting, we reinterpret it as the *grounding score*, which quantifies how grounded the textual summary is to the video. The higher this score is, the more closely the summary is anchored to the video.

However, one cannot access the joint and marginal probability distributions over the occurrences of video clips or text summaries. Therefore, we can only calculate the empirical mutual information $\log \frac{\mathbf{P}(X,Z)}{\mathbf{P}(X)\mathbf{P}(Z)}$ without the expectation operator to approximate the $\mathbf{I}(V;T)$ term. Literature refers to the empirical term without the expectation as pointwise mutual information [36], which we denote as $\mathbf{I}^P(\cdot;\cdot)$. Unlike mutual information, which must be non-negative as it is the expectation of pointwise mutual information, pointwise mutual information ranges from $[-\infty, \infty]$.

We now explain the approximation technique in VIBE to calculate the grounding score. Recall from eq. (1), we rewrite pointwise mutual information as:

$$\mathbf{I}^P(V;T) = \log \frac{\mathbf{P}(T|V)}{\mathbf{P}(T)} \approx \log \frac{\mathbf{P}(T|V, T_{\text{masked}})}{\mathbf{P}(T|T_{\text{masked}})}. \text{ (Grounding Score)} \tag{3}$$

We condition both the numerator and denominator with T_{masked}, a masked version of the text, to approximate $\mathbf{I}^P(V;T)$. The approximation holds if T_{masked} is independent of both the video V and the original text T. Masking key content makes T_{masked} effectively uninformative and thus independent. We put the detailed derivation in Appendix I. Following Jung et al. [20], we use tf-idf [37] to identify keywords and replace them with a "<MASK>" token. This masking helps ensure that eq. (3) reflects the true dependency between V and T, rather than exploiting shortcuts from the text

itself. Thus, the grounding score measures how much the TL;DR remains genuinely grounded in the video beyond what the masked text alone would suggest.

To compute eq. (3), we run two separate inferences with the same VLM to define the ratio of two conditional probabilities $\mathbf{P}(\cdot|\cdot)$. The first, $\mathbf{P}(T|V, T_{\text{masked}})$, is the product of next-token prediction probabilities when reconstructing the masked text given both the video and the masked input. The second, $\mathbf{P}(T|T_{\text{masked}})$, is computed similarly but without providing the video. Only by conditioning on the masked input can we leverage the VLM's next-token prediction probabilities to estimate these likelihoods. By dividing these two conditional probabilities and taking the logarithm, we approximate the pointwise mutual information, allowing us to estimate the grounding score.

Utility Score. The second term in eq. (2), the informativeness term, is the mutual information $\mathbf{I}(T; Y)$ between the representation and the target task. We refer to it as the *utility score*, which measures how useful the summary is for solving the downstream task. A higher utility score indicates that the summary preserves more information relevant to predicting or completing the target task. Similar to the grounding score, directly computing $\mathbf{I}(T; Y)$ is intractable because we cannot access the true probability distributions. Again, we approximate it using the pointwise mutual information with conditional probabilities:

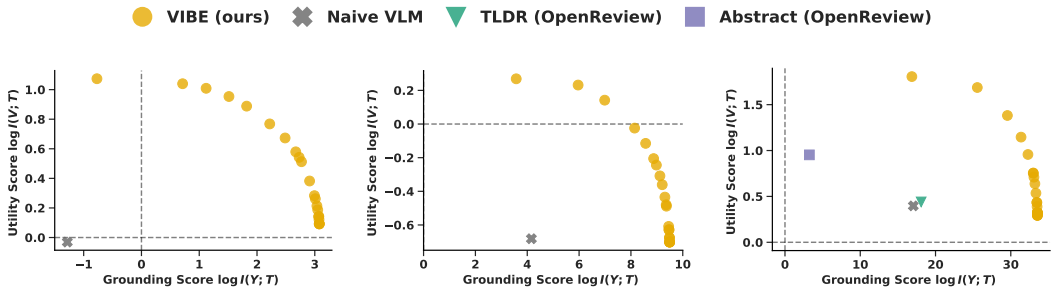
$$\mathbf{I}^P(T; Y) = \log \frac{\mathbf{P}(Y|T)}{\mathbf{P}(Y)} \approx \log \frac{\mathbf{P}(Y|T, V_{\text{masked}})}{\mathbf{P}(Y|V_{\text{masked}})}. \text{ (Utility Score)} \tag{4}$$

Similar to the grounding score in eq. (3), V_{masked} denotes video clips with key information removed to reduce their direct informativeness about both the summary T and the task label Y. To compute the utility score in eq. (4), we compare two model inferences: one conditioned on both the summary T and the masked video V_{masked}, and the other on V_{masked} alone. Since both use the same masked visual input, any improvement in predicting Y must come from the information provided by T. Intuitively, if a summary is useful, it should compensate for the missing content in the masked video and help recover task-relevant information. Thus, the utility score quantifies how well the summary restores information necessary for predicting the task, directly measuring the summary's contribution to downstream decision-making.

Rejection Sampling over Grounding and Utility Scores. We now describe how VIBE optimizes grounding and utility scores to select more informative and task-relevant video-to-text summaries for humans, as shown in Figure 1. Given multiple summaries T sampled from the VLM's output distribution $P(T|V)$, VIBE selects the candidate that maximizes a weighted sum of the two pointwise mutual information terms:

$$\arg\max_{T \sim P(T|V)} \alpha\, \mathbf{I}^P(V; T) + \beta\, \mathbf{I}^P(T; Y). \text{ (TL;DR Selection with VIBE)} \tag{5}$$

Here, α, β are hyperparameters controlling the trade-off between grounding and utility in eq. (5). A higher α emphasizes alignment with video, favoring summaries that faithfully reflect the video, while a higher β prioritizes task relevance, selecting summaries that improve downstream performance.



(a) `LongVideoBench` (b) `SUTD-TrafficQA` (c) `LearningPaper24`

Figure 3: **Pareto front of VIBE.** Summaries selected by VIBE form a Pareto frontier across different (α, β) per Equation (5), demonstrating optimal trade-offs between grounding and utility. In contrast, Naive VLM summaries, author-written TLDRs, and abstracts from OpenReview in `LearningPaper24` fall inside the frontier, indicating suboptimal scores.

Our formulation differs from traditional IB, where compression serves as a regularizer during model training to mitigate overfitting. In contrast, VIBE does not train a model—it selects a summary. Overfitting is therefore not a concern, and maximizing both mutual information terms is desirable to produce summaries that are both useful and faithful. To this end, we consider $\alpha, \beta \geq 0$ to jointly maximize both scores. Then, to explore potential trade-offs between the two scores, we perform a convex combination sweep over $\alpha \in \{0, 0.05, 0.1, \ldots, 1.0\}$, with $\beta = 1 - \alpha$. Each (α, β) pair is applied consistently across all datasets. This linear scalarization technique, commonly used in multi-objective optimization [38], produces the convex ("Pareto curve") shown in Figure 3, empirically revealing the inherent trade-off between grounding and utility scores.

Our results based on grounding and utility scores across sampled summaries from multiple datasets reveal a consistent trade-off between the two objectives. As shown in Figure 3, summaries selected by VIBE with varying α and β form a Pareto front, capturing the inherent trade-off between the two VIBE scores. In contrast, randomly sampled summaries (*Naive VLM*) consistently fall inside this Pareto curve across datasets, indicating suboptimal grounding and utility. We observe this pattern when evaluating the two VIBE scores of VLM-generated summaries on tasks from `LearningPaper24`, `SUTD-TrafficQA`, and `LongVideoBench`. Summaries with high grounding scores tend to mirror the video content closely but often include redundant or irrelevant details for the task. In contrast, summaries with high utility scores boost task accuracy and response efficiency but may overlook important visual context. This trade-off reflects the core tension in the IB framework. By tuning α and β, VIBE adjusts this balance to match the needs of different downstream applications. We detail the experimental settings in the next section.

As shown empirically in the experiments, higher utility scores correlate with higher human task accuracy. Thus, one can use VIBE to select summaries with higher utility scores and present only the most informative summaries to humans, filtering out less useful ones. Notably, calculating the utility score requires access to task labels Y, but not gold-standard, human-annotated labels for summary T as in previous works. When task labels are unavailable, VIBE can select summaries based on the highest grounding score, an unsupervised measure of alignment between the video and the generated summaries. Although the grounding score alone yields lower human performance than the utility score, our user study shows it still improves task accuracy compared to unfiltered VLM summaries.

From Evaluation to Real-World Use. VIBE evaluates the quality of summaries only from the VLM perspective, but its connection to human decision-making performance remains unanswered solely by the formulation. To bridge this gap, we conduct user studies measuring how different summary qualities, as evaluated by VIBE, impact human decision-making performance in Section 5. VIBE relies solely on text generation and next-token probability access, using black-box access to VLMs like the OpenAI API [39]. It enables VIBE to work with both closed- and open-source VLMs and scale well to real-world scenarios.

5 Experiment

We conduct user studies across three diverse datasets to evaluate how VIBE-generated summaries support human decision-making. The study measures participants' accuracy and response time in answering questions about video content. Here, we use Qwen2.5-VL-72B-AWQ [40] to evaluate VIBE for all datasets. A representative qualitative example from the study is provided in Appendix A. We also show that VIBE works across various VLMs through ablation studies.

5.1 Datasets

For evaluation, we select three datasets varying in domain, duration, and task nature.

LearningPaper24 We introduce `LearningPaper24`, a curated dataset of 2,287 video presentations from ICLR 2024 and NeurIPS 2024, filtered based on key criteria: a valid OpenReview ID, an accessible SlidesLive video, an author-provided TL;DR and abstract, and a clearly defined primary area. Details on the curation process and dataset statistics are provided in Appendix C. The task associated with this dataset is to identify the primary area of each paper from 12 options. See instruction details in Appendix E.

LongVideoBench & SUTD-TrafficQA `LongVideoBench` [10] features long instructional clips with QA pairs for extended reasoning, while `SUTD-TrafficQA` [11] contains short traffic clips with multiple-choice questions on causal and temporal understanding. Participants answer one question per each `LongVideoBench` clip; four questions per each `SUTD-TrafficQA` clip.

Preprocessing In all datasets, we mask text by removing keywords with high tf-idf scores. For `LearningPaper24`, which features slide videos, we extract and mask keywords in the slide using EasyOCR [41]. For the other two datasets, we apply random $1/16$ cropping to all video frames. We use a subset of each dataset for VIBE evaluation and user studies (details provided in Appendix D).

5.2 User Study Design

For each dataset, 10 video stimuli with multiple-choice questions are shown in randomized order. We adopt a *between-subjects design* where participants are assigned to one of the four conditions listed below, and the format in which the stimulus is presented varies by the condition.

Independent Variables. Participants are assigned to one of four conditions. In the **Video Only** condition, they watch the original video without text. In the remaining conditions, they view only a VLM-generated summary: a randomly selected VLM summary (**Naive**), the top-ranked summary from k response candidates[3] by utility score (**Max-U**) or by grounding score (**Max-G**).

Dependent Measures. We report three metrics: **accuracy**, the proportion of correct responses across 10 stimuli; **response time**, the time (in seconds) spent reading or watching each stimulus and answering its corresponding questions; and **inverse efficiency score (IES)**, the ratio of response time to accuracy, to account for speed–accuracy trade-offs [12].

Hypotheses. We evaluate the following hypotheses: (**H1**) Participants in the Max-U and Max-G will achieve higher accuracy than those in Naive VLM. (**H2**) Participants in the Max-U and Max-G will respond more quickly than those in the Video Only. (**H3**) Max-U and Max-G will yield lower (i.e., more efficient) IES scores than Video Only, reflecting a better speed–accuracy trade-off.

Participants. We recruit 243 participants across three datasets: 92 for `LearningPaper24`, 82 for `SUTD-TrafficQA`, and 69 for `LongVideoBench`. For `LearningPaper24`, participants are primarily CS graduate students or prescreened degree holders on Prolific [42]; participants for the other datasets are recruited generally. The average age is 37.59 ± 11.06 years, with a gender distribution of 63.37% male, 35.80% female, and 0.82% non-binary. Further details are provided in Appendix F.

5.3 User Study Results and Analysis

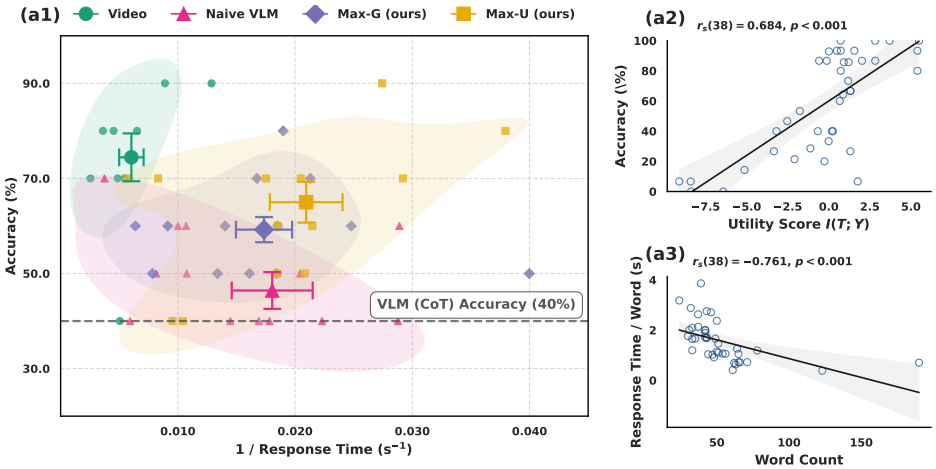
We assess statistical significance using the independent t-test, reporting the t-statistic $t(df)$ (with degrees of freedom df), the significance level p, and the effect size measured by Cohen's d.

On H1 (Accuracy). As shown in Table 1 and (a1, b1) of Figure 4 (with Max-U in yellow and Max-G in purple above Naive VLM in pink), both Max-U and Max-G consistently outperform the Naive VLM across all datasets, confirming **H1**. Max-U achieves the largest gains, with statistically significant improvements over Naive VLM in `LearningPaper24` ($t(36) = 2.486, p = 0.009, d = 0.806$), `LongVideoBench` ($t(26) = 3.215, p = 0.002, d = 1.261$), and `SUTD-TrafficQA` ($t(25) = 3.311, p = 0.001, d = 1.325$), likely due to its focus on task-relevant content that helps users quickly locate key information. However, computing the utility score requires task labels, which may limit its general applicability and scalability. Max-G, while yielding smaller gains, still significantly outperforms the baseline (e.g., `LearningPaper24`: $t(33) = 1.750, p = 0.045, d = 0.594$; `LongVideoBench`: $t(25) = 2.691, p = 0.006, d = 1.077$; `SUTD-TrafficQA`: $t(28) = 1.759, p = 0.045, d = 0.667$). Crucially, calculating the grounding score is fully self-supervised and does not require task labels, making it an alternative that still delivers meaningful accuracy gains.

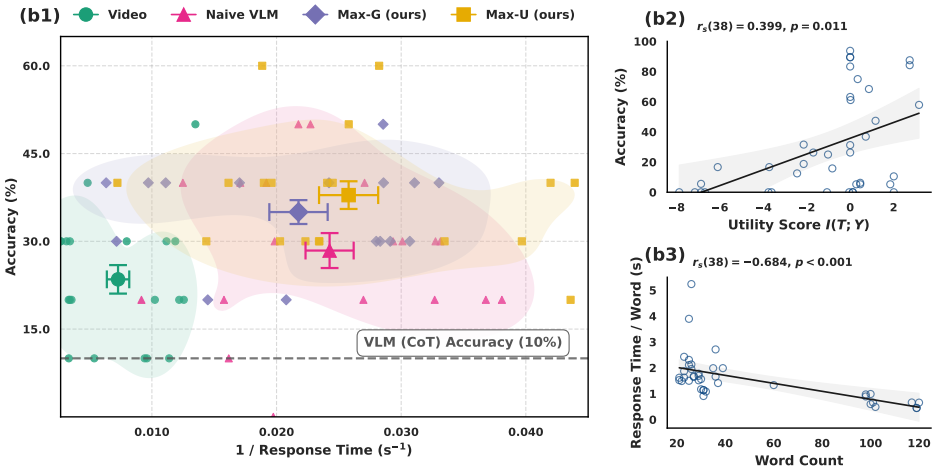
[3] $k = 5$ in this study, and the k responses are generated with various temperatures.

Dataset	Duration (s)	Metric	IV Conditions			
			Video Only	*Naive*	*Max-G (ours)*	*Max-U (ours)*
`LearningPaper24`	250–325	Acc ↑	23.50 ± 10.62	28.42 ± 12.68	35.00 ± 7.91	**37.89 ± 10.04**
		RT ↓	192.73 ± 108.01	47.48 ± 20.31	61.13 ± 39.33	**46.69 ± 25.06**
		IES ↓	9.85 ± 6.63	1.81 ± 1.51	1.85 ± 1.20	**1.28 ± 0.66**
`LongVideoBench`	221–489	Acc ↑	**74.44 ± 14.23**	46.43 ± 13.94	59.23 ± 9.17	65.00 ± 15.47
		RT ↓	202.35 ± 87.26	86.50 ± 63.33	71.93 ± 35.96	**65.86 ± 41.34**
		IES ↓	2.93 ± 1.48	1.83 ± 1.01	1.25 ± 0.67	**1.14 ± 0.80**
`SUTD-TrafficQA`	2–10	Acc ↑	82.86 ± 5.08	76.96 ± 6.89	80.47 ± 3.21	**84.81 ± 4.65**
		RT ↓	**48.63 ± 21.30**	79.46 ± 25.77	80.97 ± 41.11	137.27 ± 81.64
		IES ↓	**0.59 ± 0.27**	1.05 ± 0.37	1.01 ± 0.52	1.66 ± 1.02

Table 1: Human performance (mean ± standard deviation) across three datasets under different IV conditions (detailed in Section 5.2). Metrics: Accuracy (Acc, %, higher is better), Response Time (RT, seconds, lower is better), and Inverse Efficiency Score (IES = RT/Acc, lower is better). Bolded values indicate the best performance among the IV conditions for each metric.



(a) `LongVideoBench`



(b) `LearningPaper24`

Figure 4: **(a1, b1)** Accuracy versus inverse response time. Each point represents an individual participant; large markers indicate group means with standard error of the mean. Shaded areas denote 2D kernel density estimates (threshold = 0.45). **(a2, b2)** Correlation between accuracy and utility score. **(a3, b3)** Correlation between response time per word and word count.

On H2 (Response Time). Table 1 shows that both Max-U and Max-G significantly reduce response times compared to the Video baseline in `LearningPaper24` ($t(37) = -5.599, p < .001, d = -1.794$) and `LongVideoBench` ($t(34) = -4.503, p < .001, d = -1.510$). However, this trend does not extend to `SUTD-TrafficQA`, where response time shows no consistent improvement. It is not surprising, as the significantly shorter video durations (2–10 seconds, compared to over 3 minutes in the other datasets) inherently limit the benefits of summarization. These results align with the expectation that VIBE is most effective in reducing cognitive load for long video clips.

On H3 (IES). As shown in Table 1, Max-U significantly outperform the Video Only baseline in `LearningPaper24` ($t(37) = -5.458, p < .001, d = -1.749$) and `LongVideoBench` ($t(21) = -3.592, p = .001, d = -1.617$), with Max-G showing similar gains (`LearningPaper24`: $t(34) = -4.631$, $p < .001$, $d = -1.553$; `LongVideoBench`: $t(20) = -3.421$, $p = .001$, $d = -1.567$). Furthermore, Max-U also outperform Naive VLM ($t(26) = -1.940, p = .032, d = -0.761$) in `LongVideoBench`. These results highlight VIBE's effectiveness in settings where processing full video clips is costly. In contrast, `SUTD-TrafficQA` shows limited IES improvement: despite Max-U's accuracy improvements over Naive VLM and Video Only, the brevity of the clips and the visual nature of fine-grained actions reduce the benefit of text-based summaries.

Correlation Analyses. We explore the correlation between utility score, grounding score, word count, and human performance (accuracy and response time) using Spearman's rank coefficient [43], which is robust to outliers and non-linear relationships. Full plots are provided in Appendix G. Two consistent trends hold across datasets: First, utility score positively correlates with accuracy: strongly in `LongVideoBench` ($r_s = 0.684$, $p < .001$), and moderately in `LearningPaper24` ($r_s = 0.399$, $p = .011$) and `SUTD-TrafficQA` ($r_s = 0.463$, $p = .004$). Second, word count is strongly negatively correlated with response time per word (`LongVideoBench`: $r_s = -0.761$, `LearningPaper24`: $r_s = -0.684$, `SUTD-TrafficQA`: $r_s = -0.658$, all $p < .001$), implying that longer summaries lead to faster processing per word, possibly due to reduced engagement and more shallow reading.

Key Takeaways. Max-U and Max-G outperform Naive VLMs in accuracy, response time, and inverse efficiency score, with Max-U providing the most significant improvements, especially for longer video clips. Max-G, though offering smaller gains, remains a self-supervised alternative that improves accuracy without task labels. The benefits of summarization are less pronounced in shorter video clips, like those in `SUTD-TrafficQA`, where brief durations limit gains. Correlation analyses show the VIBE utility scores align with higher accuracy, while longer summaries tend to increase unit response time, highlighting the importance of concise, relevant information.

5.4 Ablation over VLM Variants

VIBE scores can inherit bias from the underlying VLMs due to their training data and conditional probability estimates used in mutual information computation. To assess robustness under such model-induced bias, we compare VLMs of varying size and source. As shown in Appendix B, different backbones produce consistent IB score trends and scales. This consistency suggests that lightweight models can efficiently verify outputs from larger ones. To further reduce bias, one can evaluate a single summary using multiple VLMs—an ensemble-style strategy aligned with mixture-of-experts [44, 45] and model selection techniques [46]. This reflects the intuition that good outputs are often hard to generate, but easy to verify.

Limitations. Despite being training-free, VIBE requires multiple VLM inferences for grounding score evaluation, with the number of masked tokens influencing inference steps. Its pointwise mutual information estimate may be biased by two main factors: (a) the text and video masking strategy, and (b) the conditional probability modeling of the VLMs. For (a), the tf-idf score we use to select keywords for text masking requires hyperparameter tuning, and the optimal masking strategy remains an open question. For (b), VLMs introduce bias through their training data and modeling assumptions in computing eq. (5). Using separate models for generation and evaluation, as discussed in Section 5.4, can help mitigate this bias.

6 Conclusion

In this work, we introduce VIBE, an annotation-free framework for evaluating and selecting video-to-text summaries to support human decision-making. Unlike traditional caption evaluation metrics that rely on human-written references, VIBE uses information-theoretic scores—utility and grounding—to assess how well a summary supports a downstream task and aligns with video evidence. This enables scalable, task-aware summary selection without the need for retraining or annotations. Through a large-scale user study spanning three diverse datasets, we show that summaries selected by VIBE significantly improve both task accuracy and response time, especially for longer video clips where information overload is more likely.

Acknowledgement

This work was supported in part by the National Science Foundation grants No. CNS-1836900, 2148186, the Office of Naval Research (ONR) under Grant No. N00014-22-1-2254, N00014-24-1-2097, the Defense Advanced Research Projects Agency (DARPA) contract FA8750-23-C-1018, and DARPA ANSR: RTXCW2231110. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] S. Amershi, D. Weld, M. Vorvoreanu, et al. "Guidelines for Human-AI Interaction". In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. CHI '19. Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–13. ISBN: 9781450359702.

[2] A. Holzinger. "Interactive machine learning for health informatics: when do we need the human-in-the-loop?" In: *Brain informatics* 3.2 (2016), pp. 119–131.

[3] D. Xin, L. Ma, J. Liu, et al. "Accelerating Human-in-the-loop Machine Learning: Challenges and Opportunities". In: *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*. DEEM'18. Houston, TX, USA: Association for Computing Machinery, 2018. ISBN: 9781450358286.

[4] A. Holzinger, K. Zatloukal, and H. Müller. "Is human oversight to AI systems still possible?" In: *New Biotechnology* 85 (2025), pp. 59–62. ISSN: 1871-6784.

[5] J. Xu, T. Mei, T. Yao, and Y. Rui. "Msr-vtt: A large video description dataset for bridging video and language". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 5288–5296.

[6] X. Wang, J. Wu, J. Chen, et al. "VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research". In: *The IEEE International Conference on Computer Vision (ICCV)*. Oct. 2019.

[7] W. Chai, E. Song, Y. Du, et al. "AuroraCap: Efficient, Performant Video Detailed Captioning and a New Benchmark". In: *The Thirteenth International Conference on Learning Representations*. 2025.

[8] M. Monfort, S. Jin, A. Liu, et al. "Spoken moments: Learning joint audio-visual representations from video descriptions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021, pp. 14871–14881.

[9] N. Tishby, F. C. Pereira, and W. Bialek. *The information bottleneck method*. 2000. arXiv: physics/0004057 [physics.data-an].

[10] H. Wu, D. Li, B. Chen, and J. Li. "Longvideobench: A benchmark for long-context interleaved video-language understanding". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 28828–28857.

[11] L. Xu, H. Huang, and J. Liu. "Sutd-trafficqa: A question answering benchmark and an efficient network for video reasoning over traffic events". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2021, pp. 9878–9888.

[12] J. T. Townsend and F. G. Ashby. *Stochastic modeling of elementary psychological processes*. CUP Archive, 1983.

[13] D. Xu, Z. Zhao, J. Xiao, et al. "Video Question Answering via Gradually Refined Attention over Appearance and Motion". In: *ACM Multimedia*. 2017.

[14] Y. Zhong, J. Xiao, W. Ji, et al. *Video Question Answering: Datasets, Algorithms and Challenges*. 2022. arXiv: 2203.01225 [cs.CV].

[15] P.-h. Li, Y. Yang, M. Omama, et al. "Any2Any: Incomplete Multimodal Retrieval with Conformal Prediction". In: *arXiv preprint arXiv:2411.10513* (2024).

[16] M. Omama, P.-h. Li, and S. P. Chinchali. "Exploiting Distribution Constraints for Scalable and Efficient Image Retrieval". In: *The Thirteenth International Conference on Learning Representations*. 2025.

[17] C.-Y. Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004, pp. 74–81.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. "Bleu: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[19] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. "Cider: Consensus-based image description evaluation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.

[20] J. Jung, X. Lu, L. Jiang, et al. "Information-Theoretic Distillation for Reference-less Summarization". In: *First Conference on Language Modeling*. 2024.

[21] S. Chen, D. Fried, and U. Topcu. "Human-agent cooperation in games under incomplete information through natural language communication". In: *International Joint Conference on Artificial Intelligence, Human-Centred AI* (2024).

[22] A. Belz and E. Reiter. "Comparing automatic and human evaluation of NLG systems". In: *11th conference of the european chapter of the association for computational linguistics*. 2006, pp. 313–320.

[23] Y. Graham, T. Baldwin, A. Moffat, and J. Zobel. "Can machine translation systems be evaluated by the crowd alone". In: *Natural Language Engineering* 23.1 (2017), pp. 3–30.

[24] A. Nenkova, K. McKeown, et al. "Automatic summarization". In: *Foundations and Trends® in Information Retrieval* 5.2–3 (2011), pp. 103–233.

[25] X. Pu, M. Gao, and X. Wan. "Is summary useful or not? an extrinsic human evaluation of text summaries on downstream tasks". In: *arXiv preprint arXiv:2305.15044* (2023).

[26] R. P. Heitz. "The speed-accuracy tradeoff: history, physiology, methodology, and behavior". In: *Frontiers in neuroscience* 8 (2014), p. 150.

[27] M. Kleinman, T. Wang, D. Xiao, et al. "A cortical information bottleneck during decision-making". In: *bioRxiv* (2023).

[28] P. West, A. Holtzman, J. Buys, and Y. Choi. "Bottlesum: Unsupervised and self-supervised sentence summarization using the information bottleneck principle". In: *arXiv preprint arXiv:1909.07405* (2019).

[29] C. Zhang, X. Zhou, Y. Wan, et al. "Improving the adversarial robustness of NLP models by information bottleneck". In: *arXiv preprint arXiv:2206.05511* (2022).

[30] Gordon, Greenspan, and Goldberger. "Applying the information bottleneck principle to unsupervised clustering of discrete and continuous image representations". In: *Proceedings Ninth IEEE International Conference on Computer Vision*. 2003, 370–377 vol.1.

[31] N. Tishby and N. Zaslavsky. "Deep learning and the information bottleneck principle". In: *2015 IEEE Information Theory Workshop (ITW)*. 2015, pp. 1–5.

[32] K. Kawaguchi, Z. Deng, X. Ji, and J. Huang. "How does information bottleneck help deep learning?" In: *International Conference on Machine Learning*. PMLR. 2023, pp. 16049–16096.

[33] Z. Goldfeld and Y. Polyanskiy. "The information bottleneck problem and its applications in machine learning". In: *IEEE Journal on Selected Areas in Information Theory* 1.1 (2020), pp. 19–38.

[34] S. Mai, Y. Zeng, and H. Hu. "Multimodal Information Bottleneck: Learning Minimal Sufficient Unimodal and Multimodal Representations". In: *IEEE Transactions on Multimedia* 25 (2023), pp. 4121–4134.

[35] R. Islam, H. Zang, M. Tomar, et al. "Representation Learning In Deep RL Via Discrete Information Bottleneck". In: *AISTATS 2023*. May 2023.

[36] G. Bouma. "Normalized (pointwise) mutual information in collocation extraction". In: *Proceedings of GSCL* 30 (2009), pp. 31–40.

[37] G. Salton and C. Buckley. "Term-weighting approaches in automatic text retrieval". In: *Information Processing & Management* 24.5 (1988), pp. 513–523. ISSN: 0306-4573.

[38] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[39] OpenAI. *OpenAI Python Library*. https://github.com/openai/openai-python. Accessed: 2025-04-28. 2025.

[40] S. Bai, K. Chen, X. Liu, et al. *Qwen2.5-VL Technical Report*. 2025. arXiv: 2502.13923 [cs.CV].

[41] JaidedAI. *EasyOCR: Ready-to-use OCR with 80+ supported languages*. https://github.com/JaidedAI/EasyOCR. Accessed: 2025-05-15. 2020.

[42] S. Palan and C. Schitter. "Prolific. ac—A subject pool for online experiments". In: *Journal of Behavioral and Experimental Finance* 17 (2018), pp. 22–27.

[43] C. Spearman. "The proof and measurement of association between two things." In: (1961).

[44] I. Ong, A. Almahairi, V. Wu, et al. "RouteLLM: Learning to Route LLMs from Preference Data". In: *The Thirteenth International Conference on Learning Representations*. 2025.

[45] Y. Zhou, T. Lei, H. Liu, et al. "Mixture-of-experts with expert choice routing". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 7103–7114.

[46] P.-h. Li, O. S. Toprak, A. Narayanan, et al. *Online Foundation Model Selection in Robotics*. arXiv:2402.08570 [cs]. Feb. 2024.

[47] Z. Chen, W. Wang, Y. Cao, et al. *Expanding Performance Boundaries of Open-Source Multimodal Models with Model, Data, and Test-Time Scaling*. 2025. arXiv: 2412.05271 [cs.CV].

[48] J. Zhu, W. Wang, Z. Chen, et al. *InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models*. 2025. arXiv: 2504.10479 [cs.CV].

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Yes.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the last page in the main paper. There is a section of limitations.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

Justification: We do not have theoretical results in this work. We cited all previous theoretical work we gained insights from.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide all details in the appendix and the main paper.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Refer to the supplemental material for code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See the user study section and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: See the user study section and appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The setup is mentioned in the Appendix H.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, we do.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: There is no societal impact of the work performed.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Yes, we will provide comments and datasets to reproduce our results.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Yes. Please refer to Appendix E.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: Yes, we do have the IRB's approval.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: Yes, we use VLMs as a part of our proposed method.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

Appendix

A Qualitative Result

To illustrate differences across summary generation methods, we present a qualitative example. In Table 3, each condition—Naive, Max-U, Max-G, and CoT—produces a distinct response based on the same video input (keyframe thumbnails shown in Figure 5). For better comparison, we also provide the original TL;DR and abstract from the authors on OpenReview in Table 2. We highlight key terms derived from OpenReview metadata and tf-idf analysis by masking them in gray.

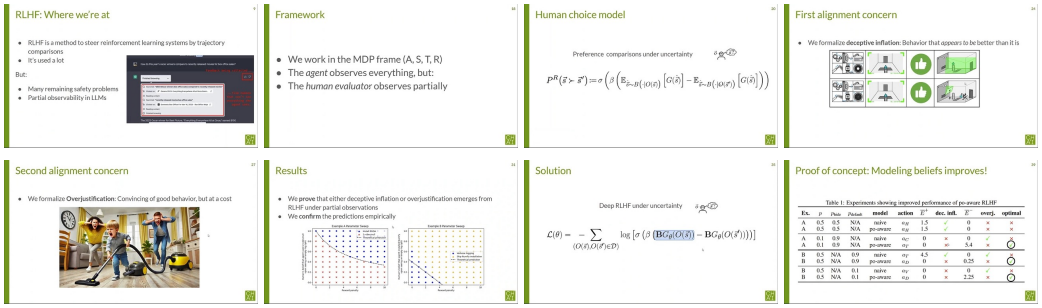


Figure 5: Keyframe thumbnails from the qualitative example SlidesLive talk. The correct answer is *(I) Human-AI Interaction and Ethics (Privacy, Fairness & Safety)*, while the VLM prediction is *(F) Reinforcement Learning & Robotics*.

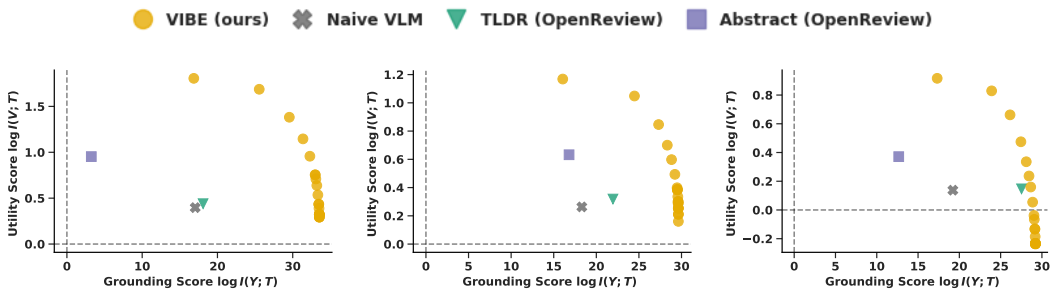
Type	Paper Summary
TL;DR	We study the challenges that arise when learning reward functions with human feedback from partial observations
Abstract	Past analyses of reinforcement learning from human feedback (RLHF) assume that the human evaluators fully observe the environment. What happens when human feedback is based only on partial observations? We formally define two failure cases: deceptive inflation and overjustification. Modeling the human as Boltzmann-rational w.r.t. a belief over trajectories, we prove conditions under which RLHF is guaranteed to result in policies that deceptively inflate their performance, overjustify their behavior to make an impression, or both. Under the new assumption that the human's partial observability is known and accounted for, we then analyze how much information the feedback process provides about the return function. We show that sometimes, the human's feedback determines the return function uniquely up to an additive constant, but in other realistic cases, there is irreducible ambiguity. We propose exploratory research directions to help tackle these challenges and experimentally validate both the theoretical concerns and potential mitigations, and caution against blindly applying RLHF in partially observable settings .

Table 2: Author-provided TL;DR and abstract on OpenReview.

Condition	Summary Response
Naive	This paper explores the challenges of partial observability in reinforcement learning from human feedback, focusing on deceptive inflation and overjustification in AI systems.
Max-U	This research investigates challenges in ensuring AI safety and addressing partial observability through advanced frameworks in human-machine interactions and reinforcement learning.
Max-G	The presentation discusses the challenges of partial observability in reinforcement learning from human feedback, focusing on deceptive inflation and overjustification in AI systems.
CoT	The slides discuss challenges in reinforcement learning from human feedback, particularly focusing on partial observability and the issues of deceptive inflation and overjustification. The content delves into the framework of Markov Decision Processes (MDPs) and the human choice model, which are central to reinforcement learning. The presentation also addresses safety problems and alignment concerns, which are critical in the context of human-AI interaction. Given the focus on reinforcement learning and the interaction between AI and human evaluators, the most suitable area for this paper is one that encompasses both reinforcement learning and the ethical considerations of human-AI interaction.

Table 3: Responses under each condition for the qualitative example, keywords from OpenReview and tf-idf masked in grey.

B Ablation Study on Evaluating with Various VLMs



(a) `Qwen-2.5-72B-AWQ` (b) `InternVL-2.5-8B-MPO` (c) `InternVL-3-38B`

Figure 6: **VIBE generalizes to various VLMs.** Summaries selected by VIBE form a Pareto frontier not only across various (α, β) but also across various VLMs from different sources.

We conduct an ablation study to examine how different VLM variants affect the reliability of VIBE. Specifically, we compare QwenVL-2.5-72B-AWQ [40] (used in the main paper) with InternVL2.5-8B-MPO [47] and InternVL3-38B [48]. These models vary in size, architecture, and pretraining sources. We apply VIBE to the same set of summaries to compute their utility and grounding scores. Results in Figure 6 show consistent VIBE score trends across models, indicating that smaller or alternate VLMs yield similar evaluation signals while reducing inference cost and potential bias. Notably, various VLMs give similar ranges of both utility and grounding scores. The consistency across models further supports the generalizability of VIBE as a training-free framework.

C `LearningPaper24` Dataset Curation

We source the dataset from the public PaperCopilot PaperLists repository, which indexes all accepted papers at ICLR 2024 and NeurIPS 2024. To ensure data quality and completeness, we apply the

following filtering criteria: (1) the paper must have a valid OpenReview ID to guarantee access to full metadata and public reviews; (2) the associated SlidesLive presentation video must be accessible and playable, verified via a headless browser; (3) a TL;DR summary must be present on OpenReview; and (4) a primary area must be specified. After applying these filters, we retain 2,287 entries from an initial set of 2,556.

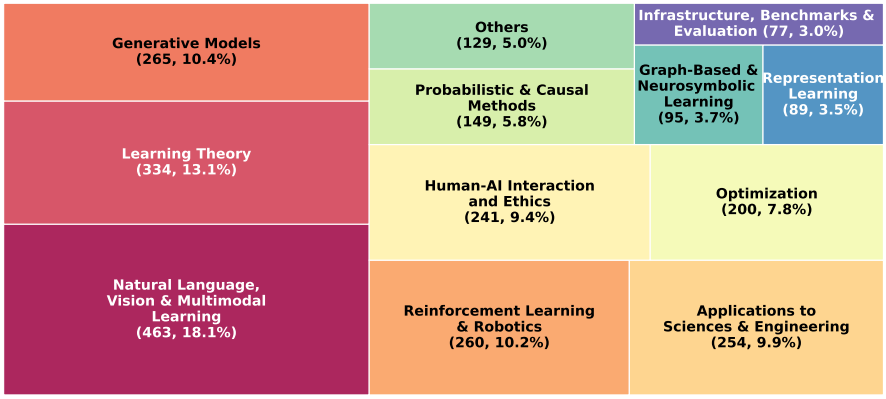


Figure 7: Distribution of papers in `LearningPaper24` by consolidated primary area.

OpenReview's primary areas are often too specific or overlapping, so we remap them into 12 broader and semantically coherent categories. For example, "self-supervised learning" and "representation learning for vision" become *Representation Learning*, while "probabilistic methods" and "causal inference" are grouped as *Probabilistic & Causal Methods*. The final taxonomy is shown in Figure 7. The final dataset includes the following required fields: OpenReview ID, SlidesLive talk ID and URL, TL;DR, abstract, and primary area.

D VIBE Masking and Evaluation

We now report how we select data to calculate the VIBE score in Figure 3. We further select 10 stimuli for our user study.

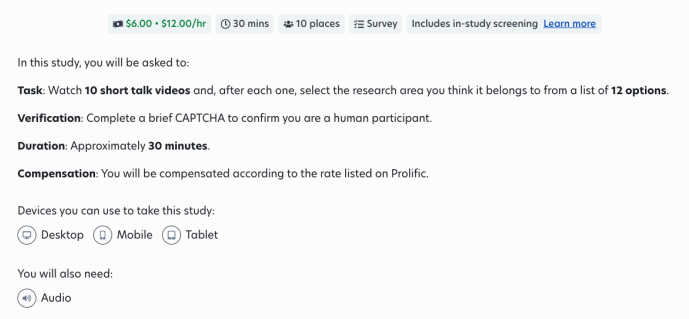
LearningPaper24 We randomly sample 80 papers each from ICLR 2024 and NeurIPS 2024 following the curation process in Appendix C. For each presentation, we generate 5 VLM responses and 1 CoT response. We compute the tf-idf score separately for each of the six response corpora. We discard words and phrases (n-grams with range 1–3) appearing in over 10% of responses and retain only those with tf-idf scores above 0.0025. We also leverage the keywords of papers provided by the authors on OpenReview for masking. We uniformly sample 20 frames from each video as VLM input, with or without random cropping applied.

SUTD-TrafficQA We randomly sample 100 video clips from the dataset. For each video clip, we generate 5 VLM responses and 1 CoT response. All responses, including CoT, form the corpus for tf-idf computation. We discard words and phrases (n-grams with range 1–3) appearing in over 30% of responses and retain only those with tf-idf scores above 0.01. We uniformly select 20 frames from each video as the input of VLM with or without random cropping. We uniformly sample 20 frames from each video as VLM input, with or without random cropping applied.

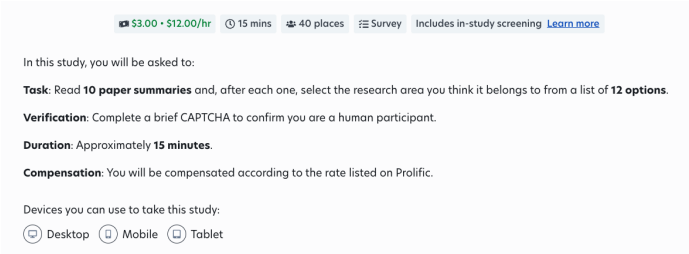
LongVideoBench We sample 150 video clips, each 30 to 500 seconds long, from categories `["E2O", "E2E", "O3O", "S2A", "S2E", "S2O", "S0S"]`. For each clip, we generate 5 VLM responses and 1 CoT response. All responses form the tf-idf corpus. We discard words and phrases (n-grams with range 1–3) appearing in more than 50% of responses and keep only those with tf-idf scores above 0.006. We uniformly select 20 frames from each video as the input of VLM with or without random cropping. We use the LongVideoBench package to obtain 32 frames, which is also uniformly sampling the video, from each video as VLM input, with or without random cropping applied.

E User Study Instruction and Interfaces

All participants provide informed consent before participation and are compensated at a rate consistent with Prolific guidelines and institutional standards. The user interface is presented in Figure 8. Below are the instructions provided to participants for each dataset. Text in parentheses indicates variations between video and text conditions.



(a) Video Only



(b) Text-based (Naive, Max-U, Max-G, CoT)

Figure 8: Prolific recruitment interfaces for `LearningPaper24` dataset showing video condition (top) and text conditions (bottom). Similar interfaces are used for the other two datasets.

Instructions for `LongVideoBench`

In this study, you will (watch/read) **10 (short videos/short summaries)** featuring a variety of everyday content. Topics may include:

- Cooking
- STEM education
- Art and creativity
- Daily vlogs and lifestyle scenes

After each (video/summary), you will answer **one multiple-choice question** to assess your recall and understanding of the content.
Please **avoid leaving the page idle**, as we are also measuring your response time.

Instructions for `SUTD-TrafficQA`

In this study, you'll (watch/read) **10 (short videos/sets of descriptions)** of real traffic videos, which may involve an accident.

After each (video/description), you'll answer **4 multiple-choice questions**. These questions may involve:

- Understanding what happened in the video
- Considering what might have happened under different conditions
- Reflecting on your own interpretation or reasoning

Please **avoid leaving the page idle**, as we are also measuring your response time.

Instructions for `LearningPaper24`

In this study, you will (watch/read) **10 (short research talk videos/paper summaries)**.
For each (video/summary), select the **research area** you think it belongs to from:

(A) Learning Theory
(B) Representation Learning
(C) Generative Models
(D) Optimization
(E) Probabilistic & Causal Methods
(F) Reinforcement Learning & Robotics
(G) Graph-Based & Neurosymbolic Learning
(H) Natural Language, Vision & Multimodal Learning
(I) Human-AI Interaction and Ethics (Privacy, Fairness & Safety)
(J) Applications to Sciences & Engineering
(K) Infrastructure, Benchmarks & Evaluation
(L) Others

Please **avoid leaving the page idle**, as we are also measuring your response time.

F Participant Recruitment and Demographics

Table 4 summarizes demographic information for participants recruited across the three datasets.

Dataset	# Participants	Gender Distribution	Age (years)
`LearningPaper24`	92	70.65% male, 28.26% female, 1.09% non-binary	32.86 ± 8.03
`SUTD-TrafficQA`	82	59.76% male, 40.24% female	40.67 ± 12.69
`LongVideoBench`	69	57.97% male, 40.58% female, 1.45% non-binary	40.22 ± 12.40
Total	243	63.37% male, 35.80% female, 0.82% non-binary	37.59 ± 11.06

Table 4: Demographic breakdown of participants by dataset.

Participants were assigned to IV conditions as shown in Table 5. In addition to the four primary conditions—Max-U, Max-G, Naive, and Video Only (see Section 5.2)—each dataset also includes a group evaluating Chain-of-Thought (CoT) responses. CoT responses are excluded from the scatter plots in the main text due to their fundamentally different generation process: unlike other conditions, CoT has access to answer options of the task, making direct comparisons unfair and potentially misleading. However, we include CoT responses in the correlation analyses, as their grounding and utility scores remain valid for assessing alignment with human performance.

Dataset	Max-U	Max-G	Naive	Video Only	CoT
`LearningPaper24`	19	16	19	20	18
`SUTD-TrafficQA`	14	17	15	16	20
`LongVideoBench`	15	14	15	10	15

Table 5: Number of participants per IV condition across datasets.

G Additional User Study Plots

Figure 9 presents the full set of scatter plots and Spearman correlation results for all datasets. The left panels show scatter plots of accuracy versus inverse response time, while the right panels display correlations between accuracy and utility score, response time and utility score, accuracy and grounding score, response time and grounding score, accuracy and word count, and response time per word and word count.

H Computation Resource

All experiments were conducted on four NVIDIA RTX 6000 Ada GPUs (48GB VRAM) using the vLLM backend. The system was equipped with an Intel(R) Xeon(R) Gold 6346 CPU @ 3.10GHz, featuring 64 cores (x86_64, 64-bit). This setup handled all models—Qwen-2.5-72B-AWQ, InternVL-2.5-8B-MPO, and InternVL-3-38B—efficiently for both generation and evaluation.

I Derivation of Mutual Information Approximation

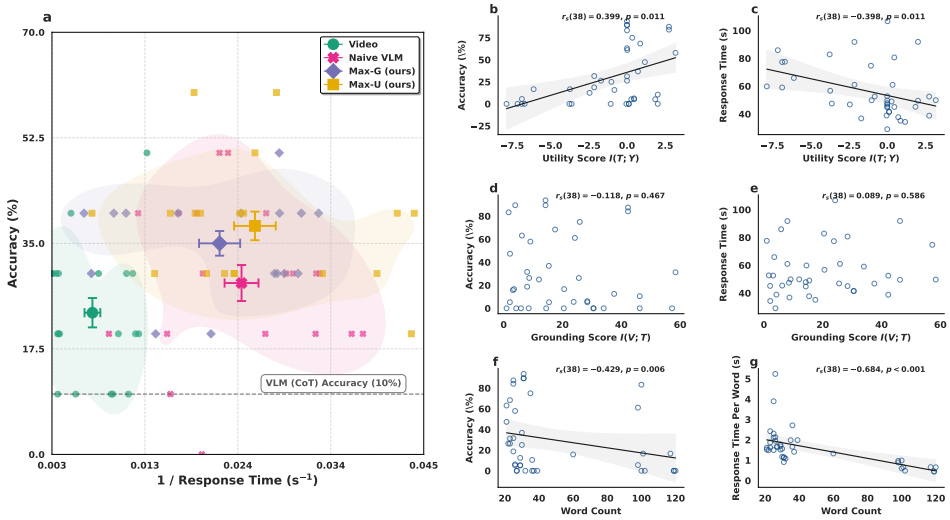
We derive the case where the approximation holds under the stated assumption, starting from the right-hand side of Eq. 3 and Eq. 4. Since both approximations follow nearly identical steps, we present the derivation of grounding score as an example:

$$\log \frac{\mathbf{P}(T \mid V, T_{\text{masked}})}{\mathbf{P}(T \mid T_{\text{masked}})}.$$

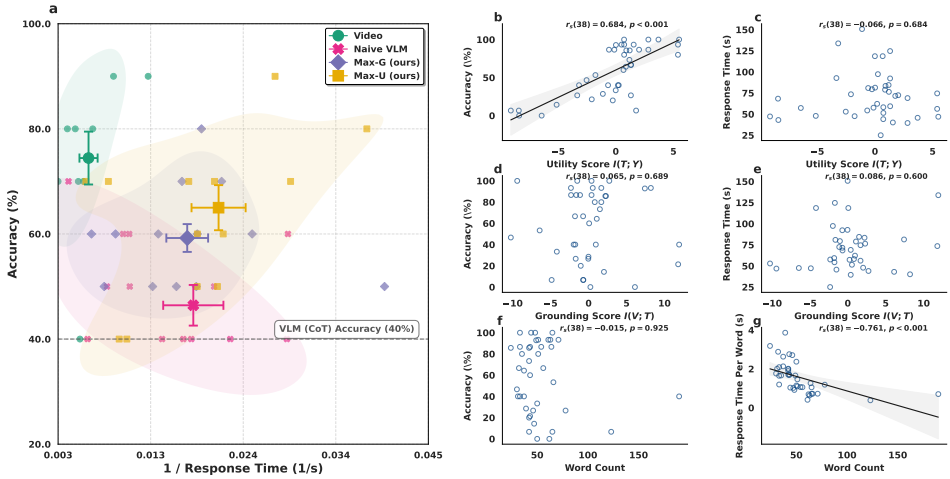
Since T_{masked} is a masked sentence with all keywords removed, we assume it contains no information and is independent from any other random variables. Thus, we can rewrite the expression as:

$$\begin{aligned}
\log \frac{\mathbf{P}(T \mid V, T_{\text{masked}})}{\mathbf{P}(T \mid T_{\text{masked}})} &= \log \frac{\mathbf{P}(T, V, T_{\text{masked}})/\mathbf{P}(V, T_{\text{masked}})}{\mathbf{P}(T, T_{\text{masked}})/\mathbf{P}(T_{\text{masked}})} \\
&= \log \frac{\mathbf{P}(T, V) \cdot \mathbf{P}(T_{\text{masked}})/(\mathbf{P}(V) \cdot \mathbf{P}(T_{\text{masked}}))}{\mathbf{P}(T) \cdot \mathbf{P}(T_{\text{masked}})/\mathbf{P}(T_{\text{masked}})} \\
&= \log \frac{\mathbf{P}(T, V)/\mathbf{P}(V)}{\mathbf{P}(T)} \\
&= \log \frac{\mathbf{P}(T \mid V)}{\mathbf{P}(T)}.
\end{aligned}$$

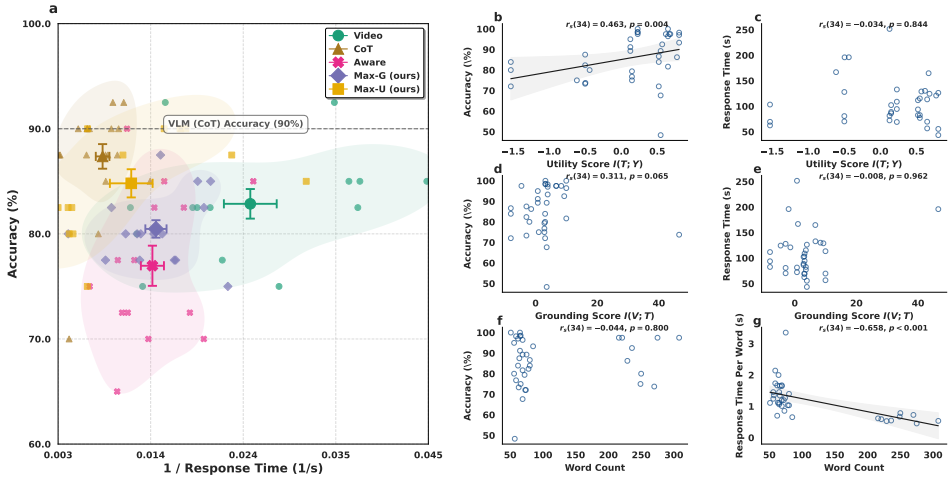
All steps follow from Bayes' theorem and the assumption of independence.



(a) `LearningPaper24`



(b) `LongVideoBench`



(c) `SUTD-TrafficQA`

Figure 9: Scatter plots and Spearman correlations across all datasets. Trendlines shown for $p < 0.05$.