# SUMREASON: A BENCHMARK FOR REASONING-DRIVEN QUERY-CONDITIONED AND STANDARD VIDEO SUMMARIZATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Video summarization benchmarks have largely evolved along two disconnected tracks: standard summarization datasets that assume a single notion of importance, and query/text-conditioned datasets whose queries often reduce to simple object or concept retrieval/matching. As a result, current benchmarks do not adequately test whether models can adapt summaries to user intent or perform the temporal, semantic, and knowledge-based reasoning required in realistic settings. We introduce *SumReason*, a unified benchmark for both standard and query-conditioned video summarization that explicitly targets this gap of reasoning-intensive video understanding. The dataset contains 37 videos, 60 video-query pairs, and 25 query-independent annotation sets, with 850 total annotations from 10 annotators. Its query hierarchy spans Direct Visual, Temporal Reasoning, High-Level Semantic Understanding, and External Knowledge Reasoning, and is designed to shift focus toward both salient and secondary events. SumReason combines strong annotation reliability with increased task difficulty: it achieves higher inter-annotator agreement than legacy benchmarks in the standard setting, while query-conditioned summaries overlap with standard summaries by only 30–40% on average, confirming substantial query-induced semantic shifts. Across extensive evaluations of $> 15$ video summarization methods and frontier open- and closed-source MLLMs, current systems exhibit substantial limitations on SumReason: standard summarization methods show markedly smaller gains than on established datasets, while MLLMs reveal a persistent reasoning gap in the query-conditioned setting. We position SumReason as a benchmark for studying intent-aware summarization moving beyond surface-level visual matching.

## 1 INTRODUCTION

The massive proliferation of video content has fundamentally strained traditional methods of digital media storage, processing, and consumption (Ma et al., 2002; Money & Agius, 2008; Truong & Venkatesh, 2007). Powered by the ubiquity of mobile hardware and the boom of social media, video data is now produced at a staggering volume (Apostolidis et al., 2021a). This explosion has resulted in chronic information overload, where identifying valuable insights within vast quantities of raw, unstructured footage becomes a significant hurdle.

**Video summarization.** In response to these difficulties, *video summarization* has become a cornerstone of computer vision research (Rochan et al., 2018; Zhang et al., 2016a). The goal of this task is to distill lengthy videos into brief, high-fidelity representations that retain the core narrative. By filtering for the most salient segments, summarization enables viewers to efficiently understand a video's substance without the need for full-length playback (Hussain et al., 2021; Otani et al., 2017). This process not only streamlines storage and bandwidth usage (Almeida et al., 2012; Muhammad et al., 2019) but also empowers diverse applications, ranging from forensic investigation and surveillance monitoring (Antani et al., 2002; Zhang et al., 2016b) to the curation of educational resources and sports highlights (Mahapatra et al., 2018; Alrumiah & Al-Shargabi, 2022).

**Limitations of existing datasets.** Despite the proliferation of video data, existing benchmarks suffer from four critical bottlenecks that hinder the development of human-centric summarization systems.



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.

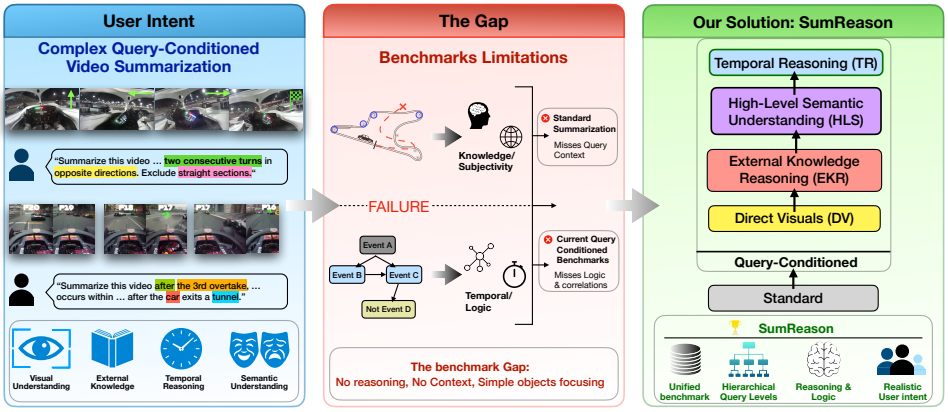



Figure 1: **`SumReason` benchmark.** While standard video summarization is query-agnostic and current query-conditioned models are restricted to simple object retrieval, `SumReason` introduces a hierarchical benchmark. Our framework evaluates a model's capacity for complex reasoning, including temporal dependencies, external knowledge, and high-level semantic intent.

**The assumption of static importance.** First, there is a general lack of query/text-conditioned benchmarks. The majority of established datasets (Gygli et al., 2014; Song et al., 2015; De Avila et al., 2011; Chu et al., 2015; Zeng et al., 2016) are designed for generic summarization, which assumes that a video has a single, static set of important segments. This ignores the reality that "importance" is a function of user intent; without query-based frameworks, models cannot learn to adapt their output to specific information needs.

**Retrieval vs. reasoning benchmarks.** Second, among the few available query-focused datasets, there is a distinct lack of diversity in query hardness. Existing benchmarks (Sharghi et al., 2016) typically rely on simple, explicit visual cues (e.g., "Focus on *CAR/SKY/TREE*"), which we categorize as *low-level retrieval* rather than true summarization. They (Vasudevan et al., 2017; Huang & Worring, 2020) lack a graded spectrum of "hardness" that challenges a model's ability to perform long-term temporal reasoning, such as identifying cross-scene correlations where the importance of a segment depends on events occurring minutes apart. Furthermore, they provide little room for high-level semantic comprehension or causal understanding, where the model must infer why an event is significant based on the narrative flow rather than just what is appearing on screen. Without such diversity, current benchmarks cannot distinguish between a model that simply matches keywords to pixels and one that understands the underlying video logic.

**Task mismatch in modern benchmarks.** Third, recent multimodal video datasets (Grauman et al., 2022) as well as narrative-oriented resources (Caba Heilbron et al., 2015) provide rich language annotations for long-form video understanding, yet they are fundamentally not designed for video summarization. Formulated primarily for long-form understanding, they lack explicit summary targets, standardized evaluation protocols, and supervision for selecting compact video subsets. Consequently, they serve as auxiliary training sources for representation learning (Bain et al., 2021; Miech et al., 2019) rather than as benchmarks to evaluate summarization quality, particularly in reasoning-heavy and query-conditioned settings, specifically for evaluating current MulitModal-LLMs (Zhu et al., 2025; Yang et al., 2025; OpenAI, 2026; Google DeepMind, 2026a) .

**Multi-perspective absence.** Finally, many of the current benchmarks (Fu et al., 2017; Potapov et al., 2014; Sharghi et al., 2016) are limited by single-annotation ground truths. This "single notion of importance" prevents the dataset from capturing multiple, equally valid interpretations that different users might have for the same video and query. Consequently, there is an urgent need for a unified benchmark that supports multi-perspective, standard, and query-driven summarization across multiple levels of difficulty—from basic visual elements to complex temporal dependencies and correlations—to serve as a testbed for the video summarization methods.

**One dataset to rule them all.** To address these limitations, we introduce `SumReason`, a unified benchmark designed to catalyze the next generation of video summarization models. By constructing a sophisticated query hierarchy that transcends simple object retrieval, our benchmark bridges the

divide between query-independent and query-conditioned summarization. Unlike existing datasets that treat these as isolated challenges, `SumReason` provides a holistic framework where importance and relevance are evaluated simultaneously within identical video contexts. To capture the inherent subjectivity of human perception, we provide multiple independent annotations per query, moving beyond the restrictive single-consensus paradigm. Furthermore, our queries are explicitly structured according to a hardness hierarchy—specifically targeting long-term temporal reasoning, cross-scene correlation, and causal understanding—as illustrated in Figure 1.

**Our primary contributions are summarized as follows:**

**(i) A unified summarization benchmark.** We propose the first dataset that provides a common ground for evaluating both query-independent and query-conditioned summarization, allowing for a comparative analysis of "general importance" versus "specific relevance." **(ii) Reasoning-driven query hierarchy.** We introduce a diverse set of text queries categorized by hardness levels, moving beyond simple object retrieval to include complex tasks requiring high-level semantic comprehension, temporal dependencies, and external knowledge. **(iii) Multi-annotator ground truths.** To reflect the subjective nature of the task, we provide multiple, non-aggregated ground truths per video and video/query pair, enabling a more robust evaluation of model diversity and alignment with varied human interpretations. **(iv) Comprehensive experimental evaluation.** We provide an extensive benchmark of over 15 state-of-the-art video summarization methods and state-of-the-art MultiModal-LLMs. Our results demonstrate that modern models have a significantly harder time improving upon the random baseline when compared to legacy benchmarks, confirming that our dataset successfully surpasses the performance ceiling reached by previous data and identifying key challenges for future development.

## 2 Related Work

**Query-independent video summarization benchmarks.** Two datasets currently dominate the standard video summarization landscape: SumMe (Gygli et al., 2014) and TVSum (Song et al., 2015). *SumMe* provides multiple user-generated summaries as key-shots, typically constrained to 5–15% of the video length. *TVSum* offers shot- and frame-level importance scores on a 1–5 scale, collected from multiple annotators. Other datasets, such as *OVP* and *YouTube* (De Avila et al., 2011), utilize sets of key-frames for evaluation. Subsequent efforts, including CoSum (Chu et al., 2015), MED-Summaries (Potapov et al., 2014), VTW (Zeng et al., 2016), and FPVSum (Ho et al., 2018), expanded content diversity and annotation formats. However, these benchmarks share a fundamental limitation: they assume a *single, static notion of importance* per video, ignoring the inherent subjectivity of relevance driven by user intent. Consequently, standard video summarization remains anchored to legacy datasets like TVSum and SumMe, while recent extensions are often restricted by domain-specific scopes (Kontostathis et al., 2024) or reliance on proxy annotations (Caba Heilbron et al., 2015).

**Query-conditioned video summarization benchmarks.** To address the lack of flexibility in standard benchmarks, QFVS (Sharghi et al., 2016) was the first to introduce textual queries by annotating 5-second shots with semantic labels. However, its formulation often reduces summarization to a binary highlight detection task based on simple visual entities (e.g., *car*, *tree*). RAD (Vasudevan et al., 2017) and RAD-Based (Huang & Worring, 2020) extended this with graded relevance scores, yet queries remained generic and concept-driven (e.g., *Audi, Barbie Girl*). Crucially, such datasets mostly reduce the task to matching basic visual entities, overlooking the semantic and narrative structure that determines importance in holistic summarization, as well as temporal, reasoning-based, and context-dependent relevance.

**Modern landscape.** Large-scale multimodal datasets such as Ego4D (Grauman et al., 2022) and ActivityNet (Caba Heilbron et al., 2015) Captions have advanced temporal grounding and captioning through dense natural language supervision. Similarly, SynopGround (Tan et al., 2024) introduces narrative-level alignment to capture higher-level semantics. However, these benchmarks are not designed for video summarization, lacking keyshot-based summaries, standardized evaluation protocols, and direct supervision for content selection.

**Video summarization methods.** Standard methods identify representative segments under a query-independent prior. Supervised approaches span LSTM-based models (Zhang et al., 2016a; Zhou et al.,

2017), encoder-decoders (Mahasseni et al., 2017), attention-based architectures (Fajtl et al., 2018; Hsu et al., 2023), and CNN-based methods (Rochan et al., 2018; Zhu et al., 2020; Son et al., 2024). Unsupervised methods exploit diversity and representativeness through reinforcement learning (Apostolidis et al., 2020a), temporal segmentation (Bang & Viet, 2025), and restorative scoring (Abbasi & Saeedi, 2023; Xin et al., 2025; Chen et al., 2022). More recently, zero-shot approaches using LLMs and video-language models (VidLMs) have emerged (Barbara & Maalouf, 2025; Saito et al., 2025). However, most remain *intent-agnostic*. Although language-guided frameworks have been explored (Barbara & Maalouf, 2025; Narasimhan et al., 2021), limited reasoning-oriented datasets have largely restricted evaluation to simpler concept-matching tasks. Additional comparisons with legacy benchmarks are provided in Appendix A.

# 3 SUMREASON DATASET

**Dataset overview.** In this section, we introduce `SumReason`, a unified benchmark designed for both standard video summarization and the relatively underexplored task of query-conditioned video summarization: It consists of 37 videos, with an average duration of approximately 4 minutes, covering a diverse set of categories, including sports, vlogs, how-to, CGI, and first-person videos. Each video is paired with 1–3 user queries, resulting in a total of 60 video-query pairs. Of the 37 videos, 25 are additionally annotated for standard (query-independent) video summarization. Unlike existing datasets, `SumReason` design makes the benchmark substantially more challenging and better aligned with real-world, user-driven information needs, where relevance is conditioned on specific intents and targets scenarios where queries often refer to *secondary or non-salient events* rather than to the dominant moments.

## 3.1 DATA COLLECTION

**Videos.** Videos were collected from YouTube and the YouTube-8M (Abu-El-Haija et al., 2016) dataset. We ensured that each selected video satisfied the following criteria: (i) Publicly available for research use, (ii) Duration between 1 and 10 minutes, (iii) Contains a clear core message, theme, or main sequence of events, and (iv) Includes well-defined important events with minimal ambiguity in their relative importance.

**Queries.** Queries were constructed based on one or more of the following aspects: (i) **Visual Content**: focuses on explicit actions or entities (e.g., *"Focus on scenes where the score changes"*, for a basketball highlight video that includes the final minutes of the game); (ii) **Underlying Message**: requires reasoning about the emotional arc or narrative progression (e.g., *"Highlight scenes that show a clear emotional shift"*, for a video of a one-legged boy being gifted a one-legged puppy, where the boy jumps through a range of different emotions—from initial rejection and frustration to a moment of realization and final bonding); and (iii) **Overall Theme**: based on the unique atmosphere of the footage (e.g., *"Focus on scenes that demonstrate the out-of-the-ordinary dining experience"*, for a video of a theatrical, high-energy meal where chefs perform showmanship—like fire flares and spatula tricks—cooking directly in front of the customer). For more query examples see Appendix B.2.

Queries were iteratively refined to ensure precision, reduce ambiguity, and align each query with specific target events. For example, *"Focus on Halal-friendly recipes containing the most visible vegetables"* was revised to *"Focus on Halal-friendly recipes containing at least two types of vegetables"* to remove the subjective notion of "most visible." Queries were also designed hierarchically to capture both salient and secondary events, spanning visual, temporal, spatial, semantic, and knowledge-based reasoning.

## 3.2 ANNOTATION PROTOCOL

Because video summarization is inherently subjective, a single unambiguous ground truth is difficult to define. Following established datasets (Gygli et al., 2014; Song et al., 2015), we therefore rely on human importance scores assigned to temporal segments rather than predefined summaries. This fine-grained annotation supports multiple valid summaries and flexible benchmarking.

**Task setup.** As a result we've collected a total of 850 annotations from 10 annotators, including 600 query-conditioned annotations and 250 standard (query-independent) annotations, i.e., 60 and

Table 1: **Descriptive analysis and annotation consistency across reasoning domains.** We report the scale of each category and the consensus reached during the majority-vote classification of queries (Cat.Agr.). Inter-annotator consistency is evaluated using summary-level metrics ($\kappa$, $\alpha$, $F_1$) and frame-level correlations ($\rho$, $\tau$), both reported as mean $\pm$ standard deviation.

| Cat. | Descriptive Analysis | | | Summary Consistency | | | Frame Score Consistency | |
|---|---|---|---|---|---|---|---|---|
| | #Qry | #Vid | Cat.Agr. (%) | Fleiss' $\kappa\uparrow$ | Cron. $\alpha$ (%) $\uparrow$ | $\overline{\text{Pair.F}_1}\uparrow$ | Spear. $\rho\uparrow$ | Kend. $\tau\uparrow$ |
| DV | 18 | 16 | 88.30 | $0.57 \pm 0.17$ | $92.0 \pm 4.7$ | $0.63 \pm 0.15$ | $0.608 \pm 0.184$ | $0.579 \pm 0.181$ |
| TR | 20 | 15 | 86.50 | $0.58 \pm 0.17$ | $92.3 \pm 4.0$ | $0.64 \pm 0.15$ | $0.578 \pm 0.196$ | $0.539 \pm 0.191$ |
| HLS | 9 | 11 | 83.30 | $0.59 \pm 0.14$ | $93.0 \pm 4.7$ | $0.65 \pm 0.12$ | $0.594 \pm 0.161$ | $0.556 \pm 0.159$ |
| EKR | 13 | 8 | 76.15 | $0.59 \pm 0.15$ | $93.0 \pm 4.0$ | $0.65 \pm 0.13$ | $0.570 \pm 0.138$ | $0.537 \pm 0.134$ |
| **Overall** | **60** | **37** | **83.56** | $\mathbf{0.58 \pm 0.16}$ | $\mathbf{92.5 \pm 4.2}$ | $\mathbf{0.64 \pm 0.14}$ | $\mathbf{0.588 \pm 0.173}$ | $\mathbf{0.553 \pm 0.169}$ |

25 for each of the 10 annotators. During the annotation process, annotators were instructed to follow these steps: (i) Carefully read the query and ensure full understanding. For example, for the query: "Focus on Halal-friendly recipes containing at least two types of vegetables," annotators must correctly interpret both dietary constraints (e.g., pork is not Halal) and ingredient categories (e.g., garlic is considered a vegetable). (ii) Watch the entire video from start to end (without audio). (iii) Following the protocol established in Song et al. (2015), videos were partitioned into 2-second fixed-length shots. This granularity provides a balanced trade-off between local temporal context and the maintenance of visual coherence. (iv) Assign an importance score ranging from 1 (not important) to 5 (very important) to each shot, based on its relevance to the query. (v) Annotators were explicitly instructed **not to treat the task as binary highlight detection**, but instead to assign scores that reflect the *relative importance of each shot within the full video context*. (vi) Contextual importance was emphasized: shots surrounding key events may receive non-zero scores if they contribute to the buildup or interpretation of important moments.

**Query categories.** In addition to assigning shot-level importance scores, annotators were asked to categorize each query into a predefined semantic category. These categories allow us to systematically analyze how different reasoning requirements affect both annotation consistency and model performance. **The query categories are defined as follows:**

**(i) Direct visual (DV).** Queries that can be answered by observing objects, actions, attributes, or events within short video segments, without requiring temporal reasoning or external knowledge. **(ii) Temporal reasoning (TR).** Queries that require understanding event order, duration, repetition, or long-range temporal dependencies across the video. **(iii) High-level semantic understanding (HLS).** Queries that rely on abstract interpretation, such as emotions, intent, narrative importance, social context. **(iv) External knowledge reasoning (EKR).** Queries that require knowledge beyond what is explicitly shown in the video, including cultural, technical, ethical, or domain-specific understanding.

**Annotation protocol. (i) Standard importance scoring.** In the query-independent setting, annotators assigned each shot an importance score from 1–5 based on its relevance to the video's main theme, message, and primary events, without conditioning on an external query. **(ii) Score-distribution regularization.** To support concise summary generation, annotators were encouraged to maintain a skewed score distribution in which only a small subset of shots received high importance. Specifically, scores of 4–5 were assigned to no more than 20% of the segments, while most segments received lower scores, following prior annotation protocols Song et al. (2015); Vasudevan et al. (2017); Huang & Worring (2020). **(iii) Validation.** All annotations were manually reviewed by three evaluators. Annotations that violated the predefined guidelines, including insufficient score regularization or inconsistent temporal scoring, were rejected and re-annotated until they satisfied the acceptance criteria. Further details on the annotation and validation procedures are provided in Appendix B.4.

## 4 Experimental Study

**Frame-, shot-, and fragment-level scores.** Annotations are first collected at the shot level, using fixed 2-second shots. Each frame then inherits the score assigned to its corresponding shot. Finally, frame-level scores are aggregated into the dataset-specific predefined fragments used for summary construction. Given a fixed summary-length budget, the final summary is obtained by selecting the set of fragments that maximizes the total fragment importance score using a Knapsack algorithm. Additional details are provided in Section 4.3 and in Section B.5 of the Appendix.

Table 2: **Comparative analysis of annotation consistency on standard summarization benchmarks.** We evaluate SumReason against SumMe and TVSum using summary-level ($\kappa$, $\alpha$, $F_1$, PoH, PoR) and frame-level ($\rho$, $\tau$) metrics. Our dataset demonstrates superior inter-annotator agreement and internal consistency across all indicators.

| Dataset | Summary Consistency | | | | | Frame Score Consistency | |
|---|---|---|---|---|---|---|---|
| | Fleiss' $\kappa \uparrow$ | Cron. $\alpha$ (%) $\uparrow$ | $\overline{\text{Pair.F}_1} \uparrow$ | PoH $\uparrow$ | PoR $\downarrow$ | Spear. $\rho \uparrow$ | Kend. $\tau \uparrow$ |
| SumMe | 0.210 | 74 | 0.31 | 55.30 | 44.90 | *n/a* | *n/a* |
| TVSum | 0.458 | 81 | 0.538 | 54.80 | 58.60 | 0.204 | 0.177 |
| **Ours** | 0.540 | **92** | 0.610 | **61.53** | **54.50** | 0.394 | 0.348 |

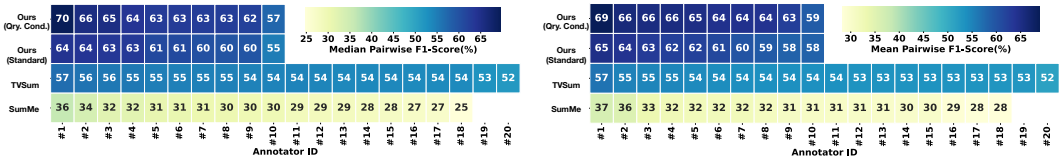



Figure 2: **Inter-annotator agreement (mean and median pairwise F1-score).** Each cell represents the agreement of a single annotator with all other annotators within the same dataset, measured as F1-score(%), rounded to the nearest integer. Left: median pairwise F1-score. Right: mean pairwise F1-score. Notably, the lowest-performing annotator in our dataset still achieves higher agreement than annotators from other datasets.

### 4.1 Inter-Annotator Agreement.

To assess the quality of the SumReason annotations, Table 1 reports both dataset composition and inter-annotator consistency. At the summary level, Fleiss' $\kappa$ measures agreement beyond chance, Cronbach's $\alpha$ measures internal consistency, and mean pairwise $F_1$ quantifies overlap between annotators' selected summaries, following prior work Gygli et al. (2014); Song et al. (2015); Apostolidis et al. (2020b). At the frame level, Spearman's $\rho$ and Kendall's $\tau$ measure rank correlation between annotators' importance scores.

**Query-conditioned agreement.** Across reasoning categories, Cat.Agr. ranges from $88.30\%$ for Direct Visual (DV)to $76.15\%$ for External Knowledge Reasoning (EKR). Summary-level consistency remains stable, with overall Fleiss' $\kappa = 0.58$, Cronbach's $\alpha = 92.5\%$, and mean pairwise $F_1 = 0.64$. Frame-level scores also show consistent positive agreement, with overall Spearman's $\rho = 0.588$ and Kendall's $\tau = 0.553$. Together, these results indicate that annotators maintain broadly consistent summary selections and frame-level rankings across the categories.

**Inter-annotator agreement for query-independent summaries.** Table 2 compares SumReason with established benchmarks using summary- and frame-level agreement metrics. Mean pairwise $F_1$ measures summary overlap, while Spearman's $\rho$ and Kendall's $\tau$ measure frame-level rank agreement. We also report **PoR**, a random baseline obtained by sampling frame-importance scores from a Gaussian prior, and **PoH**, a leave-one-out human ceiling averaged over 100 runs. SumReason achieves Fleiss' $\kappa = 0.54$ and Cronbach's $\alpha = 92$, exceeding SumMe ($0.21/74$) and TVSum ($0.45/81$). PoH reaches $61.53\%$, while frame-level agreement reaches $\rho = 0.394$ and $\tau = 0.348$, with Spearman's $\rho$ nearly twice that of TVSum. The lower PoR ($54.5\%$ vs. $58.6\%$ for TVSum) further indicates a more discriminative evaluation setting. We restrict this comparison to datasets with multiple independent annotations (Chu et al., 2015; Fu et al., 2017; Zeng et al., 2016).

**Visualization of inter-annotator agreement.** As shown in Figure 2, the inter-annotator agreement trends demonstrate that the proposed dataset provides a clearer ground-truth anchor compared to legacy benchmarks, successfully surpassing the performance ceiling reached by current datasets. By calculating the mean F1-score of each individual annotator against the remaining pool, both the **Query-Conditioned** and **Standard** versions of our dataset achieve superior mean and median stability. Notably, our least agreed-with annotator maintains a higher agreement score of ≈58% —surpassing even the most agreed-with annotators in legacy datasets like **TVSum** (≈56%) and **SumMe** (≈36%). For the full pair-wise F1-score analysis for all the datasets, see Figure 7 in Appendix Section A.3.

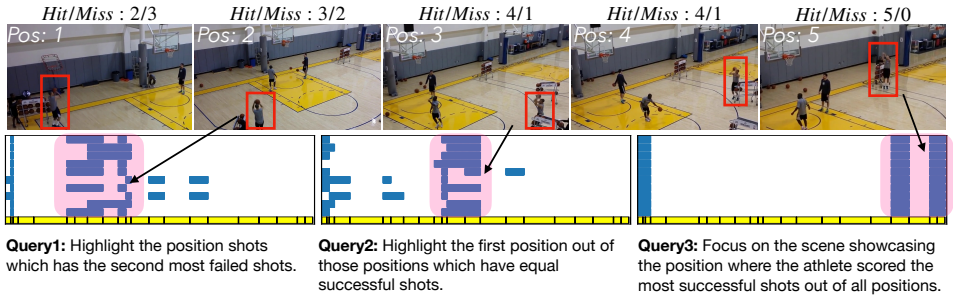
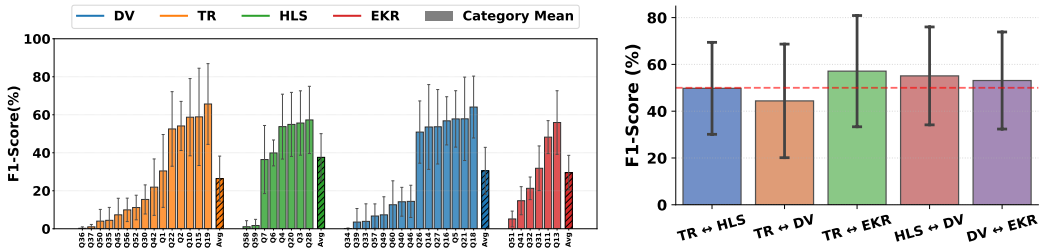



Figure 3: **Visualization of query-induced summary shift**. The yellow baseline represents the underlying temporal fragments of the source video. Each subsequent row displays the summary generated by an individual annotator (blue segments). The strong vertical alignment across rows confirms high inter-annotator consensus for each specific task. Conversely, the horizontal variation between queries, such as "**position with the second most failed attempts**" (Query 1) versus "**position with the most successful shots**" (Query 3), demonstrates a distinct shift in semantic focus.




(a) Impact of query conditioning

(b) Inter-category semantic shift

Figure 4: **Quantitative analysis of query-induced summary variance.** (a) We quantify the query-induced shift by computing the $F_1$-score between query-conditioned and standard summaries for each annotator; lower scores indicate a larger deviation from the standard summary. (b) We measure the variation between distinct query types by computing the $F_1$-score between different query-conditioned summaries for the same video, showing an average semantic shift of 50%.

### 4.2 Query-Induced Shift

We measure how user queries alter video summaries both qualitatively and quantitatively.

**Qualitative example.** Figure 3 shows that the same video yields distinct summaries under different queries. In the basketball example, the standard summary emphasizes successful shots, while query-conditioned summaries shift attention toward failed attempts, shot counts, and order- or location-dependent events. The vertical alignment across annotators (blue fragments) further indicates consistent query-driven changes in focus.

**Quantitative analysis.** For each video–query pair, we compute the $F_1$ overlap between query-conditioned and standard summaries across annotators. Figure 4a shows only 30–40% average overlap, with TR and EKR often approaching zero, indicating that queries substantially redefine relevant content rather than merely refine global saliency. Figure 4b further shows that summaries of the same video differ by roughly 50% across query categories. Additional fragment-level analysis is provided in Appendix B.3.

### 4.3 Benchmarking Existing Video Summarization Methods

**Evaluation setup.** To ensure a fair and reproducible comparison across all evaluated methods, we follow established video summarization protocols (Apostolidis et al., 2020b; Song et al., 2015), using shot-based segmentation and constrained optimization for summary generation. All methods operate under the same summary budget of $L = 15\%$. In addition to fixed 2-second segments, we evaluate a KTS-based setting, where Kernel Temporal Segmentation (KTS) partitions each video into content-

Table 3: **Benchmarking query-conditioned video summarization methods.** We report F1 scores (%) across query categories and the overall average. Underlined values indicate the lowest-performing category for each method.

| Model | DV | TR | HLS | EKR | Avg. |
|---|---|---|---|---|---|
| **Baselines** | | | | | |
| PoR | 50.0 | 47.2 | 53.4 | 49.1 | **49.9** |
| PoH | 63.6 | 63.8 | 65.4 | 65.5 | **64.6** |
| RS | 23.1 | 22.6 | 22.7 | 24.4 | **23.2** |
| CLIP | 28.4 | 30.3 | 31.7 | 26.6 | **29.2** |
| **Prior SOTA** | | | | | |
| LGRLN | 47.36 | 43.57 | 51.01 | 51.68 | **47.58** |
| FCSNA | 56.24 | 50.43 | 59.11 | 50.62 | **53.52** |
| **P2S (LLM)** | | | | | |
| GPT-5.6-Sol | 55.47 | 49.00 | 58.98 | 53.80 | **53.82** |
| Claude-Fable-5 | 50.89 | 46.24 | 58.93 | 52.22 | **50.94** |
| Qwen3.6 27B | 50.34 | 50.50 | 52.41 | 47.20 | **50.02** |
| **MLLM-BL-2s (MLLM)** | | | | | |
| Qwen3-VL-7B | 59.08 | 47.50 | 59.07 | 44.26 | **52.01** |
| InternVL3-8B | 59.49 | 48.61 | 61.16 | 57.93 | **55.78** |
| GPT 5.6 Luna | 59.98 | 50.98 | 59.88 | 54.36 | **55.75** |
| Gemini 3.1 Pro | 54.47 | 51.54 | 57.90 | 57.53 | **54.67** |
| **MLLM-BL-KTS (MLLM)** | | | | | |
| Qwen3-VL-7B | 52.84 | 50.46 | 61.14 | 47.13 | **52.05** |
| InternVL3-8B | 57.52 | 46.52 | 58.36 | 55.72 | **53.59** |
| GPT 5.6 Luna | 56.08 | 48.90 | 61.25 | 51.06 | **53.37** |
| Gemini 3.1 Pro | 55.14 | 53.05 | 55.89 | 57.17 | **54.99** |

adaptive shots according to visual changes, providing semantically coherent temporal units of variable duration. A detailed description of fragment-level scoring, 0/1 Knapsack summary generation, and the complete evaluation protocol is provided in Appendix Section B.5. All experiments were conducted on a pc equipped with 1 NVIDIA A100-SXM4-40GB GPU and AMD EPYC 7742 64-Core CPUs.

**Methods.** We evaluate `SumReason` in both query-conditioned and standard video summarization settings. For query-conditioned evaluation, we include classical baselines—random and uniform fragment sampling (RS/US), K-means clustering (CK), and PoR (Apostolidis et al., 2020b)—together with a CLIP similarity baseline and recent methods including Prompt2Summs (P2S) (Barbara & Maalouf, 2025), LGRLN (Li et al., 2025), and FCSNA (Patel et al., 2024).

**Zero-shot MLLM evaluation.** Our main evaluation targets state-of-the-art Multimodal LLMs (Google DeepMind, 2026b; Anthropic, 2026; Zhu et al., 2025), including GPT-5.6 (OpenAI, 2026), Gemini 3.1 Pro (Google DeepMind, 2026a), Qwen 3.0 (Yang et al., 2025), and InternVL3 (Zhu et al., 2025). We introduce a zero-shot scene-scoring interface over both uniform 2-second segments and KTS-derived semantic shots. For each scene, sampled frames together with the query are provided to the MLLM, which outputs a structured relevance score from 1–5. These scores are mapped to frame-level importance values and converted into final summaries using the same Knapsack selection procedure under the $15\%$ budget. Full prompts, model-specific settings, and implementation details are provided in Appendix D.

For standard video summarization, we additionally compare against representative supervised and unsupervised methods spanning recurrent, convolutional, reinforcement-learning, generative, and attention-based architectures.

**Results: query-conditioned summarization.** As shown in Table 3, results reveal a clear category-dependent performance gap. Direct Visual (DV) and High-Level Semantics (HLS) are generally the strongest categories, whereas Temporal Reasoning (TR) is frequently the most challenging, with External Knowledge Reasoning (EKR) also limiting performance for several models. Classical baselines (RS, and CLIP) remain substantially below stronger learned methods, while supervised approaches and zero-shot MLLMs achieve considerably higher F1 scores. Notably, MLLMs perform competitively on visually and semantically grounded queries but show a persistent degradation on TR across both fixed 2-second and KTS-based segmentation. This suggests that reasoning over temporally distributed events remains a key bottleneck for current query-conditioned video

Table 4: **SOTA video summarization performance relative to PoR.** F1 improvements ($\Delta$) are measured against the random baseline (PoR). Red and yellow mark the lowest and second-lowest gains for each method, respectively. SumReason yields the smallest average gains despite the highest human agreement (PoH), highlighting a performance gap for current methods.

| Supervision | Method(Year) | Type | SumReason (Ours) | | | SumMe | TVSum |
|---|---|---|---|---|---|---|---|
| | | | Prec. | Rec. | $F_1(\Delta)$ | $F_1(\Delta)$ | $F_1(\Delta)$ |
| Baselines | PoR (Apostolidis et al., 2020b) | Random | 54.5 | 54.5 | 54.5 | 44.9 | 58.6 |
| | PoH (Apostolidis et al., 2020b) | Human | 61.55 | 61.51 | 61.53 | 55.3 | 54.8 |
| Heuristics | RS | Sampling | 22.77 | 25.02 | 23.82 | 36.97 | 18.63 |
| | US | Sampling | 22.93 | 13.26 | 16.23 | 29.99 | 12.01 |
| | CK | Clustering | 15.48 | 16.45 | 15.95 | 31.58 | 5.02 |
| Supervised | DR-DSN$_{sup}$(2017) (Zhou et al., 2017) | RL + LSTM | 32.34 | 32.22 | 32.28(-22.2) | 41.4(-3.5) | 57.6(-1.0) |
| | VASNet(2018) (Fajtl et al., 2018) | Self-Attention | 52.02 | 52.42 | 52.22(-2.3) | 49.63(+4.7) | 61.43(+2.83) |
| | SUM-FCN(2018) (Rochan et al., 2018) | CNN | 53.48 | 59.78 | 56.42(+1.9) | 47.5(+2.6) | 56.8(-1.8) |
| | DSNet(2020) (Zhu et al., 2020) | CNN + Attention | 63.0 | 55.8 | 59.2(+4.7) | 51.18(+6.3) | 62.5(+3.9) |
| | PGL-SUM(2021) (Apostolidis et al., 2021b) | PE + Attention | 54.86 | 53.97 | 54.4(-0.1) | 55.6(+10.7) | 61.0(+2.4) |
| | STVT(2023) (Hsu et al., 2023) | Attention | 56.9 | 62.74 | 59.64(+5.1) | 55.1(+10.2) | 67.1(12.3) |
| | CSTA(2024) (Son et al., 2024) | CNN + Attention | 54.4 | 54.6 | 54.5(+0.0) | 51.5(+6.6) | 53.8(-4.8) |
| Unsupervised | SUM-GAN-AAE(2020) (Apostolidis et al., 2019) | VAE + Attention | 53.7 | 55.3 | 53.76(-0.7) | 48.9(+4.0) | 58.3(-0.3) |
| | AC-SUM-GAN(2021) (Apostolidis et al., 2020a) | GAN + RL | 55.9 | 56.0 | 55.95(+1.5) | 50.8(+5.9) | 60.6(+2.0) |
| | CA-SUM(2022) (Apostolidis et al., 2022) | Contrastive + Attention | 57.4 | 57.0 | 57.2(+2.7) | 51.1(+6.2) | 61.1(+2.5) |
| | Uformer(2022) (Chen et al., 2022) | U-Transformer | 63.22 | 56.14 | 59.43(+4.9) | 53.9(+9.0) | 63.0(+4.4) |
| | RS-SUM(2023) (Abbasi & Saeedi, 2023) | Self-Attention | 55.55 | 55.56 | 55.55(+1.1) | 52.0(+7.1) | 61.4(+2.8) |
| | SegSum(2025) (Bang & Viet, 2025) | Temp Seg + Attn | 58.62 | 58.46 | 58.52(+4.0) | 54.0(+9.1) | 62.0(+3.4) |
| | mpA_uvS(2025) (Xin et al., 2025) | Attention | 60.4 | 59.76 | 60.08(+5.6) | 52.8(+7.9) | 63.0(+4.4) |
| Zero-shot | *Prompts2Summs*(2025) (Barbara & Maalouf, 2025) | LLM+VidLM | 58.07 | 58.14 | 58.1(+3.6) | 56.84(+11.9) | 62.22(+3.6) |

summarization systems. Additional category-level results, paired significance tests against random baseline (PoR), results and temporal-granularity robustness analysis in Appendix C.

**Results: standard summarization.** As shown in Table 4, we benchmark a selection of supervised and unsupervised methods, alongside classical baselines and modern zero-shot vision-language models. Comparatively, our dataset proves to be the most challenging benchmark for existing models, as evidenced by the **lowest average improvement over the random baseline (PoR)**. While state-of-the-art methods typically achieve gains of +10 to +12 points on legacy datasets like SumMe and TVSum, these margins are significantly curtailed on our benchmark, with average improvements falling below +5 points. This reduced margin—despite a lower random floor and higher human consensus—highlights a significant generalization gap, particularly for CNN and RL-based architectures, which frequently underperform a random frame-level scores. Conversely, the relative success of attention-based models suggests that global temporal modeling is a prerequisite for navigating the high visual entropy and multi-domain diversity inherent in our benchmark.

## 5 Limitations and Conclusion

**Limitations.** SumReason is designed as a dense evaluation benchmark for intent-aware video summarization.In constructing the dataset, we intentionally prioritized annotation quality, multi-annotator coverage, and reasoning-oriented query design over raw scale. As a result, although the current release already spans multiple domains and query types, it does not yet cover the full diversity of real-world video sources, user intents, or deployment settings. We view scaling the benchmark in both video count and query breadth as a natural direction for future work.

**Conclusion.** We introduced SumReason, a benchmark for standard and reasoning-intensive query-conditioned video summarization. Its query hierarchy spans Direct Visual Understanding, Temporal Reasoning, High-Level Semantic Understanding, and External Knowledge Reasoning, enabling controlled evaluation of how user intent changes which content should be summarized. Our experiments show that modern MLLMs perform comparatively well on visually and semantically grounded queries, while Temporal Reasoning remains a persistent bottleneck across both fixed 2-second and KTS-based temporal segmentation. The large divergence between standard and query-conditioned summaries further demonstrates that user queries can fundamentally redefine temporal relevance rather than simply refine generic saliency. We hope SumReason supports the development of video summarization systems that better align with complex user intent.

## Reproducibility Statement

The manuscript and appendix specify the annotation protocol, preprocessing steps, query construction, scoring conventions, temporal segmentation, fragment-level aggregation, summary generation, evaluation metrics, and model-specific inference settings. We additionally document the prompts and zero-shot evaluation pipelines used for the MLLM baselines. Aggregate results and statistical analyses are reported in the manuscript and appendix. Upon release, we will provide the annotations, benchmark splits, queries, evaluation code, and scripts required to reproduce the benchmark and its evaluation protocol. The released dataset contains only our derived annotations and metadata; raw videos, frames, and other media assets are not redistributed. Instead, scripts will allow researchers to retrieve publicly available videos directly from their original sources.

## Ethics Statement

**Annotator safety and human annotation.** The annotation task involved standard, benign video content and excluded violent, explicit, graphic, traumatic, or otherwise sensitive material. Annotators were therefore not intentionally exposed to content expected to pose psychological or physical risk, and we regard the annotation procedure as minimal risk. Annotators received task-specific training, were instructed to prioritize annotation quality over speed, and all annotations were reviewed through the quality-control procedure described in the manuscript.

**Dataset use and media distribution.** `SumReason` is intended solely as a non-commercial academic benchmark for research on video summarization and related video-understanding methods, following the distribution practice of established benchmarks such as SumMe and TVSum. Our release contains only the annotations we created, including importance scores, queries, metadata, and evaluation resources. We do not host, store, or redistribute the underlying videos, extracted frames, or other copyrighted media. Instead, researchers will be provided with scripts and source identifiers for retrieving videos directly from their original YouTube locations. This design preserves the original hosting platform and content creators' control over availability while avoiding redistribution of copyrighted media. Videos that are removed or made unavailable by their original owners are not redistributed by us.

**Intended use and limitations.** The benchmark is designed for evaluating video summarization systems and should not be interpreted as assessing the safety, reliability, or broader reasoning ability of a model beyond the tasks represented in `SumReason`. Agreement with human annotations indicates consistency with the benchmark's summarization criteria, not general factual correctness or safety.

## Statement of Generative AI Use

Generative AI tools assisted with language and manuscript editing and with identifying potentially relevant related work. The authors conceived the research questions, designed the benchmark and annotation protocol, conducted the experiments, performed and verified the analyses, and determined the interpretations and claims presented in the paper. AI-assisted text and suggested references were reviewed and verified by the authors before inclusion. The authors take full responsibility for the content of the manuscript and the reported results.

## References

Mehryar Abbasi and Parvaneh Saeedi. Adopting self-supervised learning into unsupervised video summarization through restorative score. In *Proceedings of IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023.

Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

Jurandy Almeida, Neucimar J Leite, and Ricardo da S Torres. Vison: Video summarization for online applications. *Pattern Recognition Letters*, 33(4):397–409, 2012.

Sarah S Alrumiah and Amal A Al-Shargabi. Educational videos subtitles' summarization using latent dirichlet allocation and length enhancement. *Computers, Materials & Continua*, 70(3), 2022.

Sameer Antani, Rangachar Kasturi, and Ramesh Jain. A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video. *Pattern recognition*, 35(4): 945–965, 2002.

Anthropic. Claude Fable 5 and Claude Mythos 5. `https://www.anthropic.com/news/claude-fable-5-mythos-5`, June 2026. Accessed: 2026-09-25.

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Unsupervised video summarization via attention-driven adversarial learning. In *International Conference on multimedia modeling*, pp. 492–504. Springer, 2019.

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Ac-sum-gan: Connecting actor-critic and generative adversarial networks for unsupervised video summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8): 3278–3292, 2020a.

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Performance over random: A robust evaluation protocol for video summarization methods. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 1056–1064, 2020b.

Evlampios Apostolidis, Eleni Adamantidou, Alexandros I Metsai, Vasileios Mezaris, and Ioannis Patras. Video summarization using deep neural networks: A survey. *Proceedings of the IEEE*, 109 (11):1838–1863, 2021a.

Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. Combining global and local attention with positional encoding for video summarization. In *2021 IEEE International Symposium on Multimedia (ISM)*, pp. 226–234, December 2021b.

Evlampios Apostolidis, Georgios Balaouras, Vasileios Mezaris, and Ioannis Patras. Summarizing videos using concentrated attention and considering the uniqueness and diversity of the video frames. ICMR '22, pp. 407–415, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392389. doi: 10.1145/3512527.3531404. URL `https://doi.org/10.1145/3512527.3531404`.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

Vo Quoc Bang and Vo Hoai Viet. Integrate the temporal scheme for unsupervised video summarization via attention mechanism. *IEEE Access*, 13:38147–38162, 2025. doi: 10.1109/ACCESS.2025.3546149.

Mario Barbara and Alaa Maalouf. Prompts to summaries: Zero-shot language-guided video summarization. *arXiv preprint arXiv:2506.10807*, 2025.

Michele Basseville, Igor V Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice hall Englewood Cliffs, 1993.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pp. 961–970, 2015.

Yaosen Chen, Bing Guo, Yan Shen, Renshuang Zhou, Weichen Lu, Wei Wang, Xuming Wen, and Xinhua Suo. Video summarization with u-shaped transformer. *Applied Intelligence*, pp. 1–17, 2022.

Wen-Sheng Chu, Yale Song, and Alejandro Jaimes. Video co-summarization: Video summarization by visual co-occurrence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3584–3592, 2015.

Sandra Eliza Fontes De Avila, Ana Paula Brandao Lopes, Antonio da Luz Jr, and Arnaldo de Albuquerque Araújo. Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern recognition letters*, 32(1):56–68, 2011.

Xianzhi Du, Yeqing Li, Yin Cui, Rui Qian, Jing Li, and Irwan Bello. Revisiting 3d resnets for video recognition. *arXiv preprint arXiv:2109.01696*, 2021.

Jiri Fajtl, Hajar Sadeghi Sokeh, Vasileios Argyriou, Dorothy Monekosso, and Paolo Remagnino. Summarizing videos with attention, 2018.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*, 2022.

Cheng-Yang Fu, Joon Lee, Mohit Bansal, and Alexander Berg. Video highlight prediction using audience chat reactions. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 972–978, 2017.

Google DeepMind. Gemini 3.1 Pro: Model Card. `https://deepmind.google/models/model-cards/gemini-3-1-pro/`, February 2026a. Accessed: 2026-09-25.

Google DeepMind. Gemini 3.6 Flash: Model Card. `https://deepmind.google/models/model-cards/gemini-3-6-flash/`, July 2026b. Accessed: 2026-09-25.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18995–19012, 2022.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating summaries from user videos. In *European conference on computer vision*, pp. 505–520. Springer, 2014.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6546–6555, 2018.

Hsuan-I Ho, Wei-Chen Chiu, and Yu-Chiang Frank Wang. Summarizing first-person videos from third persons' points of view. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 70–85, 2018.

Tzu-Chun Hsu, Yi-Sheng Liao, and Chun-Rong Huang. Video summarization with spatiotemporal vision transformer. *IEEE Transactions on Image Processing*, 32:3013–3026, 2023.

Jia-Hong Huang and Marcel Worring. Query-controllable video summarization. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pp. 242–250, 2020.

Tanveer Hussain, Khan Muhammad, Weiping Ding, Jaime Lloret, Sung Wook Baik, and Victor Hugo C De Albuquerque. A comprehensive survey of multi-view video summarization. *Pattern Recognition*, 109:107567, 2021.

Mohammad Jalali, Azim Ospanov, Amin Gohari, and Farzan Farnia. Conditional vendi score: An information-theoretic approach to diversity evaluation of prompt-based generative models. *arXiv preprint arXiv:2411.02817*, 2024.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Ioannis Kontostathis, Evlampios Apostolidis, and Vasileios Mezaris. A human-annotated video dataset for training and evaluation of 360-degree video summarization methods. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences Workshops*, pp. 71–79, 2024.

Wenrui Li, Wei Han, Hengyu Man, Wangmeng Zuo, Xiaopeng Fan, and Yonghong Tian. Language-guided graph representation learning for video summarization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li. A user attention model for video summarization. In *Proceedings of the tenth ACM international conference on Multimedia*, pp. 533–542, 2002.

Debabrata Mahapatra, Ragunathan Mariappan, Vaibhav Rajan, Kuldeep Yadav, and Sudeshna Roy. Videoken: Automatic video summarization and course curation to support learning. In *Companion Proceedings of the The Web Conference 2018*, pp. 239–242, 2018.

Behrooz Mahasseni, Michael Lam, and Sinisa Todorovic. Unsupervised video summarization with adversarial lstm networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 202–211, 2017.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2630–2640, 2019.

Arthur G Money and Harry Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of visual communication and image representation*, 19(2):121–143, 2008.

Khan Muhammad, Tanveer Hussain, Javier Del Ser, Vasile Palade, and Victor Hugo C De Albuquerque. Deepres: A deep learning-based video summarization strategy for resource-constrained industrial surveillance scenarios. *IEEE Transactions on Industrial Informatics*, 16(9):5938–5947, 2019.

Medhini Narasimhan, Anna Rohrbach, and Trevor Darrell. Clip-it! language-guided video summarization. *Advances in neural information processing systems*, 34:13988–14000, 2021.

OpenAI. GPT-5.6 System Card. `https://deploymentsafety.openai.com/gpt-5-6`, July 2026. Accessed: 2026-09-25.

Azim Ospanov and Farzan Farnia. On the statistical complexity of estimating vendi scores from empirical data. *arXiv preprint arXiv:2410.21719*, 2024.

Mayu Otani, Yuta Nakashima, Esa Rahtu, Janne Heikkilä, and Naokazu Yokoya. Video summarization using deep semantic features. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13*, pp. 361–377. Springer, 2017.

Nirav Patel, Payal Prajapati, and Maitrik Shah. Your interest, your summaries: Query-focused long video summarization, 2024. URL `https://arxiv.org/abs/2410.14087`.

Danila Potapov, Matthijs Douze, Zaid Harchaoui, and Cordelia Schmid. Category-specific video summarization. In *European conference on computer vision*, pp. 540–555. Springer, 2014.

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

Mrigank Rochan, Linwei Ye, and Yang Wang. Video summarization using fully convolutional sequence networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 347–363, 2018.

Shumpei Saito, Hiroyuki Ueda, Yosuke Ito, and Kazuyoshi Yoshii. Narrativity-aware video summarization based on vision and language foundation models. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1991–1996. IEEE, 2025.

Aidean Sharghi, Boqing Gong, and Mubarak Shah. Query-focused extractive video summarization. In *European conference on computer vision*, pp. 3–19. Springer, 2016.

Jaewon Son, Jaehun Park, and Kwangsu Kim. Csta: Cnn-based spatiotemporal attention for video summarization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18847–18856, 2024.

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. Tvsum: Summarizing web videos using titles. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5179–5187, 2015.

Chaolei Tan, Zihang Lin, Junfu Pu, Zhongang Qi, Wei-Yi Pei, Zhi Qu, Yexin Wang, Ying Shan, Wei-Shi Zheng, and Jian-Fang Hu. Synopground: A large-scale dataset for multi-paragraph video grounding from tv dramas and synopses. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 8383–8392, 2024.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 6450–6459, 2018.

Ba Tu Truong and Svetha Venkatesh. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, 3(1):3–es, 2007.

Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

Arun Balajee Vasudevan, Michael Gygli, Anna Volokitin, and Luc Van Gool. Query-adaptive video summarization via quality-aware relevance estimation. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 582–590, 2017.

Chaoqun Xin, Mingyang Wang, and Xianhao Zhao. mpa _ uvs: multi-perspective attention for unsupervised video summarization—capturing global, local, and spatiotemporal context: C. xin et al. *The Visual Computer*, 41(13):11505–11518, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. Title generation for user generated videos. In *European conference on computer vision*, pp. 609–625. Springer, 2016.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *ECCV*. Springer, 2016a.

Shu Zhang, Yingying Zhu, and Amit K Roy-Chowdhury. Context-aware surveillance video summarization. *IEEE Transactions on Image Processing*, 25(11):5469–5478, 2016b.

Bin Zhao and Eric P Xing. Quasi real-time summarization for consumer videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2513–2520, 2014.

Kaiyang Zhou, Yu Qiao, and Tao Xiang. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. *arXiv:1801.00054*, 2017.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Wencheng Zhu, Jiwen Lu, Jiahao Li, and Jie Zhou. Dsnet: A flexible detect-to-summarize network for video summarization. *IEEE Transactions on Image Processing*, 30:948–962, 2020.

# A LEGACY BENCHMARKS COMPARISON

## A.1 DESCRIPTIVE COMPARISON

Table 5: **Comprehensive comparison of video summarization benchmarks**. Our dataset SumReason provides a unique balance of diverse content and query-conditioned annotations.

| Modality | Dataset | #Videos (#Queries) | Dur. (min) | Content | Annotation Type | # Annot. |
|---|---|---|---|---|---|---|
| Video | SumMe | 25 | 1–6 | Holidays, Events | Key-fragments | 15–18 |
| | TVSum | 50 | 2–10 | News, How-to | Shot scores | 20 |
| | OVP | 50 | 1–4 | Documentary | Key-frames | 5 |
| | YouTube | 50 | 1–10 | Cartoons, Sports | Key-frames | 5 |
| | CoSum | 51 | ∼4 | Sports, Events | Key-fragments | 3 |
| | MED | 160 | 1–5 | Various genres | Importance scores | 1–4 |
| | VTW | 2000 | 1.5 (avg) | Highlights | Highlight scores | – |
| | LoL | 218 | 30–50 | Gaming (eSports) | Key-fragments | 1 |
| | **SumReason (Std)** | **25** | **4 (avg)** | **Multi-domain*** | **Shot scores** | **10** |
| Video + Text | QFVS | 4 (180) | 180–300 | Vlogs | Shot labeling | 1 |
| | RAD | 200 (200) | 2–3 | Various genres | Shot scores | 5 |
| | RAD-based | 4 (190) | 2–3 | Various genres | Shot scores | 5 |
| | **SumReason (Query)** | **37 (60)** | **4 (avg)** | **Multi-domain*** | **Shot scores** | **10** |

*Includes Sports, Vlogs, How-to's, CGI, First-person, and Drone POV.

**Comparison of video summarization benchmarks.** Table 5 provides a comparative overview of existing video summarization benchmarks against our proposed SumReason dataset. While classical datasets such as SumMe and TVSum offer high annotator density, they lack the textual modality required for query-conditioned tasks. Conversely, existing query-focused benchmarks like QFVS and RAD are limited by either a very small number of unique videos or a narrow domain of content (primarily egocentric vlogs). Our dataset bridges this gap by offering a diverse range of categories—including drone footage, CGI, and first-person perspectives—which are not present in legacy benchmarks. Furthermore, by providing 10 annotators per video, we ensure a significantly higher level of ground-truth reliability compared to the single-annotator labels found in QFVS, making our benchmark a more rigorous testbed for modern vision and large-language models .

## A.2 SUMREASON DIVERSITY ANALYSIS

To evaluate the robustness and breadth of our dataset SumReason, we perform a multi-faceted diversity analysis comparing our benchmark against established video datasets, namely SumMe and TVSum. A diverse dataset is critical for ensuring that models learn generalized reasoning patterns rather than over-fitting to narrow visual domains.

Our analysis is structured across three primary dimensions:

- **Visual and semantic coverage**: We utilize t-SNE projections of CLIP-based (Radford et al., 2021) video and query embeddings to visually inspect the distribution of our data within the feature space.
- **Effective complexity**: We employ the **Vendi Score** (Friedman & Dieng, 2022) to calculate the number of effectively independent semantic concepts present in each benchmark.
- **Distributional Shift:** We compute the **Fréchet Video Distance (FVD)** to mathematically quantify the novelty of SumReason relative to legacy temporal-visual priors.

**Video semantic representation.** To analyze the thematic distribution of the datasets, we first map each video into a high-dimensional semantic space. Given that SumReason focuses on complex reasoning, we require a representation that captures more than just low-level textures. We utilize the **CLIP (ViT-L/14)** (Radford et al., 2021) model as a feature extractor. Specifically, we extract frame-level embeddings and apply **temporal mean-pooling** to generate a single, global feature vector for each video, thus capturing the intersection of visual content and high-level semantic concepts.

**Visualizing videos semantic manifold.** To qualitatively assess the coverage of these embeddings, we project the high-dimensional CLIP vectors onto a 2D plane using **t-SNE**. As illustrated in Figure 5, the distribution of SumReason is notably more expansive and uniform than that of legacy benchmarks.




(a) SumReason (Ours) (b) SumMe

(c) TVSum (d) Shared Semantic Space

Figure 5: **Visual semantic analysis.** Subfigures (a-c) show individual CLIP embedding distributions. Subfigure (d) projects all datasets into a shared manifold, highlighting that SumReason spans a significantly larger and more diverse area of the feature space than legacy benchmarks.

While SumMe and TVSum (Figures 5b and 5c) exhibit dense, localized clusters—suggesting a repetition of specific visual themes—SumReason spans multiple disjoint regions across the feature space. This broad dispersion confirms the high semantic variance of our data, stemming from our diverse category selection, which includes fast-paced sports, narrative-driven vlogs, and structured CGI environments. Furthermore, the shared projection in Figure 5d highlights that our dataset does not merely replicate existing visual distributions but significantly expands the semantic frontier of video summarization research.

**Visualizing queries semantic manifold.** To investigate the latent structure and diversity of the textual queries in SumReason, we analyze their distribution in a high-dimensional semantic manifold. We utilize the CLIP text encoder to extract embeddings from the queries and apply t-SNE to project them into a two-dimensional space.

As shown in Figure 6, queries are color-coded by their respective reasoning categories: Direct Visual (DV), Temporal Reasoning (TR), High-Level Semantic (HLS) and External Knowledge Reasoning (EKR). The absence of tight, isolated clusters across these categories indicates high semantic variance and suggests that the reasoning challenges are not tied to specific linguistic templates. This broad distribution ensures that models must understand the underlying semantics of the query rather than relying on shallow textual correlations.

**Quantitative diversity analysis.** To provide a rigorous, objective measure of the diversity of SumReason relative to legacy benchmarks, we evaluate three key metrics: Vendi score, average Cosine Similarity, and Fréchet Video Distance (FVD).

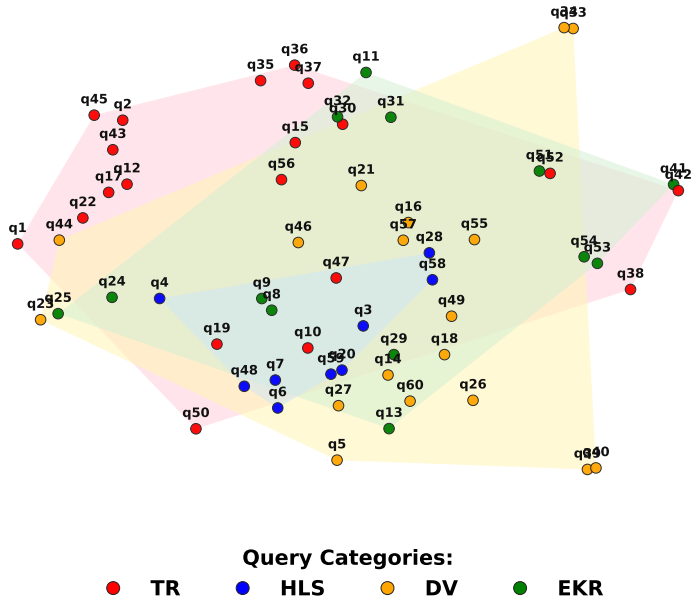



Figure 6: **T-SNE projection of query embeddings.** Points represent individual queries color-coded by reasoning category. The broad dispersion across categories highlights the semantic complexity and variety of the textual instructions in our benchmark.

**Effective diversity via Vendi score.** Vendi score (Friedman & Dieng, 2022) is an unsupervised metric used in (Jalali et al., 2024; Ospanov & Farnia, 2024) that measures "unique" elements in a dataset.

Given a positive semi-definite similarity matrix $K \in \mathbb{R}^{n \times n}$ representing the pairwise cosine similarities of our $n$ videos (where $K_{i,i} = 1$), we normalize $K$ to obtain a density matrix $P = K/n$. The Vendi Score ($VS$) is then defined as the exponential of the *von Neumann* entropy of $P$:

$$VS(K) = \exp\left(-\sum_{i=1}^{n} \lambda_i \ln \lambda_i\right) \tag{1}$$

where $\lambda_i$ are the eigenvalues of $P$. In this framework, if all videos in a dataset are identical, the Vendi Score is 1; if all videos are perfectly orthogonal (sharing zero similarity), the score is $n$.

For the 37 videos in `SumReason`, we achieve a Vendi Score of **7.1**. This score is particularly significant given the scale of the dataset; it implies that our benchmark is composed of roughly seven entirely independent semantic clusters. In contrast to legacy datasets that often suffer from high redundancy, this result confirms that `SumReason` spans a broad, multi-modal manifold rather than repeating a few narrow visual themes. As shown in Table 6, our dataset provides a higher concentration of unique semantic concepts compared to both SumMe and TVSum, establishing it as a more diverse and challenging benchmark for video reasoning.

**Embedding dispersion via cosine similarity.** To complement the Vendi Score, we evaluate the intra-dataset dispersion by computing the average pairwise cosine similarity between all video embeddings within each benchmark. For a dataset with $n$ videos, we calculate the mean similarity as:

$$\text{Sim}_{avg} = \frac{1}{n(n-1)} \sum_{i \neq j} \frac{v_i \cdot v_j}{\|v_i\| \|v_j\|} \tag{2}$$

As presented in Table 6, `SumReason` exhibits the lowest average cosine similarity (**0.59** $\pm$ **0.10**), significantly outperforming SumMe (0.65) and TVSum (0.71). This lower similarity threshold implies that the videos in our benchmark are semantically more distant from one another, covering a broader range of visual and narrative concepts. This statistical evidence aligns with our t-SNE visualizations, confirming that `SumReason` provides a more heterogeneous and challenging environment for evaluating video reasoning models.

Table 6: **Effective dataset diversity.** The Vendi Score represents the effective number of unique, orthogonal concepts within each benchmark based on global embeddings. FVD is calculated relative to `SumReason`.

| Dataset | Vendi Score (↑) | Cosine Similarity (↓) | FVD |
|---|---|---|---|
| SumMe | 5.0 | $0.65 \pm 0.07$ | 186.44 |
| TVSum | 4.7 | $0.71 \pm 0.10$ | 120.59 |
| **`SumReason` (Ours)** | **7.1** | $\mathbf{0.59 \pm 0.10}$ | — |

**Distributional novelty via FVD.** To quantify the novelty of `SumReason` relative to existing benchmarks, we compute the **Fréchet Video Distance (FVD)** (Unterthiner et al., 2018). This metric measures the 2-Wasserstein distance between two multivariate Gaussians, $(\mu_1, \Sigma_1)$ and $(\mu_2, \Sigma_2)$, representing the feature distributions of the respective datasets. Unlike spatial-only metrics, we utilize a **3D-ResNet (R3D)** (Hara et al., 2018; Du et al., 2021; Tran et al., 2018) backbone pre-trained on **Kinetics-400** (Kay et al., 2017) to extract comprehensive spatio-temporal features:

$$d^2 = \underbrace{\|\mu_1 - \mu_2\|_2^2}_{\text{Centroid Shift}} + \underbrace{\text{Tr}(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})}_{\text{Geometric Deformation}} \tag{3}$$

The *Centroid Shift* represents the average difference in semantic and motion content, while the *Geometric Deformation* accounts for variances in temporal correlation and feature-space orientation. Given our sample size ($N = 37$), we apply Tikhonov regularization ($\Sigma + \epsilon I$ with $\epsilon = 10^{-6}$) to ensure numerical stability during the computation of the covariance square roots.

The FVD scores shown in Table 6 demonstrate that `SumReason` is not a redundant subset of legacy benchmarks:

- **OOD Challenge:** The significant distance from SumMe (186.44) confirms a major out-of-distribution (OOD) shift in motion dynamics and narrative structure.
- **Feature Space Expansion:** The divergence from TVSum (120.59) mathematically justifies the necessity of this benchmark for evaluating model robustness beyond the biases inherent in legacy visual-temporal priors.

### A.3 Higher-Resolution Inter-Annotator Agreement Comparison

**Higher-resolution analysis.** We present a pairwise inter-annotator $F_1$-score comparison in Figure 7. In both the query-conditioned and standard settings, our dataset exhibits superior $F_1$ scores, averaging 65% and 61% respectively (see Figures 7(a) and (b)). These results represent a significant improvement in consensus compared to the well-established TVSum and SumMe datasets (see Figures 7(c) and (d)).

## B Additional details about our dataset

### B.1 Descriptive Statistics

**The statistical properties.**, As summarized in Table 7, underscore the distinct temporal and linguistic profiles of its constituent reasoning domains. The **External Knowledge Reasoning (EKR)** category represents the most computationally demanding subset, featuring the longest average video duration ($252.3$s) and the highest mean frame count ($8,245$). In terms of linguistic structure, the **Temporal Reasoning (TR)** domain stands out with the longest average queries ($16.4$ words), as a result of the detailed chronological descriptors required to ground temporal requests. Conversely, the **High-Level Semantic (HLS)** domain exhibits a high Type-Token Ratio (TTR) of $0.604$ despite a relatively small vocabulary size of $58$. This suggests that while HLS queries are concise, they utilize a highly specific and diverse set of terms to convey nuanced narrative intent. Collectively, these metrics demonstrate that `SumReason` provides a balanced yet challenging benchmark, spanning a wide range of video lengths and query complexities.

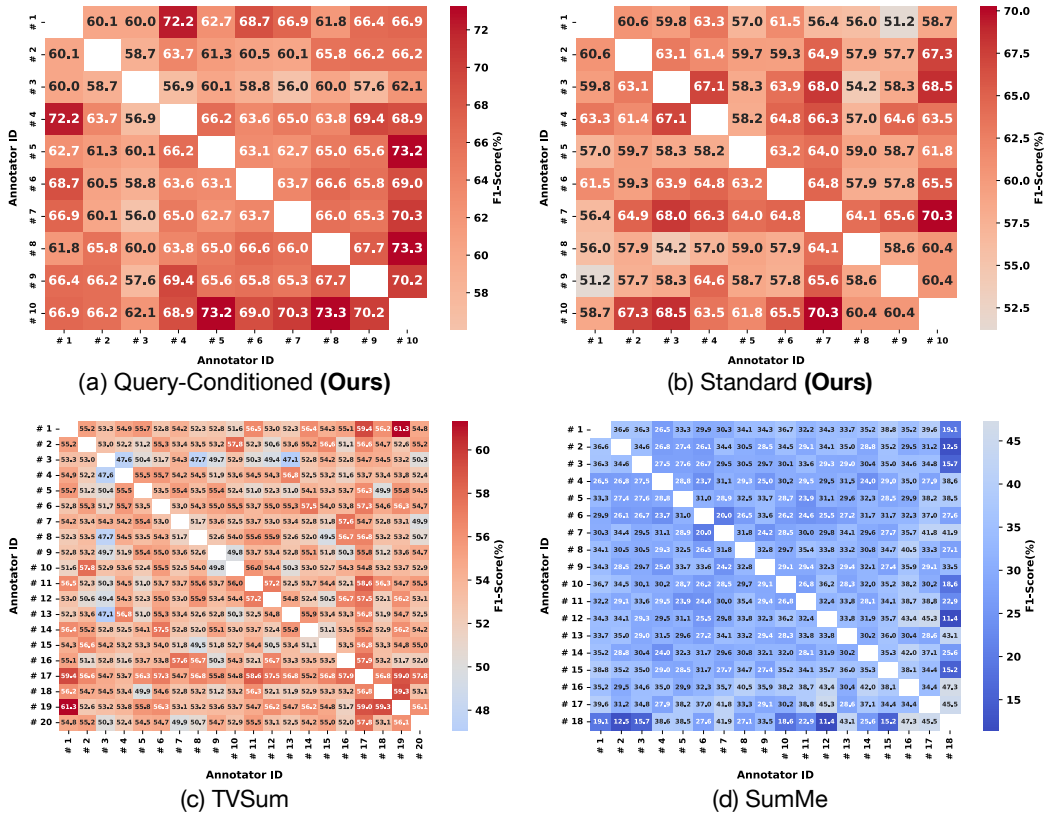



Figure 7: **Inter-annotator pair f1-score comparison.** A comparison of annotators agreement across 4 datasets. Our dataset `SumReason` exhibits a superior $F_1$ scores in Query-conditioned (a) and Standard (b) settings, **averaging above 60%** relative to the TVSum (c) and SumMe (d) datasets, **averaging 53% and 31%** respectively, suggesting a higher quality annotations.

Table 7: **Linguistic and temporal statistics of `SumReason`**. We report mean values for temporal metrics and linguistic diversity across Direct Visual (DV), Temporal Reasoning (TR), High-Level Semantic (HLS), and External Knowledge Reasoning (EKR) categories. The Type-Token Ratio (TTR) represents the mean lexical density per category.

| Category | Avg. #frm | FPS | Avg. Dur. (s) | Avg. Qry Len. | Vocab. Size | TTR (↑) |
|---|---|---|---|---|---|---|
| DV | 5,856 | 27.5 | 214.4 | 10.9 | 113 | 0.577 |
| TR | 4,863 | 27.8 | 183.1 | 16.4 | 174 | 0.530 |
| HLS | 6,138 | 26.8 | 229.5 | 10.7 | 58 | 0.604 |
| EKR | 8,245 | 30.7 | 252.3 | 12.8 | 100 | 0.602 |
| **Overall** | **6,083** | **28.2** | **214.4** | **13.1** | **353** | **0.578** |

### B.2 Additional Query Examples

**Cognitive demands across categories.** The cognitive requirements of each category follow directly from the **query construction** and the specific video events they are designed to capture. The examples below illustrate the visual reasoning, temporal understanding, semantic interpretation, and external knowledge required by representative queries.

**Direct Visual Understanding (DV).**

**Query 1:** "Highlight segments where the wingsuit pilot navigates extremely close to terrain or other objects during flight."

**Requirement:** Requires continuous spatial and depth perception to assess proximity on a graded scale, distinguishing ordinary distant flight from increasingly close and high-risk interactions with surrounding terrain or objects.

**Query 2:** "Highlight scenes where only the shuttle's main engines are firing (visible fire is seen)."

**Requirement:** Requires fine-grained visual discrimination across the 1–5 relevance scale, separating initial ignition or smoke from scenes with clearly visible main-engine combustion.

**Temporal Reasoning (TR).**

**Query 1:** "Highlight the segments where the race winner takes the lead, and where he crosses the finish line till the 2nd place winner crosses it as well."

**Requirement:** Requires global temporal tracking across multiple events and subjects, including the buildup to the overtake, the lead transition, and the interval between the winner and second-place finisher crossing the finish line.

**Query 2:** "Focus on the scenes which directly lead to the referee making a call or signal."

**Requirement:** Requires causal temporal reasoning, linking the preceding foul or violation to the referee's subsequent decision and assigning relevance across the full causal sequence.

**High-Level Semantic Understanding (HLS).**

**Query 1:** "Highlight scenes where the protagonist expresses sadness, grief, or regret."

**Requirement:** Requires interpreting subtle affective and behavioral cues and mapping them onto an emotional-intensity scale, rather than relying only on explicit actions or objects.

**Query 2:** "Highlight scenes that show a clear emotional shift, such as a transition between different emotions."

**Requirement:** Requires tracking character-level emotional trajectories over time and identifying transitional moments where one emotional state develops into another.

**External Knowledge Reasoning (EKR).**

**Query 1:** "Highlight every scene where non-power (hand) tools are used."

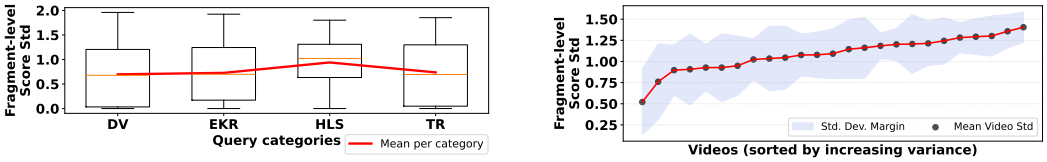
**Requirement:** Requires external knowledge of tool categories to distinguish hand tools from powered tools and to score active use more highly than ambiguous or incidental appearances.

**Query 2:** "Highlight drinks with semi-iced or slushy consistency or texture; do not include fully liquid drinks."

**Requirement:** Requires domain knowledge of physical consistency and texture to distinguish fully liquid drinks from partially frozen, semi-iced, or slushy states along a graded relevance scale.

### B.3 Fragment-Level Variability

Following the methodology of Song et al. (2015); Apostolidis et al. (2020b), we partition the videos into temporal segments using Kernel Temporal Segmentation (KTS) (Potapov et al., 2014). These segments serve as the fundamental units from which the final summaries are constructed. Detailed implementation specifics are provided in Appendix Section B.5. In Figure 8, we analyze fragment-level score variability to assess consistency between annotators. As shown in Figure 8a, disagreement in query-focused summarization varies significantly across reasoning domains, differing from traditional importance-based benchmarks. High-Level Semantic (HLS) queries lead in mean disagreement, highlighting the subjective nature of interpreting intent or narrative. In contrast, the lower medians for Direct Visual (DV) and Temporal Reasoning (TR) tasks point to a stronger consensus on physical and chronological cues. The wider variance in External Knowledge Reasoning (EKR) further underscores how annotator background knowledge influences perceived fragment relevance. In addition, Figure 8b illustrates fragment-level variability between user annotations for the standard summary task—focusing on the main events, actions and core messages. On average, the fragment-level std. is $\approx 0.9$ with a 0.2 margin. These values remain relatively low considering




(a) Query-conditioned annotations

(b) Standard annotations

Figure 8: **Fragment-level variability**. We report fragment-level std. across annotations. (a) **Query-conditioned annotations**: High level semantics (HLS) queries exhibit the highest subjectivity, while Direct visual (DV) and Temporal Reasoning (TR) show greater consensus. (b) **Standard annotations**: videos sorted by increasing variance; the red line tracks the mean fragment-level std.

the 1–5 scoring scale, suggesting that despite the inherent subjectivity of summarization, there is a consistent underlying consensus among annotators.

### B.4 Additional Annotation Details

**Preprocessing.** All videos were annotated without audio, ensuring that annotation decisions relied exclusively on visual information. During the pilot annotation phase, we observed that introductory and concluding segments (e.g., title screens and credits) exhibited high variability in importance scores across annotators. Because these segments were typically weakly related to the core video content and introduced unnecessary annotation noise, they were systematically trimmed from all videos prior to annotation.

Additionally, a subset of videos containing relatively static content was temporally downsampled (e.g., $\times 2$ playback speed). In these videos, individual events extended over substantially longer durations than average, resulting in considerable temporal redundancy. Increasing the temporal density of these low-paced segments reduced redundancy while preserving their semantic content. Crucially, temporally downsampled videos were never paired with Temporal Reasoning (TR) queries, thereby preserving the original motion dynamics and duration cues for queries requiring complex temporal dependencies.

**Annotator training.** A cohort of 10 annotators participated in the study, with all annotators completing an interactive, in-person, multi-stage training program before beginning annotation. Training first introduced the broader video summarization task, annotation objectives, evaluation protocol, and detailed annotation guidelines. Annotators then completed practice video sets spanning the same domains, query types, and core event themes represented in SumReason. Through discussions with peer annotators and instructors, they were calibrated on fine-grained temporal scoring and, in particular, on distinguishing core events, indirectly query-relevant events, contextual or narrative buildup, and irrelevant or filler content.

A central component of training was emphasizing that annotation is a *fine-grained, non-binary* scoring task rather than binary highlight detection. Annotators were explicitly trained to distinguish core events, which should generally receive the highest scores, from relevant contextual buildup or transitions, which may receive intermediate scores, and filler content, which should receive low scores. Pilot examples were used to calibrate this interpretation and reduce inter-annotator variability before the main annotation phase.

**Annotation protocol and task guidelines.** Annotators were instructed to watch each video in its entirety from start to finish before assigning scores, with explicit emphasis on annotation quality rather than annotation speed. We considered two annotation settings:

**(A) Standard Video Summarization.** Annotators rated the overall importance of each video segment using a five-point importance scale: 1 = not important at all, 2 = slightly important, 3 = moderately important, 4 = important, and 5 = very important. Annotators were instructed to consider whether a segment contained an important action or event, contributed meaningfully to the main narrative, or provided essential context for understanding subsequent events.

**(B) Query-Conditioned Video Summarization.** Annotators rated each segment strictly according to its relevance to the specified query, independently of its general importance to the overall video. We used a five-point relevance scale: 1 = not relevant at all, 2 = slightly relevant, 3 = moderately relevant, 4 = relevant, and 5 = very relevant. This distinction ensured that segments important to the video's general narrative were not automatically assigned high scores when they were unrelated to the query.

**Explicit annotation constraints.** Annotators were required to understand and verify any definitions, rules, or domain-specific context implied by a query before assigning scores. For both standard and query-conditioned annotation, instructors repeatedly emphasized non-binary scoring and temporal context. In particular, segments surrounding a core event could receive meaningful intermediate scores when they provided necessary buildup, transitions, or contextual information, even when they did not contain the focal event itself. Conversely, high scores were reserved for segments containing the central event or information directly relevant to the annotation objective.

**Quality control and re-annotation.** All query-conditioned annotations satisfied our quality-control criteria during the initial validation pass and therefore required no re-annotation. For standard video summarization, fewer than $2\%$ of annotations were rejected during quality control: 15 out of 850 individual annotations, spanning four videos and four annotators. Rejected annotations primarily exhibited one of two issues: (i) failure to capture the complete temporal extent of a core event together with its relevant contextual buildup, or (ii) inconsistent, isolated scoring patterns resulting from a misunderstanding of the fine-grained annotation protocol. These annotations underwent a targeted second-round annotation procedure. As detailed in the corresponding agreement analysis, this targeted re-annotation improved overall annotation consistency across the primary agreement metrics.

### B.5 Evaluation Setup

**Fragment segmentation.** Rather than adopting the rigid constraints of uniform segmentation (Zhao & Xing, 2014), our framework leverages Kernel Temporal Segmentation (KTS) (Potapov et al., 2014) to partition videos into semantically grounded fragments. We formalize the segmentation process as a change-point detection task (Potapov et al., 2014), a statistically principled method for capturing temporal shifts in the data distribution (Basseville et al., 1993). This approach ensures that the resulting shots are visually consistent, providing a stable basis for evaluating complex temporal dependencies and external knowledge integration.

**Fragment-based scoring.** Initially, shot-level scores (each 2 consecutive non-overlapping seconds) are projected into frame-level importance scores, and the importance score of each fragment is then computed as the mean of the frame-level scores within its temporal boundaries.

**Summary generation.** Following prior work (Song et al., 2015; Apostolidis et al., 2020b), and given fragment-level importance scores,we generate a summary by solving a 0/1 knapsack problem to maximize the captured importance within a fixed time budget:

$$\max_{u} \sum_{i=1}^{N} u_i v_i \;\; \text{subject to} \;\; \sum_{i=1}^{N} u_i w_i \leq L, \; u_i \in \{0, 1\}.$$

where $v_i$ and $w_i$ denote the importance score and duration of fragment $i$, respectively, and $u_i$ is a binary indicator of whether fragment $i$ is included in the summary, and $N$ is the total number of fragments, respectively. The final summary is constructed by concatenating selected fragments in chronological order, and the summary budget $L$ is set to 15% of the original video duration.

**Evaluation metrics and protocol.** Following the established evaluation protocol in (Apostolidis et al., 2020b), we generated five randomly-generated, non-overlapping train/test splits with a ratio of 80/20. We report the average performance metrics across these five splits for all evaluated models. We assess the quality of the automatically generated summary $\mathcal{S}$ by measuring the average agreement between it and every single gold standard $\mathcal{GT}_i \in \mathcal{GT}$ using the **F1-score**.

### B.6 Results Reproducibility

All experiments are fully reproducible using the provided dataset and publicly available implementations of baseline methods. For all standard (non query-conditioned) settings, we directly

used the official codebases and followed their original training and evaluation protocols without modification, ensuring faithful reproduction of reported results. For query-conditioned settings, we adapted LGRLN (Li et al., 2025) and Prompts2Summs (P2S) (Barbara & Maalouf, 2025) to our benchmark while preserving their original architectures and optimization procedures. The adaptations were limited to ensuring compatibility with the `SumReason` evaluation setting and did not modify the underlying learning objectives or model design. For FCSNA (Patel et al., 2024), we used the publicly available implementation based on the CHAN architecture originally proposed for the QFVS (Sharghi et al., 2016) dataset, and adapted it to support our dataset format and evaluation pipeline.

## C Statistical analysis and robustness

**Temporal Reasoning remains the primary bottleneck.** Table 8 reveals a consistent category-dependent gap across both open-weight and closed-source MLLMs. Under the 2-second evaluation, models achieve substantially stronger performance on visually or semantically grounded queries: DV ranges from $54.47$–$59.98\%$ and HLS from $57.90$–$61.16\%$. In contrast, TR remains between $47.50$ and $51.54\%$. The same pattern persists with KTS segmentation, where HLS reaches up to $61.25\%$, while TR remains within $46.52$–$53.05\%$. Thus, increasing model capability alone does not eliminate the degradation on queries requiring information to be integrated across temporally separated events.

The magnitude of this effect is particularly clear relative to the reference baselines. PoR achieves $47.2\%$ F1 on TR, meaning that several MLLM configurations operate only a few points above the random-prior baseline on this category, despite substantially outperforming it on DV and HLS. By comparison, the human ceiling (PoH) reaches $63.8\%$ on TR. Even the strongest MLLM TR result, Gemini 3.1 Pro with KTS at $53.05\%$, therefore remains more than $10$ F1 points below the human reference. This gap is considerably larger than on HLS, where the strongest MLLM result ($61.25\%$) approaches the PoH value of $65.4\%$. These results suggest that the principal limitation is not general visual relevance estimation, but reasoning over events whose relevance depends on their temporal relationships.

**The bottleneck is robust to temporal granularity.** To test whether the TR degradation is an artifact of using fixed 2-second segments, we evaluate the same MLLMs using KTS-derived semantic shot boundaries. The change in segmentation does not produce a consistent improvement. For TR, moving from 2-second segments to KTS changes performance from $47.50$ to $50.46\%$ for Qwen3-VL-7B, $48.61$ to $46.52\%$ for InternVL3-8B, $50.98$ to $48.90\%$ for GPT-5.6 Luna, and $51.54$ to $53.05\%$ for Gemini 3.1 Pro. The corresponding changes range from approximately $-2.1$ to $+3.0$ F1 points and vary in direction across models.

A similar lack of systematic improvement is visible in the overall scores. Qwen3-VL-7B (Yang et al., 2025) is essentially unchanged ($52.01$ vs. $52.05\%$), Gemini 3.1 Pro (Google DeepMind, 2026a) changes only slightly ($54.67$ vs. $54.99\%$), while InternVL3-8B (Zhu et al., 2025) and GPT-5.6 Luna (OpenAI, 2026) decrease under KTS. Hence, semantic shot boundaries can redistribute performance across individual categories, but they do not remove the underlying TR deficit. This indicates that the observed bottleneck is robust to whether the model receives uniformly sampled temporal units or content-adaptive semantic shots.

**Comparison with the PoR baseline.** Table 9 complements the category-level results with paired comparisons against PoR. Several MLLM configurations separate clearly from the random-prior baseline, including GPT-5.6 Luna with 2-second segments ($p = 1.8 \times 10^{-5}$), Gemini 3.1 Pro with KTS ($p = 0.0006$), InternVL3-8B with 2-second segments ($p = 0.0011$), and Gemini 3.1 Pro with 2-second segments ($p = 0.0029$). GPT-5.6 Luna with KTS also exceeds PoR at the conventional $0.05$ level ($p = 0.0445$). In contrast, several other MLLM and text-LLM configurations do not show a corresponding separation from PoR, despite obtaining competitive aggregate F1 scores.

The supervised FCSNA baseline also differs from PoR ($p = 0.0034$), whereas LGRLN (Li et al., 2025) does not ($p = 0.1951$). The P2S text-LLM (Barbara & Maalouf, 2025) variants similarly yield large paired $p$-values, indicating that replacing the scoring component with a stronger text LLM does not by itself produce a consistent gain over PoR under this evaluation. Together with the category-level results, these tests show that the strongest gains arise from direct multimodal scene evaluation, while substantial room remains between current MLLMs and human agreement, particularly for Temporal Reasoning.

Table 8: **Benchmarking query-conditioned video summarization methods.** We report F1 scores (mean ± standard deviation, %) across query categories and overall. Underlined values indicate the lowest-performing category for each method; the final row reports how often each category has lowest accuracy across all methods.

| Model | DV | TR | HLS | EKR | Avg. | Lowest |
|---|---|---|---|---|---|---|
| **Baselines** | | | | | | |
| PoR | $50.0 \pm 13.29$ | $\underline{47.2} \pm 13.01$ | $53.4 \pm 11.57$ | $49.1 \pm 14.45$ | $\mathbf{49.9 \pm 13.13}$ | TR |
| PoH | $\underline{63.6} \pm 12.10$ | $63.8 \pm 11.85$ | $65.4 \pm 10.40$ | $65.5 \pm 11.20$ | $\mathbf{64.6 \pm 11.39}$ | DV |
| RS | $23.1 \pm 15.40$ | $\underline{22.6} \pm 14.80$ | $22.7 \pm 13.90$ | $24.4 \pm 15.10$ | $\mathbf{23.2 \pm 14.80}$ | TR |
| US | $14.1 \pm 12.30$ | $\underline{11.0} \pm 11.10$ | $14.5 \pm 10.80$ | $18.5 \pm 13.20$ | $\mathbf{15.2 \pm 11.85}$ | TR |
| CK | $\underline{13.0} \pm 10.50$ | $14.4 \pm 11.20$ | $15.9 \pm 9.80$ | $21.0 \pm 12.40$ | $\mathbf{16.1 \pm 10.98}$ | DV |
| CLIP | $28.4 \pm 16.10$ | $30.3 \pm 15.90$ | $31.7 \pm 14.20$ | $\underline{26.6} \pm 15.80$ | $\mathbf{29.2 \pm 15.50}$ | EKR |
| **Prior SOTA** | | | | | | |
| LGRLN | $47.36 \pm 19.30$ | $\underline{43.57} \pm 15.94$ | $51.01 \pm 14.14$ | $51.68 \pm 14.21$ | $\mathbf{47.58 \pm 16.23}$ | TR |
| FCSNA | $56.24 \pm 17.10$ | $\underline{50.43} \pm 13.21$ | $59.11 \pm 11.01$ | $50.62 \pm 13.28$ | $\mathbf{53.52 \pm 14.27}$ | TR |
| **P2S (LLM)** | | | | | | |
| GPT-5.6-Sol | $55.47 \pm 15.80$ | $\underline{49.00} \pm 13.50$ | $58.98 \pm 11.20$ | $53.80 \pm 14.10$ | $\mathbf{53.82 \pm 13.90}$ | TR |
| GPT-4o | $56.04 \pm 14.71$ | $\underline{50.06} \pm 15.02$ | $56.14 \pm 14.41$ | $53.69 \pm 15.64$ | $\mathbf{50.27 \pm 16.04}$ | TR |
| Gemini-3.6-Flush | $52.30 \pm 14.21$ | $\underline{44.51} \pm 15.70$ | $56.82 \pm 10.15$ | $49.91 \pm 13.43$ | $\mathbf{49.86 \pm 14.39}$ | TR |
| Claude-Fable-5 | $50.89 \pm 9.53$ | $\underline{46.24} \pm 17.63$ | $58.93 \pm 13.32$ | $52.22 \pm 14.55$ | $\mathbf{50.94 \pm 14.88}$ | TR |
| Qwen3.6 27B | $50.34 \pm 12.98$ | $50.50 \pm 17.78$ | $52.41 \pm 12.02$ | $\underline{47.20} \pm 13.21$ | $\mathbf{50.02 \pm 14.36}$ | EKR |
| Qwen3-VL 7B | $50.53 \pm 11.16$ | $47.28 \pm 14.40$ | $56.40 \pm 9.06$ | $\underline{47.05} \pm 14.97$ | $\mathbf{49.57 \pm 13.04}$ | EKR |
| **MLLM-BL-2s (MLLM)** | | | | | | |
| Qwen3-VL-7B | $59.08 \pm 17.69$ | $47.50 \pm 20.03$ | $59.07 \pm 14.92$ | $\underline{44.26} \pm 13.69$ | $\mathbf{52.01 \pm 18.18}$ | EKR |
| InternVL3-8B | $59.49 \pm 19.01$ | $\underline{48.61} \pm 15.95$ | $61.16 \pm 13.91$ | $57.93 \pm 13.05$ | $\mathbf{55.78 \pm 16.55}$ | TR |
| GPT 5.6 Luna | $59.98 \pm 16.11$ | $\underline{50.98} \pm 17.38$ | $59.88 \pm 14.84$ | $54.36 \pm 16.34$ | $\mathbf{55.75 \pm 16.50}$ | TR |
| Gemini 3.1 Pro | $54.47 \pm 19.11$ | $\underline{51.54} \pm 15.57$ | $57.90 \pm 19.64$ | $57.53 \pm 17.39$ | $\mathbf{54.67 \pm 17.43}$ | TR |
| **MLLM-BL-KTS (MLLM)** | | | | | | |
| Qwen3-VL-7B | $52.84 \pm 18.66$ | $50.46 \pm 19.94$ | $61.14 \pm 14.78$ | $\underline{47.13} \pm 15.53$ | $\mathbf{52.05 \pm 17.82}$ | EKR |
| InternVL3-8B | $57.52 \pm 18.59$ | $\underline{46.52} \pm 14.90$ | $58.36 \pm 16.98$ | $55.72 \pm 17.85$ | $\mathbf{53.59 \pm 17.35}$ | TR |
| GPT 5.6 Luna | $56.08 \pm 14.63$ | $\underline{48.90} \pm 13.59$ | $61.25 \pm 11.50$ | $51.06 \pm 14.88$ | $\mathbf{53.37 \pm 14.28}$ | TR |
| Gemini 3.1 Pro | $55.14 \pm 19.25$ | $\underline{53.05} \pm 18.15$ | $55.89 \pm 20.12$ | $57.17 \pm 17.76$ | $\mathbf{54.99 \pm 18.28}$ | TR |
| **Lowest frequency** | **2/22 (9.1%)** | **15/22 (68.2%)** | **0/22 (0%)** | **5/22 (22.7%)** | – | – |

Table 9: **Paired significance tests against PoR.** We report paired $p$-values comparing each method with the PoR baseline.

| Model | Category | Paired $p$-value (vs. PoR) |
|---|---|---|
| PoH (human upper bound) | Human ceiling | $5.0 \times 10^{-10}$ |
| GPT-5.6 Luna (2s) | MLLM (per-scene) | $1.8 \times 10^{-5}$ |
| Gemini 3.1 Pro (KTS) | MLLM (whole-video) | 0.0006 |
| InternVL3-8B (2s) | MLLM (per-scene) | 0.0011 |
| Gemini 3.1 Pro (2s) | MLLM (whole-video) | 0.0029 |
| FCSNA-QFVS | Supervised | 0.0034 |
| GPT-5.6 Luna (KTS) | MLLM (per-scene) | 0.0445 |
| InternVL3-8B (KTS) | MLLM (per-scene) | 0.0962 |
| LGRLN | Supervised | 0.1951 |
| Qwen3-VL-8B (KTS) | MLLM (per-scene) | 0.2108 |
| Qwen3-VL-8B (2s) | MLLM (per-scene) | 0.2858 |
| Claude-fable-5 | Text-LLM (P2S) | 0.3027 |
| Qwen3-VL-8B (P2S) | Text-LLM (P2S) | 0.7628 |
| Qwen3.6-27B | Text-LLM (P2S) | 0.7797 |
| Gemini-3.6-flash | Text-LLM (P2S) | 0.9238 |
| GPT-5.6 Sol | Text-LLM (P2S) | 0.9824 |

# D ZERO-SHOT MLLM EVALUATION SETUP

To evaluate frontier Multimodal Large Language Models (MLLMs) without task-specific fine-tuning, we developed a zero-shot scene-scoring pipeline that integrates directly into our standard Knapsack evaluation framework.

## D.1 PIPELINE ARCHITECTURE

Given a video $V$ and query $Q$, $V$ is first segmented into discrete scenes using two distinct temporal granularities: (i) fixed 2-second uniform segments (`MLLM-BL-2s`), and (ii) KTS semantic shot boundaries (`MLLM-BL-KTS`), matching ground-truth shot boundaries. For each scene independently, frames are uniformly sampled at 1 fps (capped at a maximum of 16 frames per scene). Raw sampled frames alongside the text query $Q$ are ingested in a single MLLM API call, where the model evaluates query relevance by producing an explicit integer score from 1 to 5 based on a structured rubric. Finally, these per-scene integer scores are normalized into frame-level probabilities, enabling standard 0/1 Knapsack optimization under strict length budget constraints ($\leq 15\%$) for direct $F_1$ metric computation against human ground truth.

## D.2 PROMPT TEMPLATE

The exact system prompt utilized across all MLLM evaluations is structured as follows:

> You are an expert video summarization judge specializing in visual temporal reasoning and query-conditioned relevance evaluation. **Task:** You are given a natural-language query and a temporal sequence of sampled frames from a single video scene. Judge how relevant this scene is to the query and assign an integer relevance score from 1 to 5.
>
> **Relevance Rubric:**
>
> - 1 (Irrelevant): No visual elements related to the query; background filler.
> - 2 (Marginally Relevant): Minor related visual context, misses the core event.
> - 3 (Moderately Relevant): Partially depicts the queried event, lacking key details/actions.
> - 4 (Highly Relevant): Clearly depicts the queried event with strong visual evidence.
> - 5 (Essential): A primary visual highlight directly satisfying the central query intent.
>
> **Instructions:** 1. Analyze the visual content and temporal progression across the frames. 2. Compare the visual evidence against the query. 3. Answer with a single integer from 1 to 5, nothing else. **Query:** "<QUERY>"
> [sampled scene frames]

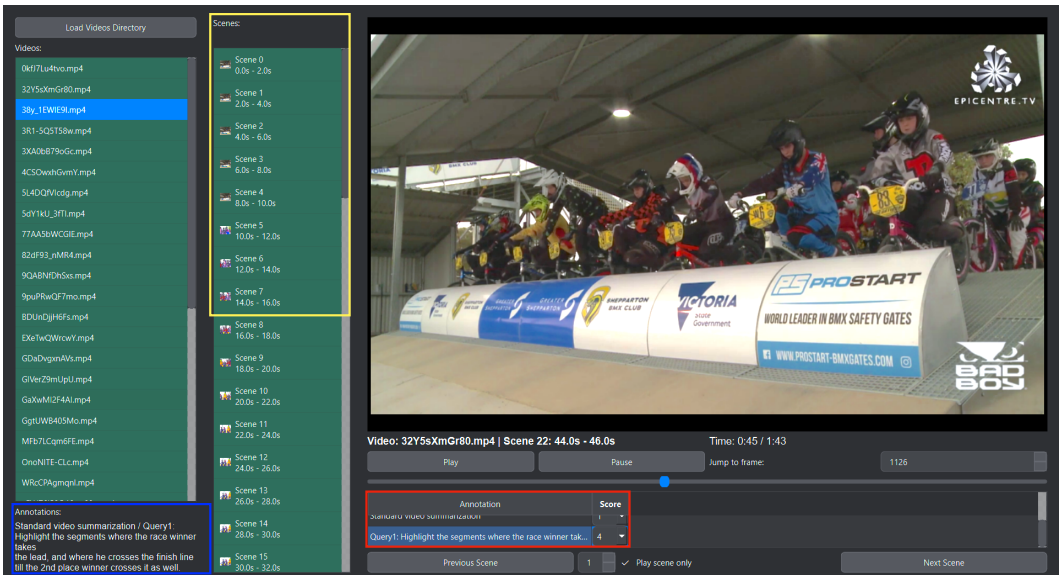



Figure 9: **Video annotation framework.** The framework presents the video player (top-right), a timeline for precise temporal navigation, and an annotation panel (bottom) where annotators assign importance scores to selected segments based on a given query. The left panel lists available videos and pre-segmented scenes, enabling efficient navigation and consistent annotation across the dataset. This design supports fine-grained, query-conditioned annotation and facilitates multi-annotator consistency.

## E Future Work

**Broader impact.** This work enable testing AI systems to summarize videos by making them better understand user intent and complex reasoning, enabling more efficient and personalized access to large-scale video content.

Finally, while `SumReason` enables consistent evaluation using multiple human annotations, summary quality is inherently multi-faceted and cannot be fully captured by overlap-based metrics alone. Measures such as F1 provide a useful and standard benchmark, but future work may further enrich evaluation by considering user intent, diversity, and human preference.

## F Video Annotation Framework

**Video annotation framework overview.** To support consistent and efficient data collection, we developed a dedicated video annotation interface tailored for video summarization. As shown in Figure 9, the framework allows annotators to view videos alongside a structured set of pre-segmented scenes, enabling precise temporal navigation and selection. Annotators are provided with a query and are asked to assign importance scores to specific segments using an intuitive annotation panel or they are asked to addign an importanse score for each shot based on the general importance of a shot withing the video (standard setting). The system includes a timeline slider for fine-grained control, real-time playback, and visual indicators of segment boundaries, ensuring accurate alignment between video content and annotations. A side panel lists all available scenes within each video, allowing annotators to systematically review and compare segments. This design promotes consistency across annotators and supports efficient collection of high-quality, and reasoning-driven annotations.