HAVEN: HIERARCHICALLY ALIGNED MULTIMODAL BENCHMARK FOR MULTI-TASK VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent research often evaluates models' video understanding abilities through tasks such as summarization, temporal reasoning, and cross-modal reasoning. Existing benchmarks typically focus on a single task and a single level of granularity for each video, limiting their ability to provide a comprehensive evaluation of video understanding. We introduce HAVEN, a benchmark that links visual and textual annotations at the frame, shot, and video levels. Built from six existing video-summarization datasets, HAVEN contains 11,055 videos and uses a 365-video subset with human-audited textual annotations for standard evaluation. Its task suite covers summarization, temporal ordering, cross-modal grounding, and saliency ranking. Empirical experiment results reveal a comprehensive and fine-grained evaluation of the video understanding capabilities of widely used vision-language models. These observations show that textual-summary scores alone do not capture the variation measured by visual selection and alignment tasks. HAVEN provides a shared annotation structure for examining these complementary aspects of video understanding.

1 INTRODUCTION

Multimodal large language models (MLLMs) have rapidly advanced video understanding, achieving strong performance on tasks such as summarization (Lin et al., 2024; Zhu et al., 2018; Han et al., 2023), captioning (Xu et al., 2016; Wang et al., 2019; Caba Heilbron et al., 2015), and question answering (Li et al., 2024; Fu et al., 2025; Liu et al., 2024b). However, these gains raise a fundamental question: *do MLLMs genuinely understand video content, or do they produce plausible outputs by exploiting language priors?* Resolving this ambiguity is challenging due to the lack of evaluation frameworks capable of rigorously assessing fine-grained, temporal alignment between visual evidence and generated language.

Current video benchmarks suffer from two systemic limitations (Li et al., 2024; Fu et al., 2025; Chandrasegaran et al., 2024; Liu et al., 2024b;a). First, they provide fragmented supervision. Most datasets operate at a single temporal granularity (e.g., keyframes, clips, or video summaries), ignoring the inherently hierarchical structure of real-world videos, where frames build shots, and shots construct coherent narratives. Second, they lack precise cross-modal grounding. Text annotations are typically weakly aligned with visual content, making it difficult to assess whether models actually localize and reason over the relevant evidence (Yuan & Zhang, 2025). Together, these limitations create an evaluation blind spot: models can achieve high scores through fluent generation while exhibiting poor multimodal alignment. As a result, current evaluation protocols systematically overestimate the capabilities of modern MLLMs, particularly on tasks that do not require temporal compositionality or saliency-aware reasoning (Liu et al., 2024b; Li et al., 2024; Fu et al., 2025; Li et al., 2023; Sun et al., 2024).

To address these limitations, we introduce HAVEN (Hierarchically Aligned multimodal benchmark for multi-task Video undErstandiNg), a hierarchically aligned multimodal benchmark that supports evaluation of video understanding across multiple temporal levels and complementary tasks. HAVEN is built upon two core principles: (1) **Hierarchical completeness**, providing unified annotations at the frame, shot, and video levels to enable evaluation across multiple temporal scales; and

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.

(2) **Cross-modal continuity**: dense, explicit alignment between visual content and textual descriptions, ensuring that language must be grounded in temporally localized evidence.

Built on this shared hierarchical annotation structure, HAVEN supports tasks for summarization, temporal ordering, cross-modality grounding, and saliency ranking. Together, these tasks enable models to be evaluated from complementary perspectives on the same video collection, including textual summarization, visual evidence selection, shot-level ordering, and sentence-to-visual grounding.

We benchmark a range of state-of-the-art MLLMs on HAVEN (Figure 1) and observe substantial differences across evaluation dimensions. Models with similar textual-summary scores can differ considerably in visual evidence selection. We further observe task-dependent model profiles across temporal ordering, saliency ranking, and multimodal grounding. Together, these results demonstrate the value of evaluating complementary aspects of video understanding within a shared hierarchical benchmark.

Our main contributions are three-fold: (1) We introduce HAVEN, a novel hierarchically structured multimodal dataset with annotations at the frame, shot, and video levels and explicit links across modalities and levels. (2) We establish a rigorous testing framework featuring tasks across four dimensions (summarization, multimodal grounding, temporal ordering, and saliency ranking), enabling a multi-faceted assessment of model capabilities. (3) We provide an extensive empirical analysis that identifies key limitations in state-of-the-art MLLMs, exposing a marked discrepancy between textual summarization performance and performance on tasks requiring explicit visual evidence selection and multimodal grounding.

2 Related Work

Video Understanding Benchmarks. Recent benchmarks have advanced the evaluation of MLLMs on video understanding, covering tasks such as event understanding, temporal reasoning, question answering, and multi-shot comprehension (Li et al., 2024; Fu et al., 2025; Liu et al., 2024b; Han et al., 2023; Caba Heilbron et al., 2015; Huang et al., 2025). Recent works have introduced benchmarks designed for long video understanding, such as Hourvideo (Chandrasegaran et al., 2024) and other Q&A-based benchmarks(Xiao et al., 2021; 2024). These benchmarks extend evaluation to more realistic long-video scenarios, but are typically organized around isolated task formats and focus on general understanding. As a result, they provide limited support for jointly evaluating model capabilities of diverse dimensions within a unified framework. Compared to existing benchmarks, we enhance the dataset with richer multimodal signals and fine-grained cross-modal alignment, which enables a more diverse set of tasks beyond traditional Q&A or coarse-grained formulations. This provides a new perspective for evaluating video understanding.

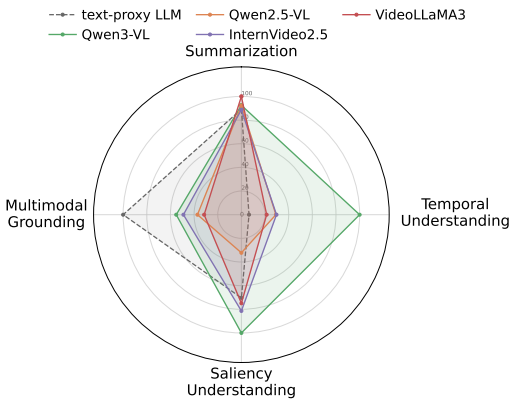

Figure 1: Comparison between different MLLMs on HAVEN across different capabilities. We include text-proxy LLMs as baselines.

Video Summarization Benchmarks. Video summarization research has been supported by datasets such as SumMe (Gygli et al., 2014), TVSum (Yale Song et al., 2015), and subsequent resources (Sul et al., 2023; Lin et al., 2024), enabling studies on key segment selection, importance estimation, and summary generation (Gong et al., 2014; Zhang et al., 2016; Rochan & Wang, 2019). Other large-scale video-text datasets (Wang et al., 2023; Chen et al., 2024; Huang et al., 2020; Zhang, 2025) support multiple tasks on video understanding evaluations. More recent work has explored multimodal understanding and evaluation, including cross-modal integration, factual consistency, and evaluation reliability (Zhu et al., 2018; Liu et al., 2025; Otani et al., 2019; Wan

Table 1: **Source datasets used to construct HAVEN Benchmark.** Numbers in parentheses under "#Videos" indicate the number of videos included in the standard evaluation subset. "Alignment" denotes aligned annotations between visual and textual data, "K_f" = key frames, "D_f" = frame descriptions, "K_s" = key shots, and "D_s" = shot descriptions.

Dataset	#Videos (Eval.)	Modalities	Alignment	K_f	D_f	K_s	D_s
HAVEN	**11055 (365)**	**V+T→V+T**	✓	✓	✓	✓	✓
↪ SumMe (2014)	25 (25)	V→V	×	✓	×	✓	×
↪ TVSum (2015)	50 (50)	V→V	×	✓	×	✓	×
↪ OVP (2011)	50 (50)	V→V	×	✓	×	×	×
↪ YouTube (2011)	40 (40)	V→V	×	✓	×	×	×
↪ VideoXum (2024)	5300 (100)	V→V+T	✓	✓	×	✓	×
↪ MR.HiSum (2023)	5590 (100)	V+T→V	×	✓	×	✓	×

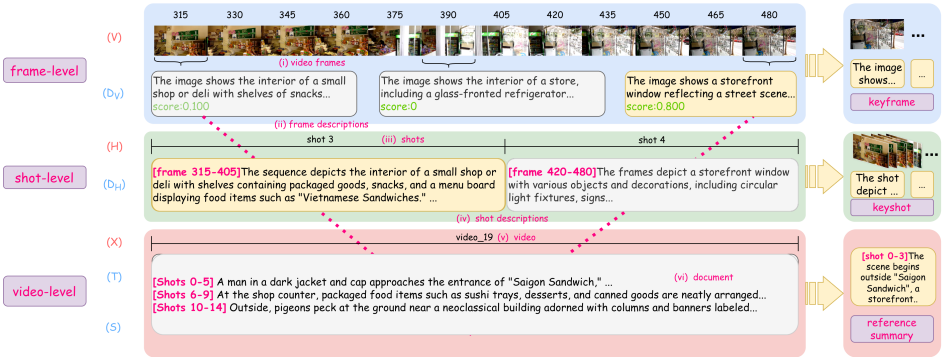

Figure 2: Example of a HAVEN data instance. Visual and textual annotations are organized at the frame, shot, and video levels, with explicit alignments between corresponding representations. Frame- and shot-level descriptions are generated from visual inputs, while the document and reference summary are composed from shot-level descriptions. Textual annotations in the evaluation subset are human-audited.

& Bansal, 2022; Li et al., 2026). In addition, prior studies highlight the importance of alignment and structured representations for understanding long or segmented content (Ernst et al., 2024; Cho et al., 2022). However, existing datasets focus on different subsets of granularities and tasks, leaving limited support for evaluating them through a shared hierarchical annotation structure.

3 HAVEN Benchmark

HAVEN is a novel, hierarchically structured multimodal benchmark with visual and textual annotations at the frame, shot, and video levels, together with explicit alignment across modalities and levels. It is built upon six video summarization datasets (see Table 1): SumMe, TVSum, OVP, YouTube, VideoXum, and Mr.HiSum. These datasets cover diverse video domains and provide complementary forms of high quality importance supervision. We map their source annotations into a common representation and equip each video with textual components initially generated by LLMs and subsequently audited and refined by humans for the evaluation subset.

3.1 Data Instance Representation

As illustrated in Fig. 2, each data instance in HAVEN comprises hierarchically structured visual and textual modalities at the frame, shot, and video levels.

At the **frame level**, a video is represented by a set of sampled frames V. Each frame is paired with a textual description D_V that provides fine-grained semantic details, along with a subset of human-annotated keyframes V_{key}.

At the **shot level**, temporally contiguous sampled frames are grouped into a set of shots H, where each shot corresponds to a subset of frames from V. Each shot is associated with a shot-level description D_H, capturing higher-level semantics and interactions across frames, as well as a subset of human-annotated key shots H_{key}.

At the **video level**, each video is paired with a document $T = \{t_j\}_{j=1}^{M}$ that provides a global description, and a summary $S = \{s_j\}_{j=1}^{P}$ that captures the key content of the video.

This hierarchical representation establishes explicit multi-level alignments between visual and textual modalities, enabling unified and consistent evaluation across tasks. An example of these hierarchical multimodal annotations is provided in Appendix A.3.

3.2 CONSTRUCTION PROCESS

The source datasets provide videos and frame importance annotations. HAVEN converts these heterogeneous annotations into a common representation, constructs a unified shot-level structure, and adds frame and shot descriptions, aligned video-level documents and summaries, and explicit cross-level links. The resulting data instances contain aligned visual and textual information at the frame, shot, and video levels.

Evaluation subset. HAVEN is constructed from six video-summarization datasets with substantially different scales. The standard evaluation subset contains all 165 videos from the four smaller datasets and 100 videos sampled from each of VideoXum and Mr.HiSum, yielding 365 videos in total. The remaining videos from the two larger datasets are released for training and further research but are excluded from aggregate benchmark evaluation.

Frame-level. We uniformly sample frames from each video at 1 or 2 FPS, following the conventions of prior datasets (De Avila et al., 2011; Zhang et al., 2016; Sul et al., 2023; Lin et al., 2024), resulting in an average of *164.9* frames per video. Each video includes source-provided HIGH-QUALITY frame importance scores, which comprises human visual-summary annotations or importance ratings for five datasets and normalized viewer-interaction signals for Mr.HiSum. The top 15% of frames (Zhang et al., 2016) are selected as reference keyframes. Additionally, each sampled frame is paired with a textual description D_V, initially generated by Video-LLaVA-7B and subsequently refined through the human-audit procedure described below.

Shot-level. Following Zhang et al. (2016), we partition each video into temporally coherent shots H using Kernel Temporal Segmentation (KTS), targeting approximately $0.1 \times |V|$ segments. Each shot is associated with a description D_H, generated by providing all frames within the shot to an MLLM under a length constraint to ensure concise yet informative summaries. Shot importance is computed as the mean importance of its constituent frames, and reference key shots are selected by 0/1 knapsack under a 15% duration budget. This level supports temporal understanding, multimodal grounding, and saliency ranking tasks.

Video-level. At the highest level, we construct a document $T = \{t_j\}_{j=1}^{M}$ by providing all shot-level descriptions to an LLM, which groups semantically related and temporally consistent shots into coherent sentences. This process ensures a fluent narrative structure while preserving detailed content. A summary $S = \{s_j\}_{j=1}^{P}$ is constructed similarly, using only key-shot descriptions. In both cases, sentences are aligned with their corresponding shot groups, establishing explicit links between textual units and visual segments. More details on the dataset construction pipeline and annotation procedures are provided in the Appendix A.

Human audit and manual refinement. For all 365 videos in the evaluation subset, we manually audit the generated annotations at all textual levels, including frame descriptions D_V, shot descriptions D_H, documents T, and summaries S. Incorrect or inaccurate annotations are manually corrected or regenerated accordingly, and the resulting annotations undergo a second round of human verification before being finalized for benchmark evaluation. Thus, the textual references used in the standard evaluation are not unverified LLM outputs, but human-audited and refined annotations. The independent quality evaluation below uses separate raters and is not used to revise the annotations.

Table 2: Human evaluation of annotation quality.

Level	Metric	Score	κ_w	AC1	Exact
Frame	Correctness	4.46 ± 0.79	0.33	0.52	58.2%
	Informativeness	4.06 ± 0.90	0.12	0.27	39.0%
Shot	Correctness	4.52 ± 0.77	0.44	0.58	62.7%
	Informativeness	4.47 ± 0.80	0.27	0.45	52.0%
Document	Correctness	4.13 ± 0.73	0.18	0.35	44.5%
	Coverage	4.37 ± 0.84	0.33	0.39	48.0%
	Readability	3.69 ± 0.87	0.10	0.20	33.5%
Summary	Correctness	3.89 ± 0.94	0.23	0.25	38.0%
	Relevance	3.19 ± 1.37	0.43	0.16	32.5%
	Conciseness	3.43 ± 1.09	0.17	0.07	24.5%
	Readability	4.37 ± 0.82	0.13	0.33	43.0%

3.3 Data Quality Assurance

We further conduct a separate human evaluation on the finalized annotations to independently assess their quality. Following prior video annotation and captioning studies that conduct human evaluation on sampled video subsets (Sul et al., 2023), (Liu & Wan, 2021), (Wu et al., 2022), we stratify 20 videos across the six source datasets and recruit 20 human evaluators to evaluate textual annotations with the corresponding visual content as a reference. Each video is rated by five annotators on a 1–5 Likert scale across level-specific criteria, including correctness and informativeness at the frame and shot levels, and correctness, coverage/relevance, conciseness, and readability at the higher semantic levels. We additionally employ behavioral quality controls to verify evaluator engagement. In addition, an optional free-form feedback field obtains detailed, criterion-specific comments, including corrections to inaccurate wording and observations of insufficient temporal coverage in summaries.

As shown in Table 2, the finalized annotations receive strong human ratings overall, particularly at the frame and shot levels (4.46 and 4.52 for correctness), while shot-level informativeness reaches 4.47. Summary relevance (3.19) and conciseness (3.43) are lower, indicating that content selection at the highest level remains more variable. This pattern is consistent with the increasing subjectivity of evaluating higher-level abstraction and content selection, where multiple descriptions or summaries may reasonably emphasize different aspects of the same video.

We report quadratically weighted pairwise Cohen's κ (Cohen, 1968), Gwet's AC1 (Gwet, 2008), and pairwise exact agreement because the ratings are concentrated near the upper end of the scale and no single agreement coefficient fully characterizes such distributions. These results support the quality of the finalized annotations while identifying summary-level content selection as the main remaining source of disagreement. Additional details of the human evaluation protocol are provided in Appendix B

3.4 Supporting Tasks

Building on the aligned dataset, we define a unified suite of tasks to evaluate multimodal understanding in MLLMs, as illustrated in Figure 3. These tasks cover complementary aspects of video understanding—including summarization, temporal reasoning, multimodal grounding, and saliency estimation.

Multimodal Summarization. HAVEN supports three multimodal summarization settings: **V2T (Video-to-Text)**, **V2V (Keyframe Selection)**, and **V2VT (Joint Video-Text) Summarization**. Given a video as input, the model is tasked with generating a textual summary, selecting a subset of keyframes, or performing both jointly.

Multimodal Grounding. Given shots H and shuffled summary sentences $\tilde{S}$, the model predicts alignments $\hat{M} = \mathcal{M}_{\text{align}}(H, \tilde{S})$, where $\hat{M} = \{(s_k, H_k)\}$ and $H_k \subseteq H$. This task evaluates whether models can associate each textual unit with its corresponding shot-level visual content. We additionally use the reference shot alignments in an oracle-conditioned diagnostic that asks models to retrieve importance-derived reference keyframes within the associated shots.

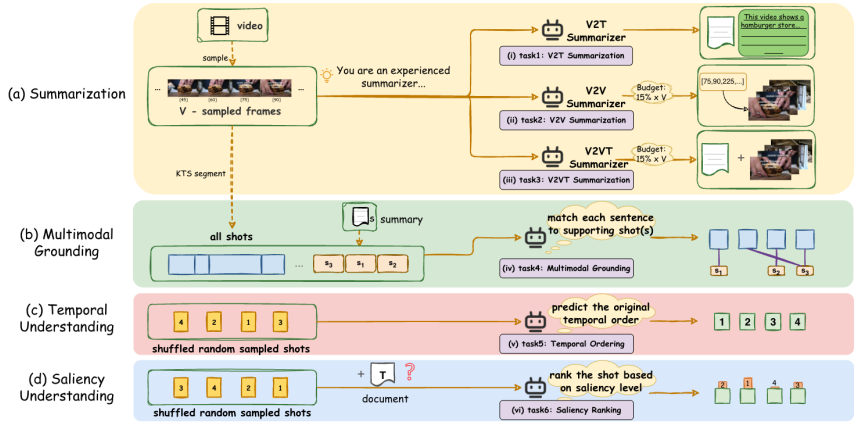

Figure 3: Overview of tasks supported by HAVEN.

Temporal Ordering. Given shuffled shots $\tilde{H}$, the model predicts the original chronological order: $(\hat{\pi} = \mathcal{M}_{\text{temp}}(\tilde{H}))$, where $\hat{\pi}$ denotes a permutation over shots. This task evaluates the ability to reason about temporal dependencies and event progression.

Saliency Ranking. Given sampled shots $\tilde{H}$, with or without document-level context T, the model predicts a saliency ranking $(\hat{\pi} = \mathcal{M}_{\text{rank}}(\tilde{H}, T))$. This task measures the ability to infer global importance and produce a coherent ranking of visual segments.

4 Experiments and Analysis

4.1 Experimental Setup

We evaluate models in a zero-shot setting on HAVEN's 365-video evaluation subset, which comprises all videos from SumMe, TVSum, OVP, and YouTube, and 100 sampled videos each from VideoXum and Mr.HiSum. The evaluation uses the human-audited annotations described in Section 3. Across the shared annotation hierarchy, We examine three questions: whether models perform consistently across textual summary evaluation and visual evidence selection; how well they reason about temporal ordering, saliency ranking and cross-modal grounding; and whether they can ground summary sentences to shots and retrieve reference keyframes from oracle shots.

Models and inputs. We evaluate GPT-5.2 (OpenAI, 2025), GLM-4.6V-Flash (Zhipu AI, 2025), InternVideo2.5-Chat-8B (Wang et al., 2025), Qwen3-VL-8B-Instruct (Qwen Team, 2025b), Qwen2.5-VL-7B-Instruct (Qwen Team, 2025a), and VideoLLaMA3-7B (Zhang et al., 2025). Summarization results are evaluated on four MLLMs (InternVideo2.5, Qwen3-VL, Qwen2.5-VL, and VideoLLaMA3) with full sampled-frame input. We additionally use GPT-5.4 as a text-proxy baseline, replacing visual inputs with frame descriptions for summarization tasks and shot descriptions for temporal ordering, summary-to-shots grounding, and saliency ranking. Appendix C provides the task prompts and input specifications.

Metrics. We organize the evaluation metrics into three task families. (1) For multimodal summarization, we evaluate textual summaries using BERTScore, ROUGE, and a GPT-based G-Eval-style score, and evaluate keyframe selection using precision, recall, and F1 under a 15% selection budget. For joint video–text summarization, CLIPScore measures the alignment between the generated summary and selected frames. CLIPScore is computed as $2.5 \max(\cos(\mathbf{t}, \mathbf{v}), 0)$, where $\mathbf{t}$ and $\mathbf{v}$ are the mean text and selected-frame embeddings. (2) For temporal and saliency reasoning, we use Spearman's rank correlation to compare the predicted and reference orderings of shots. (3) For grounding and evidence selection, we evaluate sentence-to-shot grounding using F1 over the predicted and reference shot sets, and oracle-shot keyframe retrieval using pooled precision, recall, and F1. Detailed metric definitions, aggregation procedures, and output-parsing rules are provided in Appendix C.

Table 3: Overall results across HAVEN task families. Bold indicates the best performance in each column. "–" indicates that the model is not evaluated under the corresponding setting, either because the w/ Doc setting is not applied (for text-proxy LLMs) or due to input frame limitations (e.g., GPT-5.2 and GLM in summarization). BS denotes BERTScore, and ρ denotes Spearman's rank correlation coefficient.

Model	Summarization				Temporal	Grounding	Saliency	
	V2T. BS	V2V. F1	V2VT. BS	V2VT. F1	ρ	F1	w/o Doc ρ	w/ Doc ρ
text-proxy LLM	0.856	0.171	0.858	0.171	0.019	**0.804**	0.154	–
GPT-5.2 (2025)	–	–	–	–	**0.280**	0.510	**0.210**	0.190
GLM-4.6V-Flash (2025)	–	–	–	–	0.147	0.457	0.164	0.136
InternVideo2.5 (2025)	0.847	0.273	0.859	0.065	0.083	0.394	0.169	0.178
Qwen3-VL (2025b)	0.865	0.173	0.859	0.235	0.230	0.443	0.187	**0.218**
Qwen2.5-VL (2025a)	**0.874**	0.209	0.866	0.176	0.068	0.299	0.006	0.070
VideoLLaMA3 (2025)	**0.874**	**0.293**	**0.869**	**0.267**	0.053	0.253	0.142	0.164

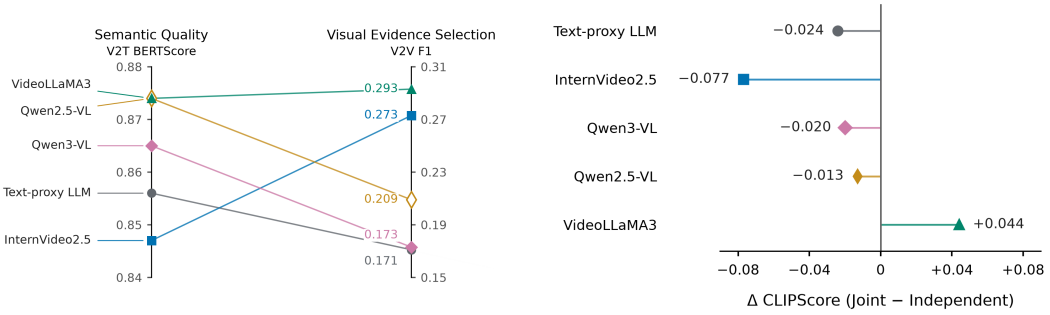

Figure 4: Relationship between semantic textual summary, keyframe selection, and cross-modal consistency. **Left:** V2T BERTScore and V2V F1 are shown on independent scales for each model. Connecting lines identify the same model across the two evaluation dimensions and do not represent absolute numerical changes. **Right:** Change in CLIPScore (Δ CLIPScore) for jointly generated V2VT outputs relative to independently generated V2T and V2V outputs.

The overall results are summarized in Table 3. The results reveal no consistently dominant model. VideoLLaMA3 performs strongly on textual summarization and independent keyframe selection but obtains lower temporal-ordering and sentence-to-shot grounding scores, whereas GPT-5.2 leads the evaluated visual models on these two shot-level tasks. The text proxy provides a strong baseline on sentence-to-shot grounding but achieves the lowest temporal ordering performance, suggesting that textual representations may preserve strong semantic importance cues while providing limited information about visual temporal progression. Overall, the results show that performance on textual summarization or any single grounding task does not reliably characterize broader video-understanding capabilities, highlighting the value of HAVEN's comprehensive evaluation.

4.2 Generation and Reasoning Capabilities

HAVEN's six tasks evaluate comprehensive model capabilities. V2T evaluates textual generation and abstraction, V2V emphasizes explicit keyframe selection, and V2VT additionally evaluates the alignment of multimodal outputs produced jointly. Temporal ordering examines reasoning over temporal progression, saliency ranking examines the relative importance of shots, and multimodal grounding evaluated cross-modal association between textual units and visual evidence. Comparing these tasks allows us to examine whether strong generation performance transfers to reasoning-oriented video tasks.

Table 4 reports the detailed scores, and Figure 4 shows how textual evaluation relates to keyframe selection and how joint generation affects cross-modal alignment.

Textual generation performance does not consistently transfer to reasoning-oriented tasks. Model rankings differ between textual summarization and keyframe selection. For example, Qwen2.5-VL and VideoLLaMA3 both obtain a BERTScore of 0.874 but differ in keyframe-selection

Table 4: Detailed results for summarization experiments under a unified evaluation setting. Δ denotes the change in CLIPScore relative to the corresponding V2T–V2V alignment. G-Eval refers to a GPT-based evaluation score following the G-Eval(2023) protocol. ROUGE denotes the aggregated score of ROUGE-1, ROUGE-2, and ROUGE-L. P and R denote precision and recall, respectively.

Model	V2T. Sum		V2V. Sum		V2VT. Sum				
	G-Eval	ROUGE	P	R	G-Eval	ROUGE	P	R	CLIPScore (Δ)
text-proxy LLM	**3.528**	0.203	0.172	0.178	3.481	**0.213**	0.173	0.178	**0.899** (-0.024)
InternVideo2.5	2.252	0.152	0.269	0.299	3.149	0.160	**0.433**	0.038	0.741 (-0.077)
Qwen3-VL	3.412	0.207	0.186	0.182	3.380	0.195	0.204	0.481	0.881 (-0.020)
Qwen2.5-VL	3.372	**0.235**	0.211	0.229	3.478	0.197	0.200	0.249	0.846 (-0.013)
VideoLLaMA3	3.353	0.223	**0.292**	**0.319**	**3.536**	0.212	0.218	**0.558**	0.850 (+0.044)

F1 (0.209 and 0.293), showing that textual-reference scores can accompany different agreement with reference keyframes. Similar differences are observed in other tasks beyond summarization. Among the four models evaluated across all task families, VideoLLaMA3 performs strongly on summarization including V2T, V2V, and V2VT, but obtains the lowest temporal-ordering and sentence-to-shot grounding scores. Qwen2.5-VL similarly achieves a high V2T BERTScore while remaining weaker on temporal ordering, saliency ranking, and grounding. In contrast, Qwen3-VL does not lead independent keyframe selection but achieves the strongest temporal-ordering and sentence-to-shot grounding among these four models. InternVideo2.5 shows another pattern, performing strongly on independent keyframe selection while remaining weaker on textual generation, temporal ordering, and joint frame selection.

Joint generation does not consistently strengthen generation–reasoning coordination. V2VT further examines whether requiring textual generation and saliency-based frame selection in the same response improves either output or their alignment. Four of the five evaluated models exhibit negative changes in CLIPScore relative to independently generated V2T and V2V outputs, ranging from -0.013 to -0.077, while VideoLLaMA3 exhibits an increase of 0.044. InternVideo2.5's BERTScore increases slightly from 0.847 to 0.859, while its frame-selection F1 decreases from 0.273 to 0.065. Qwen3-VL instead improves frame-selection F1 from 0.173 to 0.235, despite decreases in both BERTScore and CLIPScore. Thus, improving one aspect under joint generation does not necessarily improve the other aspect or strengthen their alignment.

The heterogeneous changes do not follow the models' configured context lengths. VideoLLaMA3 and InternVideo2.5 both use configured context lengths of $32{,}768$ but exhibit contrasting V2VT changes, while the longer-context Qwen-VL models show no consistent advantage across V2T, V2V, and V2VT. One possible contributor is how the models internally represent the sampled frames. The evaluated models apply different visual-tokenization and compression strategies. The present experiments do not isolate these architectural factors, but they show that context-window size alone is insufficient to characterize joint-generation behavior.

Overall, generation and reasoning profiles diverge. VideoLLaMA3 and Qwen2.5-VL are stronger on generation, Qwen3-VL on shot-level reasoning, and InternVideo2.5 on independent evidence selection. No evaluated model performs consistently across these capabilities.

4.3 Reasoning Across Temporal, Saliency, and Grounding Tasks

Following the comparison between generation and reasoning-oriented performance in Section 4.2, we examine the information required by three forms of video reasoning. Temporal ordering requires models to recover event progression and sequential relationships from shuffled shots. Saliency ranking requires models to compare the relative importance of shots to the video as a whole. Cross-modal grounding instead requires models to associate a textual unit with its corresponding visual evidence. Because these tasks differ in granularity, output structure, and evaluation metric, we compare their model capabilities rather than their absolute scores.

Textual information provides task-dependent benefits. All evaluated visual models outperform the text-proxy baseline on temporal ordering ($\rho = 0.053$–0.280 versus 0.019, Table 3). In contrast, the text proxy obtains a saliency correlation of 0.154, which exceeds only Qwen2.5-VL (0.006) and VideoLLaMA3 (0.142) and provides a strong sentence-to-shot grounding baseline, with an F1 of 0.804 compared with 0.253–0.510 for the MLLMs. This contrast is consistent with textual descrip-

Table 5: Sentence-conditioned reference-keyframe retrieval within oracle shots on 684 sentences. Candidates are sampled frames from the aligned reference shot(s), and targets are the keyframes within those shots. Bold indicates the best value in each column.

Model	Recall	Precision	F1
InternVideo2.5	0.625	0.665	0.644
CLIP	0.636	0.636	0.636
Qwen2.5-VL	0.595	**0.681**	0.635
Qwen3-VL	**0.645**	0.664	**0.655**
GLM-4.6V-Flash	0.635	0.649	0.642
VideoLLaMA3	0.303	0.619	0.406

tions preserving semantic cues that support importance estimation and within-text matching while providing less information about visual transitions, state changes, and event progression. Additional document context likewise does not uniformly improve performance on reasoning. For example, in saliency ranking it increases correlation for InternVideo2.5, Qwen3-VL, Qwen2.5-VL, and VideoLLaMA3 but decreases it for GPT-5.2 and GLM-4.6V-Flash. Thus, access to higher-level textual context alone is insufficient to ensure better saliency reasoning.

Overall, reasoning performance remains task dependent. Among the visual models, GPT-5.2 performs best on temporal ordering, sentence-to-shot grounding, and saliency ranking without document context, whereas Qwen3-VL performs best on document-conditioned saliency ranking.

4.4 Grounding and Evidence Selection Across Levels

Cross-modal localization remains challenging. As discussed in Section 4.3, the text proxy substantially outperforms the visual models on sentence-to-shot grounding, indicating that matching within HAVEN's textual representation is easier than localizing the same content in visual shots.

Sentence-conditioned keyframe retrieval within oracle shots. We next evaluate 684 summary sentences by providing the associated reference shot(s) and asking models to select $k = |G|$ reference frames. CLIP ranks candidate frames by sentence–image similarity, while generative models return frame indices. Qwen3-VL, InternVideo2.5, GLM-4.6V-Flash, CLIP, and Qwen2.5-VL obtain broadly comparable F1 scores of 0.635–0.655 (Table 5), providing no consistent advantage for the generative models over CLIP. VideoLLaMA3 obtains a lower F1 of 0.406, with precision of 0.619 and recall of 0.303.

A detailed failure analysis (Appendix Table 7) shows similar content-miss rates across models (18.3–21.1%), but substantial differences in output compliance. In particular, VideoLLaMA3 frequently returns refusal, empty, or incomplete default-order outputs, while Qwen2.5-VL encounters runtime failures. Aggregate F1 therefore reflects both evidence selection and output validity.

Together, the two grounding evaluations separate cross-modal localization from evidence selection within relevant regions, while the failure decomposition shows that structured-output compliance introduces an additional source of performance variation.

5 Conclusions

In this work, we present **HAVEN**, a novel hierarchically aligned multimodal benchmark for evaluating multimodal summarization, temporal ordering, multimodal grounding, and saliency ranking on the same video collection. Empirical experiment results reveal a comprehensive and fine-grained evaluation of the video understanding capabilities of widely used vision-language models. We hope HAVEN can serve as a standardized benchmark to facilitate the development of advanced multimodal models that not only generate coherent text, but also effectively ground, align and reason over both textual and visual information in a unified manner.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: *generating synthetic data* (the HAVEN annotations are produced by generative models, as described in Section 3); *cleaning the dataset* (regenerating malformed annotations with GPT models); *implementing methods* (most of the code for data processing, inference, and evaluation was generated by AI coding assistants from author specifications); *providing feedback on experiments and interpreting results*, in part (e.g., suggesting diagnostic analyses and auditing logs); and *translation* (the authors drafted the content and detailed explanations in their native language and used AI tools for English translation and language editing). We have not used generative AI tools to propose hypotheses, to develop theoretical models or conceptual frameworks, or to support qualitative data analysis, and mathematical claims and proofs are not applicable to this work.

Additionally, we used generative AI tools to expand author-written outlines into full prose and edit it for readability, to create or edit code and some figures, to build the annotation-verification website, and, to a limited extent, to identify relevant literature.

We have reviewed all AI-assisted work. Generated annotations were audited manually and through a independent annotation-quality evaluation (Appendix B) and all translated text, claims, and citations were checked by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

HAVEN is built on six publicly available video summarization datasets (SumMe, TVSum, OVP, YouTube, VideoXum, and Mr.HiSum). We release only derived annotations and structural metadata, not the source videos, which users must obtain under the original datasets' terms. The videos may depict identifiable people; our annotations describe visual content and do not identify individuals or infer sensitive attributes. The independent annotation-quality evaluation involved 20 volunteer doctoral researchers from diverse disciplines and countries, who rated annotation quality through a web interface without compensation. No personal information was collected beyond an anonymous identifier. Because HAVEN annotations are produced by generative models, they may contain errors or biases inherited from those models and the source data, and we assess their quality through manual auditing and human verification.

REPRODUCIBILITY STATEMENT

We provide the dataset construction and preprocessing procedures in Appendix A, the human-evaluation protocol and agreement calculations in Appendix B, and visual-input processing, task construction, prompt templates, output parsing, and evaluation metrics in Appendix C. The released benchmark will include the processed annotations, hierarchical alignment metadata, fixed evaluation subset, and reference targets. Code for data processing and evaluation will also be released to support reproduction of the reported experiments.

REFERENCES

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.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2024.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024.

Sangwoo Cho, Kaiqiang Song, Xiaoyang Wang, Fei Liu, and Dong Yu. Toward unifying text segmentation and long document summarization. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 106–118, 2022.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

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.

Ori Ernst, Ori Shapira, Aviv Slobodkin, Sharon Adar, Mohit Bansal, Jacob Goldberger, Ran Levy, and Ido Dagan. The power of summary-source alignments. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6527–6548, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-modal LLMs in Video Analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24108–24118, 2025.

Boqing Gong, Wei-Lun Chao, Kristen Grauman, and Fei Sha. Diverse Sequential Subset Selection for Supervised Video Summarization. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

Kilem Li Gwet. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48, 2008.

Michael Gygli, Helmut Grabner, Hayko Riemenschneider, and Luc Van Gool. Creating Summaries from User Videos. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (eds.), *Computer Vision – ECCV 2014*, pp. 505–520, 2014. doi: 10.1007/978-3-319-10584-0_33.

Mingfei Han, Linjie Yang, Xiaojun Chang, Lina Yao, and Heng Wang. Shot2story: A new benchmark for comprehensive understanding of multi-shot videos. *arXiv preprint arXiv:2312.10300*, 2023.

Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *European conference on computer vision*, pp. 709–727. Springer, 2020.

Yiqiao Huang, Qi He, Zhaorun Chen, Haopeng Zhang, Hanchao Yu, and Zhuokai Zhao. Autonomous multimodal reasoning via implicit chain-of-vision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 2963–2972, 2025.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22195–22206, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Zizhong Li, Haopeng Zhang, and Jiawei Zhang. Mmvir: A multi-modal and multi-granularity representation for long-range video understanding. *arXiv preprint arXiv:2601.05495*, 2026.

Jingyang Lin, Hang Hua, Ming Chen, Yikang Li, Jenhao Hsiao, Chiuman Ho, and Jiebo Luo. VideoXum: Cross-Modal Visual and Textural Summarization of Videos. *IEEE Transactions on Multimedia*, 26:5548–5560, 2024. doi: 10.1109/TMM.2023.3335875.

Hui Liu and Xiaojun Wan. Video paragraph captioning as a text summarization task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 55–60, 2021.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Yinhong Liu, Jianfeng He, Hang Su, Ruixue Lian, Yi Nian, Jake Vincent, Srikanth Vishnubhotla, Robinson Piramuthu, and Saab Mansour. Mdseval: A meta-evaluation benchmark for multimodal dialogue summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14707–14727, 2025.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024a.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024b.

OpenAI. OpenAI GPT-5 System Card. *arXiv e-prints*, 2025. doi: 10.48550/arXiv.2601.03267.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkila. Rethinking the Evaluation of Video Summaries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7596–7604, 2019.

Qwen Team. Qwen2.5-VL Technical Report. *arXiv e-prints*, 2025a. doi: 10.48550/arXiv.2502.13923.

Qwen Team. Qwen3-VL Technical Report, 2025b.

Mrigank Rochan and Yang Wang. Video Summarization by Learning From Unpaired Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7902–7911, 2019.

Jinhwan Sul, Jihoon Han, and Joonseok Lee. Mr. hisum: A large-scale dataset for video highlight detection and summarization. *Advances in Neural Information Processing Systems*, 36:40542–40555, 2023.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

David Wan and Mohit Bansal. Evaluating and improving factuality in multimodal abstractive summarization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9632–9648, 2022.

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

Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. InternVideo2.5: Empowering Video MLLMs with Long and Rich Context Modeling, 2025.

Dekun Wu, He Zhao, Xingce Bao, and Richard P Wildes. Sports video analysis on large-scale data. In *European conference on computer vision*, pp. 19–36. Springer, 2022.

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

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13204–13214, 2024.

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

Yale Song, Jordi Vallmitjana, Amanda Stent, and Alejandro Jaimes. TVSum: Summarizing web videos using titles. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5179–5187, 2015. doi: 10.1109/CVPR.2015.7299154.

Haohan Yuan and Haopeng Zhang. Understanding llm reasoning for abstractive summarization. *arXiv preprint arXiv:2512.03503*, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Haopeng Zhang. Bridging multimodal and video summarization: A unified survey. In *Proceedings of The 5th New Frontiers in Summarization Workshop*, pp. 157–171, 2025.

Ke Zhang, Wei-Lun Chao, Fei Sha, and Kristen Grauman. Video summarization with long short-term memory. In *European conference on computer vision*, pp. 766–782. Springer, 2016.

Zhipu AI. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. URL https://arxiv.org/abs/2507.01006.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. MSMO: Multimodal Summarization with Multimodal Output. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4154–4164, 2018. doi: 10.18653/v1/D18-1448.

A Dataset Construction and Representation

This appendix complements Section 3 with implementation details of the construction pipeline and an example of the released representation. HAVEN preserves two connected layers for each video: a structural layer containing sampled-frame indices, shot boundaries, source importance signals, and reference selections; and a textual layer containing frame descriptions, shot descriptions, aligned documents, and aligned summaries.

A.1 Source preprocessing and structural representation

We convert the six source datasets to a common H5 representation. Each video record stores its sampled-frame indices (`picks`), original-frame shot boundaries (`change_points`), frame-level importance or keyframe labels, the number of original frames, and the derived shot-level targets. Dataset identifiers and the association with the source benchmark are retained throughout processing.

Frames are sampled at the rate specified in the main paper. We map the source annotations from their original frame space to this sampled sequence before constructing the benchmark targets. The 15% selection budget used by the summarization tasks is defined over the sampled-frame sequence. For source datasets that provide continuous importance scores, these scores are retained for ranking and converted to budget-constrained reference selections when a binary keyframe target is required. For datasets that provide binary selections, the source selections are mapped directly to sampled-frame positions.

We obtain temporally contiguous shots using Kernel Temporal Segmentation (KTS), following the standard video-summarization preprocessing used by the source benchmarks. The target number of segments is approximately $0.1|V|$, where $|V|$ is the number of sampled frames. We use the source-provided or reconstructed frame-importance signal to compute the mean importance of the sampled

frames assigned to each shot. Reference key shots are selected with 0/1 knapsack under the same 15% duration budget used for video summarization.

A.2 Hierarchical textual annotation construction

Textual annotations are constructed progressively from the shared structural representation.

Frame descriptions. Each sampled frame is paired with a visually grounded description. The released record retains the original frame identifier, its description, its reference-selection status, and processing-status metadata.

Shot descriptions. The sampled frames assigned to a shot are provided together in temporal order to produce a concise shot description. Each released shot record stores the shot identifier, its ordered frame list, frame count, textual description, reference key-shot label, and processing status.

Aligned documents and summaries. The ordered shot descriptions are composed into a detailed document. Consecutive, semantically related shots may be grouped into one sentence, and each sentence retains the identifiers of the shots from which it was constructed. The summary is constructed from the descriptions of the reference key shots under a stronger compression constraint and retains the same sentence-to-shot links. These links support sentence-to-shot grounding without requiring post-hoc text–video alignment.

Human audit. For all 365 videos in the evaluation subset, the generated frame descriptions, shot descriptions, documents, and summaries are manually audited. Incorrect descriptions are corrected or regenerated, and the revised annotations undergo a second verification pass. This construction-stage audit is distinct from the independent quality evaluation in Appendix B; ratings collected in that evaluation are not used to revise the evaluated annotations.

During the audit, we observed recurring errors such as activity labels inferred from contextual appearance rather than the complete scene and inconsistent object labels across neighboring frames. Affected annotations were regenerated with prompts that asked the model to reconsider activity labels against the full visual evidence and used the inferred shot-level scene only as soft context for frame-level descriptions. Items that remained ambiguous were manually rewritten or excluded from the finalized evaluation subset.

A.3 Released representation and example

The textual layer is released as processed JSON containing frame descriptions, shot records, aligned documents, aligned summaries, and sentence-to-shot links. The structural and evaluation layer is released in H5 format and contains the sampled-frame indices, shot boundaries, source importance signals, video length, and derived reference selections. Figure 6 shows one instance and the links between its visual and textual units.

B Independent Evaluation of Annotation Quality

We conduct an evaluation of the finalized annotations that is independent of the construction-stage audit. The purpose is to measure perceived quality at each level of the hierarchy and to identify criteria on which judgments remain variable.

B.1 Sampling and assignment

The evaluation contains 20 videos stratified across all six source datasets: five from Mr.HiSum, six from VideoXum, three from TVSum, three from OVP, two from YouTube, and one from SumMe. Twenty evaluators participate, and each video is independently rated by five evaluators, producing 100 deduplicated video-level submissions. For each video, an evaluator judges two sampled frame descriptions, one sampled shot description, the aligned full document, and the aligned summary.

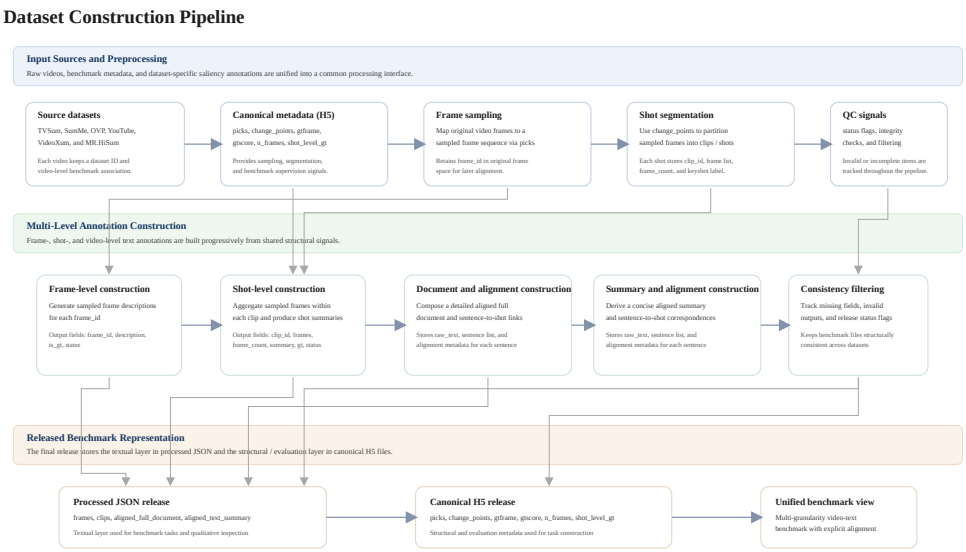

Figure 5: Dataset construction pipeline. Source-specific structural and importance annotations are converted to a common representation before frame-, shot-, and video-level textual annotations are constructed and audited.

Evaluators view the corresponding visual content while scoring the textual annotation. Frame items show the sampled frame, shot items show the associated video segment, and document and summary items show the full video. The assignment procedure balances workload and prevents an evaluator from receiving the same video twice. When duplicate submissions occur, only the most recent complete submission from that evaluator is retained.

B.2 Evaluation dimensions and rubric

Each criterion is rated on a five-point Likert scale. Frame and shot descriptions are evaluated for *correctness* and *informativeness*. Documents are evaluated for *correctness*, *coverage*, and *readability*. Summaries are evaluated for *correctness*, *relevance*, *conciseness*, and *readability*. Evaluators are instructed to judge each criterion independently and to use the visual evidence rather than stylistic preference as the basis for correctness and coverage judgments.

The common score interpretation is:

- **5 (excellent):** fully satisfies the criterion with no noticeable problem;
- **4 (good):** satisfies the criterion with only minor problems;
- **3 (acceptable):** partially satisfies the criterion but has a clear limitation;
- **2 (poor):** fails on several important aspects of the criterion;
- **1 (very poor):** largely fails to satisfy the criterion.

The evaluation interface provides the full rubric on a separate instruction page and an expandable criterion-specific guide beside each rating control. An optional free-form field allows evaluators to identify inaccurate wording, missing events, redundancy, or other criterion-specific concerns.

B.3 Behavioral quality control

The interface uses hidden control and corrupted items to check evaluator engagement. A control item preserves a manually verified annotation, whereas a corrupted item replaces a frame description with text drawn from an unrelated video. These items are visually indistinguishable from ordinary

evaluation items. The interface also records task completion behavior and requires evaluators to proceed through the visual material before submitting their ratings. Quality-control responses are examined separately from the annotation-quality judgments.

B.4 Score aggregation and inter-annotator agreement

For each video and criterion, we first average the five evaluator scores. The reported quality score is the mean and standard deviation of these video-level averages. Consequently, each evaluated video has equal weight in the final score regardless of the number of lower-level annotation items shown in its package.

We report quadratically weighted pairwise Cohen's κ (Cohen, 1968), unweighted Gwet's AC1 (Gwet, 2008), and pairwise exact agreement. Quadratic weighting treats larger distances on the five-point scale as more severe disagreement. AC1 is reported without ordinal weighting as a chance-corrected complement that is less sensitive to concentrated marginal ratings.

The coefficients are computed from pooled within-item rating pairs. For each level–criterion subset, we iterate over its annotation items, construct pairs only among ratings assigned to the same item, pool those pairs across all items in the subset, and compute one coefficient from the resulting aggregate. We do not first compute coefficients for fixed pairs of evaluators and then average them. For weighted κ, every ordered pair (a, b) with $a \neq b$ contributes to a global symmetric confusion matrix; an item with five ratings therefore contributes 20 ordered pairs. AC1 and exact agreement use the corresponding unordered pairs $a < b$, giving 10 pairs for an item with five ratings. Counting both directions for the symmetric κ matrix does not change the resulting coefficient.

Pairwise exact agreement is

$$A_{\text{exact}} = \frac{\sum_i \sum_{a<b} \mathbb{I}(y_{ia} = y_{ib})}{\sum_i \binom{m_i}{2}}, \tag{1}$$

where y_{ia} is evaluator a's rating of item i and m_i is the number of valid ratings for that item. All agreement statistics are recomputed separately for each level–criterion subset rather than copied from a coefficient pooled across different criteria.

Because 76.3% of the raw ratings fall in categories 4 or 5, chance-corrected coefficients and literal agreement provide complementary information. We therefore interpret agreement jointly with the absolute scores and do not treat any single coefficient as a complete measure of annotation quality.

B.5 LLM-judge sensitivity analysis

We additionally compare human judgments with two independent LLM judges on a stratified subset of 32 annotations from eight videos, comprising eight frame descriptions, eight shot descriptions, eight documents, and eight summaries. Each item retains ratings from five human evaluators; their mean forms the item-level human score. The LLM judges receive the same criterion definitions and rating scale. This analysis is a sensitivity check on evaluator dependence and is not used to construct or revise the benchmark annotations.

As reported in Table 6, the judges broadly agree on the high quality of lower-level annotations but differ on several higher-level criteria. The largest difference occurs for summary correctness: the human, GPT-based, and Claude-based means are 4.10, 1.88, and 4.50. The shared rubric penalizes significant omissions, a rule that is natural for local descriptions but can be overly literal for a deliberately selective summary. We therefore treat these LLM-judge results as evidence of evaluator sensitivity rather than as an additional ground-truth quality score.

C Prompt Design and Evaluation Details

This appendix reports the model implementations, visual-input processing, task construction, prompt templates, and output parsing used in the experiments. The templates below preserve the task-defining instructions and output constraints; placeholders such as N, M, K, `[SENTENCE]`, and `[DOCUMENT]` are instantiated for each example. They are therefore prompt templates rather than verbatim transcripts of every model call.

Table 6: Human and LLM-judge scores on the 32-item stratified sensitivity subset. Values are mean $\pm$ standard deviation.

Level	Criterion	Human	GPT judge	Claude judge
Frame	Correctness	4.55 ± 0.55	5.00 ± 0.00	4.88 ± 0.33
	Informativeness	4.17 ± 0.83	3.88 ± 0.33	3.88 ± 0.33
Shot	Correctness	4.58 ± 0.54	4.38 ± 0.48	4.25 ± 0.43
	Informativeness	4.55 ± 0.63	4.25 ± 0.66	4.75 ± 0.43
Document	Correctness	3.98 ± 0.79	3.50 ± 0.71	4.25 ± 0.66
	Coverage	4.00 ± 0.97	4.75 ± 0.43	4.12 ± 0.33
	Readability	3.58 ± 0.77	3.12 ± 0.60	3.62 ± 0.48
Summary	Correctness	4.10 ± 0.86	1.88 ± 0.83	4.50 ± 0.50
	Relevance	3.58 ± 1.18	2.75 ± 1.16	4.62 ± 0.48
	Conciseness	3.65 ± 1.04	4.38 ± 0.48	4.38 ± 0.48
	Readability	4.45 ± 0.67	4.75 ± 0.43	4.75 ± 0.43

C.1 Model implementations

Visual input processing. Within each task instance, models received the same ordered sampled-frame sequence before model-specific preprocessing and context limits were applied. For the three summarization settings, the Qwen processors preserved aspect ratio while constraining each image to approximately 448×448 pixels; the resulting measured visual-token counts were approximately 252–280 tokens per frame for Qwen2.5-VL and 180–204 for Qwen3-VL. InternVideo2.5 resized frames to 448×448, padded frame counts to a multiple of four where required by the implementation, and used HiCo compression to produce 16 visual tokens per frame. VideoLLaMA3 received the ordered frames as a video input and allocated an approximately 16,384-token visual budget across the sequence before its additional token-compression step; its final retained token count is content dependent and was not logged.

For the shot-level tasks, Qwen2.5-VL and Qwen3-VL used the source image resolution without an explicit `max_pixels` override. InternVideo2.5 used the same 448×448 preprocessing and four-frame grouping described above, and VideoLLaMA3 again received the ordered frames as a video input. GPT-5.2 received JPEG-encoded images through the multimodal interface without an explicit image-detail override, and GLM-4.6V-Flash used its default processor at the source resolution. Thus, identical sampled frames do not imply identical internal visual-token counts or compression.

C.2 Task construction

All reported experiments use the 365-video evaluation subset described in the main paper. The four smaller source datasets contribute all 165 videos, and VideoXum and Mr.HiSum contribute 100 videos each. The same saved subset is used across models.

Summarization. V2T, V2V, and V2VT use the sampled frames in chronological order. V2T asks for a 3–5 sentence summary. V2V requests a frame budget of $M = \lceil 0.15N \rceil$, where N is the number of sampled frames in the full video. V2VT jointly requests a summary containing the same number of sentences as the reference summary and an importance selection under the same 15% frame budget. When a model-specific input limit reduces the submitted sequence, the requested number is capped by the number of submitted frames and the prompt states both the full-video target and the submitted-frame budget.

Temporal ordering and saliency ranking. For temporal ordering, we deterministically sample up to five non-empty shots per video and shuffle their presentation order using seed 42. The main setting supplies the shuffled visual shots without document context. For saliency ranking, we deterministically sample up to eight non-empty shots using the same seed and evaluate both visual-only and document-context conditions. Each selected shot retains all of its available sampled frames. Both tasks request a complete permutation of the presented shot identifiers.

Table 7: Failure-mode breakdown for oracle-shot keyframe retrieval. Hit@k indicates that at least one reference keyframe occurs among the parsed predictions; Full-k is the percentage of outputs containing k distinct valid frame IDs. The remaining columns partition instances without a hit.

Model	Hit@k	Full-k	Runtime	Refusal/ empty	Other invalid	Default	Content miss
CLIP ViT-B/32	79.8	100.0	–	–	–	–	20.2
GLM-4.6V-Flash	79.5	96.5	0.0	0.0	0.1	0.3	20.0
InternVideo2.5	78.5	90.1	0.0	0.0	0.0	0.9	20.6
Qwen3-VL	78.2	95.9	0.0	0.0	0.6	0.1	21.1
Qwen2.5-VL	73.4	88.2	5.6	0.0	0.1	0.3	20.6
VideoLLaMA3	56.7	51.9	0.0	16.4	0.0	8.6	18.3

Sentence-to-shot grounding. For each video, we deterministically sample up to three aligned summary sentences using seed 42 and shuffle their presentation order. Every shot is supplied as a candidate clip. To control the visual input size while retaining coverage of each shot, we use the first $\min(8, \max(4, \mathrm{round}(0.3n_s)))$ sampled frames from a shot containing n_s sampled frames. The model predicts one or more clip identifiers for each sentence.

Sentence-conditioned keyframe prioritization within oracle shots. This experiment includes every aligned summary sentence that has at least one associated shot and at least one reference keyframe inside its associated shot set. This produces 684 sentence instances: 66 from OVP, 195 from TVSum, 70 from SumMe, 56 from YouTube, 169 from Mr.HiSum, and 128 from VideoXum. For a sentence associated with multiple shots, their sampled frames are concatenated in shot and temporal order. The requested budget is k=|G|, where G is set of reference keyframes within the oracle shots. The targets inherit the source datasets' saliency or keyframe annotations; they are not independently annotated, exhaustive sentence-specific evidence frames. The experiment therefore measures sentence-conditioned recovery of inherited reference keyframes within a known relevant region rather than end-to-end sentence-to-frame localization.

Failure-mode categorization. For additional diagnosis, we classify each oracle-shot retrieval instance according to whether the parsed output contains at least one reference keyframe. Table 7 reports Hit@k, complete-output rate, and the failure categories used in our analysis. The categories are mutually exclusive after accounting for successful hits.

C.3 Prompt templates

Table 8 lists the templates for the main visual tasks. Frame, shot, clip, and sentence labels are always one-indexed in the prompt.

The GPT-5.4 text proxy receives textual annotations in place of visual inputs. These comparisons are diagnostic rather than controlled modality ablations because both the model and the input representation change, and the proxy text shares annotation provenance with the reference summaries. The task instructions and output formats otherwise match their visual counterparts:

- **V2T proxy:** chronological frame descriptions are summarized into the reference number of sentences.
- **V2V proxy:** chronological frame descriptions are used to select exactly M one-indexed frame identifiers.
- **V2VT proxy:** chronological frame descriptions are used to jointly produce a numbered summary and an `IMPORTANT FRAMES` list.
- **Temporal proxy:** shuffled event sentences are reordered as `Sentence X` $\rightarrow$ `Sentence Y` $\rightarrow \cdots$.
- **Grounding proxy:** shot descriptions replace clip frames, while the model returns the same sentence-to-clip assignments.
- **Saliency proxy:** shot descriptions replace shot frames, while the model returns the same complete importance ranking.

Table 8: Prompt templates for the main visual experiments. Bracketed fields are populated for each instance.

Task	Input	Task-defining template	Required output
V2T	Chronological frames	"You are given N frames from a video shown in chronological order. Write a concise summary describing the main events, actions, scenes, and topics from beginning to end. The summary should contain 3–5 sentences."	Free-form 3–5 sentence summary.
V2V	Chronological frames	"From these N frames, select the M most important frames that best represent the key moments of the video. Output only a JSON array of one-indexed frame numbers with exactly M elements."	`[1, 5, 12]`
V2VT	Frames grouped by shot	"Based on all N frames, complete two tasks: (1) write a summary in exactly K sentences; and (2) select exactly M important frames. Use the specified two-section format and no extra text."	`SUMMARY:` numbered sentences; `IMPORTANT FRAMES: [Frame X, ...].`
Temporal ordering	Shuffled shots	"Reorder all shots to match the actual temporal sequence from first to last. Include every shot exactly once."	`Shot X` → `Shot Y` → ···
Sentence-to-shot grounding	All clips and shuffled summary sentences	"For each summary sentence, identify the clip or clips to which it corresponds. List clip numbers separated by commas."	One line per sentence: `Summary i: Clip X, Clip Y.`
Saliency ranking	Selected shots; optional document prefix	"Rank all shots from most important to least important based on their visual content and relevance to the overall video. Include every shot exactly once." In the context condition, `[DOCUMENT]` is prepended as a background description.	`Shot X` → `Shot Y` → ···
Oracle-shot keyframe prioritization	Sentence and candidate frames from its oracle shots	"Rank the k frames that show the strongest visual evidence for `[SENTENCE]`, from strongest to weakest. List k distinct frame numbers, most likely first."	`Frames: X, Y, Z`

C.4 Output parsing and aggregation

Identifier parsing. For V2V and V2VT, we extract in-range frame identifiers, remove duplicates while preserving order, and retain at most the requested budget. An output containing fewer valid identifiers is retained without padding. For sentence-to-shot grounding, each requested `Summary i` line is parsed independently; absent or invalid clip identifiers yield an empty predicted set for that sentence. For temporal ordering and saliency ranking, a ranking is valid only if it is a complete permutation containing every presented identifier exactly once. Incomplete, duplicated, or unparsable rankings are recorded as parse failures and are not assigned a correlation value.

For oracle-shot keyframe prioritization, the parser first searches for the `Frames:` field and otherwise falls back to integers in the generated text. It discards out-of-range and repeated identifiers, preserves their first-occurrence order, and stops after k valid identifiers. Outputs containing fewer than k identifiers are not padded. InternVideo2.5 pads its visual input to a multiple of four by repeating the final real frame; predictions referring to padded positions are mapped back to that final real frame. Runtime errors and empty outputs produce an empty prediction set in the pooled frame-level evaluation.

Set metrics for sentence-to-shot grounding. For a predicted clip set P and reference set G, we compute

$$\mathrm{Prec} = \frac{|P \cap G|}{|P|}, \qquad \mathrm{Rec} = \frac{|P \cap G|}{|G|}, \qquad F_1 = \frac{2\,\mathrm{Prec}\,\mathrm{Rec}}{\mathrm{Prec} + \mathrm{Rec}}, \tag{2}$$

with zero assigned when the relevant denominator is zero. We also compute $\mathrm{IoU} = |P \cap G|/|P \cup G|$ and $\mathrm{Hit@Any} = \mathbf{1}[|P \cap G| > 0]$. Aggregation is a two-level macro-average: metrics are first averaged over the sampled sentences within each video and then averaged over videos. The main table reports F_1.

Ranking metrics. Temporal ordering and saliency ranking are evaluated with Spearman's rank correlation for each valid video-level permutation and then macro-averaged over videos. Shot saliency targets are the mean source-provided frame-importance score within each shot. The document and no-document saliency settings use the same selected shots for a given video and seed.

Pooled frame-level metrics for oracle-shot prioritization. Let P_i and G_i be the predicted and reference frame sets for sentence i. We pool intersections and set sizes over all 684 sentences:

$$\mathrm{Prec}_{\mathrm{pool}} = \frac{\sum_i |P_i \cap G_i|}{\sum_i |P_i|}, \qquad \mathrm{Rec}_{\mathrm{pool}} = \frac{\sum_i |P_i \cap G_i|}{\sum_i |G_i|}, \tag{3}$$

and report their harmonic mean. Hit@k is the fraction of sentences for which $P_i \cap G_i \neq \varnothing$. CLIP ranks candidates deterministically by cosine similarity between the sentence embedding and each candidate-frame embedding and returns the top k frames. The saliency-prior diagnostic ranks candidates by their inherited frame-importance score and does not use the sentence.

CLIPScore in V2VT. For the joint summarization setting, text embeddings are averaged across generated summary sentences and visual embeddings are averaged across the selected frames before a single cosine similarity is computed. We use

$$\mathrm{CLIPScore} = 2.5 \max(\cos(\bar{e}_{\mathrm{text}}, \bar{e}_{\mathrm{visual}}), 0). \tag{4}$$

Thus, the score is not an average of independently scored sentence–frame pairs.

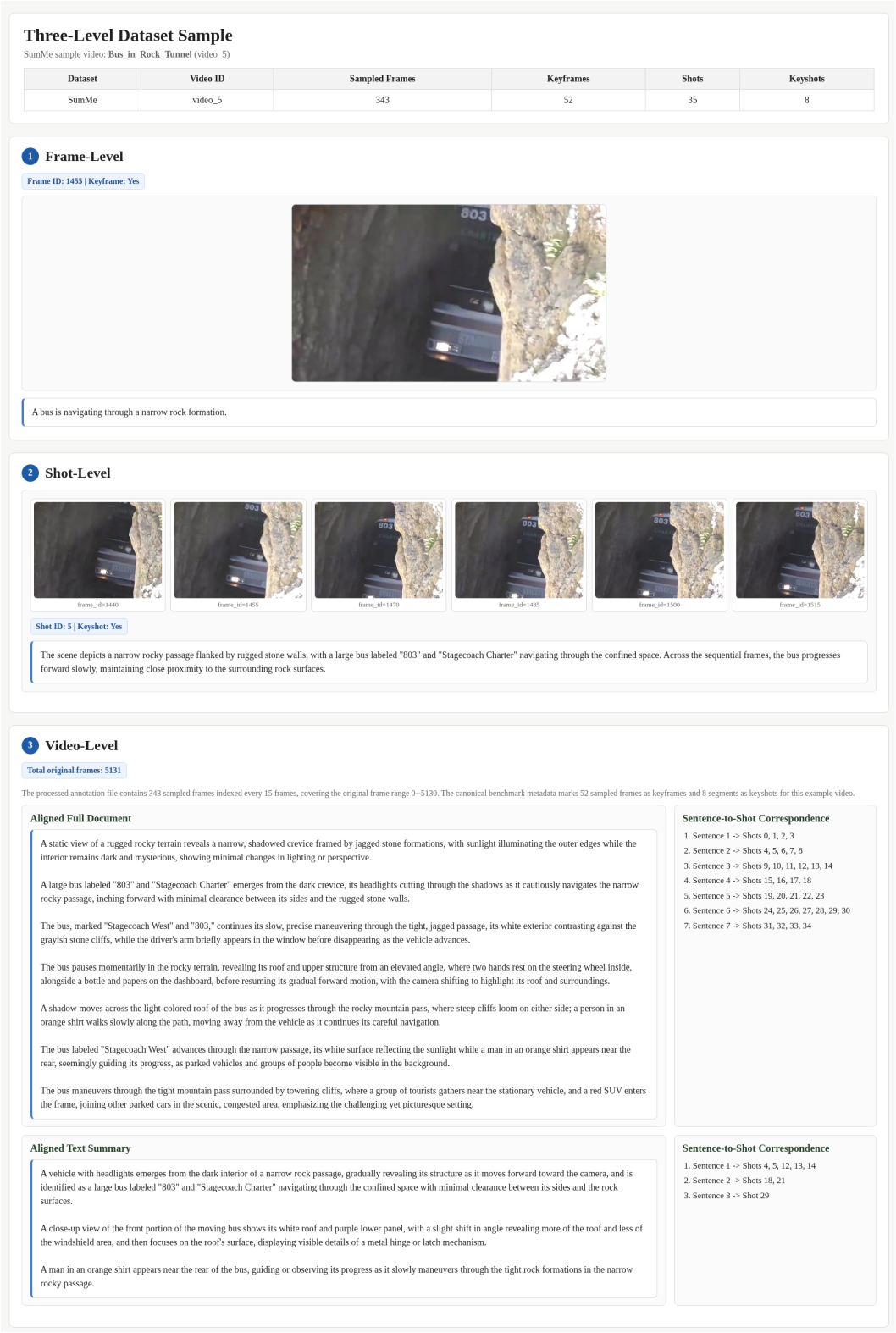

Figure 6: Example of the hierarchical representation for `video_5` (*Bus in Rock Tunnel*) from SumMe. Frame descriptions, shot descriptions, the aligned document, and the reference summary are linked to their corresponding visual units.

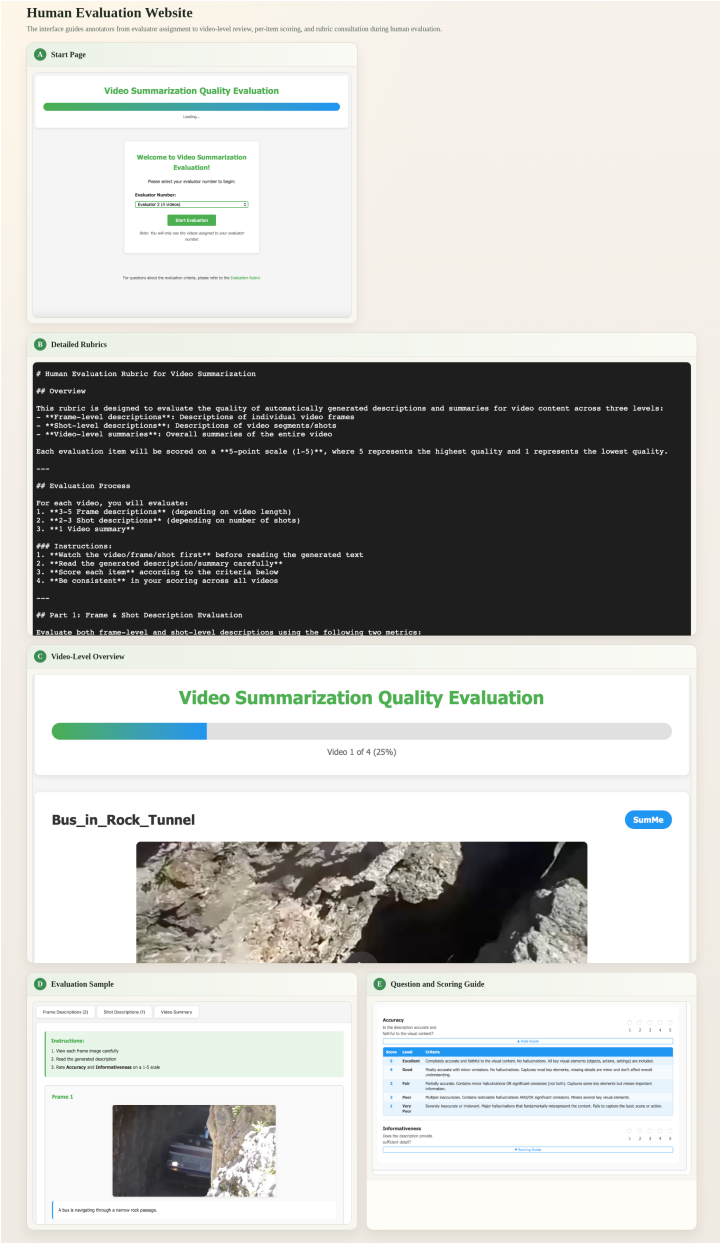

Figure 7: Human-evaluation interface. The workflow presents the visual evidence, annotation text, criterion-specific question, and expandable scoring guide separately for each judgment.