DVBench: Benchmarking MLLMs for Understanding Dynamic Charts and Narratives in Data Videos

Bomiao Wang[1]*, Zekai Shao[1]*, Jiexiang Lan[1], Xiaoliang Fu[1], Xingchen Zeng[2], Siming Chen[1]†

[1]Fudan University, [2]The Hong Kong University of Science and Technology (Guangzhou)
rimeiw66@gmail.com, zkshao23@m.fudan.edu.cn, simingchen@fudan.edu.cn

Abstract

While MLLMs have made significant strides in chart comprehension and video understanding, current evaluations largely isolate these capabilities, leaving a critical gap in understanding temporally evolving structured visual information. To address this gap, we introduce **DVBench**, a benchmark for evaluating MLLMs on data videos, a storytelling medium that integrates dynamic charts with structured narratives. We decompose data video understanding into five dimensions. DVBench comprises 300 real-world data videos and 1,000 human-verified QA pairs curated through a rigorous semi-automated pipeline. Extensive evaluations of nine MLLMs show that Gemini-3.1-Pro achieves the best overall performance, while Kimi-k2.5 is the strongest open-source model. We further identify two notable phenomena: open-source model performance does not scale strictly with parameter size, and narrative proficiency does not guarantee visual capability. Fine-grained analyses and ablation studies further reveal dimension-specific weaknesses and the effects of frame configurations and subtitle inputs, informing future MLLM development. DVBench is publicly available at https://bomiaowang.github.io/DVBench/.

1 Introduction

The continuous evolution of MLLMs has significantly advanced the frontier of visual understanding, enabling breakthroughs in cross-modal comprehension and complex reasoning (Zhang et al., 2024; DeepMind, 2025; Anthropic, 2026). To comprehensively assess the capabilities and progress of MLLMs, existing benchmarks evaluate models from multiple aspects of visual understanding (Li et al., 2024a), among which chart understanding evaluates a model's ability to decode structured, dense visual information and perform precise analytical reasoning (Masry et al., 2022; Wang et al., 2024), while video understanding assesses its capacity to track dynamic entities and comprehend sequential events (Li et al., 2024b; Fu et al., 2025).

However, these two foundational lines of evaluation have remained largely isolated, leading to a critical blind spot in current benchmarks: the perception and reasoning of dynamic, structural visual information over time. On one hand, chart understanding benchmarks are restricted to static visualizations (Wang et al., 2024; Xia et al., 2025; Masry et al., 2022). On the other hand, video understanding benchmarks focus almost exclusively on physical entities and actions within natural scenes. Even when complex narratives are introduced, they are typically confined to cinematic or real-world events (Wang et al., 2025b; Zhang et al., 2025), leaving a significant gap in the comprehension of data-driven storytelling.

This distinct fragmentation underscores the necessity of evaluating models on *Data Videos* (see Fig. 1 for an illustrative example), a storytelling medium that bridges this gap by integrating dynamic visualizations, narrative structures, and cinematic transitions (Heer and Robertson, 2007). Understanding data videos poses a comprehensive challenge for MLLMs, requiring the ability to (1) read and interpret dynamic charts, including their visual encodings, data insights, and animation semantics; (2) understand how these insights are organized and developed into a coherent narrative; and (3) integrate the correspondence between visual content and textual narration.

To this end, we introduce **DVBench**, the first comprehensive benchmark dedicated to evaluating the comprehension of data videos across five dimensions: **Narrative**, **Animation**, **Chart Perception**, **Chart Reasoning**, and **Alignment**. DVBench comprises 300 high-quality data videos,

*Equal contribution; co-first authors.
†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 9315–9338
October 24-29, 2026 ©2026 Association for Computational Linguistics

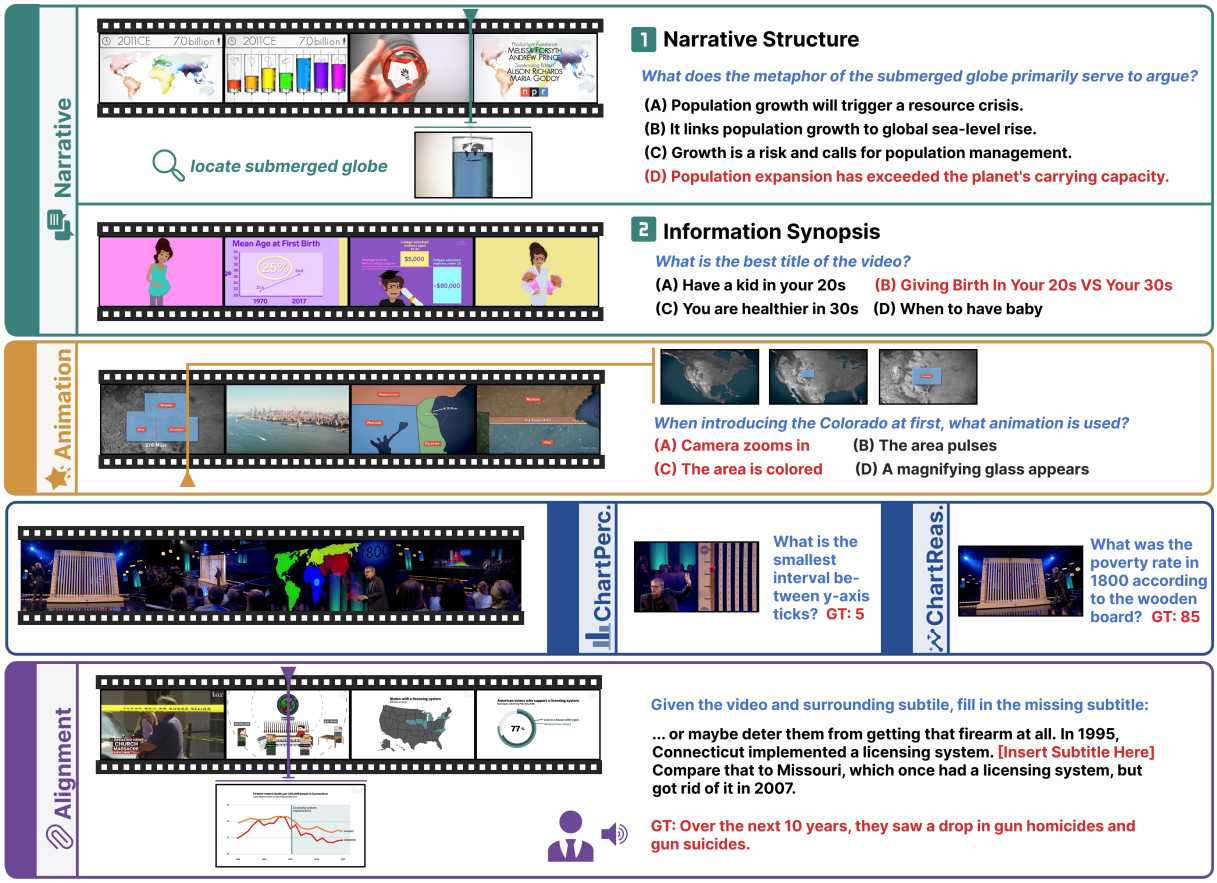

Figure 1: **Overview of the five evaluation dimensions in DVBench**. Our taxonomy systematically assesses models across: (1) **Narrative**, evaluating high-level comprehension of video themes; (2) **Animation**, focusing on the perception of dynamic transitions; (3) **Chart Perception** (denoted as ChartPerc.) and (4) **Chart Reasoning** (denoted as ChartReas.), respectively test fundamental visual identification and data insights extraction from dynamic charts in videos; and (5) **Alignment**, requiring models to infer masked subtitles. Correct options and ground truth (GT) answers are colored in red.

accompanied by 1,000 QA pairs constructed through a rigorous semi-automatic pipeline and human verification.

With this benchmark, we conduct extensive evaluations of nine prominent open-source and proprietary MLLMs. The experimental results show Gemini-3.1-Pro achieves the best performance, and Kimi-k2.5 emerges as the leading open-source model. Furthermore, we observe two notable phenomena: first, the capacity to understand data videos does not scale strictly with the model size; second, proficiency in the narrative understanding does not correlate with performance in the fine-grained visual perception.

To further characterize model capabilities, we conduct fine-grained analyses using question-level annotations derived from the design space of data videos. The results reveal model weaknesses in perceiving non-linear animation pacing and counting in dynamic charts, while robustness to increasing video length varies across models. Ablations further show that denser frame sampling does not necessarily help models, and incorporating subtitles leads to performance gains.

Our contributions are as follows: (1) we introduce a new task, data video understanding, with five dimensions to systematically evaluate MLLMs; (2) we curate a dataset of 300 data videos and 1,000 QA pairs via a rigorous semi-automatic pipeline and manual review; and (3) we conduct extensive evaluations with fine-grained analyses and ablations, yielding insights into current MLLM capabilities and limitations.

2 Related Works

We compare DVBench and other multimodal understanding benchmarks in Table 1. The following subsections discuss the current benchmarks of chart and video understanding and data video study, highlighting the necessity of our DVBench.

Benchmark	Dynamic Chart	Narratives	Annotation	Multi-dim. Eval
ChartQA (Masry et al., 2022)	✗	✗	👤 & 🤖	✗
CharXiv (Wang et al., 2024)	✗	✗	👤 & 🤖	✗
ChartX (Xia et al., 2025)	✗	✗	👤 & 🤖	✓
VisText (Tang et al., 2023)	✗	✓	👤 & 🤖	✗
ChartInsighter (Wang et al., 2025a)	✗	✓	👤	✓
ChartCap (Lim et al., 2025)	✗	✓	👤 & 🤖	✗
MVBench (Li et al., 2024b)	✗	✗	🤖	✓
MotionBench (Hong et al., 2025)	✗	✗	👤	✗
Video-MMMU (Hu et al., 2026)	✗	✗	👤	✓
LVBench (Wang et al., 2025b)	✗	✓	👤 & 🤖	✓
VRBench (Yu et al., 2025)	✗	✓	👤 & 🤖	✓
SeriesBench (Zhang et al., 2025)	✗	✓	👤 & 🤖	✓
DVBench (Ours)	✓	✓	👤 & 🤖	✓

Table 1: **Comparison of DVBench with existing multimodal understanding benchmarks.** 👤 and 🤖 respectively denote manual and automatic dataset curation. Multi-dim. Eval indicates whether the benchmark assesses multi-dimensional capabilities.

2.1 Chart Understanding Benchmarks

Current benchmarks in chart understanding mainly include chart QA and chart captioning. Early chart QA benchmarks focused on structural extraction and reasoning over scientific data within open-vocabulary scenarios (Kafle et al., 2018; Methani et al., 2020). Subsequent works introduced more sophisticated reasoning logic and expert-validated questions covering diverse real-world topics (Masry et al., 2022; Wang et al., 2024; Wu et al., 2024; Xia et al., 2025). Evaluation frameworks for chart captioning are dedicated to faithfulness, granularity, and hallucination-free generation (Tang et al., 2023; Wang et al., 2025a; Lim et al., 2025; Wang et al., 2026a). However, these benchmarks all remain confined to static visualizations. In contrast, our work investigates dynamic charts within videos, which pose a significantly higher cognitive challenge.

2.2 Video Understanding Benchmarks

The evaluation of video understanding capability is currently shifting from short clips toward long-form content and complex narratives. Early benchmarks progressed from static spatial perception to the dynamic perception of actions and object state changes (Li et al., 2024b; Hong et al., 2025; Ha et al., 2026). Subsequent research introduced reasoning tasks designed to evaluate causal relationships and temporal logic within long videos (Wang et al., 2025b; Yu et al., 2025). Regarding video caption, recent frameworks focus on fine-grained evaluation by quantifying correctness, comprehensiveness, and instruction-following capabilities (Li et al., 2025; Liu et al., 2025). Furthermore, various domain-specific video benchmarks have emerged to evaluate specialized narratives, ranging from cinematic plot progression (A Shah et al., 2026; Ataallah et al., 2025) to multi-disciplinary knowledge and scientific processes (Hu et al., 2026; Deng et al., 2025). Our DVBench establishes a framework to evaluating the understanding of data-driven narratives.

2.3 Data Video

Data videos have emerged as a prominent storytelling medium across diverse domains, demonstrating an exceptional ability to engage audiences and convey rich information through the integration of dynamic visualizations and structured narratives (Segel and Heer, 2010; Amini et al., 2015; Shao et al., 2025; Shen et al., 2025). Extensive research has investigated the foundational elements, characterizing the design space of visual-animation interplay and scene-semantic alignment (Shi et al., 2021; Cheng et al., 2022; Gao et al., 2026). Furthermore, researchers have investigated narrative strategies to better organize and convey information in data videos (Yang et al., 2022; Xu et al., 2022, 2023; Wei et al., 2025). Building on the research, we propose a systematic evaluation framework to assess MLLMs across diverse dimensions.

3 DVBench

3.1 Task Dimensions

Inspired by research in data videos, we deconstruct the understanding of data videos into five dimensions: Narrative, Animation, Chart Perception, Chart Reasoning, and Alignment in Fig 1.

For **Animation**, **Chart Reasoning**, and **Alignment**, we further divide each dimension into fine-grained categories derived from prior data video research (Wang et al., 2020; Shi et al., 2021; Cheng et al., 2022). These categories characterize different capabilities and support the analyses in Sec. 4.2.2. Definitions are introduced below.

Narrative assesses the ability to comprehend the high-level storytelling logic and communicative intent. It is further divided into two tasks: (1) *Narrative Structure* focuses on *how* the video narrates, which involves three question types: mapping visuals to narrative intent (e.g., What does the metaphor of the submerged globe primarily serve to argue?), mapping intent back to specific visuals (e.g., What visualization is presented to illustrate

the contrast between public perception and reality?) (Yang et al., 2022), and identifying the logical relationships between video clips (e.g., Why does the video introduce the income distribution chart by a world map?). (2) *Information Synopsis* focuses on *what* the video narrates, which involves tasks such as selecting the best title and identifying the video's attitude towards the topic.

Animation focuses on the perception of dynamic animation techniques and visual transitions used in charts. We characterize this dimension using four editorial layers (Shi et al., 2021): (1) *visualization elements* (e.g., transforming visual marks, axes, or legends), (2) *added elements* (e.g., adding annotations and embellishments), (3) *camera* (e.g., zooming or panning to alter viewing perspectives), and (4) *timeline* (e.g., controlling the pacing and speed of animation).

Chart Perception evaluates the recognition and direct reading of visual encodings and explicitly presented graphical information, including chart titles, axis labels, legend entries, tick marks, the number of ticks, and tick interval sizes. The target information is explicitly presented in the chart and can be obtained through direct visual reading, without requiring the derivation of higher-level data insights.

Chart Reasoning evaluates the ability to derive higher-level data insights through reasoning over the information presented in the chart. In contrast to Chart Perception, the target information is not explicitly stated as a graphical element and must be inferred by analyzing one or more visual elements, potentially across multiple frames. We systematically categorize these insights into 10 fine-grained types according to (Wang et al., 2020), ranging from localized queries (i.e., retrieving *values*, identifying *extremes*, and conditional *categorization*) and comparative statistics (i.e., measuring *proportions*, comparing *differences*, *ranking*, and *aggregation*) to broader structural analyses (i.e., tracking temporal *trends*, analyzing *distributions*, and finding *associations*).

Alignment evaluates the correspondence between dynamic charts and their narration through a subtitle cloze task, where models infer missing narrations from visual cues and surrounding subtitle context. We characterize this dimension using two semantic categories based on the role of the masked subtitle: *Data Insight* and *Data Context*. *Data Insight* refers to analytical facts, such as temporal trends or exact values, that are grounded in explicit visual evidence, whereas *Data Context* refers to background information that describes the broader narrative setting (Cheng et al., 2022).

3.2 Benchmark Construction

Our benchmark construction pipeline consists of three stages, as illustrated in Fig. 2.

Video Collection. To ensure the quality and representativeness of our dataset, we collect 300 data videos from prior research dedicated to data videos (Yang et al., 2022; Shi et al., 2021; Cheng et al., 2022; Xu et al., 2022, 2023; Gao et al., 2026; Gunturu et al., 2025). These videos cover a broad range of topics and presentation styles, while consistently featuring dynamic charts and structured narratives.

QA Pair Candidate Generation. We employ dimension-specific strategies to generate QA pair candidates.

For **Narrative** dimension, we input the video and subtitle into Qwen3-VL-Plus (Bai et al., 2025) to generate pseudo questions for the two narrative tasks following the taxonomy defined in Sec. 3.1. We further refine the distractors by, for example, mapping visual cues to irrelevant narrative strategies. Finally, we filter out trivial questions that Kimi-k2.5, Gemini-3.1-Pro, and Qwen3.5-397B-A17B can all answer correctly.

For the remaining four dimensions, we first employ Qwen3-VL-Plus to locate data clips, i.e., video segments in which charts convey data context or insights (Amini et al., 2017; Cheng et al., 2022). Based on these identified clips, we construct QA candidates for dimensions below.

For **Animation** and **Chart Perception** dimensions, we manually design questions targeting dynamic transitions and fundamental visual elements within the data clips, respectively. For **Chart Reasoning** dimension, to ensure a comprehensive extraction of the underlying data insights, we simultaneously extract visual data insights from the data clips using Qwen3-VL-Plus and textual insights from the subtitles using DeepSeek-v3.2 (DeepSeek-AI et al., 2025). After combining and formalizing these multi-modal insights into a structured schema (i.e., data fact) (Wang et al., 2020), the automated pipeline generates two distinct types of questions. The first type directly queries the extracted data insights (e.g., identifying specific values, trends, or differences). To generate the second type, we further cluster the

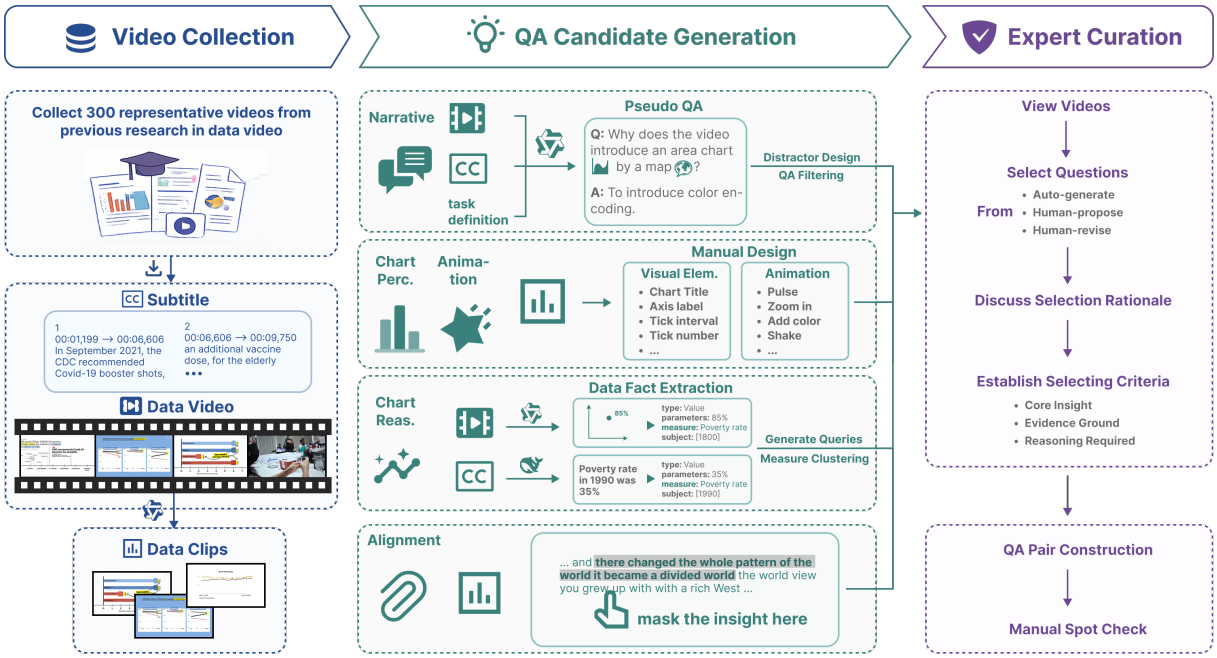

Figure 2: **Benchmark construction pipeline.** The process is divided into three main stages: (1) video collection, (2) QA candidate generation, and (3) expert curation. We collect videos from previous research and download the videos and subtitles. Then we utilize MLLM to identify data clips. During QA candidate generation, we employ dimension-specific strategies: Narrative questions are generated via MLLMs and undergo rigorous distractor design and filtering; Chart Perception and Animation questions are manually curated to target fundamental visual elements and dynamic transitions; Chart Reasoning utilizes an automated pipeline to extract, formalize, and cluster data insights for complex cross-frame queries; and Alignment questions are formulated as subtitle cloze tasks. Finally, annotators establish selection criteria and conduct benchmark curation and verification.

insights by their *subjects* and *measures*, utilizing these clusters to construct complex temporal tasks, such as tracking the change of a specific measure across different timestamps. For **Alignment** dimension, we use the identified data clips and mask narrations corresponding to either data insights or data contexts to construct subtitle cloze tasks. Detailed prompt templates used for QA candidate generation are provided in Appendix B.2.

Expert Curation. To ensure annotation quality, two annotators with visualization expertise first jointly viewed 10 videos to establish a consistent annotation protocol. They independently selected questions from an initial pool of 88 candidates and then discussed the rationale behind their selections, from which they derived a shared question selection criteria. Following these criteria, one annotator constructed the full QA set, while the other independently inspected 10% of the QA pairs, achieving 100% agreement. Details are in Appendix B.3.

3.3 Dataset Statistics

QA Pairs. Fig. 3(a) details the distribution of the 1,000 QA pairs across our five evaluation dimensions. We allocate the largest proportion to Chart Reasoning for two reasons. First, viewers predominantly focus on the underlying data insights rather than superficial graphical elements. Second, prior study (Wang et al., 2024) indicates that while models excel at basic perception, they still exhibit vulnerabilities in reasoning.

Figure 3(b) reports the chart type distribution in visual questions, which include Animation, Chart Perception, and Chart Reasoning dimensions only. Bar and line charts are the most prevalent chart types. The remaining questions cover a diverse long tail of diverse chart types, demonstrating the broad visual coverage of DVBench. Statistics on QA lengths are provided in Appendix C.

Videos. Figure 3(c) presents the video length and topic distribution. Video lengths range from under 30 seconds to approximately 37 minutes. We group them into three duration categories: Short ($\leq$ 3 mins), Medium (3–6 mins), and Long ($>$ 6 mins). The three groups are relatively bal-

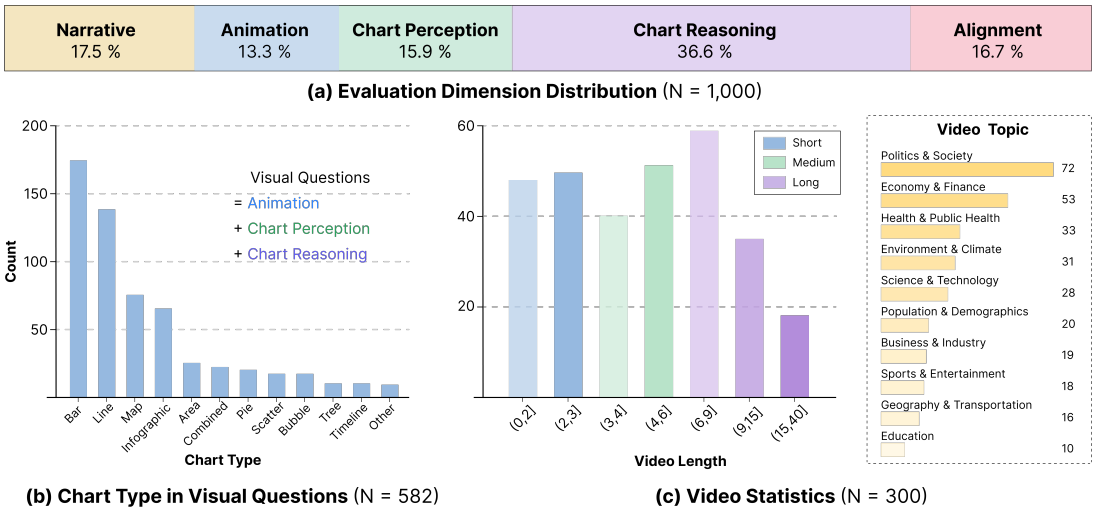

Figure 3: **Statistics of DVBench.** **(a)** Distribution of questions across the five evaluation dimensions. **(b)** Chart-type distribution of visual questions and its dimension-specific breakdown. Visual questions include only Animation, Chart Perception, and Chart Reasoning. **(c)** Video length and topic distribution of the collected data videos.

anced. The collected videos cover ten real-world topics, with Politics & Society and Economy & Finance being the two largest categories.

4 Experiment

4.1 Settings

We test a total of nine MLLMs, including both open-source (Qwen3.5 series (QwenTeam, 2026), Gemma-4 (Team et al., 2026a), Kimi-k2.5 (Team et al., 2026b)) and proprietary MLLMs (Gemini-3.1-Pro (Google DeepMind, 2026), Claude-4.6-Sonnet (Anthropic, 2026), and GPT-5.4 (OpenAI, 2026)). Model configurations, evaluation prompts and metrics are respectively detailed in Appendices D.1, D.5, D.4.

In the main experiment, each model follows its default frame-sampling configuration, with video-only input. To examine the influence of input configurations, we further conduct two ablations: a frame ablation with a unified 64-frame setting and different sampling rates, and a subtitle ablation comparing video-only input with video supplemented by subtitles.

4.2 Result

4.2.1 Main Result

Table 2 presents a comprehensive performance evaluation of 9 MLLMs across 5 dimensions.

The general superiority of proprietary models and the exceptional performance of Kimi-k2.5. The results indicate that proprietary models generally exhibit superior performance. However, Kimi-k2.5 emerges as a notable exception, achieving the second-highest average accuracy of 75.87% among all models, surpassing Claude-4.6-Sonnet and GPT-5.4. A similar pattern is observed in the Alignment dimension, where proprietary models and Kimi-k2.5 generally outperform the other open-source models across text-based, visual-grounding, and human evaluation metrics.

The capacity of open-source models to understand data videos does not scale directly with the number of parameters. Within the Qwen3.5 series, the relatively smaller dense model Qwen3.5-27B achieves 62.30% average accuracy and higher Alignment performance than its larger Mixture-of-Experts counterparts, Qwen3.5-397B-A17B and Qwen3.5-122B-A10B.

Open-source models demonstrate parity in macro-narrative understanding but exhibit disparity in fine-grained visual perception. Prominent open-source models display superiority in narrative comprehension. For example, Qwen3.5-27B achieves 74.29% accuracy in the Narrative dimension, closely trailing Claude-4.6-Sonnet. However, a significant performance gap emerges within visual dimensions. Qwen3.5-27B scores only 59.40% in Animation and 56.83% in Chart Reasoning, lagging substantially behind proprietary models such as Gemini-3.1-Pro (69.92% and 77.87%) and Claude-4.6-Sonnet (68.42% and 71.58%).

Model	Closed-ended Dimensions					Align.					
	Avg.	Narr.	Anim.	Perc.	Reas.	BL-2	MET	$F1_{BERT}$	EMS	EMS_{ref}	Human Eval.
Human	91.93	87.14	90.74	95.31	94.52	–	–	–	–	–	–
Proprietary Models											
Gemini-3.1-Pro (Google DeepMind, 2026)	**77.91**	**76.57**	69.92	**86.16**	**77.87**	**26.77**	38.20	83.02	26.49	54.83	3.95
Claude-4.6-Sonnet (Anthropic, 2026)	72.51	75.43	68.42	74.84	71.58	25.81	38.26	**83.28**	27.17	54.55	3.89
GPT-5.4 (OpenAI, 2026)	65.31	69.14	63.91	68.55	62.57	19.88	31.68	77.67	26.55	52.93	3.71
Open-source Models											
Kimi-k2.5 (Team et al., 2026b)	75.87	72.57	**72.18**	83.65	75.41	25.77	**41.22**	83.09	**27.48**	**54.89**	**3.97**
Qwen3.5-27B (QwenTeam, 2026)	62.30	74.29	59.40	64.15	56.83	17.81	27.57	79.21	26.25	52.73	3.48
Qwen3.5-397B-A17B (QwenTeam, 2026)	61.82	73.14	62.41	64.78	54.92	16.94	26.18	76.77	25.89	52.52	3.25
Qwen3.5-122B-A10B (QwenTeam, 2026)	59.30	70.86	58.65	62.26	52.73	12.96	21.74	77.61	25.78	51.96	3.00
Gemma-4-31B-It (Team et al., 2026a)	58.82	64.00	57.89	60.38	56.01	16.53	25.75	78.18	26.26	52.65	2.95
Qwen3.5-9B (QwenTeam, 2026)	52.82	55.43	54.89	61.01	47.27	8.20	16.07	73.98	25.45	50.41	2.73

Table 2: **Performance of evaluated MLLMs on DVBench.** The four closed-ended dimensions are evaluated by accuracy, where Avg. denotes their average accuracy. For the *Alignment* dimension, we report three categories: **text-based metrics** (BLEU-2, METEOR, and BERTScore F1), **visual grounding metrics** (EMScore and $EMScore_{ref}$), and **human evaluation**. EMScore and $EMScore_{ref}$ are metrics assessing the correspondence between generated text and visual content (Shi et al., 2022; Wang et al., 2026b). Human Eval. denotes the average rating on a five-point scale. The Human Baseline is evaluated on a stratified 20% subset of the closed-ended questions. Best and second-best results among MLLMs are highlighted in **bold** and underline, respectively.

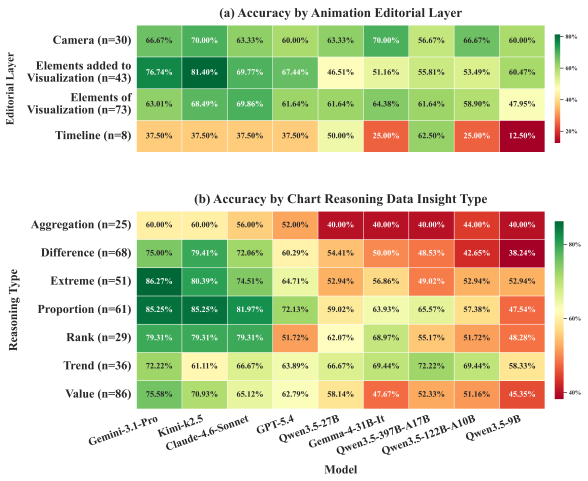

Figure 4: **Fine-grained performance across the Animation and Chart Reasoning dimensions.** (a) Accuracy distribution within the Animation task, evaluated across four editorial layers. (b) Accuracy evaluation across 10 reasoning question types. The x-axis, representing the evaluated MLLMs, is shared across both subplots.

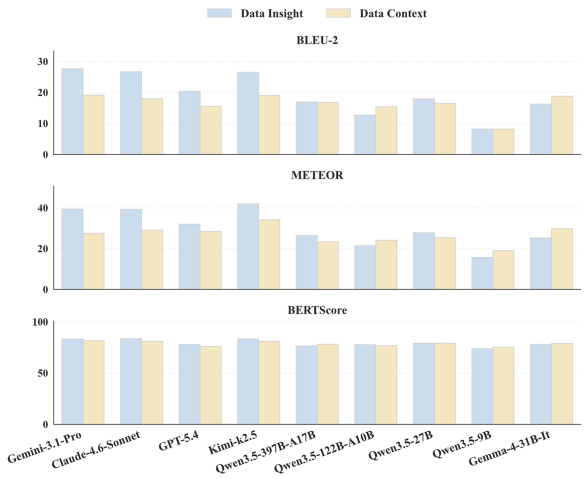

Figure 5: **Evaluation results of Alignment dimension.** We report the BERTScore, BLEU-2, and METEOR scores of 9 MLLMs on two semantic categories: *Data Insight* (blue bars) and *Data Context* (yellow bars). The x-axis, representing the evaluated MLLMs, is shared across all subplots.

4.2.2 Fine-grained Analysis

In this section, we further analyze model performance using the fine-grained categories defined in Sec. 3.1. Specifically, we examine performance across the four editorial layers of Animation, the 11 data insight types of Chart Reasoning, and the two semantic categories of Alignment. We additionally investigate how video length affects model performance across different dimensions.

Animation: Failure to Capture Non-linear Animation Pacing. We analyze the Animation dimension from the perspective of the four editorial layers (i.e., *what* is being animated) introduced in Sec. 3.1. The four layers cover changes in visualization elements, added elements, camera perspectives, and animation pacing. Examples of the four layers are provided in Appendix F.1. As shown in

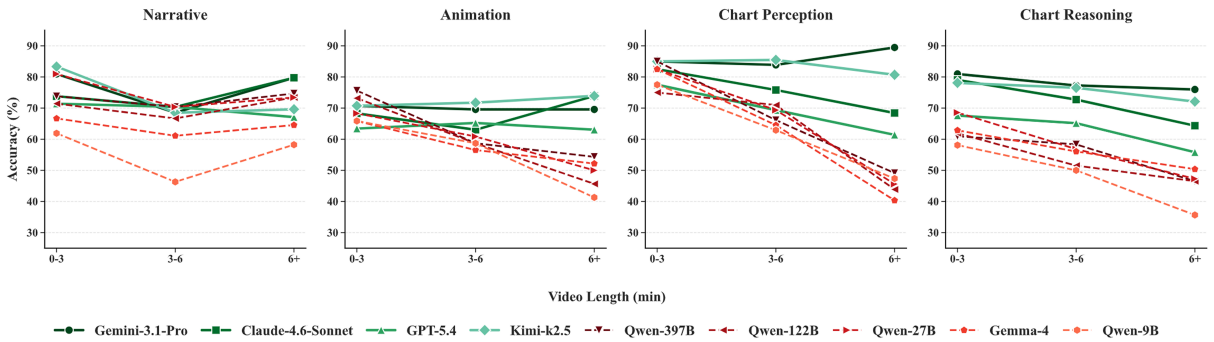

Figure 6: **Performance across varying video lengths.** The line charts illustrate the accuracy of 9 MLLMs on 4 dimensions as the video length increases from short ($\leq$ 3 mins) to long ($>$ 6 mins). Green lines represent strong models, while red lines represent weak ones.

Fig. 4(a), the Timeline layer is particularly challenging, with nearly all models achieving accuracy below 37.50%. While questions about visualization elements, added elements, and camera perspectives can often be answered by recognizing changes in visual states, Timeline questions require models to perceive how the pace of these changes evolves over time, such as acceleration and deceleration. The consistently low performance suggests that current MLLMs struggle to capture such fine-grained temporal dynamics in data videos. We further examine the effect of frame sampling on Timeline questions in Appendix E. The results suggest that sparse sampling sparse sampling is one contributor rather than the cause of the poor performance.

Chart Reasoning: Dynamic Charts Induce Severe Counting Hallucinations. Within the Chart Reasoning dimension, Aggregation evaluates the ability to count and synthesize information across visual elements. These questions require models to identify the elements that satisfy a given condition and determine their exact number, sometimes by tracking the relevant elements across multiple frames. As shown in Fig. 4(b), Aggregation is one of the most challenging reasoning categories, with several open-source models achieving accuracy around 40.00%. This suggests that current MLLMs still struggle with fine-grained counting of visual elements and condition-based aggregation in dynamic charts. Detailed cases are provided in Appendix F.2.

Alignment: The Divergence in Insight and Context Grounding. Text-based metrics, visual grounding metrics, and human evaluation reveal a similar pattern: Kimi-k2.5, Gemini-3.1-Pro, and Claude-4.6-Sonnet achieve the strongest performance. However, we observe a clear divergence in the text-based metrics. As shown in Fig 5, BLEU-2 and METEOR show noticeable differences between the two categories: proprietary models and Kimi-k2.5 generally perform better on *Data Insight* than on *Data Context*, whereas the other open-source models exhibit a larger drop on *Data Insight*. Error analysis further shows that this gap is mainly associated with visual-text grounding errors rather than phrasing differences. Detailed analysis and examples are provided in Appendix F.3.

Video Length: Strong Models Are Not Susceptible to Long Videos. We analyze model performance across short, medium, and long videos. As illustrated in Fig 6, leading proprietary models and Kimi-k2.5 demonstrate remarkable temporal resilience; their overall performance plateaus even as the video length extends beyond 6 minutes. In contrast, weaker models experience a nearly linear decline in accuracy as the video length increases. Furthermore, we observe that the impact of video length is task-dependent. For high-level Narrative dimension, which requires holistic semantic abstraction, the performance of nearly all models remains relatively stable, unaffected by increased video length, except for Qwen3.5-9B. However, in fine-grained visual tasks (illustrated in the three rightmost subfigures), weaker models struggle with long videos, with accuracy frequently plummeting below 50%.

4.2.3 Ablation Studies

Frame Ablation: More Frames Do Not Nec-

(a) Unified Frame Budget

Model	Default	64F	Δ
Gemini-3.1-Pro	$77.91_{1\ \text{fps}}$	75.15	-2.76
Claude-4.6-Sonnet	$72.51_{100\text{F}}$	69.87	-2.64
GPT-5.4	$65.31_{64\text{F}}$	65.31	0.00
Qwen3.5-397B-A17B	$61.82_{2\ \text{fps}}$	66.51	+4.69
Qwen3.5-122B-A10B	$59.30_{2\ \text{fps}}$	67.73	+8.43
Qwen3.5-27B	$62.30_{2\ \text{fps}}$	68.43	+6.13
Qwen3.5-9B	$52.82_{2\ \text{fps}}$	60.65	+7.83

(b) Effect of Sampling Rate

Model	0.5 fps	1 fps	2 fps
Qwen3.5-397B-A17B	**71.07**	70.23	61.82
Qwen3.5-122B-A10B	**70.73**	69.36	59.30
Qwen3.5-27B	**72.14**	70.93	62.30
Qwen3.5-9B	63.47	**64.64**	52.82

All values denote average accuracy over the four closed-ended dimensions.

Table 3: **Ablation of frame-sampling configurations.** (a) compares each model's default configuration with a unified 64-frame budget; subscripts indicate the default configuration. (b) reports the effect of frame sampling rate on the Qwen3.5 models.

Model	ΔAvg.	ΔNarr.	ΔAnim.	ΔPerc.	ΔReas.
Gemini-3.1-Pro	+1.80	+6.29	+2.26	-2.51	+1.36
Claude-4.6-Sonnet	+0.48	+1.14	-0.75	-2.51	+1.92
GPT-5.4	+4.80	+6.86	+2.26	+1.89	+6.01
Kimi-k2.5	+2.64	+5.14	-2.26	+1.88	+3.55
Qwen3.5-397B-A17B	+5.29	+5.72	0.00	+1.89	+8.47
Qwen3.5-122B-A10B	+5.89	+4.57	-0.76	+1.26	+10.93
Qwen3.5-27B	+4.81	+5.14	-0.75	0.00	+8.74
Qwen3.5-9B	+7.80	+12.57	-2.26	+2.51	+11.47
Gemma-4-31B-It	+6.73	+11.43	+3.76	+1.26	+7.92

Table 4: **Subtitle Ablation.** Absolute performance changes between video-only input and video input supplemented with subtitles. Positive values indicate performance gains with subtitle input, with darker blue backgrounds denoting larger improvements; non-positive values are left unshaded. The corresponding video-only results are reported in Table 2.

essarily Help. We further evaluate all applicable models using 64 uniformly sampled frames, a common setting within their supported input ranges. As shown in Table 3(a), Gemini-3.1-Pro and Claude-4.6-Sonnet remain the strongest models. Within the Qwen3.5 family, Qwen3.5-27B continues to outperform the larger models. These results show that the model ranking under a unified frame budget remains consistent with that observed in the main experiment, while the performance gaps between models become smaller.

We further evaluate the Qwen3.5 models under different frame sampling rates, as shown in Table 3(b). Performance does not improve monotonically with denser sampling: the 2 fps setting consistently underperforms 0.5 and 1 fps across all Qwen3.5 models, while 0.5 or 1 fps achieves the best performance. This indicates that increasing the number of sampled frames does not necessarily provide more useful evidence for understanding data videos.

Subtitle Input: Subtitles Primarily Benefit High-Level Semantic Understanding. We compare video-only input with video input supplemented by subtitles across all nine models. As shown in Table 4, subtitles improve the average performance of all models, with larger gains for the Qwen3.5 and Gemma models than for most proprietary models. The improvements are concentrated in Narrative and Chart Reasoning. In particular, Chart Reasoning improves by 7.92–11.47% for the Qwen3.5 and Gemma models. In contrast, gains on Chart Perception and Animation are smaller and less consistent. Overall, subtitle input provides greater benefits for semantic understanding than for low-level visual perception.

5 Conclusion

We introduce DVBench, the first comprehensive benchmark evaluating MLLMs on data video understanding. We divide the data video understanding task into 5 dimensions. We curate a dataset of 300 videos and 1,000 QA pairs via a semi-automatic pipeline and expert verification. Evaluation of 9 MLLMs demonstrates the superiority of proprietary models and exposes unexpected results in parameter scaling and cross-dimensional proficiency. Fine-grained analyses reveal weaknesses across different evaluation dimensions, while ablations demonstrate the impact of different input configurations. These findings provide valuable insights for future research.

Limitations

While our semi-automatic construction pipeline combined with expert validation ensures high data fidelity, the reliance on human review currently limits the scale of DVBench. We also observe that performance on the Chart Perception dimen-

sion is relatively saturated among frontier models, suggesting that future benchmark iterations should incorporate more challenging fine-grained perception tasks to better distinguish model capabilities. Future work could further expand the benchmark to include multilingual data videos, 3D visualizations, and AR infographics. Moreover, as native end-to-end video models advance toward processing denser audio-visual streams, they may provide stronger capabilities for the fine-grained temporal tasks that remain challenging for current MLLMs. DVBench's extensible taxonomy is designed to accommodate increasingly complex reasoning tasks and support the continued evaluation of future model architectures.

Acknowledgments

We want to thank the reviewers for their suggestions. This work was supported by the National Natural Science Foundation of China (No. 62472099), the AI for Science Program of the Shanghai Municipal Commission of Economy and Informatization (Grant No. 2025-GZL-RGZN-BTBX-02028), the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM904), and the Ji Hua Laboratory S&T Program (No. X250881UG250).

References

Nisarg A Shah, Amir Ziai, Chaitanya Ekanadham, and Vishal M. Patel. 2026. Cineaste: A fine-grained contextual movie question answering benchmark with automated data curation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4376–4386.

Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Christophe Hurter, and Pourang Irani. 2015. Understanding data videos: Looking at narrative visualization through the cinematography lens. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 14591468, New York, NY, USA. Association for Computing Machinery.

Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Andres Monroy-Hernandez, and Pourang Irani. 2017. Authoring data-driven videos with dataclips. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):501510.

Anthropic. 2026. Introducing claude 4.6 sonnet. https://www.anthropic.com/news/claude-sonnet-4-6.

Kirolos Ataallah, Eslam Mohamed Bakr, Mahmoud Ahmed, Chenhui Gou, Khushbu Pahwa, Jian Ding, and Mohamed Elhoseiny. 2025. InfiniBench: A benchmark for large multi-modal models in long-form movies and TV shows. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19485–19512, Suzhou, China. Association for Computational Linguistics.

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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Hao Cheng, Junhong Wang, Yun Wang, Bongshin Lee, Haidong Zhang, and Dongmei Zhang. 2022. Investigating the role and interplay of narrations and animations in data videos. volume 41, pages 527–539.

Google DeepMind. 2025. Gemini 3 flash: frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Andong Deng, Taojiannan Yang, Shoubin Yu, Lincoln Spencer, Mohit Bansal, Chen Chen, Serena Yeung-Levy, and Xiaohan Wang. 2025. Scivideobench: Benchmarking scientific video reasoning in large multimodal models. *arXiv preprint arXiv:2510.08559*.

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, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multimodal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24108–24118.

Lin Gao, Leixian Shen, Yuheng Zhao, Jiexiang Lan, Huamin Qu, and Siming Chen. 2026. Sceneloom:

Communicating data with scene context. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):1328–1338.

Google DeepMind. 2026. Gemini 3.1 pro: A smarter model for your most complex tasks. https://blog.google/innovation-and-ai/models-and-research/gemini-models/gemini-3-1-pro/.

Aditya Gunturu, Ben Pearman, Keiichi Ihara, Morteza Faraji, Bryan Wang, Rubaiat Habib Kazi, and Ryo Suzuki. 2025. Mapstory: Prototyping editable map animations with llm agents. In *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*, UIST '25, New York, NY, USA. Association for Computing Machinery.

Hyeonjeong Ha, Jinjin Ge, Bo Feng, Kaixin Ma, and Gargi Chakraborty. 2026. Narrativetrack: Evaluating video language models beyond the frame. *arXiv e-prints*, pages arXiv–2601.

Jeffrey Heer and George Robertson. 2007. Animated transitions in statistical data graphics. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1240–1247.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. 2025. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8450–8460.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. 2026. Video-MMMU: Evaluating knowledge acquisition from multidisciplinary professional videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27798–27828, San Diego, California, United States. Association for Computational Linguistics.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, Ying Tai, Wankou Yang, Yabiao Wang, and Chengjie Wang. 2024a. A survey on benchmarks of multimodal large language models. *Preprint*, arXiv:2408.08632.

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. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22195–22206.

Shihao Li, Yuanxing Zhang, Jiangtao Wu, Zhide Lei, Yiwen He, Runzhe Wen, Chenxi Liao, Chengkang Jiang, An Ping, Shuo Gao, Suhan Wang, Zhaozhou Bian, Zijun Zhou, Jingyi Xie, Jiayi Zhou, Jing Wang, Yifan Yao, Weihao Xie, Yingshui Tan, and 4 others. 2025. If-vidcap: Can video caption models follow instructions? *Preprint*, arXiv:2510.18726.

Junyoung Lim, Jaewoo Ahn, and Gunhee Kim. 2025. Chartcap: Mitigating hallucination of dense chart captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13171–13182.

Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Jixuan Chen, Pandeng Li, Boqiang Zhang, Nianzu Yang, Yinglu Li, Zuan Gao, Yun Zheng, and Hongtao Xie. 2025. Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness. *Preprint*, arXiv:2502.14914.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

OpenAI. 2026. Introducing gpt-5.4. https://openai.com/index/introducing-gpt-5-4/.

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

QwenTeam. 2026. Qwen3.5: Towards native multimodal agents.

Edward Segel and Jeffrey Heer. 2010. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1139–1148.

Zekai Shao, Leixian Shen, Haotian Li, Yi Shan, Huamin Qu, Yun Wang, and Siming Chen. 2025. Narrative player: Reviving data narratives with visuals. *IEEE Transactions on Visualization and Computer Graphics*, 31(10):6781–6795.

Leixian Shen, Haotian Li, Yun Wang, and Huamin Qu. 2025. Reflecting on design paradigms of animated data video tools. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25, New York, NY, USA. Association for Computing Machinery.

Yang Shi, Xingyu Lan, Jingwen Li, Zhaorui Li, and Nan Cao. 2021. Communicating with motion: A design space for animated visual narratives in data videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Yaya Shi, Xu Yang, Haiyang Xu, Chunfeng Yuan, Bing Li, Weiming Hu, and Zheng-Jun Zha. 2022. Emscore: Evaluating video captioning via coarse-grained and fine-grained embedding matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17929–17938.

Benny Tang, Angie Boggust, and Arvind Satyanarayan. 2023. VisText: A benchmark for semantically rich chart captioning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7268–7298, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Sherif El Abd, Vaibhav Aggarwal, Robin Algayres, Alek Andreev, Olivier Bachem, Ian Ballantyne, Cormac Brick, Victor Crbune, Michelle Casbon, Mayank Chaturvedi, Aditya Chawla, Victor Cotruta, Alice Coucke, Phil Culliton, Robert Dadashi, Lucas Dixon, Mohamed Elhawaty, Utku Evci, and 304 others. 2026a. Gemma 4 technical report. *Preprint*, arXiv:2607.02770.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Ziwei Chai, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, and 318 others. 2026b. Kimi k2.5: Visual agentic intelligence. *Preprint*, arXiv:2602.02276.

Fen Wang, Zekai Shao, Qiman Kang, Chunran Hu, Zhixuan Zhang, Lexu Xie, Chao Liu, and Siming Chen. 2026a. Chartfi: Benchmarking faithfulness and insightfulness of chart descriptions from multimodal large language models. *Preprint*, arXiv:2605.23694.

Fen Wang, Bomiao Wang, Xueli Shu, Zhen Liu, Zekai Shao, Chao Liu, and Siming Chen. 2025a. Chartinsighter: An approach for mitigating hallucination in time-series chart summary generation with a benchmark dataset. *IEEE Transactions on Visualization and Computer Graphics*, 31(6):3733–3745.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. 2025b. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22958–22967.

Xueyan Wang, Dingyi Yang, and Qin Jin. 2026b. HowToNarrate: A general-domain benchmark for synchronized video narration with external knowledge. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 39110–39131, San Diego, California, United States. Association for Computational Linguistics.

Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. 2020. Datashot: Automatic generation of fact sheets from tabular data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):895–905.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. In *Advances in Neural Information Processing Systems*, volume 37, pages 113569–113697. Curran Associates, Inc.

Zheng Wei, Huamin Qu, and Xian Xu. 2025. Telling data stories with the hero's journey: Design guidance for creating data videos. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):962972.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. 2024. ChartInsights: Evaluating multimodal large language models for low-level chart question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 12174–12200, Miami, Florida, USA. Association for Computational Linguistics.

Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. 2025. Chartx and chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*, 34:7436–7447.

Xian Xu, Aoyu Wu, Leni Yang, Zheng Wei, Rong Huang, David Yip, and Huamin Qu. 2023. Is it the end? guidelines for cinematic endings in data videos. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Xian Xu, Leni Yang, David Yip, Mingming Fan, Zheng Wei, and Huamin Qu. 2022. From wow to why: Guidelines for creating the opening of a data video with cinematic styles. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Leni Yang, Xian Xu, XingYu Lan, Ziyan Liu, Shunan Guo, Yang Shi, Huamin Qu, and Nan Cao. 2022. A design space for applying the freytag's pyramid structure to data stories. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):922–932.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li,

Songze Li, Zhenxiang Li, Zhongying Tu, Conghui He, Yu Qiao, Yali Wang, Yi Wang, and Limin Wang. 2025. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21655–21666.

Chenkai Zhang, Yiming Lei, Zeming Liu, Haitao Leng, ShaoGuo Liu, Tingting Gao, Qingjie Liu, and Yunhong Wang. 2025. Seriesbench: A benchmark for narrative-driven drama series understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28995–29004.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

Appendix

A Potential Risks

While DVBench is designed strictly for the academic evaluation of MLLMs, the use of data videos for model development and evaluation may still involve potential downstream risks. Data videos often present real-world statistics and analytical narratives concerning socio-economic, geopolitical, environmental, and public health topics, and may therefore contain sensitive information or reflect particular perspectives.

The potential risks are relatively constrained in DVBench because all videos are curated from previously published academic works rather than unrestricted online media. In addition, the dataset curation process includes manual review and quality control to identify and remove problematic or misleading content where possible. These measures reduce, but do not completely eliminate, the potential risks associated with the dataset.

First, there is a risk of bias inheritance. Although the videos are sourced from established prior works, their narratives may still contain implicit assumptions, framing choices, or subjective interpretations. Models trained or evaluated on such content may consequently reproduce or amplify these biases.

Second, there is a risk of misinformation generation. Our error analysis shows that current open-source MLLMs frequently exhibit numerical hallucinations and weak visual grounding when interpreting charts and animated visualizations. If such limitations are not properly mitigated, deploying these models in applications such as automated journalism, educational analytics, or data storytelling could lead to fabricated insights or visually persuasive but factually incorrect narratives.

B Dataset Construction Details

B.1 Licenses and Intended Use

DVBench is developed exclusively for non-commercial academic research and evaluation of MLLMs. The benchmark annotations, extracted structured information, QA pairs, and accompanying metadata are released under the **CC BY-NC-SA 4.0** license.

The benchmark contains 300 data videos collected from previously published research works (Yang et al., 2022; Shi et al., 2021; Cheng et al., 2022; Xu et al., 2022, 2023; Gao et al., 2026; Gunturu et al., 2025). These videos were originally made publicly available on platforms such as YouTube and Vimeo. We do not claim ownership of the original video, audio, or visual materials, whose copyrights remain with their respective creators and copyright holders. Accordingly, DVBench is intended only to facilitate standardized academic evaluation, and users are responsible for complying with applicable copyright laws and the terms of the original hosting platforms.

To respect the rights of content owners, we provide a removal mechanism for third-party content included in the benchmark. If a copyright holder believes that specific content should not be included, we will promptly review and address reasonable removal requests.

During dataset construction and manual quality control, the videos and QA pairs were also reviewed to minimize personally identifiable information (PII), offensive language, and potentially harmful content. However, given that the source videos were originally created by third parties, we cannot guarantee the complete absence of such content.

B.2 Dataset Construction Prompt

In this section, we provide the detailed prompt templates utilized in our semi-automatic benchmark construction pipeline.

Prompt for Narrative QA Generation

[Role]
You are an expert in Data Visualization Analysis and Video Understanding.

[Task Category]
Your task is to analyze the provided video content to generate QA pairs evaluating the viewer's narrative comprehension. The actual video title is: `{title}`. Your generated questions must strictly fall into one of two categories:

Category 1: Narrative Structure
Design questions focusing on the storytelling logic and visual-narrative mapping.

- **Type 1: VIS → Topic/Intent.** Ask about a specific visualization and require its narrative intent. *Distractor Requirements:* (1) Metric-focused (raw data insight); (2) Logical but Irrelevant; (3) Extraneous Insight.
- **Type 2: Topic/Intent → VIS.** Provide a narrative purpose and ask which visual method conveys it. *Distractor Requirements:* Visualizations that appear in the video but do not serve this specific purpose.
- **Type 3: Logical.** Inquire about the logical connection between two chart transitions (e.g., "The

bar chart supports the line chart") or infer the reason for the next visual content.

Category 2: Information Synopsis
Design questions focusing on the theme and subjective tone.

- **Title:** What is the best title or main topic of the video? (Use the provided title to guide the answer, and design plausible alternative titles as distractors).
- **Attitude:** What is the attitude of the video towards the main topic? (Output a single adjective like optimistic, pessimistic, or neutral).

[Output Format]

```
[
  {
    "category": "Narrative Structure",
    "type": "[VIS -> Topic, Topic -> VIS, or
Logical]",
    "question": "[Insert specific
question]",
    "answer": "[Insert specific answer]",
    "distractors": ["[Distractor 1]",
"[Distractor 2]", "[Distractor 3]"],
    "evidence_timestamp": "[Start_Time -
End_Time]"
  },
  {
    "category": "Information Synopsis",
    "type": "[Title or Attitude]",
    "question": "[Insert specific
question]",
    "answer": "[Insert specific answer]",
    "distractors": ["[Distractor 1]",
"[Distractor 2]", "[Distractor 3]"],
    "evidence_timestamp": "[Start_Time -
End_Time]"
  }
]
```

Prompt for Data Clip and Data Insight Extraction

[Role]
You are an expert in Data Visualization Analysis and Video Understanding.

[Task]
Your task is to analyze the provided video, identify video clips that includes chart, and identify data insight for each chart, such as values (e.g., the death number in 2010 was 106), trends (e.g., the birth rate is increasing from 2000 to 2005), or ranks (e.g., China's military ranks 1st).

[Constraints]

1. **Visual Grounding:** Every answer must be derived solely from the video's visual content. Avoid using external prior knowledge.
2. **Objectivity:** Questions must have a single, unambiguous, and deterministic answer.
3. **Conciseness:** Answers should be brief (a single word, a number, or a short phrase).
4. **Chart Clustering:** Organize the generated QA pairs into groups based on the specific chart or visual component they describe. Each group should represent a single, distinct chart (e.g., "World Population Line Chart" or "Income Distribution Map"). Ensure that all data points, titles, and axis information for a specific chart are clustered together in the output.

[Output Format]

```
[
  {
    "chart_group": "[Insert name of the
specific chart]",
    "insights": [
      {
        "description": "A description of the
data insight",
        "evidence_timestamp": "[Start_Time -
End_Time]"
      }
    ]
  }
]
```

Prompt for Data Insight Extraction from Subtitle

[Role]
You are an expert in Data Visualization Analysis and Video Understanding.

[Task]
Extract data insight from the subtitle of data video.

[Extraction Scope]

1. **Precise Data:** Extract explicitly mentioned percentages, monetary amounts, population figures, and other hard numbers.
2. **Qualitative Descriptions:** Even in the absence of specific numbers, you must extract descriptions of trends (growth/decline/stagnation), intensity (sharp/gradual), comparisons (surpassing/lagging), or structural changes (bridging the gap/moving into the middle class).

[Insight Definition]

- **description:** A natural language description of the data insight.
- **type:** Select the most appropriate data insight category listed below.
- **time_stamp:** The period when the insight is mentioned in the video time stamp. From when the complete sentence begins to when the sentence ends (e.g., 00:00:14,400 → 00:00:18,000).

[Insight Type Definition]

- **Value:** Retrieve exact values under specific criteria.

- **Proportion:** Measure the percentage of selected attributes within a set.
- **Difference:** Compare attributes or temporal changes.
- **Distribution:** Show the breakdown or share across attributes.
- **Trend:** Present a general tendency over time.
- **Rank:** Sort attributes based on their values.
- **Aggregation:** Calculate statistical indicators like average, sum, or count.
- **Association:** Identify correlations between multiple attributes.
- **Extreme:** Find the top/bottom cases or "-est" values.
- **Categorization:** Select attributes satisfying specific conditions.

[Output Format]

```
{
  "description": "[Insert natural language description]",
  "type": "[Insert one of the 10 fact types]",
  "time_stamp": "[Start_Time → End_Time]"
}
```

[Subtitle]
{subtitle}

Prompt for Insight Formalization

[Role]
You are a data extraction expert. Your task is to convert a list of natural language descriptions of data facts into a structured Data Fact format.

[Fact Definition]
Each Fact consists of the following components:

- **description:** A natural language description of the fact.
- **type:** Select the most appropriate fact category listed below.
- **parameters:** Specific descriptive arguments for the fact type.
- **measure(s):** Numerical fields treated as dependent variables (e.g., Sales, Units, Price), derived from functions like SUM, AVG, or COUNT. Use original words from the subtitles as much as possible. Use complete phrases (e.g., "death prevented" rather than "death").
- **subject:** Defines the scope of the fact through three factors:
  - *context:* The data subspace defined strictly by filters applied to specific table columns (dimensions). This field must only contain direct column-value filters found in the data table. Do not include narrative context, assumptions, or descriptive phrases.
  - *breakdown(s):* Dimensions used to divide the subspace into groups.
  - *focus:* Specific groups within the subspace to be highlighted or emphasized.
- **time_stamp:** The period when the fact is mentioned in the video time stamp. From when the complete sentence begins to when the sentence ends (e.g., 00:00:14,400 → 00:00:18,000).

[Fact Type Definition]

- **Value:** Retrieve exact values under specific criteria.
- **Proportion:** Measure the percentage of selected attributes within a set.
- **Difference:** Compare attributes or temporal changes.
- **Distribution:** Show the breakdown or share across attributes.
- **Trend:** Present a general tendency over time.
- **Rank:** Sort attributes based on their values.
- **Aggregation:** Calculate statistical indicators like average, sum, or count.
- **Association:** Identify correlations between multiple attributes.
- **Extreme:** Find the top/bottom cases or "-est" values.
- **Categorization:** Select attributes satisfying specific conditions.
- **Outlier:** Identify unexpected or statistically abnormal data points.

[Output Format & Examples]

```
[
  {
    "description": "Protein takes 66% in the diet on Sunday.",
    "type": "Proportion",
    "parameters": "66%",
    "measure": "Dietary Intake",
    "subject": {
      "context": ["Sunday"],
      "breakdown": "Nutrient",
      "focus": ["Protein"]
    }
  },
  {
    "description": "The Sales of iOS increased over the years.",
    "type": "Trend",
    "parameters": "increase",
    "measure": "Sales",
```

```
    "subject": {
      "context": ["iOS"],
      "breakdown": "Year",
      "focus": null
    }
  },
  {
    "description": "The national average
price for regular gas was $4.06 in July
2008.",
    "type": "Aggregation",
    "parameters": "4.06",
    "measure": "AVG(national price)",
    "subject": {
      "context": ["2008", "July"],
      "breakdown": null,
      "focus": null
    }
  },
  ... ]
```

[Input Description]
{description}

Prompt for Cross-Frame Reasoning QA Generation

[Role]
You are a data extraction and QA generation expert. Your task is to generate complex cross-frame reasoning questions for a data video based on a provided collection of structured Data Facts.

[Generation Strategies]
You will receive Data Facts grouped either by **Measure** or by **Subject**. Apply the corresponding generation logic based on the input type:

- **Strategy 1: Grouped by Measure (Cross-Attribute Derivation)**
  - *Condition:* Facts share the same Measure (dependent variable) but differ in their Subject attributes (e.g., Temporal, Spatial, or Categorical differences).
  - *Logic:* Generate questions requiring mathematical calculations (differences, multiples, ratios) or logical comparisons (ranking changes, share displacement) based on facts distributed across different video timestamps.
- **Strategy 2: Grouped by Subject (Point-to-Surface Query)**
  - *Condition:* Facts share a similar Subject (e.g., the same entity or year) but track different Measures across different timestamps.
  - *Logic:* Identify a specific context/event implicit in Fact 1 to use as a positioning anchor. Without explicitly stating this background, ask to extract the numerical value or description from Fact 2 that corresponds to the same background (e.g., "When [Event in Fact 1] occurred, what was the [Measure in Fact 2]?").

[Constraints]

1. **Mandatory Cross-Frame:** Questions must link ≥ 2 facts from different timestamps.
2. **Evidence Traceability:** The output must cite the specific atomic fact descriptions and their original timestamps.
3. **Conciseness and Directness:** Questions must be clear and focused. Answers must only contain the data fact parameters, such as specific numerical values, a delta, or a status (e.g., "50%", "200 tons", "increase"). Do not use full sentences.

[Output Format & Examples (JSON)]

```
[
  // Example for Strategy 1: Grouped by
Measure
  {
    "question": "Between 1500 and 1700, by
how many percentage points did the percentage
of death rate in the world decrease?",
    "answer": "50%",
    "evidence": [
      {"fact": "In 1800, the percentage of
death rate in the world was 85%.",
"time_stamp": "00:00:27 → 00:00:38"},
      {"fact": "In 1990, the percentage of
death rate in the world was 35%.",
"time_stamp": "00:01:23 → 00:01:30"}
    ],
    "reasoning": "Calculates the difference
between the death rate in the two
timestamps."
  },
  // Example for Strategy 2: Grouped by
Subject
  {
    "question": "At the point when Starship's
inaugural flight generated 17 million pounds
of thrust, what was its specified payload
capacity to low Earth orbit?",
    "answer": "150 metric tons",
    "evidence": [
      {"fact": "Starship's first flight
generated a record 17 million pounds of
thrust.", "time_stamp": "00:05:12 →
00:05:18"},
      {"fact": "The Starship is designed to
carry 150 metric tons to low Earth orbit.",
"time_stamp": "00:12:45 → 00:12:52"}
    ],
    "reasoning": "Fact 1 anchors the context
(thrust event). Fact 2 provides the payload
measure for that same subject."
  }
]
```

[Input Fact Clusters]
{fact_grouped_by_measure_or_subject}

B.3 Annotation Details

Annotator Background. The benchmark annotation, fact-checking, and QA verification were conducted by two authors with expertise in data

visualization and prior experience in creating and analyzing data videos. One annotator holds a Master's degree, and the other is a PhD candidate.

Question Selection Criteria. Before constructing the full benchmark, the two annotators jointly reviewed 10 data videos to establish consistent criteria for question selection. During this calibration stage, they compiled an initial pool of candidate questions from three sources: automatically extracted data insights, questions independently proposed by the annotators, and human-revised or synthesized insights based on the generated candidates. Semantically similar questions were merged and duplicates were removed, resulting in a pool of 88 candidate questions.

The two annotators then independently selected questions from this pool for inclusion in the benchmark, achieving a raw agreement rate of 94.32% and a Cohen's κ of 0.88. All disagreements were subsequently discussed and resolved. Based on this calibration process, the final selection criteria required each question to: (1) target a core insight conveyed by the video rather than an incidental detail; (2) be grounded in the original visual or narrative evidence; and (3) require sufficient understanding or reasoning, such as integrating information across multiple scene transitions, tracking dynamically changing chart elements, or performing non-trivial data interpretation.

QA Curation and Verification. Following the established protocol, one annotator constructed the complete QA set, including questions, answers, and distractors. The second annotator independently inspected 10% of the resulting QA pairs, checking question validity, answer correctness, evidence grounding, and consistency with the assigned evaluation dimension. Potential ambiguities or inconsistencies were resolved through discussion before finalizing the benchmark.

Annotation Interface. To facilitate the human-in-the-loop annotation and verification process, we developed a web-based annotation interface, as illustrated in Fig. 7. The interface integrates video playback, subtitles, extracted data insights, identified data clips, generated QA candidates, and cross-frame reasoning information within a unified workspace. The left panel contains a synchronized video player at the top and displays cross-frame reasoning questions, grouped by their measures and subjects, together with the corresponding reasoning information at the bottom. The right panel contains three vertically organized sections showing the subtitles, extracted data insights, and identified data clips with QA candidates, respectively. The interface further supports interactive temporal navigation across different annotation components. By clicking an extracted data fact, data clip, or question, annotators can directly navigate the video to its corresponding timestamp. This synchronization enables annotators to efficiently verify the relationship among visual evidence, subtitles, extracted insights, and QA candidates. Annotators can then revise candidate content, verify answers and distractors, and assign fine-grained categories according to the evaluation taxonomy.

C More Dataset Statistics

Table 5 provides a detailed statistics of question and option lengths across four core evaluation dimensions. Table 6 shows the length of the collected video subtitles and the length of missing subtitles in subtitle cloze tasks.

Dimension	Question Len.	Option Len.
Narrative	10.5 (6 – 29)	7.4 (1 – 30)
Animation	12.4 (7 – 26)	6.4 (1 – 21)
Chart Perception	12.2 (5 – 24)	2.5 (1 – 12)
Chart Reasoning	14.3 (3 – 32)	1.3 (1 – 11)

Table 5: Question and option length statistics. Values are presented as *Mean (Range: Min–Max)*.

Alignment	Max Length	Min Length	Avg Length
Video Subtitle	12035	5	1190.0
Answer	47	4	15.4

Table 6: Statistics on the lengths of full video subtitles and the length of missing subtitles that need models to predict.

D Evaluation Details

D.1 Model Configurations

Table 7 summarizes the model versions, visual input configurations, decoding hyperparameters, and inference backends used in our evaluation. To ensure deterministic generation and maximum reproducibility, we set the `temperature` to 0 and

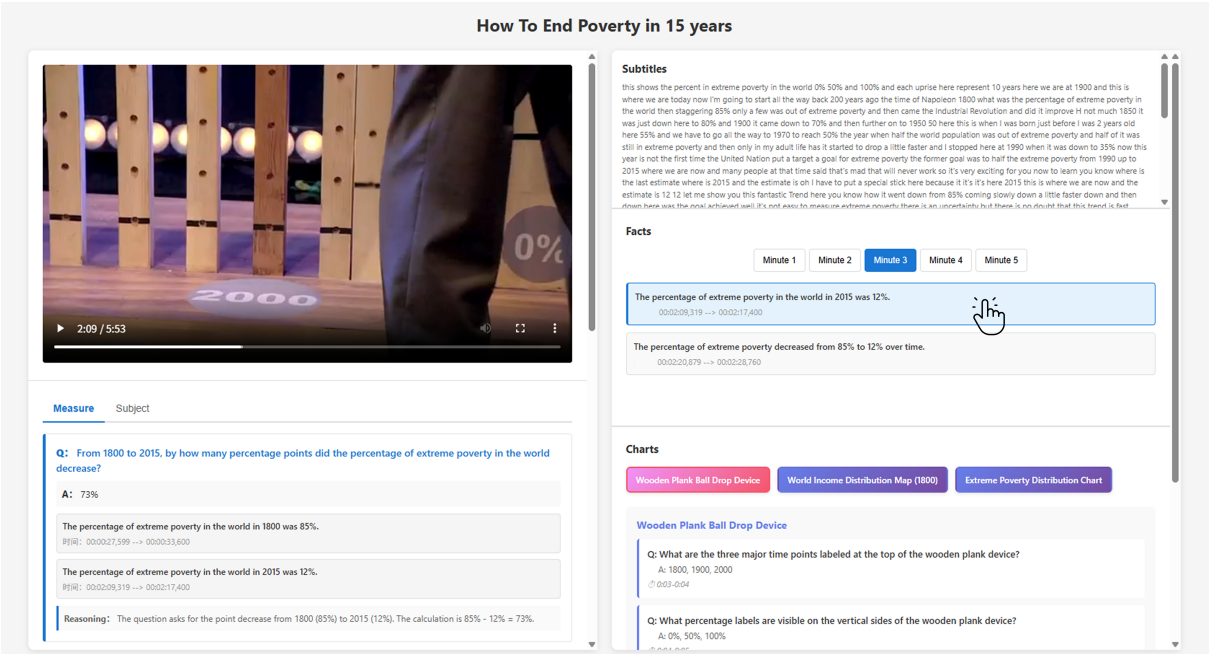

Figure 7: **Annotation interface of DVBench.** The layout is optimized for efficient verification. The left panel contains synchronized video playback (top) and cross-frame reasoning questions with their reasoning processes (bottom). The right panel displays the video subtitles (top), extracted data insights (middle), and identified data clips alongside chart perception QA pairs (bottom). An interactive click-to-seek function allows annotators to click on any insights or questions to automatically jump to the relevant timestamp for rapid visual verification.

`top-p` to 1 wherever applicable. For locally deployed open-source models running via vLLM, we explicitly enforce `do_sample=False`.

In main experiment, we adopt a default sampling rate for Gemini (1 fps), Gemma (1 fps), and Qwen series (2 fps) models. For GPT-5.4 and Claude-4.6-Sonnet, taking into account the strict payload size limitations of their API requests, we uniformly sample a fixed number of frames per video (64 frames for GPT-5.4 and 100 frames for Claude-4.6-Sonnet).

D.2 Model Licenses

Table 8 summarizes the licenses or access conditions of the models evaluated in this work. Proprietary models and Kimi-k2.5 are accessed through official APIs, while locally deployed open-source models are evaluated according to their released model and code licenses.

D.3 Computational Resources

All open-source models were deployed locally in BF16 precision on a computing node equipped with 8×NVIDIA H100 (80GB) GPUs. The evaluation of these local models required approximately 240 GPU hours. All inferences were conducted strictly for academic research and evaluation purposes.

D.4 Evaluation Metrics

Exact Match (EM) Tasks. For tasks requiring deterministic and precise answers (e.g., specific numerical values, dates, or concise data facts), we utilize the EM metric. To ensure a robust and fair evaluation against minor formatting variations, both the model's generated outputs and the ground-truth answers undergo standard text normalization (e.g., lowercasing, stripping punctuation, and removing articles) prior to the string matching process.

Choice-Based Tasks (Single and Multiple Choice). For single-choice and multiple-choice questions, we adopt standard accuracy as an evaluation metric. To rigorously evaluate multiple-choice questions, a model's prediction is considered correct if and only if it exactly matches the complete set of ground-truth options (i.e., exact subset match, with no missing or incorrect options selected).

Open-Ended Tasks (Alignment Dimension). For open-ended questions, where models are required to fill in the missing subtitles, we evaluate model outputs from three perspectives: text

Model	Version	Input Frames	do_sample	temp	top-p	Inference
Proprietary						
GPT-5.4	gpt-5.4	64	–	0	1	API
Gemini-3.1-Pro	gemini-3.1-pro-preview	1 fps	–	0	1	API
Claude-4.6-Sonnet	claude-sonnet-4-6	100	–	0	–	API
Open-Source Models						
Kimi-k2.5	moonshotai/Kimi-K2.5	–	–	–	–	API
Gemma-4-31B-It	google/gemma-4-31b-it	1 fps	False	0	1	vLLM
Qwen3.5-397B-A17B	Qwen/Qwen3.5-397B-A17B	2 fps	False	0	1	vLLM
Qwen3.5-122B-A10B	Qwen/Qwen3.5-122B-A10B	2 fps	False	0	1	vLLM
Qwen3.5-27B	Qwen/Qwen3.5-27B	2 fps	False	0	1	vLLM
Qwen3.5-9B	Qwen/Qwen3.5-9B	2 fps	False	0	1	vLLM

Table 7: **Inference configurations of the MLLMs evaluated in the main experiment**, including model versions, input frame configurations, decoding hyperparameters, and inference backends. The symbol "–" indicates that the parameter is not exposed or cannot be configured.

Model	Model License / Access	Code License / Access
Official API Models		
GPT-5.4	Official API terms	Official API terms
Gemini-3.1-Pro	Official API terms	Official API terms
Claude-4.6-Sonnet	Official API terms	Official API terms
Kimi-k2.5	Official API terms	Official API terms
Locally Deployed Open-Source Models		
Gemma-4-31B-It	Apache 2.0	Apache 2.0
Qwen3.5-397B-A17B	Apache 2.0	Apache 2.0
Qwen3.5-122B-A10B	Apache 2.0	Apache 2.0
Qwen3.5-27B	Apache 2.0	Apache 2.0
Qwen3.5-9B	Apache 2.0	Apache 2.0

Table 8: Summary of licenses and access conditions for the models evaluated in DVBench.

similarity, visual-text grounding, and human judgment.

Text-based Metrics. We evaluate the similarity between generated and reference text using **BLEU-2** (Papineni et al., 2002), **METEOR** (Banerjee and Lavie, 2005), and **BERTScore F1** (Zhang* et al., 2020). To ensure reproducibility of the lexical metrics, we compute the scores using standardized Python packages. Specifically, we utilize the NLTK library (`nltk`$\geq$`3.6.0`) for tokenization (`punkt`) and metric computation. **BLEU-2** evaluates bi-gram overlap between the generated text and the reference answer. To reduce the penalty for short responses with zero n-gram counts, we configure BLEU using NLTK's smoothing function `method1` (`SmoothingFunction().method1`). **METEOR** further incorporates stemming and synonym matching through NLTK's WordNet resources (`wordnet` and `omw-1.4`). Beyond lexical matching, **BERTScore F1** measures semantic similarity between the generated and reference text based on contextual token embeddings. We implement BERTScore using the official `bert_score` package (v0.3.13) with `roberta-large` as the contextual embedding backbone.

Visual Grounding Metrics. To further evaluate whether the generated text is grounded in the visual content of the video, we employ **EMScore** and $\mathbf{EMScore}_{\mathrm{ref}}$. EMScore measures the correspondence between the generated text and the video by jointly considering global video-text similarity and fine-grained frame-word alignment. $\mathrm{EMScore}_{\mathrm{ref}}$ further incorporates the reference text into the evaluation, assessing the generated response with respect to both the visual evidence and the reference semantics.

Human Evaluation. We additionally conduct human evaluation to directly assess the quality of the generated subtitles. Human evaluators rate each response on a five-point scale according to its correctness, consistency with the visual evidence, and faithfulness to the intended message. A score of 1 indicates a response that is incorrect or unsupported by the video, whereas a score of 5 indicates a fully correct and visually grounded response that faithfully conveys the intended meaning. We report the average human rating for each model.

D.5 Evaluation Prompt

Prompt Template: Exact Match (EM) Task

```
Based on the video, please answer the
following question:
{Question Text}
Please provide the answer directly in
<answer></answer> without any units, symbols,
or explanation. For example, if the answer
is 15% or $1, output only <answer>15</answer>
or <answer>1</answer>.
Answer:
```

Prompt for Multiple-Choice Task (Single Answer)

```
Based on the video, please answer the
following question:
{Question Text}
A. {Option 1}
B. {Option 2}
C. {Option 3}
D. {Option 4}
Output only the answer label (e.g., A, B, C,
or D) between <answer> and </answer> tags
(e.g., <answer>A</answer>).
Answer:
```

Prompt for Multiple-Choice Task (Multiple Answers)

```
Based on the video, please answer the
following question:
{Question Text}
A. {Option 1}
B. {Option 2}
C. {Option 3}
D. {Option 4}
This question has more than one correct
answer. Output answer labels of all correct
options answer label between <answer> and
</answer> tags (e.g.,
<answer>A;B;C</answer>).
Answer:
```

Prompt for Alignment Dimension

```
Please analyze the video and the surrounding
subtitle context to identify the missing
segment:
Subtitle Context:
{Subtitle}
Based on the video content and the context
of the subtitles, please fill in the
narration for [Insert Subtitle Here], and
output the filled content directly between
the <answer> and </answer> tags.
Answer:
```

E Timeline Question Analysis

To further investigate whether the difficulty of *Timeline* questions is primarily caused by sparse frame sampling, we evaluate these questions under different frame conditions available to each model.

Model	64 frames	0.5 fps	1 fps	Original setting
Gemini-3.1-Pro	37.5	-	37.5	37.5 (1 fps)
Claude-4.6-Sonnet	50.0	-	-	37.5 (100 frames)
GPT-5.4	37.5	-	-	37.5 (64 frames)
Qwen3.5-397B-A17B	50.0	50.0	50.0	62.5 (2 fps)
Qwen3.5-122B-A10B	25.0	25.0	25.0	25.0 (2 fps)
Qwen3.5-27B	37.5	37.5	37.5	50.0 (2 fps)
Qwen3.5-9B	37.5	37.5	50.0	12.5 (2 fps)

Table 9: **Timeline performance under different frame sampling conditions.** We report accuracy under different sampling rates where supported, and each model's original inference setting. The sampling rate or frame budget used in each model's original setting is shown in parentheses.

The results (Table 9) indicate that frame sampling can affect individual *Timeline* performance, but increasing the frame budget or sampling rate does not consistently resolve the difficulty. We therefore consider sparse sampling to be one possible contributor to the low performance on *Timeline* questions, rather than the sole cause.

F Qualitative Case Studies

To provide a more intuitive understanding of the models' capabilities and failure modes, we present several qualitative case studies across different evaluation dimensions.

F.1 Animation Evaluation Cases

We present qualitative examples across the four editorial layers of animation in Fig. 8. The cases highlight a critical vulnerability: while MLLMs can generally identify explicit visual transformations (e.g., color changes or added elements), they consistently struggle with the Timeline layer. Tasks requiring the perception of non-linear animation pacing, such as identifying when a chart speeds up, slows down, or pauses, prove exceptionally challenging.

F.2 Chart Reasoning Evaluation Cases

Fig. 9 illustrates qualitative examples from the Aggregation questions in Chart Reasoning.

F.3 Alignment Analysis

To investigate the specific reasons why open-source models significantly underperform proprietary models on EM metrics (BLEU-2 and METEOR), we conduct a qualitative analysis of the *low-EM / high-BERTScore* cases generated by open-source models. We define such a case at the model level as a sample whose EM scores are both

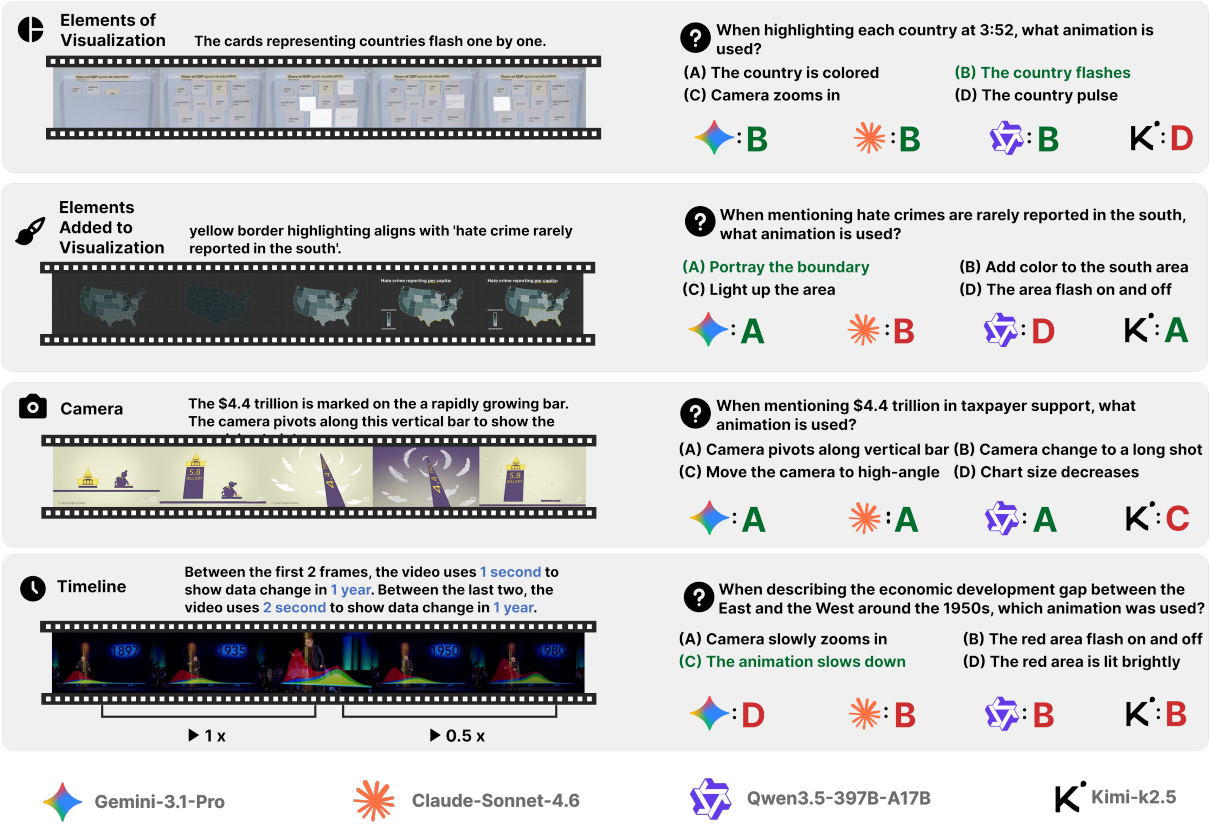

Figure 8: **Cases of Animation QA pairs.** We present questions of four editorial layers here. As we can see, questions of animation on Timeline layer (e.g., slow down, speed up, and pause in visualization animation) are most challenging for models.

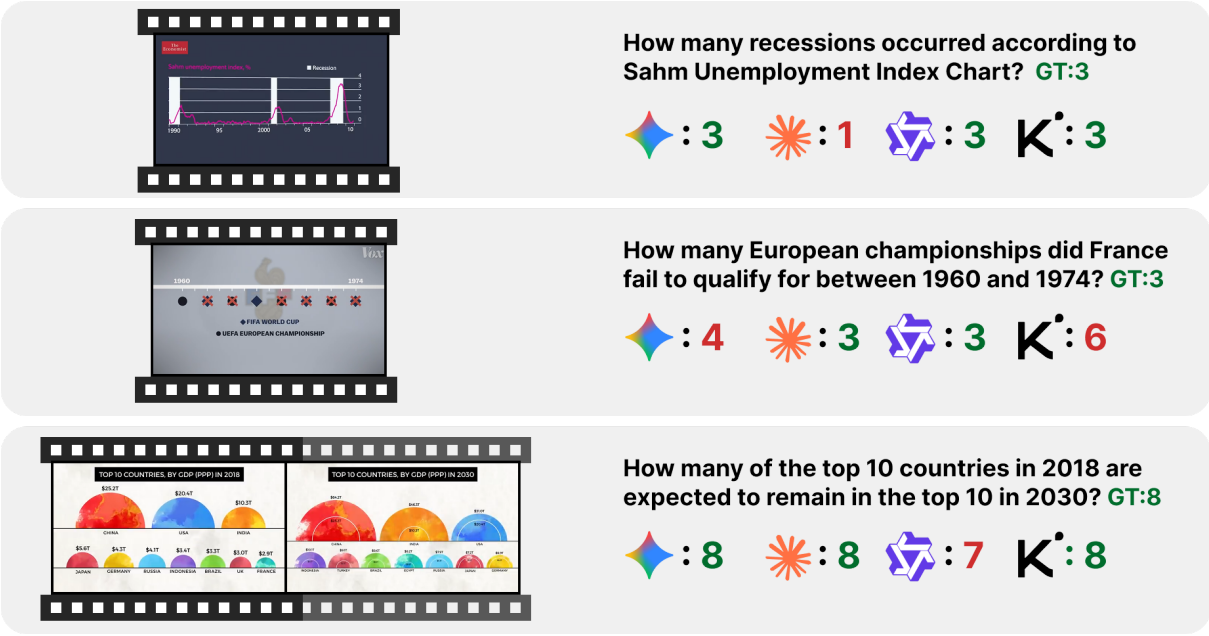

Figure 9: **Cases of Chart Reasoning QA pairs.** We illustrate three Aggregation questions. While the first two questions can be answered using a single keyframe, the third requires integrating information across two discrete frames. Successfully answering these queries requires models to accurately ground target chart elements across single or multiple frames before performing specific numerical calculations.

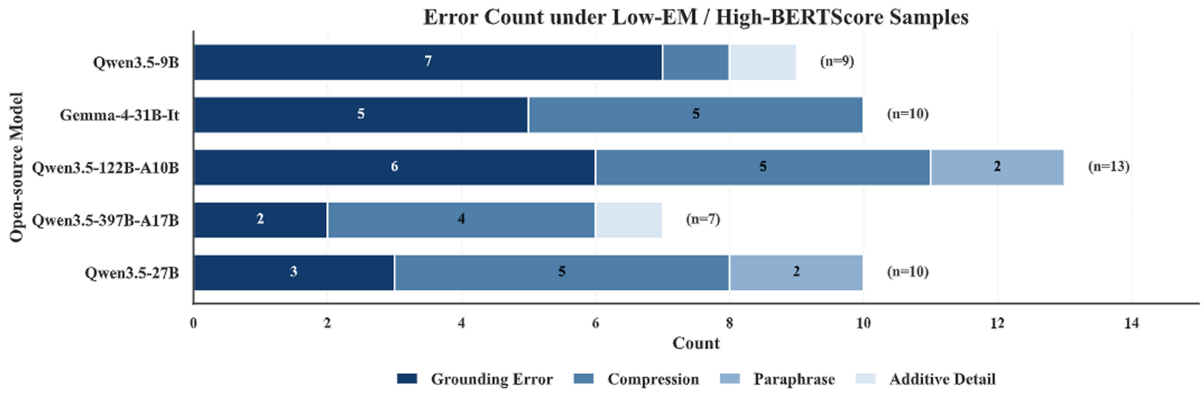

Figure 10: **Counts of error types for the low-EM / high-BERTScore answers across open-source models.** Each horizontal stacked bar shows the number of examples in the slice (defined per model as BLEU-2 model median, METEOR model median, and BERTScore model median), grouped by four error types. Per-row totals are shown to the right as (n=).

below the model's median, while its BERTScore is at or above the median. We identify four error types: **(1) Grounding Error**: the predicted answer is factually inconsistent with the gold answer or the visual evidence. **(2) Compression**: the prediction preserves the core meaning but drops important details. **(3) Paraphrase**: the prediction is semantically equivalent to the gold answer, with only surface-form variation. **(4) Additive Detail**: the prediction preserves the core meaning but adds extra details, numbers, or assertions that are not present in the gold answer.

As shown in Fig. 10, most examples fall into Grounding Error and Compression, indicating outputs that are semantically similar to references but omit or compress key information, causing a low exact-match score. Model-specific differences appear: e.g., Qwen3.5-9B has a larger share of Grounding Error, while other models show relatively higher compression. Table 10 presents representative cases, illustrating typical error patterns.

G AI Assistants Usage Declaration

AI assistants were utilized to facilitate several stages of this work, including dataset construction, coding support, figure design, and language polishing. Crucially, all AI-generated outputs underwent rigorous manual verification and extensive revision by the authors to ensure accuracy and quality.

Model	Model Response	Error Type
Case 1: there changed the whole pattern of the world it became a divided world the world view you grew up with with a rich West here and with poor Africa and Asia here		
Gemma-4-31B-It	and then in 1948 the population grew and the incomes increased even more and the world looked like this	Grounding error
Qwen3.5-397B-A17B	the shape of the world changed	Compression
Qwen3.5-122B-A10B	in the 1970s the population grew here and the incomes increased here	Grounding error
Qwen3.5-27B	but in the 1980s the growth of China they start to get out of poverty they close this Gap and they are coming over on this side and the shape of the world change again and look now Africa is following here we are today in a completely new shape of the world for those of you who still think about the world as the rich and the poor	Additive detail
Qwen3.5-9B	and the shape of the world change again	Compression
Case 2: But the IRS found that from 2008 to 2010 Americans paid only 81.7 percent of the taxes they really owed. Assuming that rate is the same today, That's about $660 billion dollars in tax dodging each year.		
Gemma-4-31B-It	But that's not all that's owed. According to an IRS study, about 18 percent of taxes owed are never paid. That's 660 billion dollars.	Paraphrase

Table 10: **Representative alignment cases illustrating different error types.** For each case, the gray-highlighted text at the top shows the gold answer, followed by model-generated responses and their corresponding error types.