DVBench: Benchmarking Multimodal Large Language Models on the Integration of Dynamic Charts and Narratives in Data Videos

Anonymous ACL submission

Abstract

While Multimodal Large Language Models (MLLMs) have made significant strides in chart comprehension and video understanding, current evaluations isolate these capabilities, leaving a critical blind spot in the perception of dynamic, structural visual information over time. To address this gap, we introduce **DVBench**, the first comprehensive benchmark dedicated to evaluating MLLMs on data videos, a storytelling medium integrating dynamic charts with structured narratives. We systematically decompose data video comprehension into 5 dimensions: Narrative, Animation, Chart Perception, Chart Reasoning and Alignment. And we curate a dataset comprising 300 real-world data videos and 1,000 QA pairs via a rigorous semi-automatic pipeline. Extensive evaluations of 9 MLLMs reveal that Gemini-3.1-Pro and Kimi-k2.5 achieve the highest overall and open-source performance, respectively. Notably, we uncover two surprising phenomena: open-source model capacity does not scale strictly with parameter size, and narrative proficiency does not guarantee visual capability. Furthermore, fine-grained analyses expose critical vulnerabilities in perceiving non-linear animation pacing and precise counting within dynamic charts. The models also exhibit a grounding gap between grasping general contexts and extracting precise insights, alongside significant performance decline in longer videos. These findings provide a crucial foundation and valuable insights for advancing multimodal reasoning. DVBench is available at anonymous repository.

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; Singh et al., 2025; Anthropic, 2026). To systematically quantify these advancements, contemporary evaluations typically assess the proficiency of models across several foundational capabilities (Li et al., 2024a; Fu et al., 2024). Among these core dimensions, 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). Concurrently, video understanding serves as the primary benchmark for spatio-temporal decoding, assessing the 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 predominantly restricted to static visualizations (Wang et al., 2024; Xia et al., 2025; Masry et al., 2022). On the other hand, mainstream 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). Data videos present a comprehensive challenge for MLLMs, as they demand a synthesis of both foundational capabilities. Comprehending it requires comprehension of both micro-level spatio-temporal details (e.g., identifying a surge in an an-

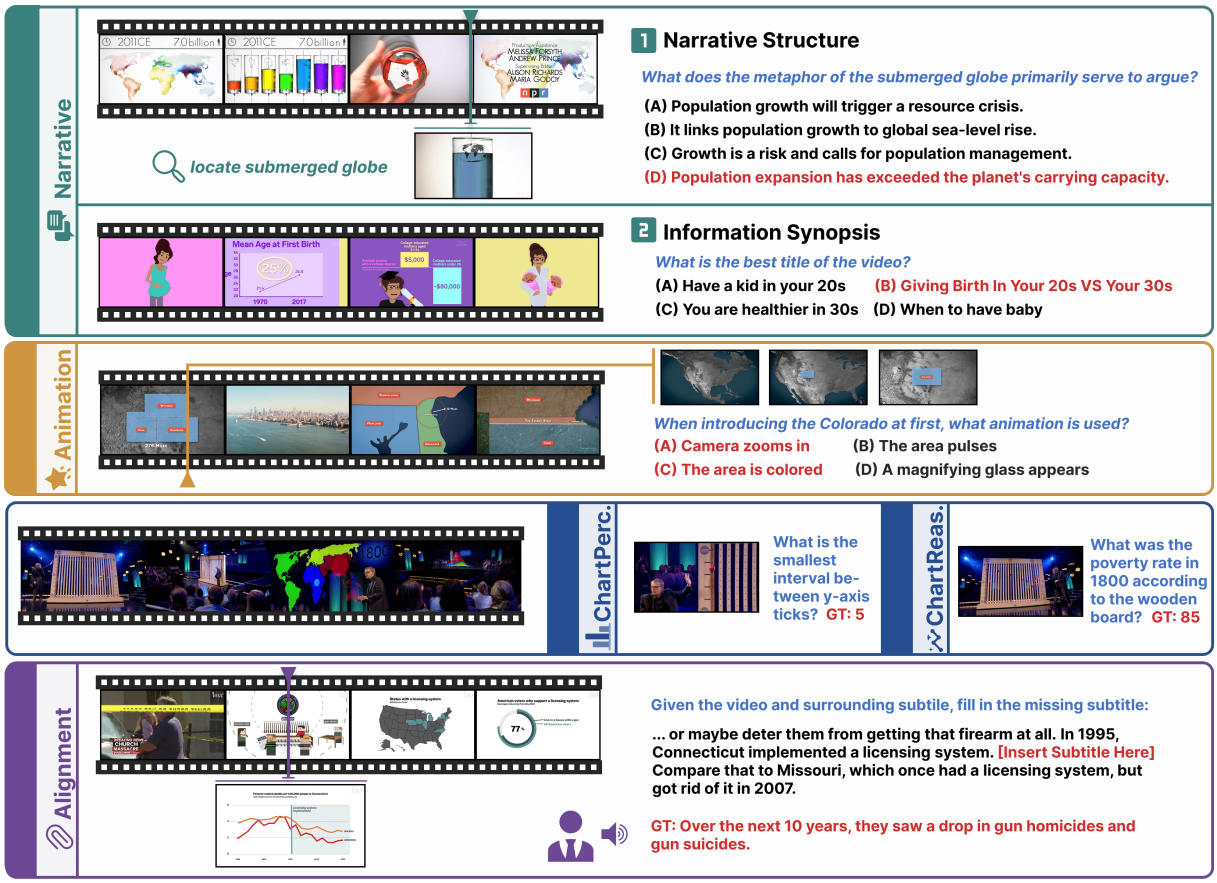

Figure 1: Overview of the five evaluation dimensions in DVBench, illustrated with representative examples. Our taxonomy systematically assesses models across: (1) **Narrative**, evaluating high-level comprehension of video themes and metaphors; (2) **Animation**, focusing on the perception of dynamic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

imated line chart) and macro-level narrative intent (e.g., summarizing the core data-driven message).

To this end, we introduce **DVBench**, the first comprehensive benchmark specifically dedicated to evaluating the comprehension of data videos across 5 critical dimensions: **Narrative**, **Animation**, **Chart Perception**, **Chart Reasoning**, and **Alignment**. DVBench comprises 300 high-quality data videos, accompanied by 1,000 QA pairs constructed through a rigorous semi-automatic pipeline and human verification.

With this benchmark, we conduct extensive evaluations of 9 prominent open-source and proprietary MLLMs. The experimental results show Gemini-3.1-Pro achieves the highest overall performance, and Kimi-k2.5 emerges as the leading open-source model. Furthermore, we observe two surprising phenomena among the open-source models: first, the capacity to understand data videos does not scale directly with the number of model parameters; second, proficiency in the Narrative dimension does not strictly correlate with performance in the fine-grained visual dimensions.

To conduct a fine-grained analysis, we systematically annotate the questions using the established design space from data video research. By evaluating the dimensions across these specific tags and varying video lengths, our deep dive reveals models' weaknesses in perceiving non-linear animation pacing and precise counting in dynamic charts. It also exposes a clear grounding gap between grasping general contexts and extracting precise insights, alongside a significant performance drop in longer videos.

Our contributions are as follows: (1) we introduce a new task, data video understanding, with 5 dimensions to systematically evaluate MLLMs;

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., 2025)	✗	✗	👤	✓
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 🤖 denote human-annotated and auto-generated data construction pipelines, respectively. Multi-dim. Eval indicates whether the benchmark assesses multi-dimensional capabilities. As shown, DVBench stands out as a comprehensive framework specifically designed to evaluate the synergy of dynamic charts and narratives.

(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 an extensive evaluation with a fine-grained analysis within the design space of data videos.

2 Related Works

We compare DVBench and other multimodal understanding benchmarks in Table 1. The following subsections detail the evolution and limitations of current evaluations across chart and video domains, highlighting the necessity of our DVBench.

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). 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 capabilities 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 specifically designed to evaluate causal relationships and temporal logic within long-form content (Wang et al., 2025b; Yu et al., 2025). Regarding video caption, recent frameworks have moved toward 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 (Shah et al., 2025; Ataallah et al., 2025) to expert-level multi-disciplinary knowledge and scientific processes (Hu et al., 2025; 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; Shen et al., 2025). Extensive research has investigated the foundational elements of this medium, characterizing the design space of visual-animation interplay and scene-semantic alignment (Shi et al., 2021; Cheng et al., 2022; Gao et al., 2025). Complementing these compositional studies, researchers have further investigated narrative strategies to better organize and convey information in data videos (Yang et al., 2021; Xu et al., 2022, 2023; Wei et al., 2024). 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 5 primary dimensions: **Narrative**, **Animation**, **Chart**

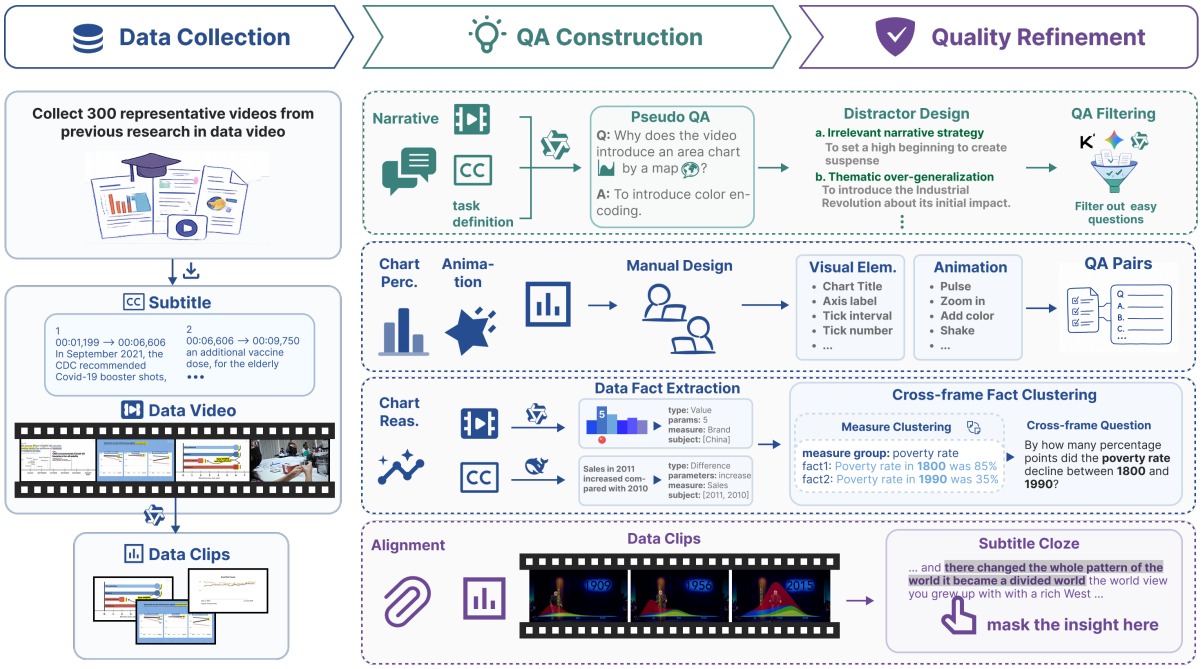

Figure 2: **Benchmark construction pipeline of DVBench.** The process is divided into three main stages: (1) data collection, (2) QA construction, and (3) quality refinement. We collect videos from previous research and download the videos and subtitles. Then we utilize MLLM to identify data clips. During QA construction, we employ dimension-specific strategies: **Narrative** questions are generated via MLLMs and undergo rigorous distractor design and filtering; **Chart Perception** (short for Chart Perc.) and **Animation** questions are manually curated to target fundamental visual elements and dynamic transitions; **Chart Reasoning** (short for Chart Reas.) utilizes an automated pipeline to extract, formalize, and cluster multi-modal data insights for complex cross-frame queries; and **Alignment** tasks are formulated as subtitle cloze tests by masking key insights within identified data clips.

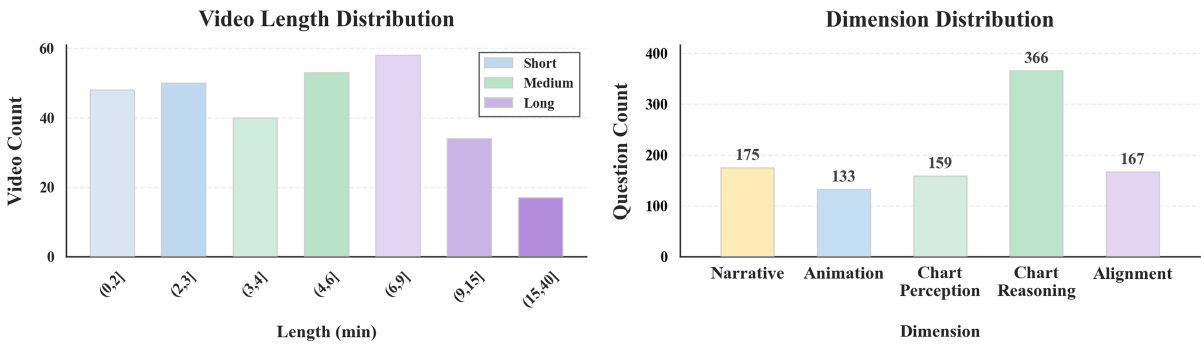

Figure 3: **Statistics analysis of DVBench**. Left subplot shows the video length distribution. Right one shows the task type distribution.

Perception, **Chart Reasoning**, and **Alignment** as shown in Fig 1.

To enable a more granular evaluation, we annotate the QA pairs within the **Animation**, **Chart Reasoning**, and **Alignment** dimensions with fine-grained tags derived from prior research. The specific definitions of these tags are detailed in the respective dimension descriptions below. Ultimately, this annotation is instrumental in facilitating the in-depth analysis of model capabilities presented later in Sec. 4.2.2.

Narrative. This dimension assesses the ability to comprehend the high-level storytelling logic and communicative intent. It is further divided into two tasks: *(i) 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.,

2021), and identifying the logical relationships between video clips (e.g., Why does the video introduce the *income distribution chart* by a *world map*?). *(ii) 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. To enable a fine-grained evaluation, each QA pair in this dimension is annotated with a specific tag denoting its target editorial layer (Shi et al., 2021). These layers encompass: (1) *visualization elements* (e.g., transforming visual marks, axes, or legends), (2) *added elements* (e.g., manipulating annotations and embellishments), (3) *camera operations* (e.g., zooming and panning to alter viewing perspectives), and (4) *timeline manipulations* (e.g., controlling the pacing and rhythm of the storytelling).

Chart Perception evaluates the capacity for fine-grained graphical perception and fundamental information extraction directly from visual elements (e.g., identifying titles, axis labels, legend entries, and tick marks).

Chart Reasoning focuses on identifying the underlying data insights conveyed by the charts within the video. We systematically categorize these insights into 11 fine-grained types according to (Wang et al., 2019), ranging from localized queries (e.g., retrieving *values*, identifying *extremes*, and conditional *categorization*) and comparative statistics (e.g., measuring *proportions*, comparing *differences*, *ranking*, and *aggregation*) to broader structural analyses (e.g., tracking temporal *trends*, analyzing *distributions*, and finding *associations*).

Alignment evaluates the cross-modal synchronization between dynamic charts and their corresponding narration. To assess this, we employ a subtitle cloze task where models must infer missing narrations by integrating visual cues with the surrounding subtitle context. Furthermore, to enable a fine-grained analysis, we annotate each QA pair with one of two semantic tags based on the role of the masked subtitle: *Data Insight* and *Data Context*. *Data Insights* represent analytical facts (e.g., temporal trends or exact values) that are predominantly anchored in explicit on-screen visual elements. While *Data Context* refers to the background information, which is inherently more abstract and requires models to holistically summarize the narrative setting (Cheng et al., 2022).

3.2 Benchmark Construction

To ensure the high quality of our dataset, we collect 300 representative data videos from research dedicated to data videos (Yang et al., 2021; Shi et al., 2021; Cheng et al., 2022; Xu et al., 2022, 2023; Gao et al., 2025; Gunturu et al., 2025), which feature rich dynamic charts and meticulously orchestrated narratives. We deploy different methods to generate QA pairs for each dimension as shown in Fig 2.

For the **Narrative** dimension, we input the video and subtitles into Qwen3-VL-Plus (Bai et al., 2025) and guide it to generate pseudo QA pairs based on the taxonomy detailed in Sec . 3.1. Specifically, the generation is strictly aligned with the two narrative tasks: *Narrative Structure* and *Information Synopsis*. To make these questions more challenging, we employ various methods to refine the distractors. For instance, we map visual cues to irrelevant narrative strategies or replace the primary insight with an over-generalized theme. Finally, we filter out trivial questions that Kimi-k2.5, Gemini-3.1-Pro, and Qwen3.5-397B-A17B can all answer correctly.

We first employ Qwen3-VL-Plus (Bai et al., 2025) to locate data clips (i.e., video segments where charts convey context or insights) (Amini et al., 2016; Cheng et al., 2022). Based on these identified clips, the QA generation is divided according to our specific dimensions:

For the **Animation** and **Chart Perception** dimensions, we manually design the questions targeting the dynamic transitions and fundamental visual elements within the data clips, respectively.

For the **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, 2025). After combining and formalizing these multi-modal insights into a structured schema (Wang et al., 2019), 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 insights by their *subjects* and *measures*, utilizing these clusters to construct complex temporal tasks, such as tracking the change of a specific measure across differ-

Model	Avg	Narr.	Anim.	Chart		Align.		
				Perc.	Reas.	BL-2	MET	$F1_{BERT}$
Proprietary Models								
Gemini-3.1-Pro (Google DeepMind, 2026a)	**77.91**	**76.57**	69.92	**86.16**	**77.87**	**26.77**	38.20	83.02
Claude-4.6-Sonnet (Anthropic, 2026)	72.51	75.43	68.42	74.84	71.58	25.81	38.26	**83.28**
GPT-5.4 (OpenAI, 2026)	65.31	69.14	63.91	68.55	62.57	19.88	31.68	77.67
Open-source Models								
Kimi-k2.5 (Team et al., 2026)	75.87	72.57	**72.18**	83.65	75.41	25.77	**41.22**	83.09
Qwen3.5-27B (QwenTeam, 2026)	62.30	74.29	59.40	64.15	56.83	17.81	27.57	79.21
Qwen3.5-397B-A17B (QwenTeam, 2026)	61.82	73.14	62.41	64.78	54.92	16.94	26.18	76.77
Qwen3.5-122B-A10B (QwenTeam, 2026)	59.30	70.86	58.65	62.26	52.73	12.96	21.74	77.61
Gemma-4-31B-It (Google DeepMind, 2026b)	58.82	64.00	57.89	60.38	56.01	16.53	25.75	78.18
Qwen3.5-9B (QwenTeam, 2026)	52.82	55.43	54.89	61.01	47.27	8.20	16.07	73.98

Table 2: Performance of evaluated MLLMs across the 5 dimensions of DVBench. Dimensions are abbreviated: Narr. (Narrative), Anim. (Animation), Perc. (Perception), Reas. (Reasoning), and Align. (Alignment).

ent timestamps.

For the **Alignment** dimension, we extract the previously identified data clips and mask the narration of data insight or context to formalize the subtitle cloze tasks.

The detailed prompt templates utilized for question generation and annotation details are respectively provided in Appendices B.2 and B.3.

3.3 Dataset Statistics

Video Length Distribution. The left subplot of Fig 3 illustrates the length distribution of the collected data videos. Specifically, the videos span a wide temporal range from under 30 seconds to up to 37 minutes. We categorize the videos into three groups: Short ($\leq$ 3 mins), Medium (3–6 mins), and Long ($>$ 6 mins). The dataset exhibits a relatively balanced distribution across these three categories. The video count peaks in the medium-length interval of (4,6] minutes. For videos exceeding 6 minutes, the quantity exhibits a gradual and continuous decline as the length increases.

Task Distribution. The right subplot of Fig. 3 details the distribution of the 1,000 QA pairs across our five evaluation dimensions: Narrative (175), Animation (133), Chart Perception (159), Chart Reasoning (366), and Alignment (167). We allocate the largest proportion to Chart Reasoning for two critical reasons. First, viewers predominantly focus on the underlying data insights rather than superficial graphical elements. Second, prior studies (Wang et al., 2024) indicate that while models excel at basic perception, they still exhibit vulnerabilities in reasoning. Thus, a denser reasoning subset is essential to rigorously benchmark future progress. Overall, this comprehensive distribution highlights that DVBench encompasses a wide spectrum of cognitive complexities, enabling a robust evaluation of dynamic multimodal reasoning. Additional plots of chart type and QA length statistics are in the Appendix C.

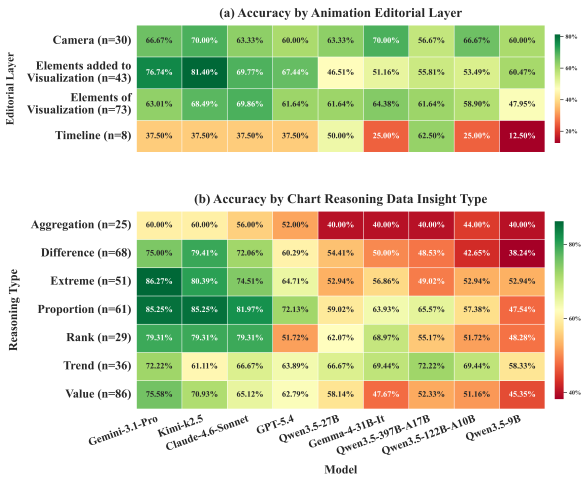

Figure 4: **Fine-grained performance across the Animation and Chart Reasoning dimensions.** (a) Accuracy distribution within the Animation task, evaluated across four editorial layers. (b) Accuracy evaluation across 10 reasoning question types. Note that the x-axis, representing the evaluated MLLMs,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

4 Experiment

4.1 Settings

We test a total of 9 MLLMs, including both open-source (Qwen3.5 series (QwenTeam, 2026),

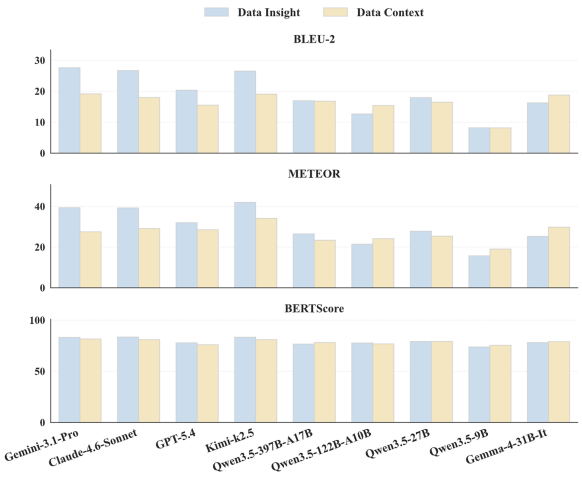

Figure 5: **Evaluation results of the Alignment dimension.** We report the BERTScore, BLEU-2, and METEOR scores of 9 MLLMs on two distinct subtitle cloze categories: *Data Insight* (blue bars) and *Data Context* (yellow bars). Note that the x-axis, representing the evaluated MLLMs, is shared across all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

Gemma-4 (Google DeepMind, 2026b), Kimi-k2.5 (Team et al., 2026)) and proprietary MLLMs (Gemini-3.1-Pro (Google DeepMind, 2026a), Claude-4.6-Sonnet (Anthropic, 2026), and GPT-5.4 (OpenAI, 2026)). Model configurations and evaluation prompts are detailed in Appendices D.1 and D.5. Our evaluation metrics include task-specific accuracy and NLP metrics, BLEU-2(Papineni et al., 2002), METEOR(Banerjee and Lavie, 2005), and BERTScore(Zhang et al., 2019), as detailed in Appendix D.4.

4.2 Result

4.2.1 Main Result

Table 2 presents a comprehensive performance evaluation of 9 MLLMs across 3 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 across the benchmark. 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.

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, outperforms its larger Mixture-of-Experts counterparts, Qwen3.5-397B-A17B (61.82%) and Qwen3.5-122B-A10B (59.30%), in overall average performance.

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 an average accuracy of 74.29% in the Narrative dimension, closely trailing Claude-4.6-Sonnet (75.43%). However, a significant performance gap emerges within the Visual dimension. 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%).

4.2.2 Analysis

In this section, we delve deeper into the evaluation by exploiting the QA tags detailed in Sec . 3.1. We systematically analyze the performance distribution across targeted dimensions: the four editorial layers within Animation, the data insight types in Chart Reasoning, and the subtitle semantic tags in Alignment. Additionally, we explore how varying video lengths affect model stability across different tasks.

Animation: Failure to Capture Non-linear Animation Pacing. We further analyze the performance in the Animation dimension from the perspective of the four editorial layers (i.e., *what* is being animated) mentioned in Sec. 3.1. Animations of Timeline layer are pacing effects, such as acceleration and deceleration. While other Animation techniques of the first three layers are primarily physical transformations, such as the alteration of visual channels, the integration of annotations, or the manipulation of shot angles. The illustration of specific examples for these four layers is presented in the Appendix E.1.

As shown in the Fig 4, accuracy on the Timeline layer is notably low, with the accuracy of nearly all models below 37.50%. Unlike standard physical transformations that models can deduce by comparing the initial and final frames, timeline tasks require the perception of the non-linear rate of change between adjacent frames. Because current multimodal models rely heavily on sparse frame sampling. This severe loss of continuous transition information renders the models incapable of accurately perceiving the speed of motion.

Chart Reasoning: Dynamic Charts Induce

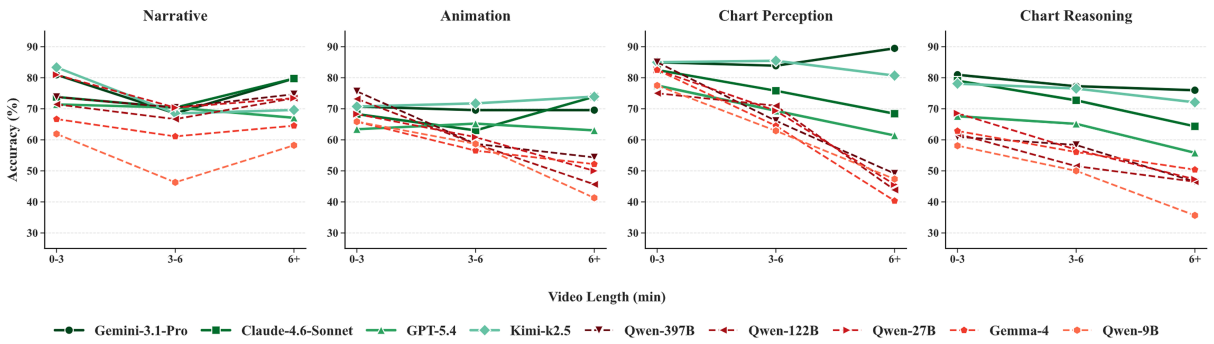

Figure 6: **Performance across varying video lengths.** The line charts illustrate the accuracy of 9 MLLMs on 4 dimensions as the video length expands from short ($\leq$ 3 mins) to long ($>$ 6 mins). Green lines represent strong models, while red lines represent weak ones.

Severe Counting Hallucinations. In the Chart Reasoning dimension, the Aggregation question evaluates the capability of models to perform counting and information synthesis. Specifically, this task requires models to conduct multi-frame reasoning to track objects over time and to identify the exact number of elements satisfying specific conditions within individual frames. However, results demonstrate that models struggle with this task, yielding the lowest accuracy across all reasoning categories. The performance of 4 open-source models drops to 40.00%. Detailed cases can be found in the Appendix E.2.

Alignment: The Divergence in Insight and Context Grounding. Overall, proprietary models still outperform open-source models across the alignment metrics, with Kimi-k2.5 remaining the strongest open-source baseline. After splitting Alignment into *Data Context* and *Data Insight*, we find that all evaluated models preserve high semantic similarity on BERTScore, indicating that coarse-grained meaning is broadly captured. However, exact-match metrics such as BLEU-2 and METEOR expose a sharper gap at the fine-grained level. Proprietary models and Kimi-k2.5 generally perform better on *Data Insight* than on *Data Context*, suggesting stronger grounding of visually salient facts into text. In contrast, open-source models show a larger drop on *Data Insight*, and the error analysis shows that this gap is dominated by visual-text grounding error, while phrasing differences account for a smaller share. Detailed analysis and examples are provided in Appendix E.3.

Video Length: Strong Models Are Not Susceptible to Long Context. We analyze model performance across short, medium, and long videos. As illustrated in Fig 6, leading proprietary models (alongside 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 temporal context expands. Furthermore, we observe that the impact of video length is task-dependent. For high-level *Narrative* tasks, which require holistic semantic abstraction, the performance of nearly all models remains relatively stable, largely 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%. This contrast highlights that the core bottleneck for current open-source MLLMs is their inability to sustain fine-grained spatial-temporal visual memory over extended contexts.

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: Narrative, Animation, Chart Perception, Chart Reasoning, and Alignment. We curate a dataset of 300 videos and 1,000 QA pairs via a semi-automatic pipeline and a subsequent manual verification. Evaluation of 9 MLLMs demonstrates the superiority of proprietary models and exposes unexpected results in parameter scaling and cross-dimensional proficiency among open-source models. Further analyses reveal bottlenecks in perceiving non-linear animation speed, the dynamic chart counting problem, fine-grained visual-textual grounding, and long-context processing for fine-grained visual tasks.

Limitations

While our semi-automatic construction ensures high data fidelity, it inherently bounds the current scale of DVBench. Also, in future work, we can integrate multilingual videos and 3D visualizations or AR infographics. Additionally, in the *Alignment* dimension, future work will integrate LLM-as-a-judge protocols to enable a more comprehensive evaluation. Finally, although our analysis reveals that current MLLMs struggle with fine-grained timeline manipulations, the evolution of native end-to-end video models capable of processing denser audio-visual streams is expected to improve baseline performances on these temporal tasks. To accommodate this progress, DVBench features an extensible taxonomy designed to seamlessly incorporate increasingly complex reasoning tasks and continuously benchmark future architectural advancements.

References

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*, pages 1459–1468.

Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Andres Monroy-Hernandez, and Pourang Irani. 2016. Authoring data-driven videos with dataclips. *IEEE transactions on visualization and computer graphics*, 23(1):501–510.

Anthropic. 2026. Introducing claude 4.6 sonnet. https://www.anthropic.com/news/claude-sonnet-4-6. Accessed: 2026-05-12.

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 19496–19523.

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.

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. In *Computer Graphics Forum*, volume 41, pages 527–539. Wiley Online Library.

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

DeepSeek-AI. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *arXiv 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, and 1 others. 2025. 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*, pages 24108–24118.

Chaoyou Fu, Yi-Fan Zhang, Shukang Yin, Bo Li, Xinyu Fang, Sirui Zhao, Haodong Duan, Xing Sun, Ziwei Liu, Liang Wang, and 1 others. 2024. Mme-survey: A comprehensive survey on evaluation of multimodal llms. *arXiv preprint arXiv:2411.15296*.

Lin Gao, Leixian Shen, Yuheng Zhao, Jiexiang Lan, Huamin Qu, and Siming Chen. 2025. Sceneloom: Communicating data with scene context. *IEEE Transactions on Visualization and Computer Graphics*.

Google DeepMind. 2026a. 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/.

Google DeepMind. 2026b. Gemma 4 model card. https://ai.google.dev/gemma/docs/core/model_card_4. Apache 2.0 License.

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*, pages 1–20.

Hyeonjeong Ha, Jinjin Ge, Bo Feng, Kaixin Ma, and Gargi Chakraborty. 2026. Narrativetrack: Evaluating video language models beyond the frame. *arXiv preprint arXiv:2601.01095*.

Jeffrey Heer and George G. 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 Computer Vision and Pattern Recognition Conference*, pages 8450–8460.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

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*, pages 5648–5656.

Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, and 1 others. 2024a. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Shihao Li, Yuanxing Zhang, Jiangtao Wu, Zhide Lei, Yiwen He, Runzhe Wen, Chenxi Liao, Chengkang Jiang, An Ping, Shuo Gao, and 1 others. 2025. Ifvidcap: Can video caption models follow instructions? *arXiv 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*, pages 13171–13182.

Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Jixuan Chen, Boqiang Zhang, Nianzu Yang, Pandeng Li, Yinglu Li, Zuan Gao, and 1 others. 2025. What is a good caption? a comprehensive visual caption benchmark for evaluating both correctness and thoroughness. *arXiv preprint arXiv:2502.14914*.

Ahmed Masry, Xuan Long Do, 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.

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*, pages 1527–1536.

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.

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.

Nisarg A Shah, Amir Ziai, Chaitanya Ekanadham, and Vishal M Patel. 2025. Cin\'{e} aste: A fine-grained contextual movie question answering benchmark. *arXiv preprint arXiv:2509.14227*.

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*, pages 1–21.

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*, pages 1–13.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

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.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, SH Cai, Yuan Cao, Y Charles, HS Che, Cheng Chen, Guanduo Chen, and 1 others. 2026. Kimi k2. 5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*.

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*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025b. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. 2019. 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, and 1 others. 2024. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697.

Zheng Wei, Huamin Qu, and Xian Xu. 2024. Telling data stories with the hero's journey: Design guidance for creating data videos. *IEEE Transactions on Visualization and Computer Graphics*.

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.

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

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*, pages 1–16.

Xian Xu, Leni Yang, David Yip, Mingming Fan, Zheng Wei, and Huamin Qu. 2022. From wowto 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*, pages 1–20.

Leni Yang, Xian Xu, XingYu Lan, Ziyan Liu, Shunan Guo, Yang Shi, Huamin Qu, and Nan Cao. 2021. 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, and 1 others. 2025. Vrbench: A benchmark for multi-step reasoning in long narrative videos. *arXiv preprint arXiv:2506.10857*.

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 Computer Vision and Pattern Recognition Conference*, pages 28995–29004.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

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

Appendix

A Potential Risks

While DVBench is designed strictly for the academic evaluation of MLLMs, we acknowledge the potential downstream risks associated with data video understanding. Since all videos in DVBench are curated from previously published academic works, the overall risk is relatively limited compared to datasets collected from unrestricted online media sources. Moreover, the dataset construction process includes manual review and quality control to reduce problematic or misleading content.

Nevertheless, data videos often present real-world statistics and analytical narratives involving socio-economic, geopolitical, environmental, and public health topics. As a result, several potential risks still remain.

First, there is a risk of bias inheritance. Although the videos are sourced from established prior works, models trained or evaluated on such content may still absorb and amplify implicit assumptions, framing choices, or subjective interpretations embedded in the original narratives.

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 abilities when interpreting charts and animated visualizations. If these limitations are not properly mitigated, deploying such models in automated journalism, educational analytics, or data storytelling systems could result in 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 to facilitate the evaluation of multimodal large language models on data video understanding tasks. All benchmark annotations, extracted structured information, QA pairs, and accompanying metadata are released under the **CC BY-NC-SA 4.0** license.

The benchmark is constructed from 300 publicly accessible data videos collected from prior research works (Yang et al., 2021; Shi et al., 2021; Cheng et al., 2022; Xu et al., 2022, 2023; Gao et al., 2025; Gunturu et al., 2025). These videos were originally published on public online platforms such as YouTube and Vimeo. The copyrights and ownership of all original video, audio, and visual materials remain with their respective creators and copyright holders.

DVBench is intended solely for research and evaluation purposes. We do not claim ownership of any third-party media content included in the benchmark. The dataset is provided only to support standardized academic evaluation of multimodal systems, and its use should comply with applicable copyright regulations and platform terms of service. If any copyright holder believes that specific content should be removed, we will promptly address reasonable removal requests.

During dataset construction and manual quality control, all videos and QA pairs were carefully reviewed to minimize the presence of personally identifiable information (PII), offensive language, or harmful 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

The data annotation, fact-checking, and QA verification tasks were conducted by the authors, who hold at least Master's degrees with specialized academic backgrounds in data visualization, human-computer interaction, and machine learning.

To ensure the high fidelity and accuracy of the QA pairs, we developed a comprehensive web-based annotation interface to facilitate our human-

in-the-loop review process. As illustrated in Fig. 7, the platform is meticulously designed for efficient cross-modal grounding and is divided into distinct functional zones. The left panel features a synchronized video player at the top, while the bottom section displays the complex cross-frame reasoning questions (aggregated by measure and subject) alongside their detailed reasoning processes. The right panel is structured into three vertical sections: the subtitles at the top, the extracted data insights in the middle, and the identified data clips with auto-generated pseudo QA pairs for chart perception at the bottom.

Crucially, the interface supports interactive temporal navigation: by simply clicking on any extracted data fact or specific question across the panels, the video automatically jumps to the corresponding timestamp. This click-to-seek synchronization significantly streamlines the review process, enabling expert annotators to quickly and accurately verify the multimodal context, refine the generated content, and assign fine-grained capability tags to support our comprehensive evaluation taxonomy.

C More Dataset Statistics

DVBench encompasses a wide spectrum of visualizations and tasks to comprehensively evaluate the cognitive capabilities of MLLMs. As illustrated in Fig. 8, we annotate the QA pairs across the Animation, Chart Perception, and Chart Reasoning dimensions with their corresponding chart types. Our dataset covers diverse chart types, ranging from standard bar and line charts to complex combined visualizations. This diversity reflects real-world data video distributions.

Table 3 provides a detailed statistics of question and option lengths across four core evaluation dimensions. Table 4 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 3: Question and option length statistics across DVBench dimensions. 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 4: 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 5 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 `top-p` to 1 wherever applicable. For locally deployed open-source models running via vLLM, we explicitly enforce `do_sample=False`.

For visual input, 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 6 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 comprehensive evaluation of these local models required approximately 60 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,

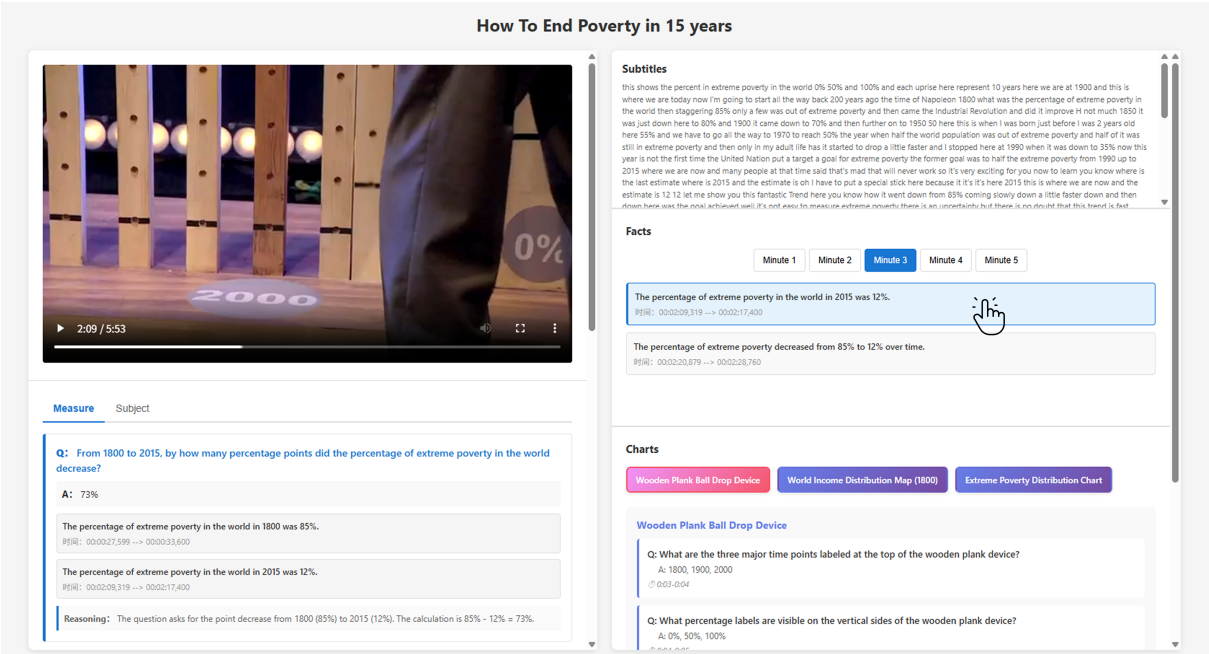

Figure 7: Annotation interface of DVBench. The layout is optimized for efficient verification. The left panel contains synchronized video playback (top) and cross-frame reasoning questions with their reasoning processes (bottom). The right panel displays the video subtitles (top), extracted data insights (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.

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 5: Summary of model versions, input frame configurations, decoding hyperparameters, and inference backends evaluated in DVBench. The symbol "–" indicates that the parameter is not exposed or cannot be configured.

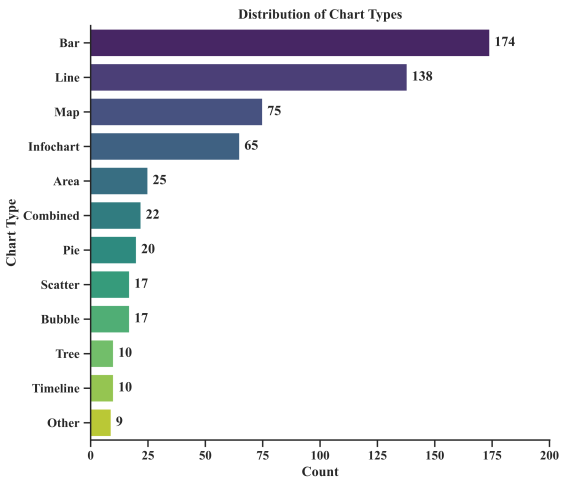

Figure 8: Distribution of chart types.

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

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 the outputs from both lexical and semantic perspectives using **BLEU-2** (Papineni et al., 2002), **METEOR** (Banerjee and Lavie, 2005), and **BERTScore F1** (Zhang et al., 2019).

To ensure full 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 the bi-gram overlap between the generated text and the reference answer. To prevent penalizing short responses with zero counts, we configure the BLEU metric using NLTK's smoothing function `method1` (`SmoothingFunction().method1`). To provide a more nuanced lexical evaluation, **METEOR** utilizes NLTK's WordNet integrations (`wordnet` and `omw-1.4`) to incorporate stemming and synonym matching algorithms.

Beyond surface-level lexical matching, **BERTScore F1** measures the semantic equivalence by calculating the cosine similarity between the embeddings of the generated and reference sentences in the BERT encoding space. We implement BERTScore using the official `bert_score` package (v0.3.13) configured with the `roberta-large` model as the default contextual embedding backbone. This provides an accurate assessment of sentence-level semantic alignment.

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 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.

E.1 Animation Evaluation Cases

To provide deeper insights into the models' capabilities in perceiving dynamic visual transitions, we present qualitative examples across the four editorial layers of animation in Fig. 9. 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 pacingsuch as identifying when a chart speeds up, slows down, or pausesprove exceptionally challenging.

E.2 Chart Reasoning Evaluation Cases

Fig. 10 illustrates qualitative examples from the Aggregation questions in Chart Reasoning dimension.

E.3 Alignment Analysis

To investigate the specific reasons why open-source models significantly underperform proprietary models on EM metrics, 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 BLEU-2 and METEOR scores are both below the model-wise 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, numbers, or qualifiers. **(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. 11, 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 7 presents representative cases, illustrating typical grounding errors, compression, paraphrase, and additive-detail cases.

F 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.

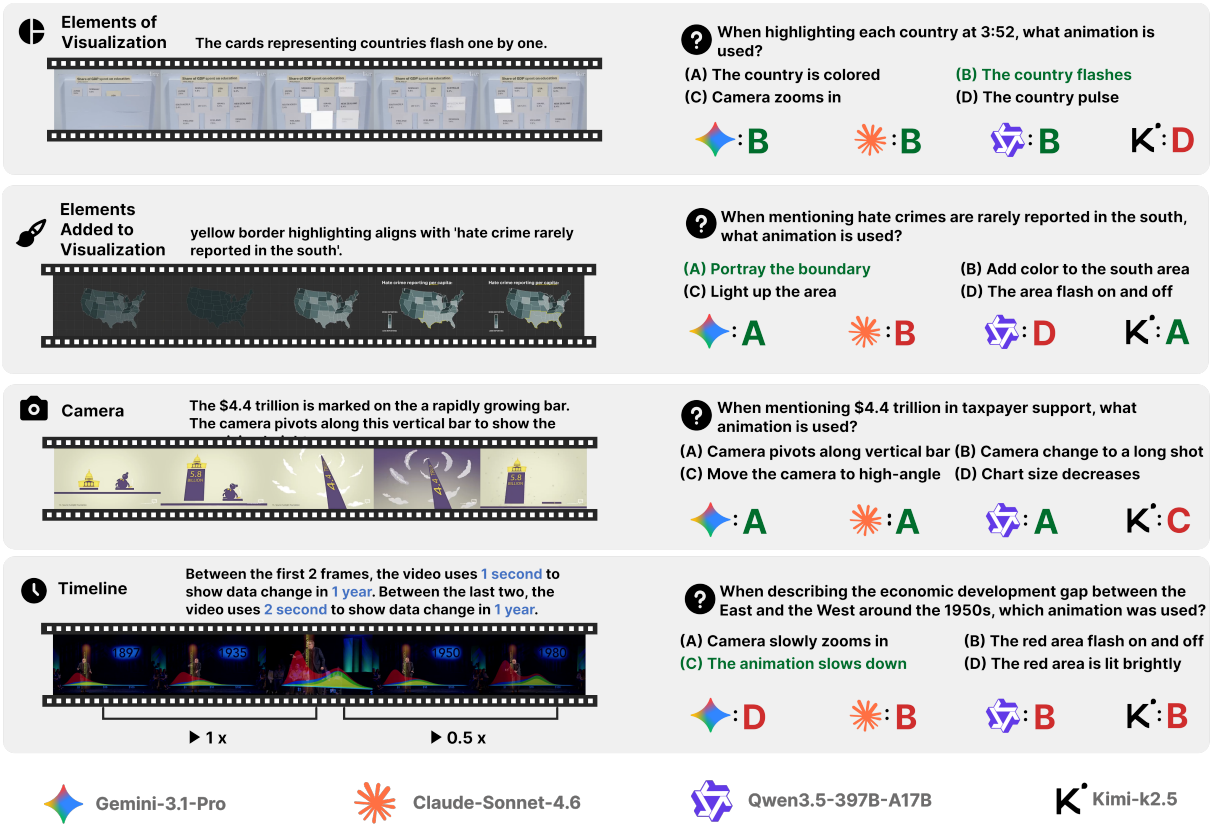

Figure 9: **Cases of Animation QA pairs.** We present questions of four editorial layers here. As we can see, questions of animation on Timeline layer (e.g., slow down, speed up, and pause in visualization animation) are most challenging for models.

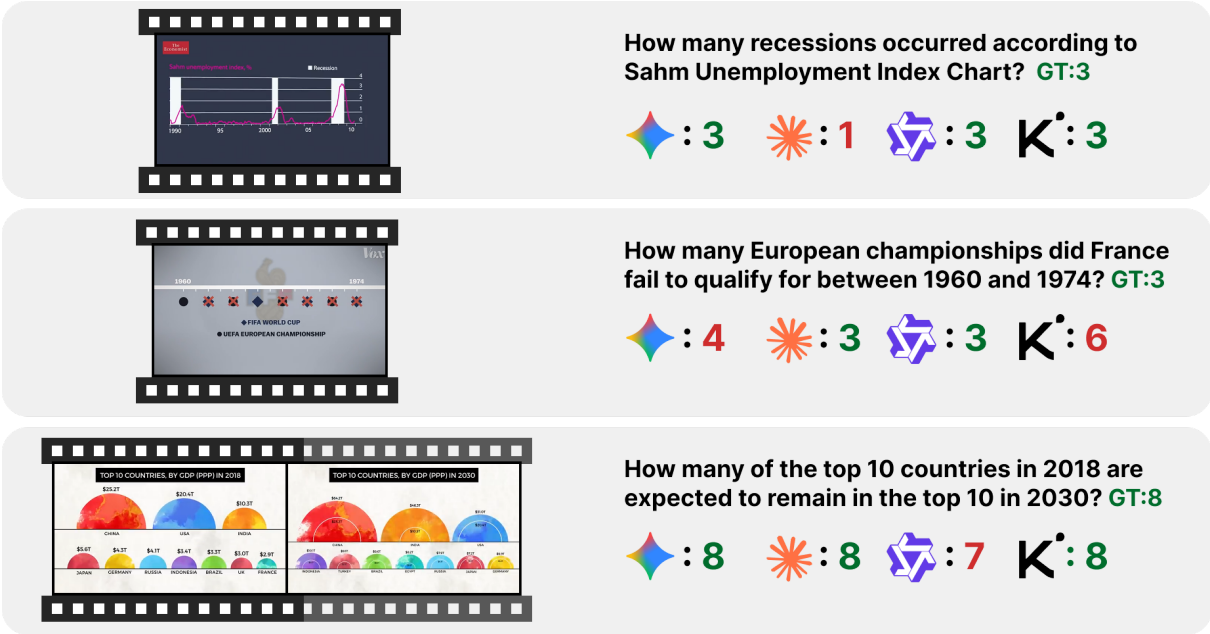

Figure 10: **Cases of Chart Reasoning QA pairs.** We illustrate three Aggregation questions. While the first two questions can be answered using a single keyframe, the third requires integrating information across two discrete frames. Successfully answering these queries requires models to accurately ground target chart elements across single or multiple frames before performing specific numerical calculations.

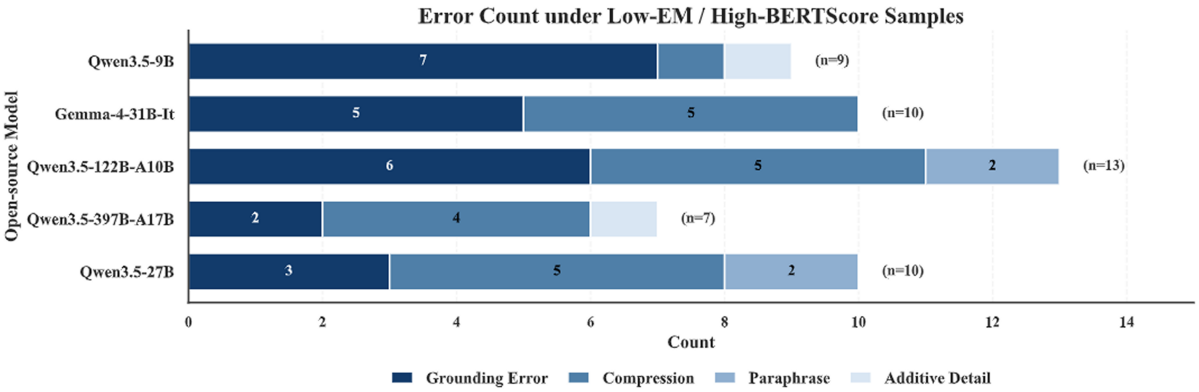

Figure 11: Counts of error types for the low-EM / high-BERTScore answers across open-source models. Each horizontal stacked bar shows the number of examples in the slice (defined per model as BLEU-2 model median, METEOR model median, and BERTScore model median), grouped by four error types. Per-row totals are shown to the right as (n=).

Model	Gold answer	Model response	Error Type
Gemma-4-31B-It	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	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	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	the shape of the world changed	Compression
Qwen3.5-122B-A10B	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	in the 1970s the population grew here and the incomes increased here	Grounding error
Qwen3.5-27B	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	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	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	and the shape of the world change again	Compression
Gemma-4-31B-It	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.	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
Gemma-4-31B-It	here come over countries Europe Brown Asia Red Middle East Queens Africa south of sahara blue and the americas yellow	Each bubble represents a country.	Compression
Gemma-4-31B-It	nuclear energy ranks last in death per energy unit produced	nuclear energy still saves more lives than it takes.	Grounding error

Table 7: Representative qualitative cases from the low-EM / high-BERTScore slice (open-source models).