DATAVISTA: DIAGNOSING MULTIMODAL LLMS ON DATA VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Data video is a media form that integrates data visualization with video narrative, widely adopted in news reporting and business analysis. Compared with general video understanding, data video understanding places greater emphasis on accurately reading data from animated charts, integrating evidence across charts and time, and understanding how narrative organization and visual design communicate information. Yet existing benchmarks target either general videos or static charts, and data video understanding has not been systematically evaluated. We present DATAVISTA, the first benchmark for data video understanding, containing 961 real-world data videos and 6,775 evaluation questions organized under a three-level progressive capability framework (data perception, temporal reasoning, narrative understanding) with 10 fine-grained question types across five topic domains. Systematic evaluation of 19 mainstream MLLMs shows that the best-performing model, Gemini-3.1-Pro, achieves 70.0% overall accuracy, far below the human expert score of 91.7%, with models performing worst on Causal Reasoning and Narrative Structure. Increasing frame counts and adding subtitles mainly benefit data perception and temporal reasoning, with limited gains in narrative understanding. Further analysis of model responses identifies typical failure modes in chart reading, evidence judgment, and instruction understanding.

1 INTRODUCTION

Data videos are a media form that integrates data visualization with video narrative, conveying data-driven stories over time through animation, narration, and scene transitions (Amini et al., 2015; Shi et al., 2021). In recent years, organizations such as Bloomberg Graphics, NYT The Upshot, and Vox have produced a growing volume of such content, establishing data videos as an important medium for data journalism, business reporting, and science communication (Yang et al., 2022; Cheng et al., 2022). Data videos provide a challenging setting for evaluating the combined capabilities of MLLMs in dynamic visual perception, cross-temporal reasoning, and narrative understanding.

However, understanding data videos goes well beyond what general video perception covers. As shown in Figure 1, charts are presented within continuous videos rather than provided as isolated images (Thompson et al., 2020); key information is often distributed across multiple time points; and visual presentation also organizes narratives and communicates information. General video understanding benchmarks (Video-MME (Fu et al., 2025), MVBench (Li et al., 2024)) focus on perception and reasoning over general video content, whereas static visualization benchmarks (ChartQA (Masry et al., 2022), VisJudge-Bench (Xie et al., 2026b)) primarily address the understanding and evaluation of static charts. Jointly assessing evidence localization, cross-chart integration, and narrative understanding within continuous videos still lacks systematic benchmark support.

To fill this gap, we present DATAVISTA, the first benchmark for data video understanding. Drawing on Bloom's cognitive taxonomy (Bloom et al., 1956), visualization literacy assessment frameworks (Boy et al., 2014; Lee et al., 2017), and design spaces for data storytelling (Segel & Heer, 2010), we systematically analyze the core abilities required for data video understanding and distill three progressive capability dimensions: data perception (accurately reading values and identifying visual elements from video frames), temporal reasoning (integrating information across time points and across charts), and narrative understanding (grasping the creator's visual design intent and argument

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

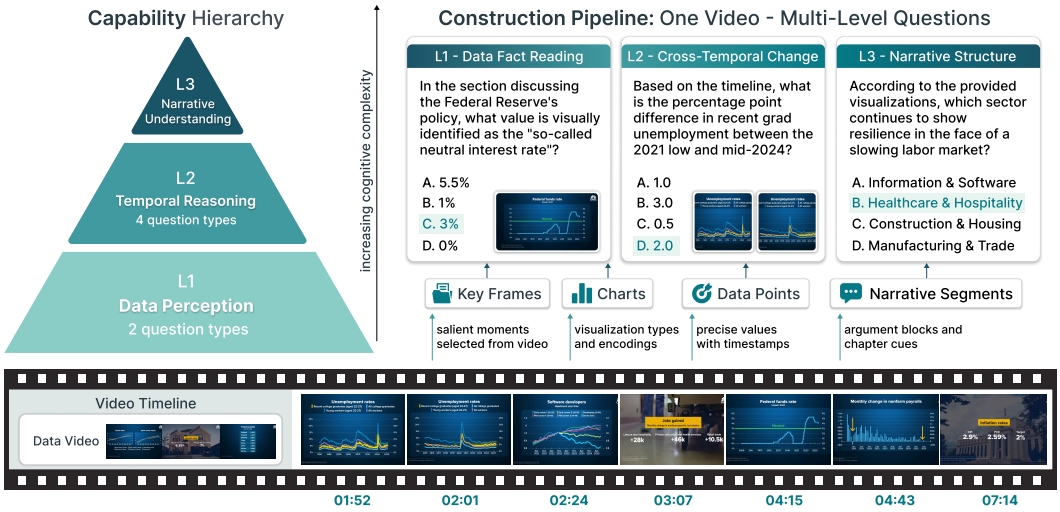


Figure 1: Three-level capability hierarchy and construction pipeline of DATAVISTA.

organization). These three levels form a progressive evaluation framework comprising 10 fine-grained question types.

DATAVISTA contains 961 professionally produced data videos from real-world sources, spanning five topic domains (Economy, Society, Science, Politics, and Culture), paired with 6,775 evaluation questions, all in multiple-choice format including both single- and multi-select items. Using DATAVISTA, we systematically evaluate 19 mainstream MLLMs. The results show that even the most advanced model, Gemini-3.1-Pro (70.0% accuracy), lags significantly behind human experts (91.7%), with Causal Reasoning and Narrative Structure being the weakest question types across all models. Furthermore, the benefits of increasing frame counts and adding subtitles vary across capability levels, with relatively limited gains in narrative understanding. Error analysis further reveals typical failure modes in chart reading, evidence judgment, and instruction understanding.

Our contributions are summarized as follows:

- We present DATAVISTA, containing 961 real-world data videos and 6,775 evaluation questions, as the first dedicated benchmark for data video understanding.
- We design a three-level progressive capability framework with 10 fine-grained question types to systematically evaluate abilities ranging from chart reading and cross-chart evidence integration to understanding narrative and visual design.
- We systematically evaluate 19 mainstream MLLMs and analyze how performance differs across capability levels, input settings, and error modes.

2 RELATED WORK

Video understanding benchmarks. Existing benchmarks evaluate multimodal understanding, temporal reasoning, and long-video comprehension. Video-MME (Fu et al., 2025) and Video-MME-v2 (Fu et al., 2026) cover diverse video content and input modalities; MVBench (Li et al., 2024), NExT-QA (Xiao et al., 2021), and EgoSchema (Mangalam et al., 2023) examine temporal, causal, and long-range reasoning. LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2024), MLVU (Zhou et al., 2025), and Neptune (Nagrani et al., 2024) further assess understanding over long videos. Specialized benchmarks address human-centric perception and cognition (Cai et al., 2025), video aesthetics (Li et al., 2026), and driving-scene understanding (Zeng et al., 2025). In contrast, DATAVISTA focuses on quantitative evidence in data videos, evaluating how models read dynamically presented chart information, integrate evidence across charts, and understand how this evidence supports the video's claims.

Visualization understanding and quality assessment. Chart understanding benchmarks have expanded from single-chart question answering to multi-chart reasoning. ChartQA (Masry et al., 2022) and PlotQA (Methani et al., 2020) assess visual and logical reasoning over static charts;

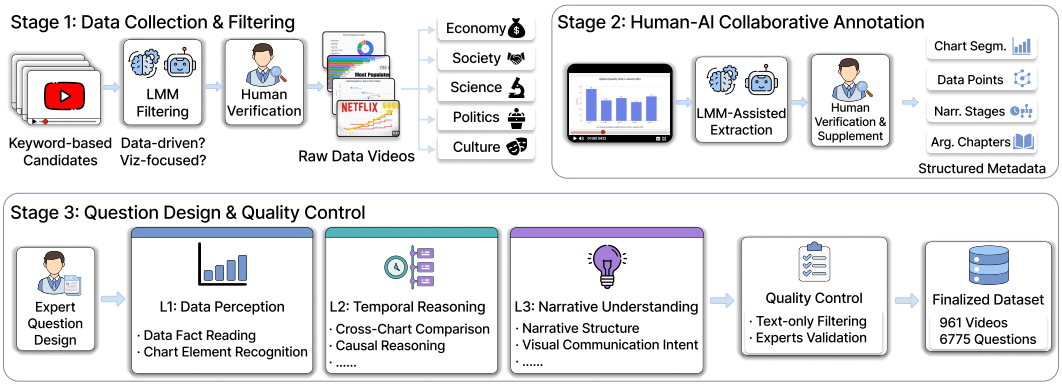


Figure 2: Overview of the DATAVISTA construction process, including data collection, human-AI collaborative annotation, and question design.

ChartQAPro (Masry et al., 2025) and RWA-ChartQA (Hutchinson et al., 2025) introduce diverse questions closer to real-world applications, while MultiChartQA (Zhu et al., 2025b) further examines multi-hop reasoning across charts. Related tasks include chart summarization and comparison (Chart-to-Text (Kantharaj et al., 2022) and ChartDiff (Ye, 2026)), as well as visualization aesthetics and infographic reliability assessment (VisJudge-Bench (Xie et al., 2026b) and IGenBench (Tang et al., 2026)). DATAVISTA extends understanding evaluation to dynamic, narrated data videos, focusing on how chart evidence supports claims, how different segments form a narrative, and how animation and visual design serve communication.

Data video narrative. Data videos use animated visualizations and narration to communicate data stories. Research has progressed from narrative visualization genres (Segel & Heer, 2010) and video clip structure and viewer engagement (Amini et al., 2017; 2018) to the narrative role of animation (Shi et al., 2021), story structure (Wei et al., 2025), and narration–animation coordination (Cheng et al., 2022). Data Player (Shen et al., 2024) and DataMagic (Xie et al., 2026a) automatically generate data videos from existing charts and raw tables, respectively. These studies focus on designing and generating data stories, while DATAVISTA examines whether models understand the data relationships and narrative expression within these stories, extending data video research from content creation to understanding evaluation.

3 DATAVISTA: DESIGN AND CONSTRUCTION

Figure 2 shows the three-stage construction pipeline of DATAVISTA: (1) video collection and filtering, (2) capability hierarchy and question design, and (3) expert annotation and quality control.

3.1 VIDEO COLLECTION AND PREPROCESSING

Sources. Drawing on prior collection practices (Yang et al., 2022; Fu et al., 2026), we collect 5,604 candidate videos from YouTube through keyword search and channel-level scraping, covering professional media, data journalism, educational, and independent creator channels. Representative channels and search keywords are listed in Appendix B.1.

Topic taxonomy. To support fine-grained topic-level evaluation, we draw on topic classifications in data journalism research (Stalph, 2018; Loosen et al., 2020) to establish a two-level taxonomy with 5 domains (Economy, Society, Science, Politics, and Culture) and 22 subcategories, as shown in Figure 3a. Videos first receive initial tags based on source channels and search keywords, then are assigned to categories using channel affiliation and title cues. The complete category mapping and subcategory distribution are provided in Appendix B.5.

Duration and recency. As shown in Figure 3b, videos range from 28 seconds to 15 minutes, with a median of ≈5.6 minutes, matching the typical narrative density reported for professional data videos (Amini et al., 2015; Yang et al., 2022). To reduce leakage risk, DATAVISTA prioritizes recently published content, with approximately 68.5% of videos published in or after June 2025

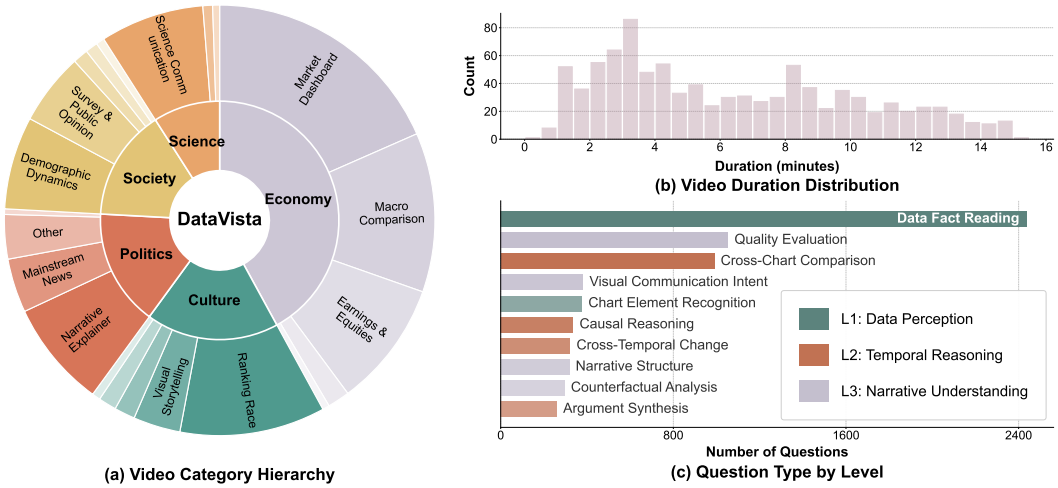


Figure 3: Overview of the DATAVISTA dataset. (a) Video category hierarchy showing the distribution across five major domains and their sub-categories. (b) Distribution of video durations in minutes. (c) Question types organized by evaluation level.

(contamination analysis in Appendix E.9). For content quality, we use source credibility as the primary gate rather than a single engagement metric, and record view count only as a supplementary indicator. We further manually remove classic films, television episodes, and flagship content of top-tier creators, following the decontamination practice of recent video benchmarks (Fu et al., 2026).

Screening pipeline. We use Gemini-3-Flash to pre-screen candidate videos (Figure 2, Stage 1), requiring: (1) core information to be data-driven rather than conveyed solely through narration; (2) at least one clearly visible and sustained data visualization element; (3) a clear information communication intent, excluding pure chart displays, software tutorials, and tool demonstrations. Annotators then watch the videos to confirm or correct each model classification. The screening prompt, output schema, and verification interface are provided in Appendix B.1. Of the 5,604 candidate videos, 967 pass screening and proceed to question construction; after question review, the retained questions cover 961 videos.

3.2 CAPABILITY HIERARCHY

Understanding a data video requires coordinating multiple abilities, from perceiving data in individual frames, to integrating information across charts over time, to interpreting the creator's narrative intent (Segel & Heer, 2010; Amini et al., 2015). To characterize strengths and bottlenecks in this process, we follow Bloom's Taxonomy (Bloom et al., 1956), which is widely used to guide assessment design and orders cognitive objectives from remembering and understanding toward analyzing and evaluating. We accordingly organize the benchmark questions into three progressive levels comprising 10 fine-grained question types (complete definitions in Appendix A).

Level 1: Data Perception. This foundational level assesses whether a model can locate relevant visual evidence in a video and read the data it contains, corresponding to the "reading" level in visualization literacy research (Boy et al., 2014; Lee et al., 2017). Unlike static chart question answering where the target chart is pre-isolated, relevant charts here appear within a continuous video containing multiple visual states, requiring the model to localize evidence before reading; answers do not depend on frame ordering. It contains two question types: *Data Fact Reading* (reading and cross-frame aggregation of values, trends, and extrema) and *Chart Element Recognition* (chart type, axis, legend, and color encoding).

Level 2: Temporal Reasoning. Building on L1, this level assesses the model's ability to integrate evidence across charts and reason about changes and relationships over time (Amini et al., 2015; Thompson et al., 2020). It contains four question types: *Cross-Temporal Change*, tracking how events evolve and extrapolate over time; *Cross-Chart Comparison*, comparing values across charts or deriving, through arithmetic, results not stated explicitly in the video (Cleveland & McGill,

1984); *Causal Reasoning*, identifying causal chains supported by evidence from multiple charts; and *Argument Synthesis*, identifying the key evidence that supports the video's central argument.

Level 3: Narrative Understanding. This highest level assesses the model's understanding of how data are organized and presented as a narrative, as well as overall video quality. It contains four question types: *Narrative Structure*, covering narrative pattern recognition, data-flow comprehension, and the roles and connections of non-adjacent segments within the overall argument (Yang et al., 2022); *Visual Communication Intent*, analyzing how chart and animation design serves narrative purposes (Shi et al., 2021; Shen et al., 2025); *Counterfactual Analysis*, reasoning about how changes in design choices would affect information delivery; and *Quality Evaluation*, making holistic judgments along the Fidelity–Expressiveness–Aesthetics dimensions (Xie et al., 2026b).

3.3 Question Design

Question format. DataVista contains both four-option single-answer and multiple-select questions. Given the inherently combinatorial nature of causal argumentation and narrative design in data videos (Yang et al., 2022), Causal Reasoning, Argument Synthesis, Narrative Structure, Visual Communication Intent, and Counterfactual Analysis often require models to identify the correct subset from multiple visual evidence items or design options, and multiple-select questions concentrate in these types. Figure 3c shows the distribution of question types across levels, with statistics on per-level question counts and the distribution of questions per video in Appendix B.10.

Distractor design. Each level follows differentiated distractor design principles. L1 numerical questions include at least one numerically close distractor to test precise reading. L2 cross-chart questions include at least one distractor that can be selected based on a single chart alone, reducing the shortcut of answering based on a single chart. L3 narrative questions draw all options from real categories in the narrative design space for data videos (Yang et al., 2022), avoiding out-of-domain fake options. Quality evaluation questions follow the Fidelity–Expressiveness–Aesthetics framework (Xie et al., 2026b): all options share the same positive tone and differ only in the specific visual mechanism identified, reducing the shortcut of answering based on sentiment polarity. The complete distractor design principles for each question type are provided in Appendix A.

3.4 Expert Annotation and Quality Control

Candidate question design. The annotation workflow is illustrated in Figure 2 (Stages 2–3). Five annotators with experience in data video production or data visualization research design the candidate questions. To assist this process, Gemini-3-Flash pre-extracts chart segments, key data points, narrative stages, and argument chapters (prompt and output schema in Appendix B.2). Annotators watch each video in full, verify and correct the extracted information segment by segment, supplement omissions, and mark timestamps for key data points. They then write question stems, options, and answers following the three-level capability framework and distractor design principles, with detailed specifications in Appendix A. Annotation time per video ranges from approximately 1 to 2 hours, depending on video duration, chart density, and narrative complexity.

Text-only filtering. To reduce shortcuts that do not require video evidence (Fu et al., 2026), we provide Gemini-3-Flash with only question stems and options, filtering out candidates whose predicted answers exactly match the references. Across five representative models from different model families, filtered-out questions are also easier to answer from text alone, with accuracy 22.9–32.2 percentage points higher than on retained questions. Screening details and cross-model analysis are provided in Appendix B.3.1.

Human review and finalization. Four reviewers who did not participate in question design review the candidates retained after text-only filtering. Tasks are assigned by video, with two reviewers separately checking each question's wording, video evidence for the answer, and distractors. Disagreements are discussed against the original video; revised questions are rechecked and those still ambiguous are removed. Among candidates retained after text-only filtering, 80.1% pass review, 15.4% pass after manual revision, and 4.5% are rejected. The final dataset contains 6,775 ques-

Table 1: Exact-match accuracy (%) on DATAVISTA. *w/o Sub.* and *w/ Sub.* use the same 50 sampled frames without audio, without and with subtitles; *Avg Acc* weights the three levels by question count, and *Overall* pools both conditions. Human evaluation uses complete videos with original audio.

Model	Level 1		Level 2		Level 3		Avg Acc		**Overall**
	w/o Sub.	w/ Sub.	w/o Sub.	w/ Sub.	w/o Sub.	w/ Sub.	w/o Sub.	w/ Sub.	
Human Baseline									
Human Expert	97.1		89.0		86.8		91.7		–
Closed-Source Models									
Gemini-3.1-Pro	**78.9**	**81.8**	**66.4**	**71.5**	55.1	57.9	**68.2**	**71.7**	**70.0**
GPT-5	76.5	78.0	61.2	66.7	55.9	58.1	66.0	68.8	67.4
Gemini-3-Flash	75.6	78.0	59.7	68.8	52.3	55.1	64.1	68.5	66.3
GPT-4o	65.0	66.4	42.5	44.8	56.6	**58.7**	56.1	58.0	57.1
Claude-Sonnet-4.6	70.3	74.5	50.2	60.7	**59.2**	58.6	61.3	65.8	63.6
Open-Source Models									
Qwen3-VL-32B-Instruct	68.5	71.6	39.0	48.5	50.5	54.2	54.7	59.8	57.3
InternVL3.5-38B	62.7	65.6	46.1	48.1	44.9	48.9	52.6	55.6	54.1
InternVL3.5-14B	54.5	60.3	36.9	42.7	40.2	41.4	45.2	49.6	47.4
InternVL3.5-8B	54.5	60.5	35.9	36.3	39.3	39.6	44.7	47.4	46.1
Qwen3-VL-8B-Instruct	59.0	64.7	28.5	33.9	38.0	38.9	44.1	48.2	46.2
InternVL3-8B	55.2	59.0	33.7	38.2	36.2	37.7	43.4	46.7	45.1
Qwen2.5-VL-32B-Instruct	63.4	65.4	46.4	47.8	48.0	48.9	54.0	55.5	54.8
Qwen2.5-VL-7B-Instruct	56.3	63.2	33.6	33.9	31.2	31.8	42.3	45.4	43.9
InternVL3.5-4B	53.7	48.3	35.9	33.6	39.9	35.8	44.5	40.4	42.5
Qwen3-VL-4B-Instruct	57.9	61.0	25.4	30.5	34.9	37.7	41.8	45.4	43.6
Qwen2.5-VL-3B-Instruct	47.7	53.0	35.3	36.9	27.1	27.7	38.0	40.8	39.4
InternVL3.5-2B	47.7	52.1	27.8	29.8	27.1	28.0	35.9	38.5	37.2
Qwen3-VL-2B-Instruct	50.3	57.4	28.8	30.5	25.2	24.9	36.7	40.0	38.4
InternVL3-2B	40.5	44.6	25.0	21.8	23.5	21.1	31.0	31.1	31.1

tions covering 961 videos. The complete review procedure, level-specific criteria, and statistics are provided in Appendix B.3.2.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Models. We evaluate 19 MLLMs, comprising five closed-source and 14 open-source models. The open-source models span the Qwen3-VL (Bai et al., 2025a), Qwen2.5-VL (Bai et al., 2025b), InternVL3.5 (Wang et al., 2025), and InternVL3 (Zhu et al., 2025a) families, with Instruct variants used for the Qwen models. Table 1 lists all model variants. Unless noted otherwise, later diagnostic analyses in this section use five representative models: Gemini-3.1-Pro, GPT-5, Claude-Sonnet-4.6, Qwen3-VL-32B, and InternVL3.5-38B.

Input settings. To compare models under a uniform visual budget, the main evaluation provides all models with question stems, answer options, and 50 frames uniformly sampled from the full video, without audio. We use the same frame set for conditions without and with subtitles, adding subtitle text in the latter. Models provide answers together with supporting reasons. The impact of varying frame budgets and native video inputs is reported in Section 4.3 and Appendix E.2, with evaluation details in Appendix D.1.

Evaluation metric. We use per-question exact-match accuracy: single-select answers must be correct, and multi-select predictions must exactly match the ground-truth option set. Overall accuracy is computed by pooling responses from the conditions with and without subtitles. For multi-select questions, we also report option-level precision, recall, and F1 to assess partially correct responses (Appendix E.5). Uncertainty estimation is described in Appendix D.3.

Human evaluation. Six evaluators watch complete videos and listen to the original audio. Questions are assigned by video, with two independent answers per question. Those who wrote or finally reviewed a question do not answer it. Scoring follows the same criterion as for the models.

4.2 Overall Performance

Human performance. Human evaluators achieve 91.7% overall accuracy, with 97.1%, 89.0%, and 86.8% at L1, L2, and L3, respectively. They perform near ceiling on L1 data reading and maintain high accuracy on temporal reasoning and narrative understanding. Exact agreement between paired responses is 88.6% overall and 95.5%, 85.7%, and 81.3% across the three levels.

Overall results. As shown in Table 1, Gemini-3.1-Pro and GPT-5 achieve overall accuracies of 70.0% and 67.4%, respectively, with the former still 21.7 percentage points below the human baseline. The strongest open-source model, Qwen3-VL-32B, reaches 57.3%, trailing Gemini-3.1-Pro by 12.7 points. Its Thinking mode raises overall accuracy to 62.9%, still below the leading closed-source models. Gains are concentrated in temporal reasoning, without corresponding improvement in narrative understanding. Per-level results and analysis of both modes are provided in Appendix E.3. With subtitles, equal weighting across the five domains leaves the ranking of the five representative models unchanged (Appendix E.1).

Subtitle gains and baselines. Adding subtitles raises overall accuracy by a median of 3.1 percentage points across the 19 models, with the largest gain of 5.1 points for Qwen3-VL-32B; Section 4.4 analyzes how subtitles affect the individual capability levels. Random selection and the most-frequent-answer baseline score 20.8% and 27.5%, respectively; their settings and per-level results are provided in Appendix D.2.

Cross-level differences. Temporal reasoning is particularly discriminative in these model comparisons. Gemini-3.1-Pro leads Qwen3-VL-32B by 25.3 percentage points on L2, compared with 10.3 points on L1 and 4.2 on L3. On L3 the gap lies mainly between models and humans: the highest accuracy reached by any model is 81.8% on L1, 15.3 percentage points below humans, but only 59.2% on L3, where the gap widens to 27.6 points.

Native video input. Given complete videos with their original audio tracks, Gemini-3.1-Pro and Gemini-3-Flash achieve overall accuracies of 73.5% and 68.9%, respectively, both below the human score of 91.7%. On narrative understanding, they score 59.4% and 55.9%, respectively, leaving a substantial gap to the human score of 86.8%. Detailed comparisons and cases are provided in Appendix E.2.

4.3 Frame Budget Sensitivity

To examine how sampling density affects performance, we compare five representative models at frame budgets of 8, 16, 32, and 50. Frames are sampled at the midpoints of equal temporal segments across the full video, without subtitles or audio.

As shown in Figure 4, increasing the frame budget improves overall accuracy and changes which models lead. GPT-5 performs best at 8 and 16 frames, whereas Gemini-3.1-Pro leads at 32 and 50 frames. Among the open-source models, InternVL3.5-38B leads at lower budgets, whereas Qwen3-VL-32B performs better at higher budgets.

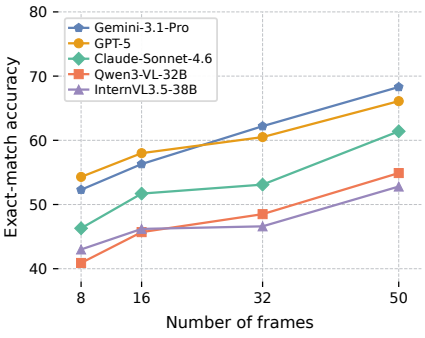


Figure 4: Exact-match accuracy across frame budgets without subtitles or audio.

Basic data perception is more sensitive to sampling density. Increasing the budget from 8 to 32 frames raises mean L1 accuracy across the five models by 11.6 percentage points, compared with 4.7 on L2 and 2.1 on L3. With 100 frames, L1 can still benefit, whereas L2 and L3 show no consistent improvement (Appendix E.4). These results indicate that increasing the number of sampled frames primarily improves chart reading, with more limited gains in temporal reasoning and narrative understanding.

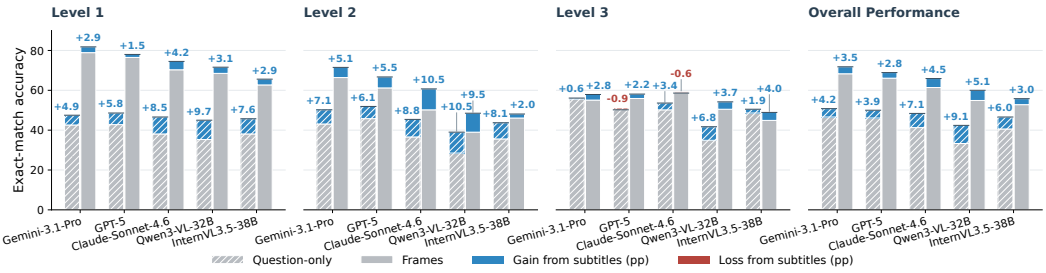


Figure 5: Exact-match accuracy (%) of five models under question-only and frame inputs, by level. Colored segments show the change after adding subtitles (percentage points).

4.4 Input Evidence Analysis

To examine the contributions of frames and subtitles, we compare the five representative models under four input conditions: questions only, questions with subtitles, questions with frames, and questions with both frames and subtitles. All conditions include answer options and exclude audio. Figure 5 reports the results.

The contribution of frames decreases across levels. Relative to questions alone, adding frames raises mean L1 accuracy across the five models from 39.4% to 71.4%, a gain of 32.0 percentage points, compared with 14.7 on L2 and 5.1 on L3 (from 48.0% to 53.1%). On L1, frames alone also outperform questions with subtitles (46.7%) by 24.7 points, highlighting the importance of visual chart information for basic data reading. Together with the frame-budget results, these findings indicate that increasing visual input primarily improves basic chart reading, with limited gains in narrative understanding.

Subtitles provide the largest supplementary gain for temporal reasoning. Adding subtitles to frames raises mean L2 accuracy from 52.6% to 59.1%, a gain of 6.5 percentage points, compared with 2.9 on L1 and 2.4 on L3. Joint input also outperforms questions with subtitles alone by 13.1 points, indicating that frames and subtitles provide complementary information for temporal reasoning.

Subtitles can also narrow performance gaps between short and long videos. For GPT-5, adding subtitles raises accuracy on videos longer than 9 minutes from 59.4% to 66.9%, reducing the gap from the shortest-video group from 8.9 to 1.4 percentage points. Duration-group results for all models are provided in Appendix E.6.

4.5 Capability Dimension Analysis

We further compare the five representative models across ten question types to examine differences within each capability level (Figure 6).

Causal and argument tasks. With both frames and subtitles, mean accuracy is 73.6% on Cross-Temporal Change and 62.6% on Cross-Chart Comparison, compared with 39.2% on Causal Reasoning and 53.9% on Argument Synthesis; even the best model reaches only 44.2% on Causal Reasoning. Unlike tracking changes or comparing values, the latter two tasks require identifying which data support a specified causal relationship or the central argument and selecting the complete set of supporting evidence. These two types also benefit most from subtitles, gaining 8.5 and 16.6 percentage points, respectively, compared with at most 4.1 for the other two types.

Narrative structure. Across the five models, accuracy on Cross-Temporal Change ranges from 56.0% to 87.8%, but all score between 46.9% and 51.0% on Narrative Structure. These questions further require understanding how data are organized into a narrative and how arguments connect across chapters. Within L3, Counterfactual Analysis averages 63.3% and Narrative Structure averages 49.6%, indicating that the difficulty of L3 concentrates on understanding narrative structure.

Answer completeness. For the multi-select questions within these task types, we supplement exact match with option-level F1 to distinguish partially correct from fully correct answers (Appendix E.5). With subtitles, Argument Synthesis reaches 88.0% F1 but 53.9% exact match, indicating that many responses are only partially correct and that fully correct answers remain challenging. Even with partial credit, Causal Reasoning has the lowest F1 among the five multi-select types.

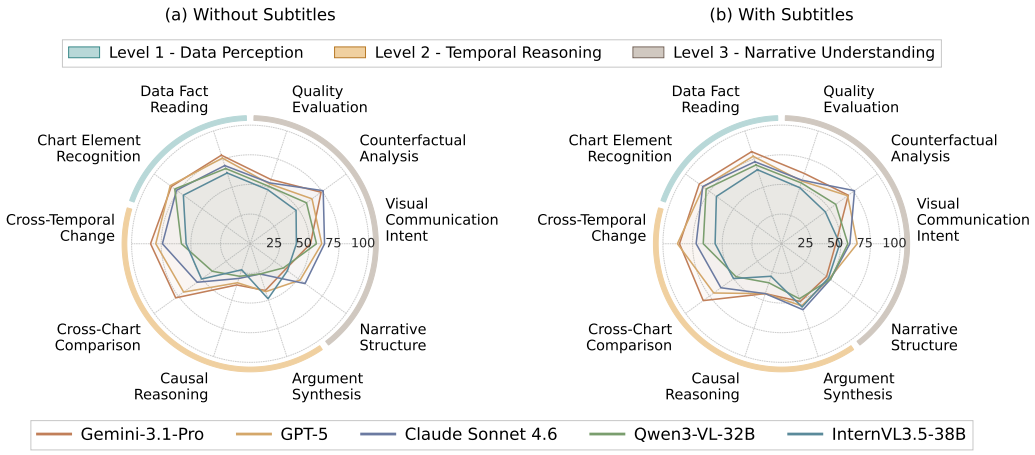


Figure 6: Per-type exact-match accuracy (%) of five representative models: (a) without subtitles and (b) with subtitles.

4.6 Error Analysis

Following diagnostic frameworks from chart question answering and video understanding (Masry et al., 2022; Zhu et al., 2025b; Fu et al., 2025), we manually inspect 120 randomly sampled incorrect responses from Gemini-3.1-Pro under 50-frame input with subtitles and group them into perception, reasoning, and instruction understanding.

Perceptual errors (47.5%). The most common cases are value and text reading errors (36.7%), such as reading the Nintendo Wii first-three-year sales total as 45 million rather than about 20 million. The remaining cases include confused visual mappings (5.0%), reversed trends or rankings (3.3%), and chart-structure errors such as misjudged axis scales (2.5%). These errors occur with the relevant frames already available, indicating that accurately extracting chart information remains a bottleneck in data video understanding.

Reasoning errors (42.5%). After chart facts have been extracted, the main difficulty lies in building logical relations. Pure arithmetic errors are rare (0.8%); more common are cross-chart linking (4.2%) and evidence judgment (37.5%). On causal reasoning and argument synthesis in particular, the model often reads the chart facts correctly but fails to judge whether a given fact supports a specified claim, indicating that it still struggles to connect observed data with the narrative argument.

Instruction understanding errors (10.0%). These errors do not arise from failing to read the chart, but from not following constraints in the question: the model shifts the specified time range or target object (5.0%), maps an otherwise reasonable conclusion to the wrong option (3.3%), or overlooks a restriction on the evidence source (1.7%), for instance selecting a narrated revenue target when the question asks for chart-specific facts.

Further analysis of Qwen3-VL-32B also identifies cases where the model correctly reads chart facts but misjudges whether they support the claim in the question; detailed diagnostics are provided in Appendix E.8. Error cases for Gemini-3.1-Pro and example questions across capability levels are provided in Appendix E.7 and Appendix C, respectively.

5 Conclusion

We present DataVista, the first benchmark for data video understanding, comprising 961 real-world data videos and 6,775 questions across three capability levels and 10 question types. Evaluation of 19 MLLMs reveals a substantial gap from human performance, with persistent difficulties in identifying evidence for causal claims and understanding narratives. DataVista provides a foundation for systematically evaluating and improving data video understanding.

AI USE STATEMENT

In this work, during dataset construction we used Gemini-3-Flash for data-video pre-screening, structured information pre-extraction, and text-only filtering based on question stems and options alone. These outputs were verified or corrected by humans, and questions were written by annotators. We used GPT-6 Astra for English grammar polishing and for suggestions on wording. We did not use generative AI to create synthetic data, propose theoretical models or mathematical claims, write proofs, or introduce novel algorithmic ideas or academic claims. The remaining required-disclosure tasks are not applicable to this work. All AI-assisted content was reviewed by the authors, who take full responsibility for the final content. Prompts contained no private or sensitive data. Large language models are not authors of this paper.

ETHICS STATEMENT

This work introduces DataVista, an academic benchmark for data video understanding. The source videos used to construct this benchmark were collected from public video platforms (e.g., YouTube). To advance research on multimodal large language models in this emerging domain, we open-source all high-quality annotations meticulously created by our team, including question-answer pairs, evidence timestamps, and structured metadata. The copyright of all source videos belongs to their respective original creators or rightsholders, and this benchmark along with its open-source data is intended solely for academic research. Furthermore, during the dataset construction process, we ensured that all participating domain experts received fair compensation. The data collection and annotation procedures of this study have been reviewed and approved by our institutional ethics board. We hope DataVista will serve as a valuable evaluation tool to promote the development of automated data visualization assessment and understanding technologies.

REPRODUCIBILITY STATEMENT

To support reproducibility, we release the question annotations and video source records. Anonymized data are available at `https://anonymous.4open.science/r/benchmark-29BB/`. Detailed descriptions of our methodology are provided throughout the paper and appendices: video collection, screening, and question design are detailed in Section 3 and Appendix B; the three-level capability framework is defined in Appendix A; evaluation prompts and input settings (including frame sampling strategies) are provided in Appendix D.1, and scoring rules in Section 4.1; and data contamination analysis is discussed in Appendix E.9. To ensure long-term consistency and standardization of future model evaluations, we will deploy an official automated evaluation service and a public leaderboard upon publication, allowing researchers to seamlessly benchmark their models and track ongoing progress in data video understanding.

REFERENCES

Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Christophe Hurter, and Pourang Irani. Understanding data videos: Looking at narrative visualization through the cinematography lens. In Bo Begole, Jinwoo Kim, Kori Inkpen, and Woontack Woo (eds.), *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI 2015, Seoul, Republic of Korea, April 18-23, 2015*, pp. 1459–1468. ACM, 2015. doi: 10.1145/2702123.2702431. URL https://doi.org/10.1145/2702123.2702431.

Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Andrés Monroy-Hernández, and Pourang Irani. Authoring data-driven videos with dataclips. *IEEE Trans. Vis. Comput. Graph.*, 23(1): 501–510, 2017. doi: 10.1109/TVCG.2016.2598647. URL https://doi.org/10.1109/TVCG.2016.2598647.

Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Jason Leboe-McGowan, and Pourang Irani. Hooked on data videos: assessing the effect of animation and pictographs on viewer engagement. In Tiziana Catarci, Kent L. Norman, and Massimo Mecella (eds.), *Proceedings of the 2018 International Conference on Advanced Visual Interfaces, AVI 2018, Castiglione della Pescaia, Italy, May 29 - June 01, 2018*, pp. 21:1–21:9. ACM, 2018. doi: 10.1145/3206505.3206552. URL https://doi.org/10.1145/3206505.3206552.

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, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Ming-Hsuan Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *CoRR*, abs/2502.13923, 2025b. doi: 10.48550/ARXIV.2502.13923. URL https://doi.org/10.48550/arXiv.2502.13923.

Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, et al. *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Longman New York, 1956.

Jeremy Boy, Ronald A. Rensink, Enrico Bertini, and Jean-Daniel Fekete. A principled way of assessing visualization literacy. *IEEE Trans. Vis. Comput. Graph.*, 20(12):1963–1972, 2014. doi: 10.1109/TVCG.2014.2346984. URL https://doi.org/10.1109/TVCG.2014.2346984.

Jia Bu, Mingwei Jiang, Shuqi Liu, Tong Lyu, Lumeng Wu, Shiqi Jiang, Boyuan Huangfu, Changbo Wang, and Chenhui Li. NewsVis: GenAI-Based visual storytelling for corporate financial news. *IEEE Trans. Vis. Comput. Graph.*, 32(7):5501–5517, 2026. doi: 10.1109/TVCG.2026.3668994. URL https://doi.org/10.1109/TVCG.2026.3668994.

Yuxuan Cai, Jiangning Zhang, Zhenye Gan, Qingdong He, Xiaobin Hu, Junwei Zhu, Yabiao Wang, Chengjie Wang, Zhucun Xue, Chaoyou Fu, et al. Humanvideo-mme: Benchmarking mllms for human-centric video understanding. *arXiv preprint arXiv:2507.04909*, 2025.

Hao Cheng, Junhong Wang, Yun Wang, Bongshin Lee, Haidong Zhang, and Dongmei Zhang. Investigating the role and interplay of narrations and animations in data videos. *Comput. Graph. Forum*, 41(3):527–539, 2022. doi: 10.1111/CGF.14560. URL https://doi.org/10.1111/cgf.14560.

William S. Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. doi: 10.1080/01621459.1984.10478080. URL https://doi.org/10.1080/01621459.1984.10478080.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24108–24118. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.02245. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, et al. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*, 2026.

Lily W. Ge, Yuan Cui, and Matthew Kay. CALVI: critical thinking assessment for literacy in visualizations. In Albrecht Schmidt, Kaisa Väänänen, Tesh Goyal, Per Ola Kristensson, Anicia Peters, Stefanie Mueller, Julie R. Williamson, and Max L. Wilson (eds.), *Proceedings of the 2023*

CHI Conference on Human Factors in Computing Systems, CHI 2023, Hamburg, Germany, April 23-28, 2023, pp. 815:1–815:18. ACM, 2023. doi: 10.1145/3544548.3581406. URL https://doi.org/10.1145/3544548.3581406.

Maeve Hutchinson, Radu Jianu, Aidan Slingsby, Jo Wood, and Pranava Madhyastha. Chart question answering from real-world analytical narratives. In Jin Zhao, Mingyang Wang, and Zhu Liu (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 760–773. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.ACL-SRW.50. URL https://doi.org/10.18653/v1/2025.acl-srw.50.

Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq R. Joty. Chart-to-text: A large-scale benchmark for chart summarization. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pp. 4005–4023. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.ACL-LONG.277. URL https://doi.org/10.18653/v1/2022.acl-long.277.

Sukwon Lee, Sung-Hee Kim, and Bum Chul Kwon. VLAT: development of a visualization literacy assessment test. *IEEE Trans. Vis. Comput. Graph.*, 23(1):551–560, 2017. doi: 10.1109/TVCG.2016.2598920. URL https://doi.org/10.1109/TVCG.2016.2598920.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. MVBench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 22195–22206. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02095. URL https://doi.org/10.1109/CVPR52733.2024.02095.

Yunhao Li, Sijing Wu, Zhilin Gao, Zicheng Zhang, Qi Jia, Huiyu Duan, Xiongkuo Min, and Guangtao Zhai. VideoAesBench: Benchmarking the video aesthetics perception capabilities of large multimodal models. *CoRR*, abs/2601.21915, 2026. doi: 10.48550/ARXIV.2601.21915. URL https://doi.org/10.48550/arXiv.2601.21915.

Wiebke Loosen, Julius Reimer, and Fenja De Silva-Schmidt. Data-driven reporting: An on-going (r)evolution? an analysis of projects nominated for the data journalism awards 2013–2016. *Journalism*, 21(9):1246–1263, 2020. doi: 10.1177/1464884917735691. URL https://doi.org/10.1177/1464884917735691.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. EgoSchema: A diagnostic benchmark for very long-form video language understanding. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/90ce332aff156b910b002ce4e6880dec-Abstract-Datasets_and_Benchmarks.html.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq R. Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022*, Findings of ACL, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/V1/2022.FINDINGS-ACL.177. URL https://doi.org/10.18653/v1/2022.findings-acl.177.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md. Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md. Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*,

Findings of ACL, pp. 19123–19151. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.findings-acl.978/.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2020, Snowmass Village, CO, USA, March 1-5, 2020*, pp. 1516–1525. IEEE, 2020. doi: 10.1109/WACV45572.2020.9093523. URL https://doi.org/10.1109/WACV45572.2020.9093523.

Arsha Nagrani, Mingda Zhang, Ramin Mehran, Rachel Hornung, Nitesh Bharadwaj Gundavarapu, Nilpa Jha, Austin Myers, Xingyi Zhou, Boqing Gong, Cordelia Schmid, Mikhail Sirotenko, Yukun Zhu, and Tobias Weyand. Neptune: The long orbit to benchmarking long video understanding. *CoRR*, abs/2412.09582, 2024. doi: 10.48550/ARXIV.2412.09582. URL https://doi.org/10.48550/arXiv.2412.09582.

Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE Trans. Vis. Comput. Graph.*, 16(6):1139–1148, 2010. doi: 10.1109/TVCG.2010.179. URL https://doi.org/10.1109/TVCG.2010.179.

Leixian Shen, Yizhi Zhang, Haidong Zhang, and Yun Wang. Data player: Automatic generation of data videos with narration-animation interplay. *IEEE Trans. Vis. Comput. Graph.*, 30(1):109–119, 2024. doi: 10.1109/TVCG.2023.3327197. URL https://doi.org/10.1109/TVCG.2023.3327197.

Leixian Shen, Haotian Li, Yun Wang, and Huamin Qu. Reflecting on design paradigms of animated data video tools. In Naomi Yamashita, Vanessa Evers, Koji Yatani, Sharon Xianghua Ding, Bongshin Lee, Marshini Chetty, and Phoebe O. Toups Dugas (eds.), *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pp. 190:1–190:21. ACM, 2025. doi: 10.1145/3706598.3713449. URL https://doi.org/10.1145/3706598.3713449.

Yang Shi, Xingyu Lan, Jingwen Li, Zhaorui Li, and Nan Cao. Communicating with motion: A design space for animated visual narratives in data videos. In Yoshifumi Kitamura, Aaron Quigley, Katherine Isbister, Takeo Igarashi, Pernille Bjørn, and Steven Mark Drucker (eds.), *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pp. 605:1–605:13. ACM, 2021. doi: 10.1145/3411764.3445337. URL https://doi.org/10.1145/3411764.3445337.

Florian Stalph. Classifying data journalism: A content analysis of daily data-driven stories. *Journalism Practice*, 12(10):1332–1350, 2018. doi: 10.1080/17512786.2017.1386583. URL https://doi.org/10.1080/17512786.2017.1386583.

Yinghao Tang, Xueding Liu, Boyuan Zhang, Tingfeng Lan, Yupeng Xie, Jiale Lao, Yiyao Wang, Haoxuan Li, Tingting Gao, Bo Pan, Luoxuan Weng, Xiuqi Huang, Minfeng Zhu, Yingchaojie Feng, Yuyu Luo, and Wei Chen. Igenbench: Benchmarking the reliability of text-to-infographic generation. *CoRR*, abs/2601.04498, 2026. doi: 10.48550/ARXIV.2601.04498. URL https://doi.org/10.48550/arXiv.2601.04498.

John Thompson, Zhicheng Liu, Wilmot Li, and John T. Stasko. Understanding the design space and authoring paradigms for animated data graphics. *Comput. Graph. Forum*, 39(3):207–218, 2020. doi: 10.1111/CGF.13974. URL https://doi.org/10.1111/cgf.13974.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark. *CoRR*, abs/2406.08035, 2024. doi: 10.48550/ARXIV.2406.08035. URL https://doi.org/10.48550/arXiv.2406.08035.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, JingJing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu,

Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing Open-Source multimodal models in versatility, reasoning, and efficiency. *CoRR*, abs/2508.18265, 2025. doi: 10.48550/ARXIV.2508.18265. URL https://doi.org/10.48550/arXiv.2508.18265.

Zheng Wei, Huamin Qu, and Xian Xu. Telling data stories with the hero's journey: Design guidance for creating data videos. *IEEE Trans. Vis. Comput. Graph.*, 31(1):962–972, 2025. doi: 10.1109/TVCG.2024.3456330. URL https://doi.org/10.1109/TVCG.2024.3456330.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. LongVideoBench: A benchmark for long-context interleaved video-language understanding. In Amir Globerson, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/329ad516cf7a6ac306f29882e9c77558-Abstract-Datasets_and_Benchmarks_Track.html.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 9777–9786. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00965. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

Yupeng Xie, Chen Ma, Zhenyang Wang, Liangwei Wang, Jiayi Zhu, Chuxuan Zeng, Zhouan Shen, Boyan Li, and Yuyu Luo. DataMagic: Transforming tabular data into data insight video. *Proceedings of the VLDB Endowment*, 19(12):4526–4529, 2026a. doi: 10.14778/3827998.3828057. URL https://doi.org/10.14778/3827998.3828057.

Yupeng Xie, Zhiyang Zhang, Yifan Wu, Sirong Lu, Jiayi Zhang, Zhaoyang Yu, Jinlin Wang, Sirui Hong, Bang Liu, Chenglin Wu, and Yuyu Luo. VisJudge-Bench: Aesthetics and quality assessment of visualizations. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=lG1HWWdEbN.

Leni Yang, Xian Xu, Xingyu Lan, Ziyan Liu, Shunan Guo, Yang Shi, Huamin Qu, and Nan Cao. A design space for applying the freytag's pyramid structure to data stories. *IEEE Trans. Vis. Comput. Graph.*, 28(1):922–932, 2022. doi: 10.1109/TVCG.2021.3114774. URL https://doi.org/10.1109/TVCG.2021.3114774.

Rongtian Ye. ChartDiff: A large-scale benchmark for comprehending pairs of charts. *CoRR*, abs/2603.28902, 2026. doi: 10.48550/ARXIV.2603.28902. URL https://doi.org/10.48550/arXiv.2603.28902.

Tong Zeng, Longfeng Wu, Liang Shi, Dawei Zhou, and Feng Guo. Are vision llms road-ready? A comprehensive benchmark for safety-critical driving video understanding. In Luiza Antonie, Jian Pei, Xiaohui Yu, Flavio Chierichetti, Hady W. Lauw, Yizhou Sun, and Srinivasan Parthasarathy (eds.), *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, V.2, KDD 2025, Toronto ON, Canada, August 3-7, 2025*, pp. 5972–5983. ACM, 2025. doi: 10.1145/3711896.3737396. URL https://doi.org/10.1145/3711896.3737396.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. MLVU: benchmarking multi-task long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 13691–13701. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01278. URL https://openaccess.thecvf.com/content/CVPR2025/html/Zhou_MLVU_

Benchmarking_Multi-task_Long_Video_Understanding_CVPR_2025_paper.html.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479, 2025a. doi: 10.48550/ARXIV.2504.10479. URL https://doi.org/10.48550/arXiv.2504.10479.

Zifeng Zhu, Mengzhao Jia, Zhihan Zhang, Lang Li, and Meng Jiang. MultiChartQA: Benchmarking vision-language models on multi-chart problems. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 11341–11359. Association for Computational Linguistics, 2025b. doi: 10.18653/V1/2025.NAACL-LONG.566. URL https://doi.org/10.18653/v1/2025.naacl-long.566.

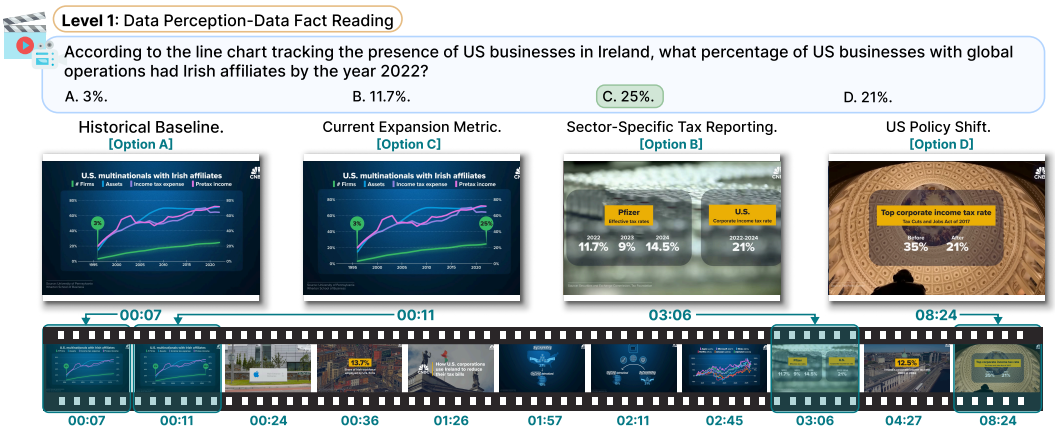


Figure 7: An example L1 question in DATAVISTA (Data Fact Reading). The question requires extracting a specific numerical value from a trend chart. The correct answer (C) is directly grounded in the visual label for the year 2022, while distractors correspond to the historical baseline (A), sector-specific tax reporting (B), and the US policy shift (D).

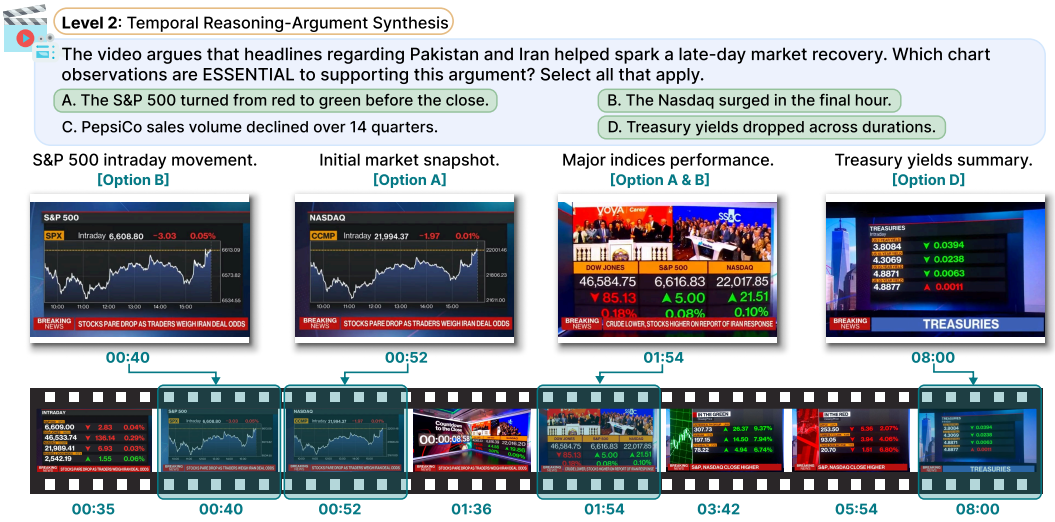


Figure 8: An example L2 question in DATAVISTA. The video argues that headlines regarding Pakistan and Iran helped spark a late-day market recovery. The task is to select all essential chart observations (A, B, D) that support the argument.

A CAPABILITY HIERARCHY AND QUESTION TYPE SPECIFICATION

This appendix provides the complete definitions of all 10 question types in DATAVISTA, organized by capability level. Each question type is specified along four aspects: (i) definition, (ii) target competency, (iii) required visual evidence, and (iv) distractor design principles. These specifications served as annotation guidelines during dataset construction and as the reference framework for the fine-grained capability analysis presented in Section 4.

A.1 DESIGN RATIONALE

The three progressive levels and ten question types in DATAVISTA are designed around capability organization and evidence requirements.

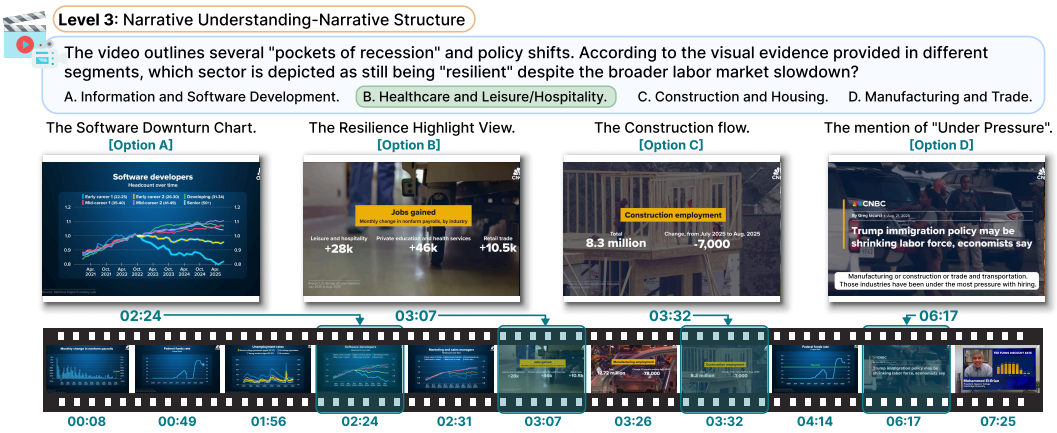


Figure 9: An example L3 question in DATAVISTA (Narrative Structure). This task evaluates the model's ability to synthesize global narrative themes. The model must distinguish the sector portrayed as still resilient (B: healthcare and leisure/hospitality) from the other narrative roles: the software-employment downturn (A), the decline in construction employment (C), and manufacturing, construction, and trade, which the video explicitly describes as under pressure (D).

Capability organization. The three levels are distinguished by the judgments required: L1 focuses on reading data facts and chart elements; L2 on data relationships and evidential support across charts and time points; and L3 on how this information is organized into a narrative and how visual design serves communication. This design draws on the visualization literacy frameworks proposed by Boy et al. (2014) and Lee et al. (2017), extending them to cross-chart relationships and narrative context in data videos. For example, Argument Synthesis focuses on identifying key data supporting the central claim, whereas Narrative Structure focuses on the roles and organization of different segments within the overall argument.

Evidence requirements. Evidence requirements depend on the specific task. Value retrieval typically relies on the relevant chart state, cross-chart comparison requires combining data from different charts, and narrative and design judgments also require the corresponding context. Questions about animation or presentation order require examining how the visual content changes over time. Multiple frames support aggregation or comparison, while order information supports judgments about change processes and narrative organization.

Section A.2 specifies each question type's definition, target competency, evidence requirements, and distractor design. Individual types can include multiple modes: Cross-Chart Comparison covers visual comparison and arithmetic derivation; Narrative Structure covers narrative patterns, data flow, and cross-section argumentation.

A.2 QUESTION TYPE SPECIFICATION

The following gives a complete specification of all 10 question types. The *Evidence* field describes the visual information and context targeted by each question type, including relevant chart states, cross-chart evidence, and temporal or narrative relationships. Among these, Data Fact Reading, Cross-Temporal Change, Cross-Chart Comparison, and Narrative Structure each include multiple assessment modes, detailed in their respective definitions.

Level 1: Data Perception

Data Fact Reading

- *Definition:* Reading data information from charts, including four modes: (i) value retrieval, extracting specific values annotated in the chart; (ii) trend identification, determining the direction and shape of variable changes; (iii) extremum detection, locating maxima, minima, peaks, or valleys; (iv) statistical aggregation, aggregating

information across multiple frames (e.g., counting the amount of distinct chart types appearing in a video), independent of frame order.

- *Target Competency:* Reading chart data and identifying visual variables; aggregating information from multiple video frames.
- *Evidence:* Single frame / multi-frame unordered (aggregation mode).
- *Distractor Design:* At least one distractor has a value within $\pm 15\%$ of the correct answer; at least one distractor is drawn from another visible data point in the same chart; aggregation-mode options are consecutive integers ($N-1, N, N+1, N+2$).

Chart Element Recognition

- *Definition:* Recognizing basic visual elements of a chart, including chart type, axis labels, units, legends, color encodings, etc.
- *Target Competency:* Chart type identification; visual encoding comprehension.
- *Evidence:* Single frame.
- *Distractor Design:* For chart-type questions, at least one distractor is a visually similar chart family member (e.g., area chart vs. line chart, stacked bar vs. grouped bar); for axis-label questions, at least one distractor is a label from another chart in the same video.

Example. Figure 7 shows a representative L1 question (Data Fact Reading, value retrieval mode). The question asks for a specific percentage from a trend chart; the correct answer is directly grounded in a visual label at a single frame, while distractors are drawn from numerically close values in the same chart or related metrics from adjacent frames. This example illustrates how L1 questions test precise data reading without requiring temporal reasoning or cross-chart integration.

Level 2: Temporal Reasoning

Cross-Temporal Change

- *Definition:* Cross-temporal data change reasoning, including two modes: (i) change tracking, recording the direction, magnitude, or timing of data changes within the visible time window; (ii) trend extrapolation, predicting future values beyond the video's time range based on observed trends.
- *Target Competency:* Temporal change tracking; trend extrapolation.
- *Evidence:* Data values or chart states at relevant time points, together with their temporal relationships.
- *Distractor Design:* At least one distractor describes a value from only one time point rather than a temporal difference; at least one distractor gives the opposite direction of change; all options in extrapolation mode must be specific numerical values or narrow ranges, not qualitative descriptions.

Cross-Chart Comparison

- *Definition:* Comparing data relationships across charts, including two modes: (i) visual comparison, determining relative relationships among data in two or more charts by visual inspection alone; (ii) arithmetic derivation, performing arithmetic operations (ratios, products, proportions, etc.) on values from multiple sources to obtain derived values not explicitly stated in the video.
- *Target Competency:* Cross-chart integration; cross-source numerical derivation.
- *Evidence:* Relevant values and visual evidence from at least two chart segments.
- *Distractor Design:* At least one distractor is correct for one chart but incorrect when combined with another (single-chart trap); arithmetic-derivation distractors correspond to distinct computational errors (wrong input value, reversed operation, using complement, etc.); all options must agree in thematic direction, differing only in precise visual details.

Causal Reasoning

- *Definition:* Identifying causal chains jointly supported by multiple charts in the video. Each option is an independently verifiable visual observation.
- *Target Competency:* Cross-chart causal evidence identification.
- *Evidence:* Observations from at least two chart segments that support the stated causal relationship, including temporal relationships relevant to the claim.
- *Distractor Design:* Correct options must span $\geq$2 chart segments; distractors fall into two types: (A) observations not shown in the video but causally plausible, and (B) details genuinely present in the video but not directly relevant to the stated causal relationship.

Argument Synthesis

- *Definition:* Identifying the key evidence in the video's core argument. Each option is a data fact from a specific chart segment.
- *Target Competency:* Full video argument synthesis; key evidence identification.
- *Evidence:* Chart evidence across multiple narrative stages, considered in the context of the video's central argument.
- *Distractor Design:* Correct options must span $\geq$2 narrative stages; distractors fall into two types: (A) facts genuinely present in the video but serving only as supporting details rather than key evidence, and (B) evidence that contradicts the argument's direction or belongs to a different claim.

Example. Figure 8 shows a representative L2 question (Argument Synthesis). The video argues that geopolitical headlines sparked a late-day market recovery; the model must select all essential chart observations (from multiple chart segments at different timestamps) that jointly support this argument. Correct options span at least two distinct chart segments, while distractors include facts genuinely present in the video but irrelevant to the stated causal claim. This example shows that Argument Synthesis requires integrating chart evidence from multiple charts at different temporal positions, going beyond single-frame perception.

Level 3: Narrative Understanding

Narrative Structure

- *Definition:* Understanding the video's narrative structure, including three modes: (i) pattern recognition, identifying narrative patterns from the design space of Yang et al. (2022); (ii) data flow comprehension, understanding the organizational order and presentation strategy of data facts; (iii) cross section argumentation, integrating data evidence from two or more non-adjacent argumentative segments to understand the video's overall argument structure.
- *Target Competency:* Narrative structure comprehension; cross section synthesis.
- *Evidence:* Multiple narrative stages or argument sections, including their presentation sequence for narrative organization and their evidence relationships for cross-section argumentation.
- *Distractor Design:* For pattern recognition and data flow modes, all options must be genuine categories from the same stage of the data stories design space (Yang et al., 2022), which organizes 14 narrative patterns and their associated data flows across the Setting, Rising-Climax, and Resolution stages (for example, Statistic Hook in the Setting stage, Showing Contrast in the Rising-Climax stage, and Recap in the Resolution stage); cross section distractors are over-generalizations of a single section's claim or misattributions of argumentative structure.

Visual Communication Intent

- *Definition:* Understanding how chart design choices and animation strategies serve data narrative intent.

- *Target Competency:* Design intent comprehension; visual mechanism attribution.
- *Evidence:* Relevant chart-design features and their narrative context, including the animation sequence for questions about animation strategies.
- *Distractor Design:* Options are drawn from the visual communication categories of the same design space; for questions about animation strategies, options are drawn from the six categories of animation techniques, namely entrance, emphasis, exit, transition, data-driven, and temporal animation (Shi et al., 2021). At least one distractor correctly identifies the visual mechanism but attributes it to the wrong narrative purpose (correct mechanism, wrong intent); at least one distractor correctly identifies the narrative purpose but attributes it to the wrong visual mechanism (correct intent, wrong mechanism).

Counterfactual Analysis

- *Definition:* Reasoning about the impact on communication goals when a specific design choice is changed.
- *Target Competency:* Design counterfactual reasoning.
- *Evidence:* The original design and its communication context, including temporal sequences when the proposed change concerns animation or narrative order.
- *Distractor Design:* Alternative designs must be genuinely feasible, not "straw man" proposals; extreme options such as "completely incomprehensible" or "no difference at all" are prohibited; the number of correct answers is not fixed to prevent models from exploiting a fixed answer count.

Quality Evaluation

- *Definition:* Making holistic quality assessments of the video's data presentation, with three variants: the Fidelity variant evaluates data presentation faithfulness; the Expressiveness variant evaluates information communication effectiveness; the Aesthetics variant evaluates visual design quality.
- *Target Competency:* Holistic assessment of fidelity, expressiveness, and aesthetics.
- *Evidence:* Chart content and visual presentation across the relevant video segments; full-video context for the Expressiveness variant.
- *Distractor Design:* All options share the same positive tone, differing only in the specific visual mechanisms identified; extreme wordings such as "severe distortion," "complete failure," or "flawless" are prohibited; distractors must point to visual elements genuinely present in the video, not fabricated issues. When a question targets potentially misleading design, the mechanisms considered are drawn from visually detectable misleader categories such as truncated axes, unconventional scale directions, and cherry-picked time ranges (Ge et al., 2023), and are used only when the issue is genuinely visible in the video.

Example. Figure 9 shows a representative L3 question (Narrative Structure, cross-section argumentation mode). The video discusses "pockets of recession" across multiple labor-market segments presented in non-adjacent sections; the model must synthesize information from these dispersed segments to identify which sector the video characterizes as "resilient." Distractors correspond to sectors that are genuinely discussed but serve a different narrative role (sectors under pressure rather than resilient ones). This example highlights how L3 questions demand understanding the creator's narrative design and integrating thematic roles across the full video structure.

B Dataset Construction and Statistics

B.1 Video Collection Pipeline

We collect videos from YouTube through two complementary strategies: channel-level crawling and keyword search. Channel-level crawling covers professional media (Bloomberg, Financial Times, The Economist, WSJ), data journalism organizations (Vox, Our World in Data, Pew Research Center),

educational channels (Kurzgesagt, TED-Ed, Khan Academy, 3Blue1Brown), and independent data visualization creators. Keyword search uses multiple query groups (e.g., "data-driven story explanation chart," "climate change graph explained," "election results map breakdown," "public health statistics video") spanning all five topic domains to supplement long-tail content that channel crawling cannot easily reach.

Candidate videos undergo automated filtering (duration range, upload date threshold) and deduplication, followed by round-robin sampling by provisional topic tags. Each video is registered with structured metadata upon ingestion, including upload date, view count, channel information, YouTube category tags, duration, and available subtitle languages, supporting the contamination analysis (Appendix E.9) and statistical analyses below.

AI screening and human verification. Data video identification follows a two-stage pipeline of AI pre-screening and human verification. In the AI pre-screening stage, we feed video frames, subtitle text, and metadata (title, channel, duration, tags) into Gemini-3-Flash, which judges whether the video qualifies as a data video according to the three inclusion criteria in Section 3.1 and outputs structured annotations: if the judgment is negative, the model provides a specific rejection reason (e.g., no visualization, decorative only, live-action dominant, narration dominant); if positive, it further outputs identified visualization types, narrative stage annotations, language prior risk assessment, and key timestamps supporting the judgment. Annotators watch each video in the review interface (Figure 10), verify the AI classification, and correct it where necessary.

The complete screening prompt is shown below:

AI Screening Prompt for Data Video Identification

You are an expert annotator for the DataVista data-video benchmark. Your task is to analyze a YouTube video and return a single JSON object with EXACTLY the schema below.
Definition of "data video": A video is a DATA VIDEO if and only if ALL three conditions hold:

1. It contains at least one visible data visualization (chart, graph, statistical map, data table, etc.) backed by real quantitative data.
2. The visualization is CENTRAL and SUSTAINED—it must drive the narrative for a meaningful portion of the video, not appear as a brief location marker, evidence overlay, or decorative insert (<5 s total is insufficient).
3. The primary storytelling mode is DATA-DRIVEN: the charts/graphs are what the audience is meant to read and reason about.

CRITICAL rejection criteria — mark "no" if ANY of the following is true:

- The video is mainly live-action footage, interviews, or on-the-ground reporting, even if it briefly shows maps or text overlays to locate events.
- The only "visualizations" are annotated photos, satellite imagery markups, crime-scene diagrams, or location/timeline graphics used as EVIDENCE in investigative journalism (not as quantitative data presentations).
- The main insight is delivered by narration/voiceover; charts merely illustrate what is already stated in words (narration-dominant).
- The video uses infographic-style text cards or flow diagrams to explain a PROCESS or POLICY (not to present statistical data trends or comparisons).
- The visualization count is 1–2 and they appear only briefly.

Output schema (abbreviated; full schema includes all fields below):

- `is_data_video`: "yes" | "no"
- `rejection_reason` (if "no"): one of {no_visualization, decorative_only, live_action_only, investigative_journalism, narration_dominant, process_explainer, advertisement, pure_text_animation, other}
- `visualization_types`: list from 18 normalized types (bar_chart, line_chart, area_chart, pie_donut_chart, scatter_plot, bubble_chart, map_choropleth,

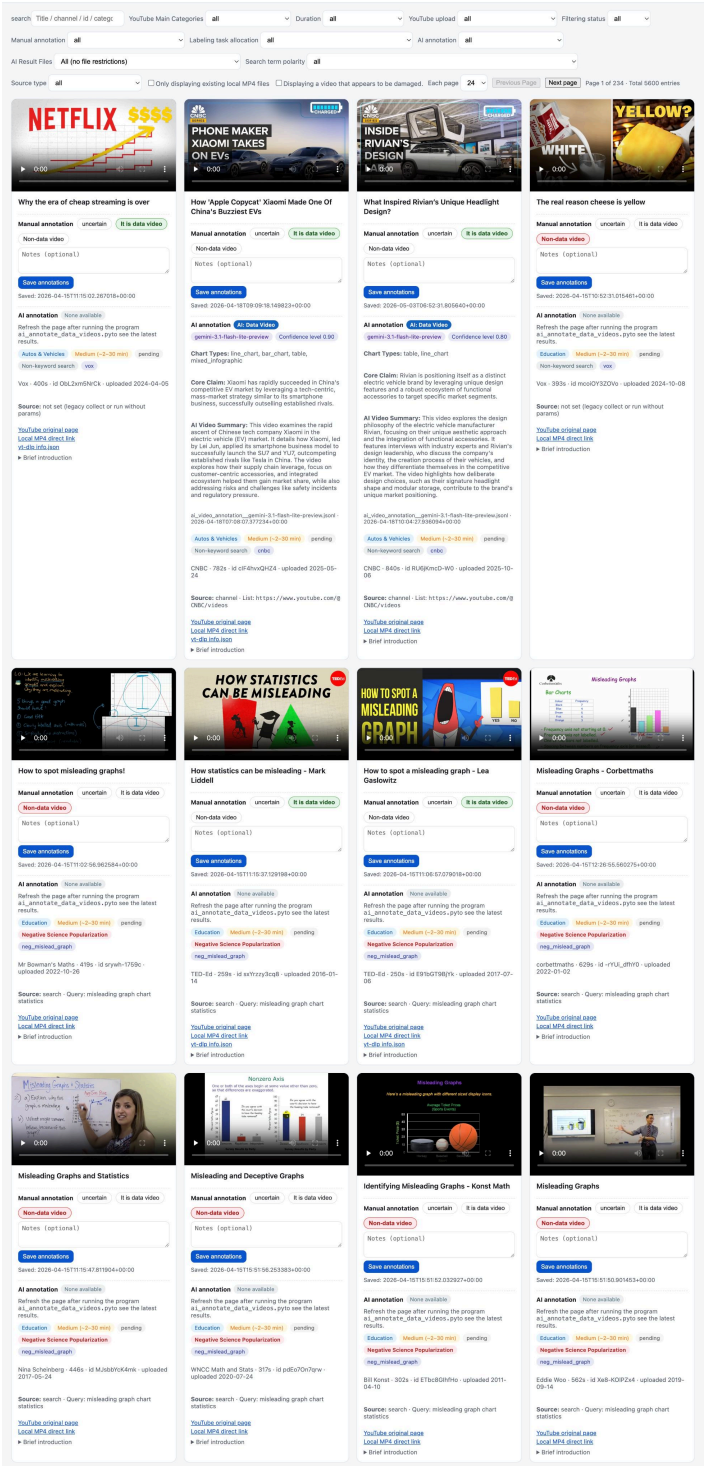


Figure 10: The human review interface for video screening. Annotators can watch the video, view AI-generated annotations, and confirm or correct the data video classification.

map_symbol, treemap, heatmap, table, flow_diagram, network_graph, timeline_visual, pictograph_isotype, annotated_image, mixed_infographic, other_chart)

- `visualization_count`: "1-2" | "3-5" | "6-10" | "10+"
- `visualization_dynamic`: "static" | "animated" | "mixed"
- `narrative_setting`, `narrative_rising_climax`, `narrative_resolution`: labels from a predefined narrative design space (e.g., statistic_hook, showing_contrast, recap)
- `data_flow`: primary data-flow strategies (e.g., comparison, ranking, time_series_progression, cause_and_effect)
- `visual_communication`: main visual-communication intents (e.g., emphasize, compare, reveal, summarize)
- `language_prior_risk`: "low" | "medium" | "high" — estimates how much key data insight can be obtained from transcript alone
- `evidence_timestamps`: 2–4 key timestamps with brief notes supporting the classification
- `confidence`: 0.0–1.0

The model receives the video (base64-encoded), the subtitle transcript when one is available, and metadata (title, channel, duration, tags) as input.

B.2 Structured Pre-extraction

Videos that pass screening enter structured pre-extraction, which compiles a timestamped evidence index of each retained video to support question design. Gemini-3-Flash receives the full video and returns chart segments, narrative stages, argument chapters, and a video digest, with exact time fields required throughout so that annotators can jump to the corresponding segment in the review interface.

Structured Pre-extraction Prompt

You are preparing an evidence map for the DataVista data-video benchmark. Watch the full video and return a single JSON object with EXACTLY the fields below. Enumerate every distinct data visualization, partition the video into narrative stages, and surface the moments worth examining. Timestamps must be exact.

Output schema:

- `video_description`: a 2–4 sentence summary of the entire video.
- `chart_segments`: every distinct data visualization in the video, each with a stable `id` (`c1`, `c2`, ...), a `label` describing the chart form and its content, a `question_reference` phrase safe to reuse in question stems, and a `time_range` in `MM:SS-MM:SS`.
- `narrative_stages`: the time ranges of `setting`, `rising_climax`, and `resolution`, together with a content-specific reference phrase for each stage.
- `argument_chapters`: 3–8 data-argument units subdividing the narrative stages, each with a `time_range`, its parent `stage`, a `question_reference` phrase, and the `key_chart_ids` anchoring that unit. Required when the video contains at least four distinct chart segments or exceeds three minutes.
- `video_digest`: a reviewer-facing briefing containing `distinct_viz_count` and a one-line inventory of the chart families present; `quiz_hooks`, listing examinable moments with a `time_hint`, the associated `chart_ids`, and `suggested_levels`; `cross_modal_hooks`, recording moments where a narration claim can be compared against the value the chart actually shows; and `stage_notes`, summarizing the visualization activity within each stage.

Table 2: Text-only screening results by capability level.

Level	Screened	Filtered	Sent to review	Filtered (%)
L1	9,306	6,369	2,937	68.4
L2	9,876	7,845	2,031	79.4
L3	9,715	7,590	2,125	78.1
Total	28,897	21,804	7,093	75.5

Consistency requirements: `distinct_viz_count` must equal the length of `chart_segments`; every `chart_ids` entry must reference a declared segment id; `video_digest` must remain consistent with the segments, stages, and chapters.

Annotators then verify this output against the video segment by segment, correcting inaccurate time ranges and chart descriptions and supplementing omitted visualizations, before the metadata is used for question design.

B.3 Question Screening and Quality Review

B.3.1 Text-Only Filtering

Screening rule. We use Gemini-3-Flash for single-round text-only answering of candidate questions. The input contains question stems and options, without video frames, subtitles, audio, or reference answers. Predictions are matched against the reference letter for single-answer questions and the complete reference set for multiple-select questions. Matching candidates are filtered out; the rest proceed to human review.

Text-Only Screening Prompt

You are an answering module for benchmark screening. You are given a list of multiple-choice questions with NO video or image input. Answer each question using only the question text and options.

Hard constraints:

- Output ONLY a JSON array of answer strings, one per question, in the same order as the input.
- For single-choice questions: one uppercase letter (A/B/C/D).
- For multi-select questions: all applicable uppercase letters in alphabetical order, comma-separated, no spaces (e.g., "A,C"). The number of letters is not fixed.
- No explanation, no extra fields, no markdown. Just the JSON array.

Example output for 3 questions: ["B", "A,C", "D"]

Questions:
Q1. {question stem}
A. {option A} B. {option B} C. {option C} D. {option D}

...

Output (JSON array only):

Screening statistics. Of the 28,897 screened candidates, 21,804 (75.5%) are filtered out and 7,093 proceed to human review. Table 2 reports the results by capability level.

Table 3: Text-only accuracy (%) on retained and filtered-out questions. Differences are filtered-out minus retained, in percentage points, with 95% video-cluster bootstrap confidence intervals.

Model	Retained	Filtered out	Difference [95% CI]
Gemini-3.1-Pro	55.6	85.6	+30.0 [25.2, 35.1]
GPT-5	51.5	82.0	+30.5 [23.6, 37.5]
Claude-Sonnet-4.6	43.9	76.0	+32.2 [24.9, 39.5]
Qwen3-VL-32B-Instruct	36.5	60.0	+23.4 [15.0, 31.8]
InternVL3.5-38B	42.5	65.4	+22.9 [16.6, 29.3]

Cross-model filtering analysis. To examine whether the filtering effect extends to other models, we sample equally sized sets of retained and filtered-out questions. We prioritize pairs from the same video with matching capability level, question type, and answer format, relaxing these criteria in a predefined order when necessary. Gemini-3.1-Pro, GPT-5, Claude-Sonnet-4.6, Qwen3-VL-32B-Instruct, and InternVL3.5-38B receive only question stems and options and are scored by exact match. We estimate 95% confidence intervals for the filtered-out-minus-retained accuracy differences using 10,000 video-cluster bootstrap resamples.

As shown in Table 3, text-only accuracy is 22.9–32.2 percentage points higher on filtered-out questions across all five models. This consistent difference indicates that the filtered-out questions are more readily answerable from question stems and options across model families.

B.3.2 Human Review

Reviewers and procedure. Four reviewers who did not participate in question design review the candidates retained after text-only filtering. Tasks are assigned by video, and each question is checked separately by two reviewers against the original video for clear wording, video-supported reference answers, and plausible distractors.

Level-specific criteria. L1 checks numerical values, units, and the corresponding data entities. L2 checks the completeness of evidence across charts and time points, and the validity of comparisons, calculations, and support relations. L3 checks the evidence underlying narrative and visual-design judgments, including whether alternative answers are equally reasonable.

Disagreement resolution and outcomes. Reviewers resolve disagreements through discussion grounded in the original video evidence. Questions requiring revision are edited and their answers and options rechecked; those with unresolved ambiguity are removed. Among candidates retained after text-only filtering, 80.1% pass review, 15.4% pass after manual revision, and 4.5% are rejected.

B.4 Video Source Distribution

We summarize the most prolific source channels within each topic domain. In the Economy domain, CNBC is the largest contributor (153 videos), followed by Half as Interesting (18) and Financial Times (16). The Society domain draws primarily from Pew Research Center (53), Happy Statistics (15), and Our World in Data (6). The Science domain is mainly from popular science channels such as Two Minute Papers and Kurzgesagt – In a Nutshell. The Politics domain concentrates on Vox (73) and Bloomberg Television (19), while the Culture domain features independent creator channels focused on statistics and rankings, such as World Statistic and BarChart Race. Overall, the source channels span four broad categories: professional media outlets, research institutions, educational science-communication channels, and independent data creators, ensuring diversity in both content expertise and visual style.

B.5 Topic and Subcategory Distribution

Table 4 shows the two-level topic taxonomy of DataVista, comprising five top-level domains further divided into 22 subcategories. The video counts for the five domains are: Economy (404), Culture (173), Politics (152), Society (145), and Science (87). Economy is the largest domain,

Table 4: Distribution of 5 major themes and 22 subclass topics in DATAVISTA.

Theme	Subclass Topic	Videos	Pct.(%)
Economy (404)	Market Dashboard	177	18.4
	Macro Comparison	115	12.0
	Earnings & Equities	91	9.5
	Ranking & Country Race	16	1.7
	Other	5	0.5
Culture (173)	Ranking Race	105	10.9
	Visual Storytelling	34	3.5
	Sports Analytics	15	1.6
	Film & Entertainment	13	1.4
	Other	6	0.6
Politics (152)	Narrative Explainer	77	8.0
	Mainstream News	39	4.1
	Other	32	3.3
	Event & Governance	4	0.4
Society (145)	Demographic Dynamics	67	7.0
	Survey & Public Opinion	52	5.4
	Relational Patterns	11	1.1
	Geo-social Analysis	9	0.9
	Other	6	0.6
Science (87)	Science Communication	75	7.8
	Educational Explainer	5	0.5
	Geospatial Analysis	7	0.7
Total		**961**	**100.0**

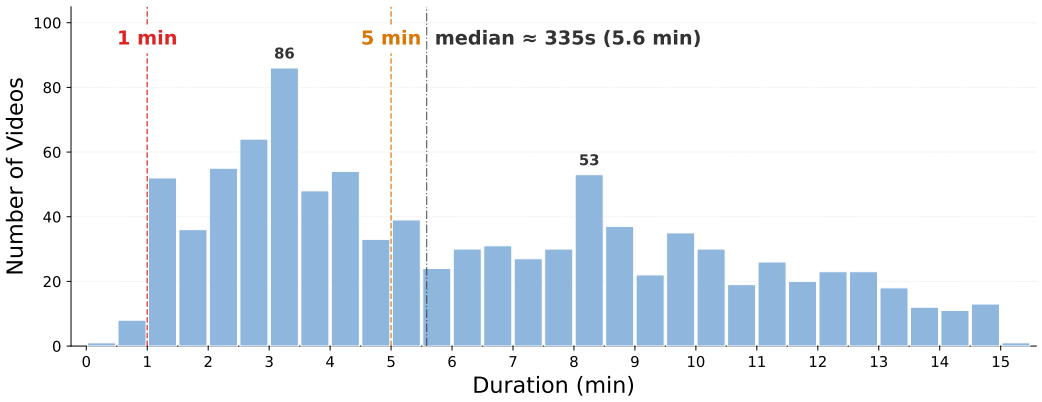


Figure 11: Distribution of video durations (30-second bins).

accounting for 42.0% of videos. This is consistent with economic reporting's emphasis on quantitative indicators, trends, and comparisons, which lends itself to video narratives organized around data visualizations (Bu et al., 2026). At the subcategory level, *Econ: Market Dashboard* (177), *Econ: Macro Comparison* (115), and *Culture: Ranking Race* (105) are the three largest subcategories, whereas *Politics: Event & Governance* (4) and *Science: Educational Explainer* (5) are among the smallest. Per-domain evaluation and equally weighted domain results are provided in Appendix E.1.

B.6 VIDEO DURATION DISTRIBUTION

Figure 11 shows the histogram of video durations (binned at 30-second intervals). The median duration is approximately 335 seconds (≈5.6 minutes), with about 45.1% of videos falling within the 60–300 second (1–5 minute) range. The shortest video is 28 seconds and the longest is 900 seconds

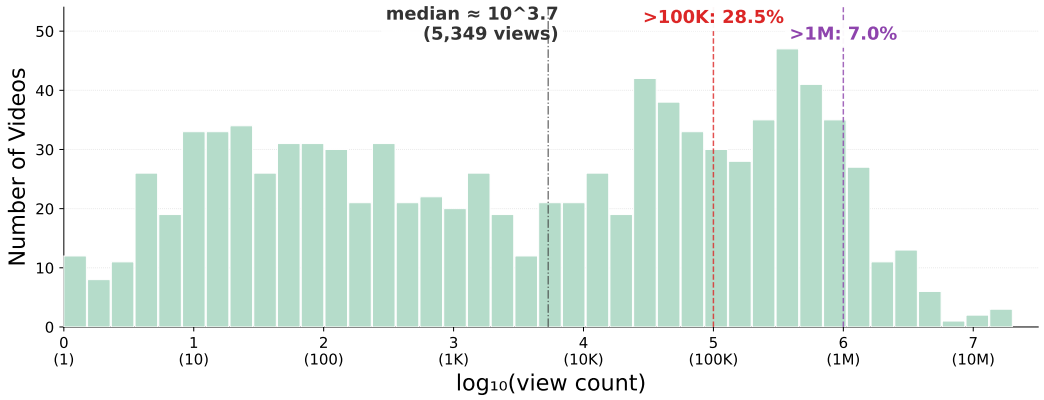


Figure 12: Distribution of video view counts ($\log_{10}$ scale).

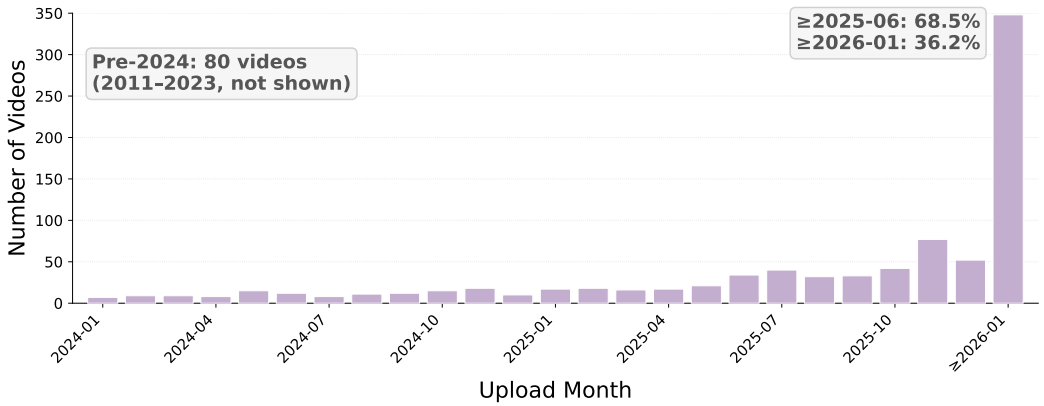


Figure 13: Monthly distribution of video upload dates.

(15 minutes). The distribution is not unimodal and symmetric: the primary peak is located at the 180–210 second (3–3.5 minute) bin, but secondary bumps appear near 240 seconds and 480 seconds. This reflects the coexistence of two narrative modes: short commentary clips and medium-length explanatory data videos. This distribution is consistent with observations by Amini et al. (2015) and Yang et al. (2022) regarding the typical narrative density of professional data videos.

B.7 Video View Count Distribution

Figure 12 presents the $\log_{10}$-scaled histogram of video view counts (40 equal-width bins; 945 videos have view count records). The median view count is approximately 5,349. About 28.5% of videos have more than 100K views, and approximately 7.0% exceed 1M views. The distribution exhibits a bimodal pattern: the first peak is located near $\log_{10} \approx 1.6$ (≈40 views), corresponding to recently collected videos still in their cold-start phase; the second, broader bump spans $\log_{10} \approx 2$–5 (thousands to hundreds of thousands), corresponding to the typical reach of mainstream professional media content. This bimodal structure corroborates DataVista's simultaneous attention to both *timeliness* and *influence* during data collection.

B.8 Publication Date Distribution

Figure 13 shows the monthly distribution of video publication dates. Approximately 68.5% of videos were published in or after June 2025; by calendar year, 399 videos (41.5%) were collected from 2025 and 348 videos (36.2%) were published in or after January 2026. A ramp-up trend is visible

Table 5: Distribution of chart types across 961 videos (2,210 total entries; ≈2.30 per video). **Count** is the occurrence count for each type; **%** is its share of all entries. "Other" subsumes miscellaneous and unclassified types.

Chart Type	Count	%	Chart Type	Count	%	Chart Type	Count	%
Bar Chart	624	28.2	Pictograph / Isotype	59	2.7	Bubble Chart	17	0.8
Line Chart	409	18.5	Flow Diagram	59	2.7	Other	16	0.7
Infographic	266	12.0	Timeline	56	2.5	Heatmap	15	0.7
Table	228	10.3	Annotated Image	42	1.9	Treemap	11	0.5
Pie / Donut	112	5.1	Scatter Plot	33	1.5	Network Graph	6	0.3
Choropleth Map	99	4.5	Area Chart	26	1.2			
Symbol Map	96	4.3	Unknown	36	1.6			
						Total	**2,210**	**100**

from mid-2024 through the beginning of 2026, reflecting the dataset's deliberate strategy of skewing toward more recent content to reduce data contamination risk(see Appendix E.9 for details).

B.9 Chart Type Distribution

Table 5 summarizes the distribution of chart types. We annotated each video with structured visualization type labels; videos containing multiple chart types receive separate entries for each type, so the total number of entries can exceed the number of videos. A total of 18 normalized chart types were identified, yielding 2,210 entries across 961 videos (an average of ≈2.30 types per video). The distribution is highly concentrated: bar charts appear in at least 64.9% of videos (624 entries) and line charts in about 42.6% (409 entries); bar, line, and area chart entries together account for approximately 47.9% of all entries. Infographics (≈12.0% of entries), tables (≈10.3%), and two map variants (choropleth maps at ≈10.3% and symbol maps at ≈10.0%, both measured by video coverage rate) also appear frequently. Pie/donut charts, timelines, pictograms, and flow diagrams each fall in the moderate range. Scatter plots, bubble charts, heatmaps, treemaps, and network diagrams belong to the long tail, collectively covering fewer than about 10% of videos when counted by whether at least one such type appears, yet they serve as important probes for model comprehension on less conventional chart types. This distribution aligns well with the content characteristics of professional data videos, which emphasize temporal and categorical comparisons.

B.10 Question Count Distribution

After text-only filtering and human review, DataVista retains 6,775 questions covering 961 videos. The median number of questions per video is 7, with a mean of 7.0, a range of 1–19, and first and third quartiles of 5 and 9. Videos with 5–10 questions account for 63.1% of the dataset; those with 1–4, 11–15, and at least 16 questions account for 22.3%, 13.4%, and 1.2%, respectively.

L1, L2, and L3 contain 2,817, 1,908, and 2,050 questions, respectively, with per-video medians of 3, 2, and 2. The three levels cover 93.5%, 83.4%, and 88.4% of videos, respectively, so most videos have questions at each level.

C Case Studies

This appendix presents representative questions from DataVista as a supplement to Section 4. The cases are organized by capability level (L1, L2, and L3), using video evidence to explain what each question requires and to analyze differences in the answers of representative models.

Each case presents the question and options, timestamped key frames, the basis for the answer, and model responses side by side, with correct answers highlighted in green. The analysis connects evidence to responses: we explain how the key frames support the answer, compare correct selections, omissions, and incorrect choices, and discuss how these differences affect numerical judgments, argument completeness, and design understanding.

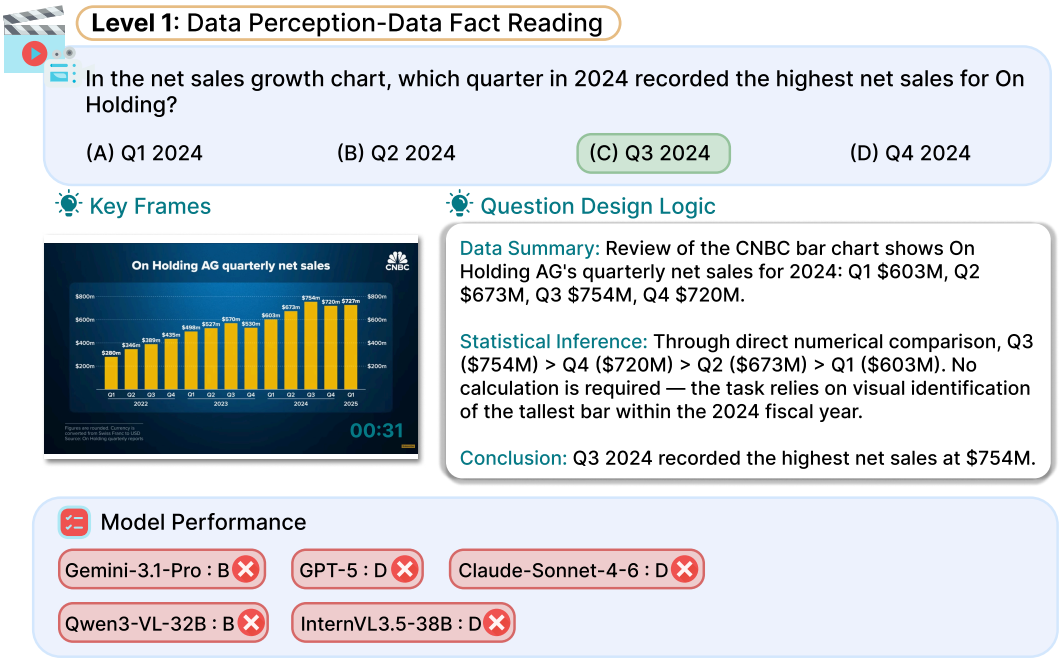


Figure 14: An example L1 question in DATAVISTA (Data Fact Reading)

C.1 LEVEL 1: DATA PERCEPTION

This subsection examines model responses on numerical comparison and visual encoding recognition through quarterly peak identification and legend mapping. The cases require comparing quarterly sales within a specified year and identifying the meaning of a line from its color and style, respectively. We analyze response differences using the displayed values, visual marks, and model selections.

Case 1: Quarterly Peak Identification (L1, Data Fact Reading). As shown in Figure 14, the question asks which quarter in 2024 recorded the highest net sales for On Holding. The correct answer (C, Q3 2024) can be determined by visual comparison of bar heights within the 2024 fiscal year, where Q3 ($754M$) is the tallest bar. However, all five representative models answer incorrectly: Gemini-3.1-Pro and Qwen3-VL-32B select Q2, while GPT-5, Claude-Sonnet-4.6, and InternVL3.5-38B select Q4. The fourth quarter, selected by three models, records sales of $720M$, approximately 4.5% below the third quarter. This case illustrates that even a basic task such as identifying an annual peak can involve difficulty distinguishing closely valued alternatives, with models selecting the second-highest quarter as the highest.

Case 2: Legend-Element Mapping (L1, Chart Element Recognition). As shown in Figure 15, the question asks what the solid light-blue line represents on a federal funds rate projection chart. The correct answer (C, the historical lower limit) is directly readable from the chart legend, which explicitly maps the solid light-blue line to "Lower limit." Only Gemini-3.1-Pro answers correctly; all other models select B (the projected median rate), assigning the meaning of the grey projection dots to the solid light-blue line. Although the legend explicitly distinguishes the two types of marks, the four models fail to match the specified line to its meaning, revealing difficulty in recognizing the mapping between color, mark shape, and data meaning.

C.2 LEVEL 2: TEMPORAL REASONING

This subsection uses argument synthesis and cross-temporal arithmetic cases to illustrate the requirements of cross-segment evidence integration in L2 temporal reasoning and the corresponding model performance. Argument synthesis requires selecting all key facts from different charts that support the central claim, while cross-temporal arithmetic requires reading data at specified times

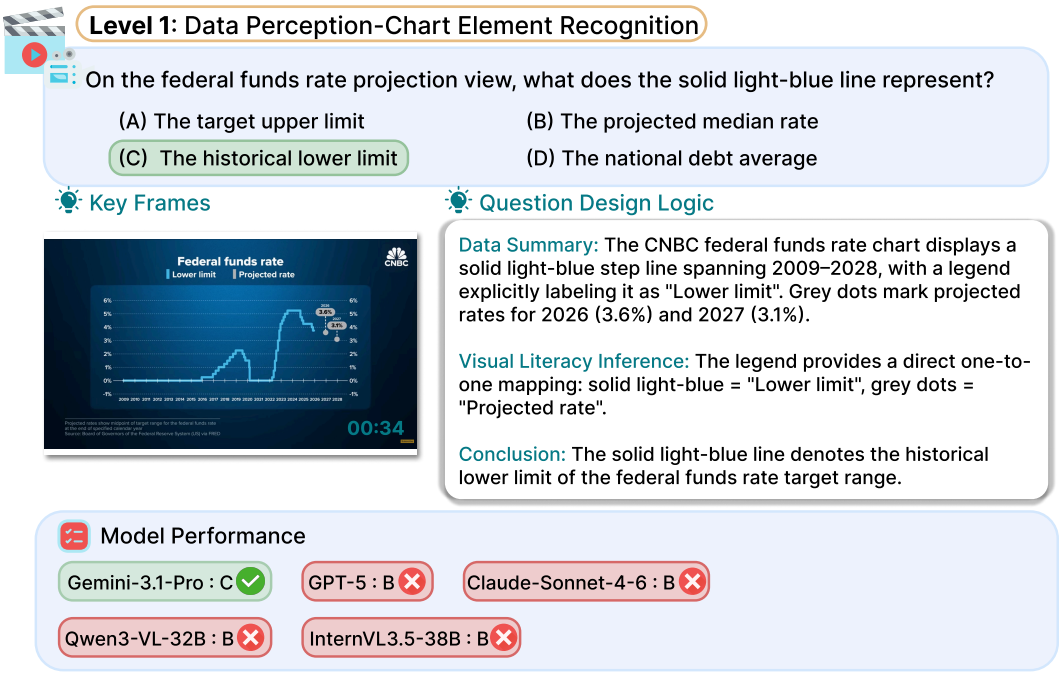


Figure 15: An example L1 question in DATAVISTA (Chart Element Recognition).

and calculating the change. The case analyses examine whether models select the complete set of supporting evidence and correctly perform numerical comparisons across time, respectively.

Case 1: Cross-Chart Evidence Integration (L2, Argument Synthesis). Figure 16 shows the video argues that remittances have become a "default pathway" for Nepal's economy. The model must select all chart-specific data facts that are essential to supporting this argument from across three non-adjacent segments: a geographic map (03:37), an age demographics chart (04:27), and a remittance GDP-share comparison (07:03). The correct answer (B, C, D) requires synthesizing structural push factors (young workforce, landlocked terrain) with the economic outcome (remittance dominance), while option A (a Discord poll result) is factually present but irrelevant to the causal chain. GPT-5 and Gemini-3.1-Pro correctly identify all three essential pieces; the other three models select B and C but omit D. Options B and C concern the population structure and the economic importance of remittances, while D adds the geographic constraints on domestic industrial development. All three answers omit the explanatory link concerning limited domestic development opportunities, leaving the selected evidence incomplete in explaining why remittances have become a "default pathway" for Nepal's economy.

Case 2: Cross-Temporal Arithmetic (L2, Cross-Temporal Change). As shown in Figure 17, the question asks for the total increase in box office earnings between two specific milestones in an animated ranking. The correct answer (D, $274.85M$) requires subtracting the value at the first milestone ($122.53M$ at 01:13) from the peak value ($397.38M$ at 01:39). Only Gemini-3.1-Pro answers correctly; the remaining four models all select B ($294.85M$). This shared incorrect answer exceeds the correct increase by $20M$. The question combines basic subtraction with matching data across time: models must locate two specified moments in a continuously updated ranking, read the respective values, and calculate the change. All four models fail to obtain the correct result on this combined task, illustrating that basic numerical operations in dynamic videos also depend on accurate temporal localization and data reading.

C.3 LEVEL 3: NARRATIVE UNDERSTANDING

This subsection uses cases of visual communication intent, counterfactual design, and narrative structure to illustrate the task requirements and model performance in L3 narrative understanding.

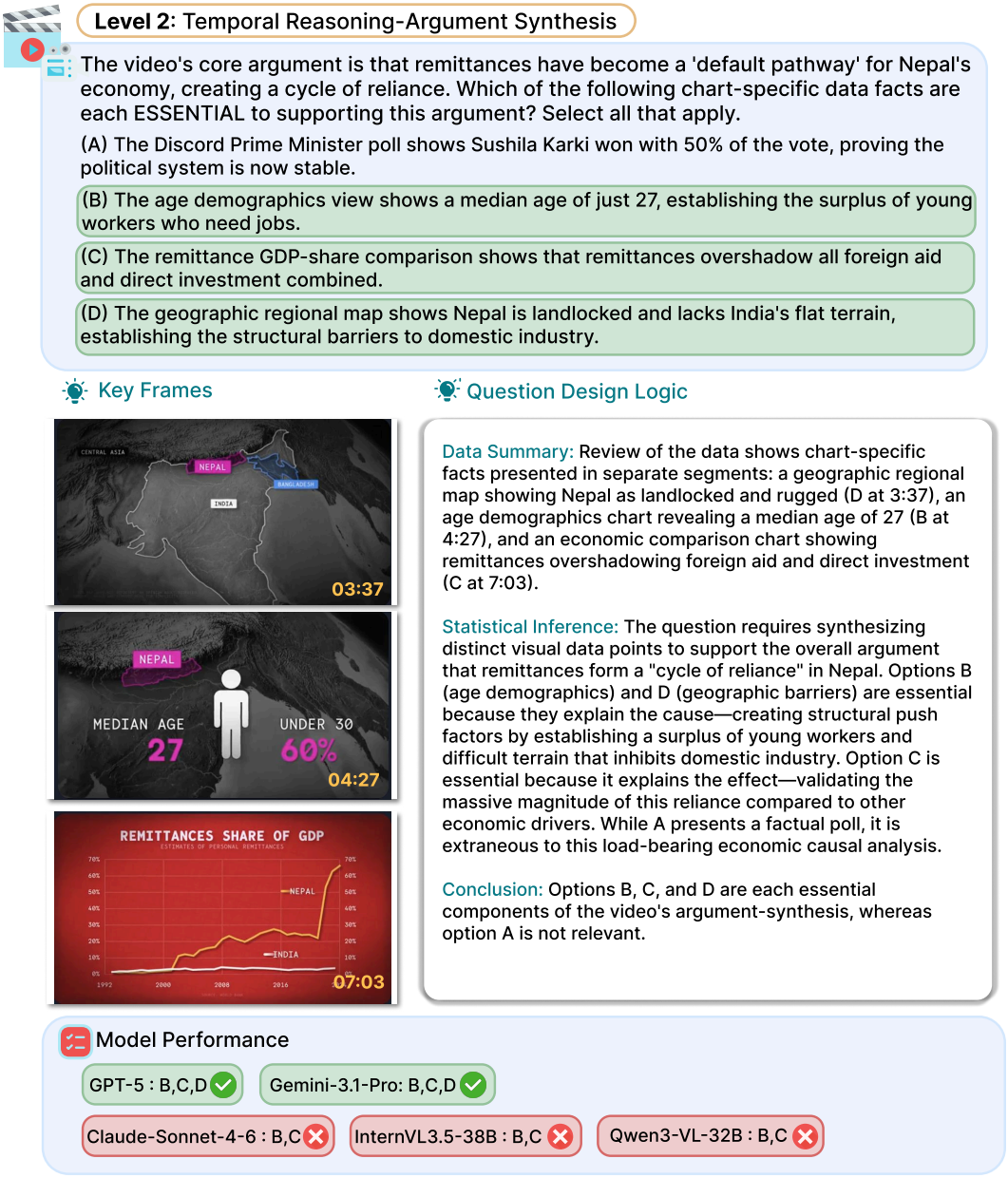


Figure 16: An example L2 question in DATAVISTA (Argument Synthesis).

The three cases examine whether models can identify the visual techniques used in a video, assess the benefits and costs of changing where information is presented, and understand how different charts are organized within a specified segment, respectively. The analyses focus on which design elements models omit or misidentify and how these answer discrepancies affect their interpretation of information emphasis, narrative pacing, and argument structure.

Case 1: Visual Communication Technique Identification (L3, Visual Communication Intent). As shown in Figure 18, the question asks which visual communication techniques the video uses to emphasize demographic statistics. The correct answer (A, D) involves recognizing full-screen bold typography for the "60%" figure and a side-by-side geographic juxtaposition of Nepal and Bhutan populations. These techniques communicate information by emphasizing an individual value and establishing a comparison between countries, respectively. Only Gemini-3.1-Pro correctly identifies both techniques. GPT-5 selects only A, omitting the juxtaposition in D; Claude-Sonnet-4.6,

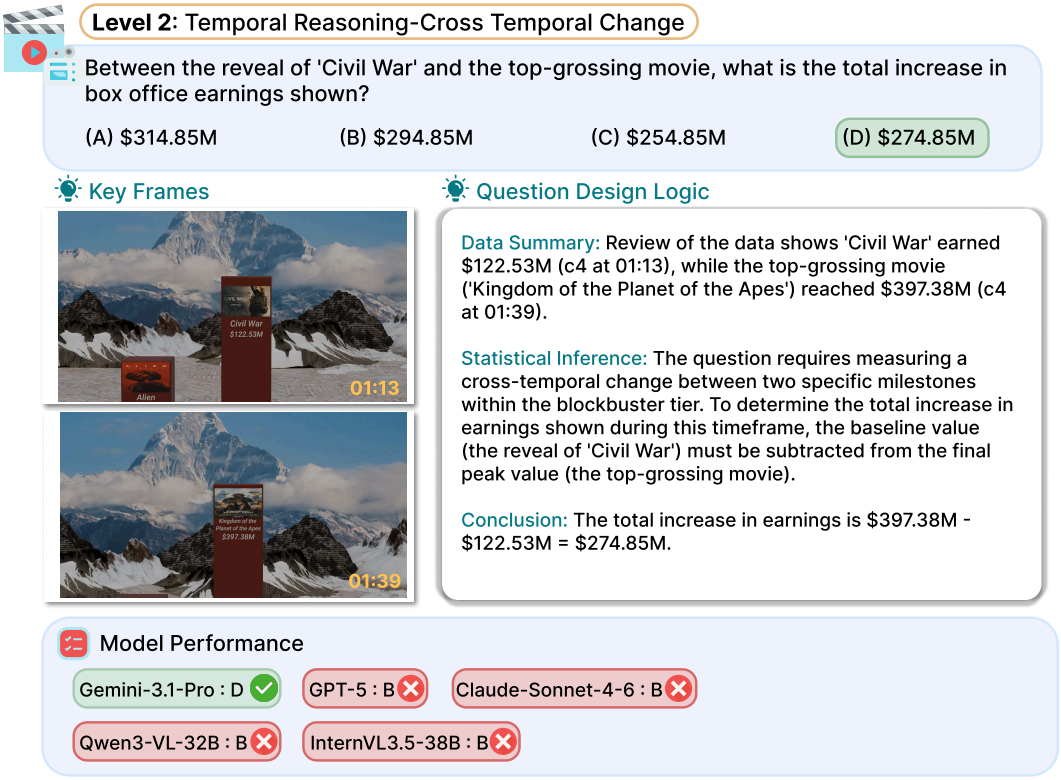


Figure 17: An example L2 question in DATAVISTA (Cross-Temporal Change).

Qwen3-VL-32B, and InternVL3.5-38B select A and B, both omitting D and incorrectly selecting the cut-out people and icons in B. However, the job categories are presented through bar charts and text labels, without these human figures. All models identify the emphasis on an individual value, but all four incorrect answers omit the comparison established through spatial juxtaposition, and three also attribute absent graphical elements to the video design. These responses reflect incomplete understanding of visual communication: they miss the comparative relationship conveyed by the layout and include design judgments unsupported by the displayed content.

Case 2: Counterfactual Design Reasoning (L3, Counterfactual Analysis). As shown in Figure 19, this multiple-select question asks how the narrative impact would change if a Hepatitis B risk table were shown alongside the opening vaccination schedule instead of later. The correct answers are A, B, and D: presenting risk data earlier helps explain vaccination at birth and makes the schedule easier to understand, but may also interrupt the opening overview and increase cognitive load. No model gets this fully correct: Gemini-3.1-Pro and Claude-Sonnet-4.6 select only A, while the others select A and B but miss D. All five models recognize the explanatory role of the risk data but omit the potential information overload from introducing clinical details earlier. This shared omission indicates that their assessments of the design change in this case emphasize local explanatory benefits without fully accounting for its effect on the overall narrative pacing.

Case 3: Organizing Evidence Across Charts (L3, Narrative Structure). As shown in Figure 20, the question asks which descriptions capture how the video organizes the visual evidence in the House and Senate control-math section. The correct answer is A, B, C. A corresponds to the two seat graphics: 35 Senate seats are up for re-election, and the House stands at 214 Democrats to 218 Republicans. B corresponds to target states marked successively on the map: Maine, North Carolina, and Georgia at 02:29, with Ohio and Michigan added at 02:34. C corresponds to the Cook Report table, which groups House races by competitiveness. D concerns the Texas candidates introduced at the opening and does not belong to this section. Only Gemini-3.1-Pro selects the full set A, B, C. GPT-5 and Claude-Sonnet-4.6 select A and B, omitting the Cook grouping; Qwen3-VL-32B and

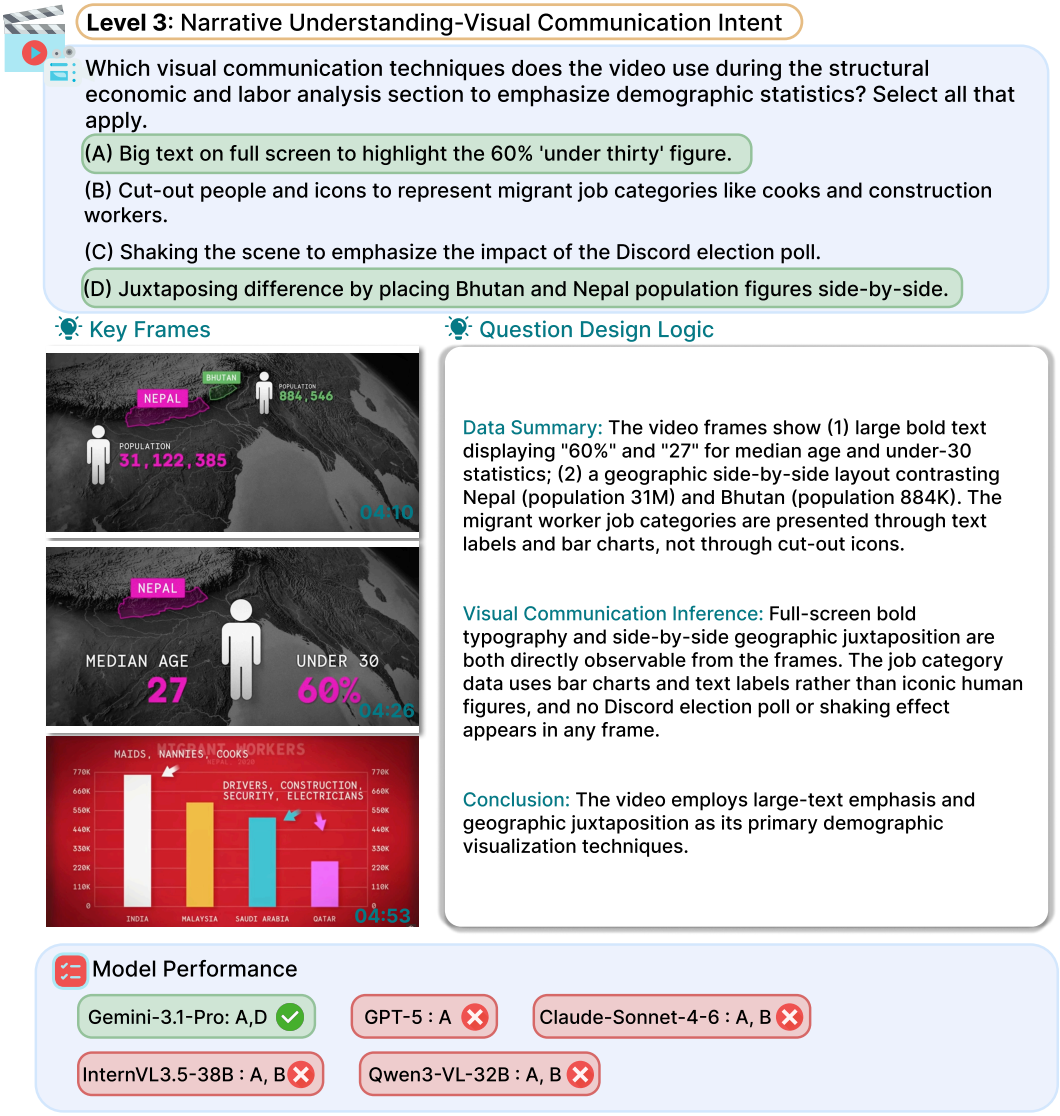


Figure 18: An example L3 question in DATAVISTA (Visual Communication Intent).

InternVL3.5-38B select A and C, omitting the states marked successively on the map. The first pair misses how the races are organized by competitiveness, while the second misses the progressive introduction of key states. All four models identify the seat comparison and exclude material from outside the section, but fail to fully recognize how the video combines comparison, progressive highlighting, and categorization to organize its account of the electoral landscape.

Level 3: Narrative Understanding-Counterfactual Analysis

If the Hepatitis B chronic infection risk table (showing 90% risk) had been shown alongside the opening childhood vaccination schedule instead of much later, how would the narrative impact most likely change? Select all that apply.

(A) It would provide immediate biological justification for the 'birth dose' shown in the schedule.

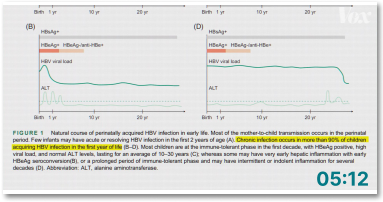
(B) It would make the opening schedule look less overwhelming by explaining one row immediately.

(C) It would weaken the later argument about general infant vulnerability.

(D) It would likely confuse the viewer by introducing clinical data before the schedule overview was finished.

Key Frames





Question Design Logic

Data Summary: Review of the two visual sources reveals: (1) the CDC 2025 childhood vaccination schedule shows Hepatitis B Dose 1 administered at birth — the earliest vaccine on the entire chart; (2) Figure from a clinical reference states that chronic infection occurs in more than 90% of children acquiring HBV infection in the first year of life.

Narrative Inference: Earlier placement would simultaneously provide immediate biological justification for the unique birth dose and reduce the schedule's perceived density by anchoring one row with a concrete rationale. However, it would also risk cognitive overload by introducing granular clinical data (viral load curves, HBeAg markers) before the viewer has finished processing the schedule overview.

Conclusion: Showing the risk data upfront strengthens comprehension of one specific vaccine row at the cost of disrupting the viewer's initial orientation to the full schedule.

Model Performance

Gemini-3.1-Pro : A ✗ | GPT-5 : A, B ✗ | Claude-Sonnet-4-6 : A ✗

Qwen3-VL-32B : A, B ✗ | InternVL3.5-38B : A, B ✗

Figure 19: An example L3 question in DATAVISTA (Counterfactual Analysis).

Level 3: Narrative Understanding-Narrative Structure

In the House and Senate control math section, which descriptioons best capture how the video organizes the visual evidence? Select all that apply.

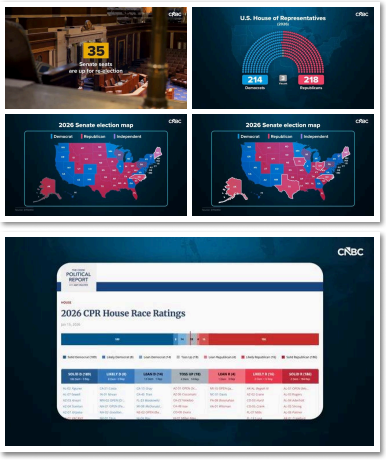
(A) It contrasts the different flip counts needed for the House and the Senate.

(B) It lists target Senate states one by one on the map.

(C) It orders House races by how competitive they are.

(D) It asks why individual candidates are running in Texas.

Key Frames



Question Design Logic

Data Summary: The House opens with a semicircular 214D/218R seat graphic, then cuts to a Cook table ranked by competitiveness with Toss Up boxed; the Senate shifts to a blue-red map, first circling ME, NC, GA at 02:29, then adding OH, MI at 02:34. The Texas candidate talk belongs to an earlier primary segment.

Narrative Inference: A captures the contrasting visual grammar across chambers; B captures the cumulative, state-by-state map highlighting; C captures the Cook table as a competitiveness ranking rather than a list. D is a distractor—why candidates run in Texas is earlier primary commentary, not a visual organizing pattern.

Conclusion: The item checks whether viewers can distinguish contrast, cumulative highlighting, and competitiveness ranking as visual rhetorics while dismissing unrelated spoken detail.

Model Performance

Gemini-3.1-Pro: A, B, C ✓

GPT-5 : A, B ✗

Claude-Sonnet-4-6 : A, B ✗

Qwen3-VL-32B : A, C ✗

InternVL3.5-38B : A, C ✗

Figure 20: An example L3 question in DATAVISTA (Narrative Structure).

D Evaluation Details

This appendix provides the detailed evaluation settings as a supplement to Section 4.

D.1 Input Organization and Prompt Templates

The main evaluation uses shared answer requirements, with questions from the same video submitted in one request. Closed-source models are queried through official APIs; open-source models use official instruct/chat checkpoints. Decoding is deterministic with temperature 0. Below, we summarize these requirements and the input organization for the conditions without and with subtitles.

Answer Requirements

Answer every question using the supplied evidence. Return only a JSON array in question order, with one object per question containing `qid`, `answer`, `reason`, and `evidence`. Copy each question ID exactly; do not skip, merge, or reorder questions.

- **Answer:** For single-choice questions, return one letter: A, B, C, or D. For multiple-select questions, return comma-separated option letters, such as A,C. Every question requires an option answer.
- **Reason:** Provide the reason supporting the answer. Connect the cited observations to the answer and include key calculations when relevant. For multiple-select questions, explain which options satisfy the question and any important exclusions. Do not repeat the question.
- **Evidence:** Provide an array of source references, each containing `source_id`, `timestamp`, and `observation`. Use only supplied source IDs and keep observations concise and specific to the cited source.

Do not use Markdown fences or add text outside the JSON array.

The condition without subtitles provides 50 uniformly sampled video frames, each labeled with a source ID and timestamp. The subtitle condition adds subtitle text with source IDs and time information to the same frames. Frames and subtitles form the evidence list, followed by the question stems, question types, and options. Both conditions use the same answer requirements. Each question is labeled in the prompt as single-choice or multiple-select.

Subtitles are obtained from YouTube. English subtitles are used for 921 videos, of which 386 have a creator-provided English track and 535 have only an automatically generated one. The remaining 40 videos have no English track and fall back to another language, most often Hindi.

Input Organization

Available evidence:
{Frame source IDs and timestamps; corresponding images are attached.}
{Subtitle source IDs, time information, and text; included only with subtitles.}

Question 1 | qid={question ID} | SINGLE-CHOICE / MULTIPLE-SELECT
{question stem}
A) {option A}
B) {option B}
C) {option C}
D) {option D}

...

Now output the JSON array of exactly N answer objects.

D.2 Random Selection and Most-Frequent Answers

We use two baselines without video content to provide random-guessing and answer-frequency references for model performance. Table 6 reports the results.

Table 6: Exact-match accuracy (%) of the statistical baselines. Random selection reports the theoretical expectation; most-frequent answers use video-level five-fold cross-validation.

Baseline	L1	L2	L3	Overall
Random selection	24.8	19.5	16.4	20.8
Most-frequent answers	30.8	18.3	31.5	27.5

Table 7: Exact-match accuracy (%) and 95% intervals.

Model	w/o Sub.	w/ Sub.	Overall
Closed-Source Models			
Gemini-3.1-Pro	68.2 [67.0, 69.7]	71.7 [70.5, 73.1]	70.0 [68.8, 71.3]
GPT-5	66.0 [64.5, 67.5]	68.8 [67.5, 70.2]	67.4 [66.1, 68.8]
Gemini-3-Flash	64.1 [62.6, 65.7]	68.5 [67.2, 70.0]	66.3 [65.0, 67.7]
GPT-4o	56.1 [54.9, 57.7]	58.0 [56.8, 59.4]	57.1 [56.0, 58.4]
Claude-Sonnet-4.6	61.3 [60.0, 62.8]	65.8 [64.5, 67.2]	63.6 [62.4, 64.9]
Open-Source Models			
Qwen3-VL-32B-Instruct	54.7 [53.4, 56.4]	59.8 [58.6, 61.4]	57.3 [56.1, 58.8]
InternVL3.5-38B	52.6 [51.3, 54.2]	55.6 [54.3, 57.2]	54.1 [52.9, 55.6]
InternVL3.5-14B	45.2 [44.0, 46.7]	49.6 [48.4, 51.2]	47.4 [46.3, 48.9]
InternVL3.5-8B	44.7 [43.5, 46.1]	47.4 [46.2, 48.8]	46.1 [44.9, 47.4]
Qwen3-VL-8B-Instruct	44.1 [42.9, 45.7]	48.2 [47.2, 49.8]	46.2 [45.1, 47.6]
InternVL3-8B	43.4 [42.2, 45.0]	46.7 [45.5, 48.4]	45.1 [43.9, 46.6]
Qwen2.5-VL-32B-Instruct	54.0 [52.6, 55.6]	55.5 [54.1, 57.0]	54.8 [53.4, 56.3]
Qwen2.5-VL-7B-Instruct	42.3 [41.1, 43.8]	45.4 [44.3, 47.0]	43.9 [42.9, 45.3]
InternVL3.5-4B	44.5 [43.1, 46.1]	40.4 [38.7, 42.3]	42.5 [41.2, 43.9]
Qwen3-VL-4B-Instruct	41.8 [40.6, 43.4]	45.4 [44.1, 47.0]	43.6 [42.4, 45.1]
Qwen2.5-VL-3B-Instruct	38.0 [36.7, 39.4]	40.8 [39.5, 42.4]	39.4 [38.3, 40.8]
InternVL3.5-2B	35.9 [34.7, 37.4]	38.5 [37.3, 40.1]	37.2 [36.2, 38.6]
Qwen3-VL-2B-Instruct	36.7 [35.5, 38.2]	40.0 [38.8, 41.6]	38.4 [37.2, 39.8]
InternVL3-2B	31.0 [29.8, 32.5]	31.1 [29.9, 32.6]	31.1 [30.0, 32.4]

Random selection. For single-choice questions, we select uniformly from the four options. For multiple-select questions, we select uniformly from the 15 nonempty option subsets. The exact-match probabilities are therefore $1/4$ and $1/15$, respectively. Averaging these per-question probabilities gives the expected accuracy of random selection.

Most-frequent answers. We partition videos into five folds and predict each fold using the most frequent answer in the other four folds. Frequencies are computed separately for single-choice and multiple-select questions, treating each complete option set as an answer for the latter. All questions from the same video remain in one fold. We pool predictions across the five folds to compute exact-match accuracy.

D.3 Statistical Reliability

To account for dependence among questions from the same video, we perform 10,000 video-level bootstrap resamples with replacement (seed 42). Each resample retains all available responses from the selected videos and recomputes accuracy using the main-table scoring rule. The 2.5th and 97.5th percentiles form the 95% confidence interval.

As shown in Table 7, the interval half-widths for overall accuracy are approximately 1.2–1.4 percentage points. Gemini-3.1-Pro outperforms the best open-source model, Qwen3-VL-32B, by 12.7 percentage points, well beyond this margin, and the two 95% intervals do not overlap, indicating that this gap reflects a genuine performance difference rather than sampling noise. The same resolu-

Table 8: Per-domain exact-match accuracy (%) with subtitles. Domain mean gives equal weight to the five domains.

Model	Economy	Society	Science	Politics	Culture	Avg Acc	Domain mean
Gemini-3.1-Pro	**72.1**	78.2	**77.1**	**67.9**	**65.9**	**71.7**	**72.2**
GPT-5	69.2	**78.8**	71.9	62.8	63.1	68.8	69.2
Claude-Sonnet-4.6	67.7	68.6	66.7	66.0	58.1	65.8	65.4
Qwen3-VL-32B-Instruct	61.9	67.3	57.3	62.8	47.5	59.8	59.4
InternVL3.5-38B	58.5	54.5	53.1	60.3	46.9	55.6	54.7

Table 9: Exact-match accuracy (%) across input conditions. The frame conditions use the same 50 frames, without audio; audiovisual input includes the complete video and original audio.

Model	Input	L1	L2	L3	Overall
Gemini-3.1-Pro	50 frames	78.9	66.4	55.1	68.2
	50 frames + subtitles	81.8	71.5	57.9	71.7
	Video + audio	84.6	72.3	59.4	73.5
Gemini-3-Flash	50 frames	75.6	59.7	52.3	64.1
	50 frames + subtitles	78.0	68.8	55.1	68.5
	Video + audio	80.3	66.1	55.9	68.9

tion holds across the full capability range: overall accuracy among the 19 models spans 31.1% for InternVL3-2B to 70.0% for Gemini-3.1-Pro, a 38.9-percentage-point spread that is about 30 times the interval half-width, indicating that the benchmark yields discriminative estimates from the smallest open-source models to the strongest closed-source models.

E Supplementary Experimental Analysis

E.1 Domain Performance and Weighting

To assess how domain proportions affect model comparisons, we compute subtitle-condition accuracy for five representative models in Economy, Society, Science, Politics, and Culture, and compare question-weighted accuracy with an equally weighted mean across the five domains.

As shown in Table 8, equal domain weighting changes each model's accuracy by at most 0.9 percentage points and leaves the five-model ranking unchanged, indicating that their aggregate comparison is stable under this reweighting. Relative performance within individual domains can differ: although Claude-Sonnet-4.6 has lower overall accuracy than GPT-5, it scores 66.0% in Politics compared with GPT-5's 62.8%.

E.2 Native Video Input Comparison

To examine how input format affects model performance, we compare Gemini-3.1-Pro and Gemini-3-Flash with 50 frames, 50 frames with subtitles, and complete audiovisual input. All three conditions include the question stems and answer options. The audiovisual condition retains the original audio track without additional subtitle text. Table 9 reports the results.

Compared with 50 frames alone, complete audiovisual input improves overall accuracy by 5.3 percentage points for Pro and 4.8 for Flash, with the largest gains on L2: 5.9 and 6.4 points, respectively. Inspection of sampled frames, source videos, and model responses links some improvements to briefly displayed evidence. For example, a basketball question asks for the total number of successful shots across two shooting styles, but the results table appears between two sampled frames. Pro answers incorrectly with frame input; with audiovisual input, it reads 18 and 14 successful shots and correctly sums them to 32. Similarly, a nuclear-power question asks about the initial state in 1960, but the sampled frames have already advanced to 1961. Both models answer with the wrong country

Table 10: Exact-match accuracy (%) of Qwen3-VL Instruct and Thinking variants.

Model	Level 1		Level 2		Level 3		Avg Acc		**Overall**
	w/o Sub.	w/ Sub.	w/o Sub.	w/ Sub.	w/o Sub.	w/ Sub.	w/o Sub.	w/ Sub.	
Qwen3-VL-8B-Instruct	59.0	64.7	28.5	33.9	38.0	38.9	44.1	48.2	46.2
Qwen3-VL-8B-Thinking	62.5	69.6	51.5	53.2	42.7	45.8	53.4	57.8	55.6
Qwen3-VL-32B-Instruct	68.5	71.6	39.0	48.5	50.5	54.2	54.7	59.8	57.3
Qwen3-VL-32B-Thinking	72.3	77.4	55.3	64.4	48.0	50.2	60.2	65.5	62.9

count under frame input and correctly under audiovisual input. These cases show that sampling the relevant chart does not necessarily capture the values and temporal states required by the question.

Subtitle supplementation. What subtitles can recover depends on what the narration states. For example, the 50% increase needed for a rainfall calculation is absent from the sampled frames but retained in the subtitles. With subtitles, Pro correctly computes 3,000 mm, matching its audiovisual response. In contrast, the basketball narration does not state the successful-shot counts, and Pro remains incorrect after subtitles are added. At the aggregate level, adding subtitles narrows the overall gap between frame-based and audiovisual input to 1.8 points for Pro and 0.4 for Flash.

Reading available evidence. Complete audiovisual input does not improve every response: relative to frames with subtitles, Pro gains 0.8 points on L2, whereas Flash loses 2.7 points. Our case inspection also identifies errors despite the relevant evidence being available. For example, an industry-wealth ranking chart is visible in both inputs. Both models correctly identify Fashion & Retail as the second-ranked industry with frame input, but answer Manufacturing with audiovisual input. Here, the error concerns reading available chart evidence rather than missing it during sampling.

E.3 Test-Time Reasoning

The Qwen3-VL entries in Table 1 are Instruct variants. To examine test-time reasoning, we evaluate the corresponding Thinking variants of Qwen3-VL-8B and Qwen3-VL-32B under the same 50-frame input, with and without subtitles. Other settings follow Section 4, and the results are in Table 10.

Thinking raises overall accuracy at both scales, from 46.2% to 55.6% for 8B and from 57.3% to 62.9% for 32B, though the gains are uneven across levels. L2 improves the most: without and with subtitles, 8B rises from 28.5% and 33.9% to 51.5% and 53.2%, and 32B from 39.0% and 48.5% to 55.3% and 64.4%. L1 also gains 3.5 to 5.8 percentage points. L3 changes inconsistently: 8B gains 4.7 and 6.9 points, whereas 32B drops from 50.5% and 54.2% to 48.0% and 50.2%.

As with subtitles (Section 4.4), Thinking also yields its largest gain on L2. The drop for the 32B model on L3 indicates that additional reasoning time does not necessarily produce a more accurate narrative judgment. Even with Thinking enabled, the strongest open-source result, 65.5% for Qwen3-VL-32B with subtitles, remains below Gemini-3.1-Pro and GPT-5 under the same condition (71.7% and 68.8%) and well below the human score of 91.7%.

E.4 Higher Frame Budgets

To examine performance beyond 50 frames, we extend the inputs of Qwen3-VL-32B and InternVL3.5-38B to 100 frames. Commercial APIs limit the number of images per request, so this comparison focuses on open-source models. Both budgets include subtitles. The 100-frame input retains the original 50 frames and adds samples between them; Table 11 compares the two budgets.

Increasing the budget from 50 to 100 frames raises Qwen's overall accuracy by 2.6 percentage points and lowers InternVL's by 0.6. L1 improves for both models, whereas L2 and L3 do not move together: Qwen's L3 stays at 54.2%, while InternVL declines on L2 and L3. This extends the trend in Section 4.3. Denser sampling can add visual detail for basic reading, but the gain does not reliably carry over to temporal reasoning or narrative understanding.

Table 11: Exact-match accuracy (%) at 50 and 100 frames with subtitles.

Model	Frames	Overall	L1	L2	L3
Qwen3-VL-32B	50	59.8	71.6	48.5	54.2
	100	62.4	76.3	50.8	54.2
	Change	+2.6	+4.7	+2.3	0.0
InternVL3.5-38B	50	55.6	65.6	48.1	48.9
	100	55.0	67.0	46.1	46.7
	Change	−0.6	+1.4	−2.0	−2.2

Table 12: Multi-select performance (%). Exact match is computed on multi-select items only. Gold and Selected denote the mean numbers of gold and predicted options. Empty or invalid predictions receive zero scores.

Task	EM	Precision	Recall	F1	Gold	Selected
Without Subtitles						
Causal Reasoning	27.9	94.3	71.8	79.9	2.7	2.1
Argument Synthesis	37.3	93.5	73.8	80.6	3.0	2.3
Narrative Structure	42.7	92.2	78.3	82.7	2.4	2.1
Visual Communication Intent	53.9	95.0	81.0	85.3	2.2	1.9
Counterfactual Analysis	62.6	93.1	86.6	88.1	2.3	2.1
With Subtitles						
Causal Reasoning	37.5	94.1	75.7	82.2	2.7	2.2
Argument Synthesis	53.9	96.0	83.8	88.0	3.0	2.6
Narrative Structure	49.0	93.8	82.4	86.0	2.4	2.1
Visual Communication Intent	54.6	95.4	81.4	85.8	2.2	1.9
Counterfactual Analysis	61.7	94.2	86.3	88.4	2.3	2.1

E.5 Multi-Select Diagnostics

Evaluation. To distinguish option identification from complete answers on multi-select questions, we report option-level precision, recall, and F1 alongside exact match. The statistics cover Causal Reasoning, Argument Synthesis, Narrative Structure, Visual Communication Intent, and Counterfactual Analysis. We compare Gemini-3.1-Pro, GPT-5, Claude-Sonnet-4.6, Qwen3-VL-32B-Instruct, and InternVL3.5-38B under both subtitle conditions.

We compute each metric per question, average within each model, and then average equally across the five models.

Answer completeness. Table 12 shows incomplete coverage of correct options across all five multi-select types: under both input conditions, precision exceeds recall, and models select fewer options on average than the gold answers contain. With subtitles, Argument Synthesis reaches 96.0% precision but 83.8% recall and 53.9% exact match. Models identify most of the correct content, but often fail to include all required options. Improving these tasks therefore calls for more complete coverage of the answer set while maintaining accurate option selection.

Task-specific performance. Because the number of required options varies across task types, we further group questions by gold answer-set size. As shown in Table 13, with subtitles, Causal Reasoning has F1 scores of 74.1% and 85.3% in the two- and three-option groups, respectively, lower than the other four task types. Argument Synthesis reaches 90.1% F1 in the three-option group, second only to Counterfactual Analysis, but its exact-match accuracy is 57.7%. These tasks highlight different priorities for improvement: Causal Reasoning requires better identification and coverage of supporting options, whereas Argument Synthesis often yields many correct selections without a complete answer set.

The remaining groups contain one Visual Communication Intent question with a single correct option, for which EM and F1 are both 100% under both conditions, and five Argument Synthesis questions

Table 13: Multi-select EM and F1 (%) by the number of correct options, using the same aggregation as Table 12.

	Two Correct Options		Three Correct Options	
Task	EM	F1	EM	F1
Without Subtitles				
Causal Reasoning	40.0	74.8	23.4	81.8
Argument Synthesis	37.1	69.8	40.3	83.9
Narrative Structure	50.4	81.7	32.9	84.1
Visual Communication Intent	58.1	85.3	38.1	84.2
Counterfactual Analysis	61.9	87.1	65.0	91.1
With Subtitles				
Causal Reasoning	36.9	74.1	37.7	85.3
Argument Synthesis	54.3	83.0	57.7	90.1
Narrative Structure	53.3	84.1	43.4	88.4
Visual Communication Intent	57.6	85.3	42.4	86.3
Counterfactual Analysis	61.3	87.2	63.2	91.6

with four correct options. For the latter, EM is 20.0% without subtitles and 32.0% with subtitles; F1 is 76.5% and 83.4%, respectively.

E.6 Video Duration Analysis

To examine the relationship between video duration and model performance, we group evaluation videos into four duration intervals: at most 3 minutes, 3–6 minutes, 6–9 minutes, and more than 9 minutes. Table 14 reports accuracy for all 19 models under the *w/o sub* and *w/ sub* conditions.

Duration-related declines without subtitles. Models exhibit different patterns of performance decline on longer videos. GPT-5 achieves approximately 68% accuracy in the first three duration groups, dropping to 59.4% in the longest group. Claude-Sonnet-4.6 shows a gradual decline from 64.1% in the shortest group to 59.0% in the longest. Among open-source models, Qwen3-VL-32B and InternVL3.5-38B score 6.4 and 5.8 percentage points lower on the longest videos than on the shortest. GPT-4o, by comparison, stays between 54.8% and 57.8% across the four groups. These results distinguish accuracy from consistency across durations: GPT-5 outperforms GPT-4o in every group while showing a larger short-to-long gap.

Subtitles narrow short-to-long performance gaps. With subtitles, GPT-5's accuracy on the longest videos rises from 59.4% to 66.9%, narrowing its short-to-long gap from 8.9 to 1.4 percentage points. Qwen3-VL-32B improves from 53.1% to 60.4% on the longest videos, approaching its 60.7% accuracy on the shortest; Qwen3-VL-8B's gap similarly shrinks from 6.0 to 0.9 points. Claude-Sonnet-4.6's gap decreases from 5.1 to 1.2 points. This pattern across closed- and open-source models shows that subtitle information helps these models maintain performance on longer data videos. Gemini-3.1-Pro's gains are concentrated in the middle duration groups: 4.1 and 9.2 points for 3–6 and 6–9 minutes, respectively, compared with 0.3 points for videos longer than 9 minutes.

E.7 Representative Error Cases

To complement the error-type analysis of Gemini-3.1-Pro in Section 4.6, we present two representative failures in the model's actual responses. In both cases the model reads the numbers printed on the chart correctly. The failure is that it does not follow a constraint in the question: one shifts the specified time range, and the other ignores the requirement that the evidence come from the chart.

Figure 21 is the time-range case. The question asks for the increase from the trough around January 2026 to the peak in February 2026. The model locates the relevant chart, but it is distracted by the more salient one-year increase of 111.91% printed on the chart and answers with that value, rather than computing the trough-to-peak gain over the requested interval (approximately 88%). With multimodal input, the model can be drawn to salient on-chart text and pass over an explicit time window in the question.

Table 14: Model accuracy (%) across four video-duration intervals, reported separately under *w/o sub* and *w/ sub*.

Model	*w/o sub*				*w/ sub*			
	≤3 min	3–6 min	6–9 min	>9 min	≤3 min	3–6 min	6–9 min	>9 min
Closed-Source Models								
Gemini-3.1-Pro	70.2	70.8	64.5	66.7	71.4	74.9	73.7	67.0
GPT-5	68.3	68.6	67.9	59.4	68.3	72.4	67.0	66.9
Gemini-3-Flash	64.1	68.3	60.4	62.6	65.3	72.1	69.6	67.0
GPT-4o	57.1	57.8	54.8	54.9	54.4	58.4	59.9	59.7
Claude-Sonnet-4.6	64.1	62.2	59.9	59.0	64.9	67.6	67.3	63.7
Open-Source Models								
Qwen3-VL-32B	59.5	54.6	52.1	53.1	60.7	60.0	58.5	60.4
InternVL3.5-38B	54.2	52.7	56.7	48.4	55.0	57.8	59.0	51.6
InternVL3.5-14B	46.2	48.3	45.2	41.4	47.7	54.3	50.7	45.8
InternVL3.5-8B	47.3	45.7	45.2	41.0	47.3	48.9	49.3	44.7
Qwen3-VL-8B	49.2	41.6	43.3	43.2	49.6	47.9	47.5	48.7
InternVL3-8B	49.8	40.0	45.6	40.1	50.2	44.8	51.6	42.3
Qwen2.5-VL-32B	57.3	55.9	47.9	53.8	56.5	56.8	55.8	53.1
Qwen2.5-VL-7B	44.7	44.1	40.6	39.9	45.4	46.0	48.4	43.2
InternVL3.5-4B	44.3	46.3	45.2	42.5	43.5	39.7	40.1	38.8
Qwen3-VL-4B	45.4	45.1	39.2	37.4	46.9	47.0	44.2	43.6
Qwen2.5-VL-3B	38.9	38.4	38.2	36.6	38.2	41.0	46.1	39.6
InternVL3.5-2B	38.2	37.5	33.6	34.1	41.6	40.6	37.8	34.4
Qwen3-VL-2B	37.8	38.1	37.8	33.7	41.2	39.4	42.9	38.1
InternVL3-2B	32.4	29.7	31.3	31.5	31.2	30.0	34.1	30.4

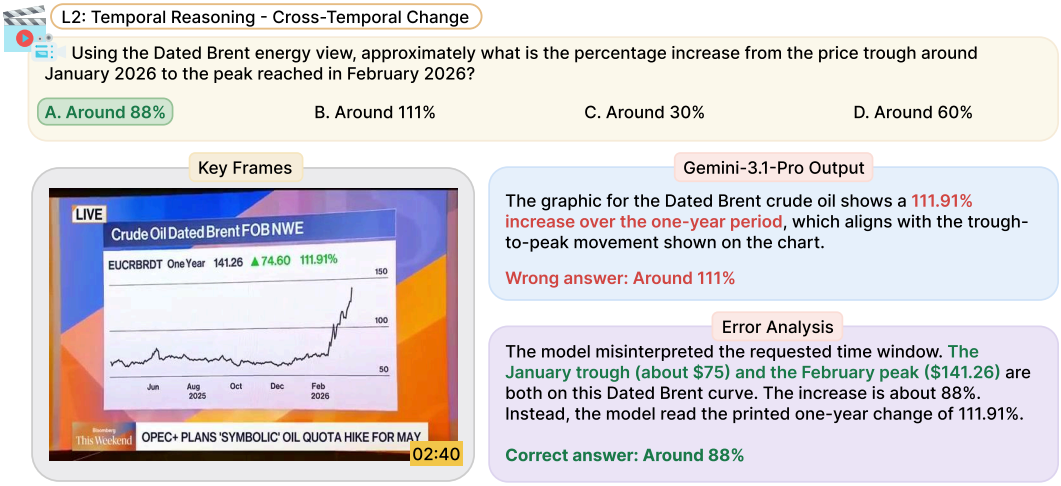


Figure 21: An instruction-understanding error of Gemini-3.1-Pro. The model is distracted by on-chart text and substitutes the printed one-year change for the requested time-window increase.

Figure 22 is the evidence-source case. The question asks for chart-specific facts that bear on the strategic value of the deal. The frames show Eli Lilly's intraday gain of 2.52% at 00:42 and Centessa's (CNTA) intraday jump of 45.58% at 00:54, both printed on the performance cards. The 2034 revenue target of about $2 billion is only spoken around 00:35 and does not appear on any chart. The model treats this narrated figure, together with the two on-chart gains, as supporting evidence and answers A, B, D; the correct answer is A, B. The model can read the chart correctly and still place a number from the narration into an answer that admits only chart facts.

L2: Temporal Reasoning - Argument Synthesis

To explain the strategic value of the deal, the video synthesizes immediate market data with future R&D potential. Which chart-specific facts are load-bearing for this synthesis? Select all that apply.

A. Centessa's intraday price jump of ~45% (CNTA) as seen in the stock performance view.

B. Eli Lilly's intraday gain of ~2.5% (LLY) as seen in the intraday performance card.

C. The 10.02% one-year performance shown in the Eli Lilly one-year trend chart.

D. The specific revenue consensus of $2 billion narrated as being Centessa's 2034 target.

Key Frames





Gemini-3.1-Pro Output

The synthesis relies on Centessa's ~45% jump and Eli Lilly's ~2.5% gain, combined with the narrated 2034 revenue target of $2 billion.

Wrong answer: A, B, D

Error Analysis

The question asks for chart-specific facts only.

The frames show Eli Lilly's intraday gain of 2.52% and Centessa's intraday jump of 45.58%.

The $2 billion 2034 revenue target is spoken around 00:35 and does not appear on any chart. The model included that narrated $2 billion figure as supporting evidence.

Correct answer: A, B

Figure 22: An instruction-understanding error of Gemini-3.1-Pro. The question admits only facts shown on the chart, yet the model still treats the narrated 2034 revenue target of $2 billion as supporting evidence.

E.8 Diagnostic Analysis of an Open-Source Model

We manually inspect 120 randomly sampled incorrect responses from a representative open-source model, Qwen3-VL-32B, with 50-frame input and subtitles. The three categories identified in Section 4.6 are also prominent for this model, though they take somewhat different forms.

Perceptual errors. At the basic visual extraction stage, the main bottleneck for the open-source model lies in misreading temporal and structural information. (1) Temporal anchor confusion: when computing the decline in a brand's sales, the model reads the historical peak value shown in the chart (from 2016) as the current-year figure (2024), causing the subsequent derivation to deviate entirely. (2) Chart structure misjudgment: the model often fails to correctly identify the chart's axis configuration. For example, given a bar chart whose y-axis clearly starts at 0%, the model nonetheless claims in its reasoning text that the chart uses a truncated y-axis to exaggerate the visual difference.

Reasoning errors. Most errors that can be localized at the reasoning stage occur when linking facts to a claim. On causal-reasoning and narrative-argument questions, the model typically extracts the correct values and trends from the relevant frames; however, when judging whether these visual facts support a given option's claim, it frequently reaches the opposite conclusion, failing to translate locally correct facts into valid support for the overall option.

Instruction and mapping errors. This category is particularly prominent in the open-source sample and mostly reflects a mismatch between the model's internal reasoning and its final output. (1) Option-mapping inconsistency: the model reaches the correct conclusion during its textual reasoning but disconnects from it when producing the final predicted letter. For example, when comparing per-capita GDP between two countries, the model correctly computes a roughly 20-fold difference and states in its rationale that this matches option C, yet its final prediction field contains D. (2) Constraint-scope drift: when a question explicitly restricts the comparison to a specific year, the

model disregards this temporal constraint during extraction and mixes in projected or other-year values from the same chart.

Overall, although the open-source model demonstrates some competence in extracting data from individual frames, it continues to face challenges in integrating information into a coherent overall picture and maintaining consistency over long videos. Both the difficulty of linking facts to claims and the disconnect between correct internal derivations and final outputs point to the importance of maintaining cross-frame coherence and closely following instructions over long contexts.

E.9 Upload-Time and View-Count Analysis

Publicly available data videos may be included in model pretraining corpora and thereby affect evaluation results. To examine potential data contamination, we group videos by upload time and view count and compare the overall accuracy of the 19 models evaluated in Section 4 across groups, examining whether older or more-viewed videos show a consistent performance advantage.

Upload time. We first examine whether model performance varies with video upload time. We divide videos into three groups by YouTube upload date: before 2025 (214 videos), within 2025 (399 videos), and on or after 2026-01-01 (348 videos). The results show no consistent accuracy advantage for older videos. Gemini-3.1-Pro achieves 68.9%, 68.8%, and 71.9% across the three groups, all close to its overall accuracy of 70.0% in Table 1; GPT-4o achieves 56.1%, 57.3%, and 57.5%, and InternVL3-8B achieves 45.1%, 44.1%, and 46.5%. These three models perform similarly across year groups and attain their highest accuracy on the most recent group. The difference is larger for GPT-5: accuracy on videos published in 2026 is 69.9%, exceeding its 63.2% on pre-2025 videos by 6.7 percentage points. Overall, earlier publication does not correspond to higher accuracy, and models maintain comparable or better performance on recent videos.

View count. We next examine whether model performance improves with video view count. We divide the 961 videos into four equal-frequency groups by view count, from the lowest (Group 1) to the highest (Group 4). Accuracy does not increase monotonically with view count. GPT-5 achieves 69.1%, 70.7%, 68.4%, and 60.9% across the four groups; Gemini-3.1-Pro achieves 70.7%, 75.7%, 70.6%, and 62.2%; and Qwen3-VL-32B achieves 53.3%, 57.8%, 64.0%, and 54.4%. GPT-5 and Gemini-3.1-Pro both attain their highest accuracy in Group 2 and their lowest in the most-viewed group. Qwen3-VL-32B performs best in Group 3, with a 9.6-percentage-point decline in Group 4. These results show no consistent performance advantage for highly viewed videos: more widely viewed content is not necessarily easier for models to answer correctly.

Neither analysis shows a systematic performance advantage for older or more widely viewed content.

F Limitations and Future Work

DataVista has the following limitations: (1) As this work focuses on the foundational capability of data video comprehension, data video generation and editing are not yet covered. (2) To ensure objective, standardized, and reproducible evaluation at scale, the benchmark currently relies on multiple-choice questions and does not yet include open-ended subjective items. (3) The benchmark currently consists primarily of English data videos, which limits the evaluation of multilingual data video comprehension.

Future work will extend the evaluation to data video generation and editing, introduce open-ended items along with automated scoring schemes, and gradually expand to multilingual datasets. Building on the rich structured annotations already collected, we plan to construct a fine-tuning dataset for data video understanding. Furthermore, to foster a dynamic research ecosystem, we will establish and maintain an official evaluation website and public leaderboard. This platform will periodically incorporate newly released data videos to expand the benchmark, continuously evaluate the latest models, and support researchers in submitting their custom models for evaluation, thereby driving long-term progress in the field.