TOWARDS ROBUST AND COHERENT DATA VISUALIZATION GENERATION WITH VISION–LANGUAGE MODELS

RIDWAN MAHBUB

A THESIS
SUBMITTED TO THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE

GRADUATE PROGRAM IN
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
YORK UNIVERSITY
TORONTO, ONTARIO
MAY 2026

Abstract

Visual data stories have emerged as an effective medium for communicating data by integrating visualizations with coherent narratives. Data videos, a popular format of visual data stories, combine animated, chart-centric visualizations with synchronized narration. In this thesis, we introduce a novel task of data video generation for Vision Language Models (VLMs), along with a benchmark containing 328 real world data videos. We propose a multi-agent framework for generating better data videos, composed of planning and critic agents, designed to mimic the real-world process of data video generation. While these results highlight the potential of VLMs for generating coherent visual data stories, they also underscore the need to examine their robustness. Accordingly, we investigate how misleading visual designs as input can affect VLM behavior and performance, and ways to mitigate this effect. Together, these contributions not only extend the capabilities of VLMs into a new dimension of visual data storytelling, but also provide a systematic understanding of their robustness in general.

Dedication

I dedicate this thesis to my parents, whose sacrifices over the years have shaped who I am today, and to my beloved wife, whose constant encouragement, patience, and support have been my source of motivation and inspiration.

Acknowledgements

I am deeply grateful to many individuals who have supported and guided me throughout this journey. First and foremost, I would like to express my heartfelt thanks to my supervisor, **Dr. Enamul Hoque Prince**, for giving me the opportunity to work under his kind and insightful supervision. His unwavering support, mentorship, and encouragement have been instrumental in my progress, enabling me to publish and present my work at major venues such as ACL, EMNLP, and IEEE VIS over the past two years.

I would also like to extend my sincere appreciation to my thesis committee member, **Dr. Laleh Seyyed-Kalantari**, for her timely help and generous commitment of time throughout the thesis process.

It has been a privilege to be part of the *Intelligent Visualization Lab* at York University. I thank my lab members for their help and collaboration. I also acknowledge Natural Sciences and Engineering Research Council (NSERC) and Compute Canada for supporting this research.

Table of Contents

List of Tables

List of Figures

Preface

This thesis is submitted to the Faculty of Graduate Studies in partial fulfillment of the requirements for a Master of Science Degree in Computer Science. The work presented here is done by the author **Ridwan Mahbub** under the supervision of **Dr. Enamul Hoque Prince**. Some parts of this thesis have been published or submitted as:

- **Ridwan Mahbub**, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mizanur Rahman, Mir Tafseer Nayeem, Enamul Hoque "The Perils of Chart Deception: How Misleading Visualizations Affect Vision-Language Models"

  - Accepted at IEEE VIS 2025. **Awarded Best Short Paper Award**

- **Ridwan Mahbub**, Syem Aziz, Mahir Ahmed, Shadikur Rahman, Mizanur Rahman, Shafiq Joty, Enamul Hoque "DATAREEL: Automated Data-Driven Video Story Generation with Animations"

  - Submitted to EMNLP 2026.

- **Ridwan Mahbub**, Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Mizanur Rahman, Mir Tafseer Nayeem, Enamul Hoque "Chart Deception in Vision–Language Models: From Vulnerability to Mitigation"

  - Submitted to IEEE Transactions on Visualization and Computer Graphics.

1 Introduction

Visualizations play a crucial role in translating complex data into accessible narratives, enabling a wide range of audiences to identify patterns, trends, and anomalies. Across domains such as journalism, public policy, healthcare, finance, and social media, they serve as key drivers of storytelling and support high-stakes decision-making [85]. Visual data stories extend this by integrating elements of communicative and exploratory visualization to convey a clear, intended narrative [45]. While static visual data stories remain widely used, the rapid growth of video-sharing platforms has led to increasing interest in video-based formats, commonly known as data videos [7]. Data videos are widely adopted across domains, particularly in digital journalism. Compared to static formats, they are generally more accessible and easier to follow, and they often achieve higher levels of user engagement through the use of animation, narration, and temporal sequencing [9]. A typical data video consists of multiple short, chart-centric animation clips, often accompanied by voiceover narration and interleaved with real-world images or footage [21]. One such

clip is shown in Fig. 1.1, where we observe that the clip is illustrating China's share in global arms trades. In the first scene, an arrow animation shows that the arms imports of China are reducing on a yearly basis, highlighting it reached it's lowest in the timeline of 2020-2024. In the third scene, with new animations, the blue bars appear to illustrate the increase in arms export in the same time. Finally in the last scene, the bar chart transitions into a pie chart where the position of China in the current top exporters are highlighted. With the recent rise of short-form videos and reels on social media platforms [84], short-form data videos have also begun to emerge. In this paper, we refer to these chart-centric animated clips as *datareels*, which can function either as standalone videos or as building blocks within longer videos.

Despite their impact, generating *datareels* remains challenging. Unlike sourcing real-world imagery, creating animated data visualizations like *datareels* requires careful coordination of visual encoding, temporal progression, and narration, and demands substantial expertise in visualization design, animation, and video-editing tools. This combination of data interpretation, narrative planning, animation design, and audiovisual alignment makes the task quite challenging even for humans. In practice, authors often rely on low-level professional tools such as timeline-based animation and video editors, resulting in significant manual effort and a high barrier

[1] `https://youtu.be/M6CNlvXkR7k?si=05OMIyLBj2PdeQmv&t=97`

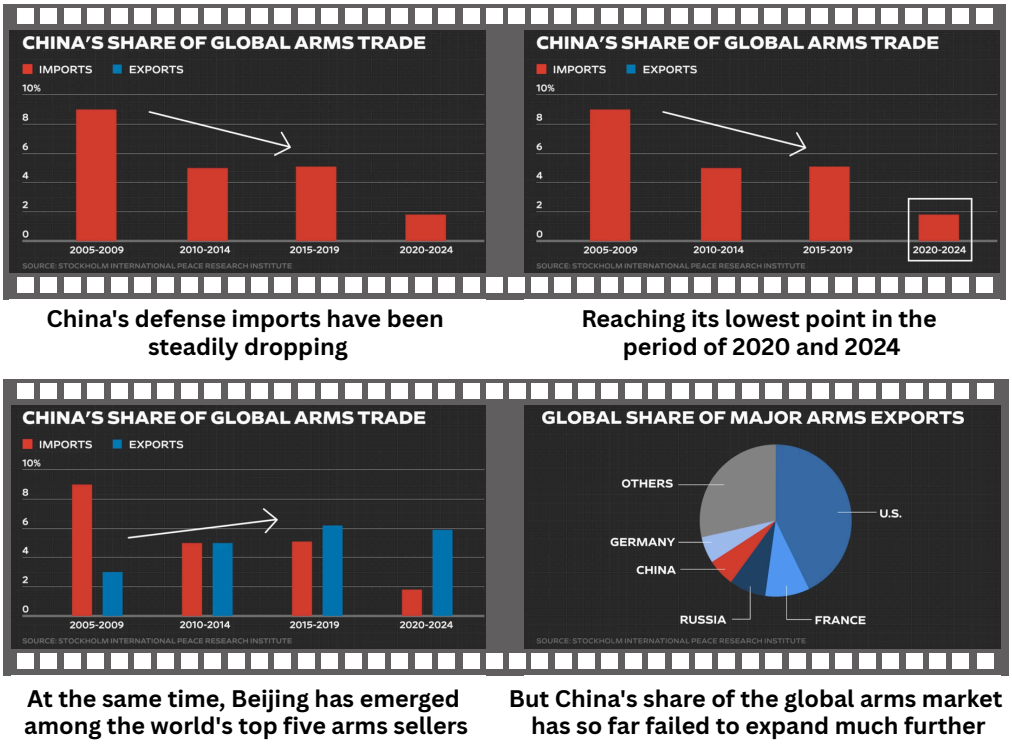


Figure 1.1: A *datareel* extracted from the YouTube channel Wall Street Journal[1].

to entry, particularly for non-experts [108]. This challenge is further amplified by the need to ensure tight alignment between animation and narration, where even small inconsistencies can undermine narrative clarity.

Prior research on data-driven storytelling has primarily focused on static visual stories, introducing narrative design spaces, visual encodings, and authoring tools to support effective communication [85, 47, 56, 77, 94, 95]. While these studies provide valuable guidance, manually creating high-quality data stories remains time-consuming and cognitively demanding. Automated approaches have also

been explored [91, 93, 107], but they typically generate isolated facts or simple annotations, falling short of producing coherent, engaging stories—let alone animated video narratives like *datareels*.

Recent advances in Large Language Models (LLMs) and Vision Language Models (VLMs) have prompted growing interest in their potential for data-centric tasks, including chart summarization [52, 82], chart question answering [68, 51], and natural language story generation [124, 116]. More recent work has begun to explore VLMs for static data story generation [49, 105], as well as human–AI collaboration for data video authoring [90]. However, the standalone capability of VLMs to *datareels* remains largely underexplored, in part due to the absence of appropriate benchmarks and evaluation methodologies.

To address this gap, we introduce DATAREEL, a benchmark for automated data-driven video story generation comprising 328 real-world *datareels* collected from diverse sources. Given a data table and a user-defined analytical intent, VLMs are required to generate code-based animated visualizations—using popular chart animation libraries such as D3.js—along with subtitles that remain coherent with the narrative. This task requires the model to jointly handle data interpretation, synchronized narrative generation, animation planning, and code implementation, making the generation of *datareels* a particularly challenging problem. Our experiments also indicate that even state-of-the-art models and agents, such as Claude

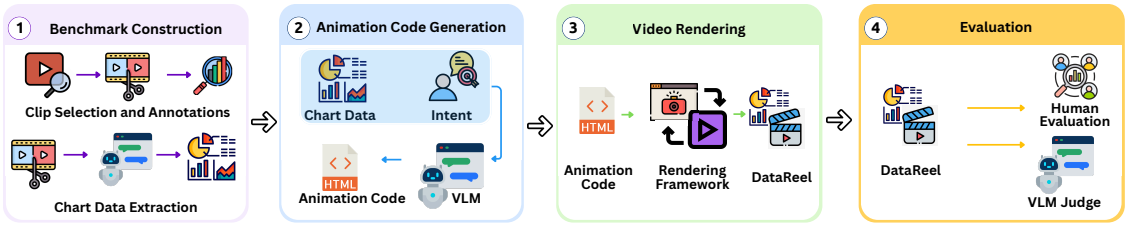


Figure 1.2: Overview of our benchmarking process: (1) We construct DATAREEL, our data video generation benchmark by collecting real-world *datareels*, annotating them, and extracting their underlying chart data. (2) Given chart data and user intent, one or multiple agentic VLMs generate the animation code. (3) This code is executed within a rendering framework to produce the final data video. (4) The resulting videos are evaluated through both human assessment and a VLM-based judge.

Opus 4.5 and Gemini 2.5 Pro, encounter challenges on this benchmark, frequently producing outputs with issues such as unstable animations, improper chart positioning, and poor adherence to reference specifications. Fig. 3.1 give an overview of the DATAREEL benchmarking process.

Motivated by the strong planning and code-generation capabilities of LLMs [50] and recent successes of agentic VLM frameworks in complex tasks [36, 119, 104, 78, 16, 114], we further propose a multi-agent framework that decomposes *datareel* generation into planning, generation, and verification stages, mirroring key aspects of the human storytelling process. Through extensive automatic and human evaluations,

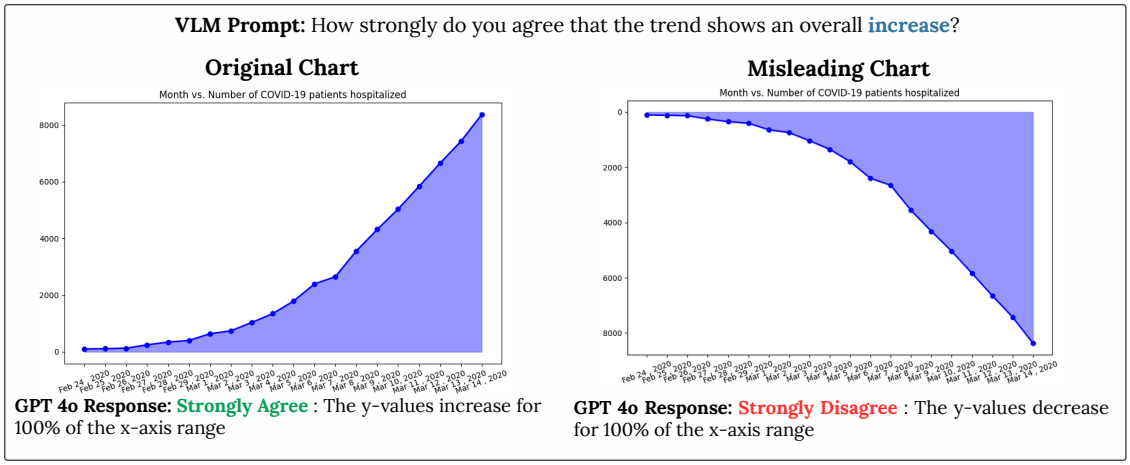


Figure 1.3: An examples of the effect of misleading chart on VLM responses. Here, inverting the y axis makes GPT4o change it's answer to the question, even though the underlying data is the same.

we show that this structured, multi-step approach outperforms direct prompting baselines.

As we extend VLMs to author data videos, ensuring their robustness becomes equally critical. While these models are typically trained on well-formed visualizations, they can be sensitive to subtle design manipulations such as truncated axes, skewed aspect ratios, or unnecessary 3D embellishments that introduce misleading visual cues without altering the underlying data. Such vulnerabilities can lead to incorrect or biased outputs, especially in complex tasks like data video generation. This motivates our complementary focus on evaluating and improving VLM robustness to deceptive visual designs. Importantly, these manipulations do

not fabricate data but instead reshape its visual representation to amplify or suppress perceived differences, thereby influencing interpretation and decision-making [44, 15, 64]. These practices don't fabricate facts; instead, they subtly alter their visual representation to amplify or minimize perceived differences, influencing narratives and decisions [44, 15, 64]. For instance, in Figure 1.3, an inflation trend is made to look significantly flatter by changing the aspect ratio—an alteration that may go unnoticed but can fundamentally mislead the viewer. Such distortions are dangerous when they influence public discourse or policy understanding without the audience being aware [57].

The risks are even greater when these visualizations are used in real-world systems. In finance, healthcare, policymaking, and enterprise reporting, visualizations interpreted by AI systems help automate investment decisions, risk assessment, strategic planning, and public communication [97, 103]. On social media, they influence public discourse [18]. If these systems rely on Vision-Language Models (VLMs) that are susceptible to misleading visuals, the consequences can be serious, leading to poor financial decisions, regulatory mistakes, or the spread of misinformation.

As VLMs become increasingly embedded in data analysis workflows, their ability to interpret visualizations faithfully is both essential and urgent. VLMs have shown remarkable promise in tasks such as chart summarization [53], question answering [43], visualization generation [41], and improving chart accessibility for individuals

with visual impairments or low visual literacy [22, 39]. Given their growing role in analysis and accessibility, it is essential that VLMs can accurately interpret data, even when visualizations are designed to deceive. Previous research has explored VLMs' capacity to characterize misleading charts, but often under idealized conditions where prompts explicitly hint at the deception [6, 65]. In everyday use, however, people usually engage with visualizations through neutral prompts, unaware of potential misrepresentation [43]. This gap between experimental conditions and real-world use leads to a key question: Can VLMs accurately detect and resist misleading charts in the absence of any explicit signals?

In order to access the reliability of VLMs for different visualization tasks, we introduce a taxonomy of eight widely used deceptive chart tactics and build a benchmark of 1,600 charts—originals and misleading variants—spanning diverse visual distortions. We evaluate cutting-edge VLMs (proprietary and open-source), generating over 16,000 responses to assess model susceptibility. Our results reveal that even top models are vulnerable to subtle visual manipulations, highlighting serious risks for real-world deployment.

This thesis brings together the challenges of generative visual storytelling and robustness to misleading designs, laying the foundation for VLM-based systems that are not only capable of producing coherent and engaging data narratives but also robust to real-world misleading visual designs as inputs.

1.1 Motivation

The growing role of automated systems in creating and interpreting visual data calls for models that can not only generate coherent and engaging data stories but also remain robust to imperfect or misleading visual inputs. This thesis is motivated by two converging challenges in visual data storytelling: (1) how to enable Vision Language Models (VLMs) to generate high-quality, structured data videos, and (2) how to ensure that these models remain reliable when faced with misleading visual designs that distort perception without altering the underlying data. These two interconnected challenges form the foundation of this thesis.

Automated data video generation offers strong potential for domains such as journalism, education, and public communication, where complex data needs to be conveyed clearly and effectively. It is particularly valuable for users with limited experience in video editing, a process that typically has a steep learning curve due to the need to coordinate multiple components such as data interpretation, visual encoding, temporal progression, and narration. As a result, professional data video production often relies on multiple tools and teams with diverse expertise to handle these different aspects.

On the other hand, as VLMs take on a greater role in interpreting and generating data stories, they become vulnerable to the same perceptual pitfalls as humans.

Subtle design choices such as truncated axes, skewed aspect ratios, or unnecessary visual embellishments can significantly influence how data is perceived, even when the underlying values remain unchanged. Without robustness to such misleading designs, the adoption of VLMs in mainstream applications remains limited despite their potential.

By addressing both the generative and interpretive aspects of VLMs in data visualization, this thesis aims to advance the development of systems that not only produce compelling visual data videos but also maintain robustness in the presence of misleading visual inputs, supporting more dependable data communication in real-world settings.

1.2 Problem Statement and Research Questions

Although VLMs are increasingly being used for various visualization tasks, their adoption for fully autonomous data video generation remains limited. Current systems face 2 main limitations:

1. *The potential of VLMs for multi-modal data video generation remains underexplored due to the lack of standardized benchmarks and evaluation frameworks, which hinders systematic development and comparison of approaches.*

   To address this challenge, we investigate the following **Research Questions**:

   - **RQ1**: How effective are current state-of-the-art VLMs at generating multi-modal, non-static visualizations, such as data videos?
   - **RQ2**: Can agentic capabilities be leveraged to improve the performance of VLMs in generating multi-modal visualizations, compared to single-model prompting?
   - **RQ3**: What are the common issues and challenges prevalent in data videos generated by VLMs?

2. *While VLMs are known to be vulnerable to some misleading visual designs, existing benchmarks do not support controlled studies in realistic settings, lack large-scale datasets and standardized metrics for rigorous evaluation.*

To investigate and mitigate this problem, we try to answer the following **Research Questions**:

- **RQ4** : Do subtle misleading visualization designs influence the responses of vision-language models, and are certain types of visual deception more effective than others?
- **RQ5** : Are vision-language models equally susceptible to visual deception, or do some models demonstrate greater robustness?
- **RQ6** : Can inference-time mitigation approaches improve the robustness of these VLMs to misleading chart designs?

These challenges and research questions are addressed in this thesis through the development of benchmarks for both automated data video generation and model robustness to misleading visual designs, along with a multi-agent framework designed to provide more effective solutions to both problems.

1.3 Contributions

This thesis makes the following core contributions:

1. We curate a novel benchmark dataset for the task of fully automatic data video generation using Vision Language Models (VLMs). Our benchmark is curated from real world sources of data videos, contains high variations and pose significant challenge to existing VLMs.

2. We propose a multi-agentic data video generation setup consisting of multiple agents, where the complex task of data video generation, consisting of planning, critiquing and coding is divided among a number of agents. Our results using both human and VLM judges indicate that data videos generated by the agentic frameworks is indeed better.

3. We analyze the data videos generated by both closed source and open source models and highlight the limitations and recurring common errors associated with using VLMs for fully autonomous data video generation.

4. We introduce a taxonomy of eight widely used deceptive chart tactics and build a benchmark of 1,600 charts. Using this benchmark, we evaluate cutting-edge VLMs (both proprietary and open-source) and analyze how misleading designs affects models of different types.

5. We further extend the taxonomy and benchmark with variations inspired by real-world cases and propose a metric called deception score to quantify the degree to which model responses are affected by misleading chart designs.

6. We evaluate the effectiveness of inference-time mitigation strategies in reducing the impact of misleading visual design choices. We show that using multi-agent VLM for visualization analysis can to some extent reduce the impact of misleading design choices.

As a secondary contribution, we release the datasets and code used in this thesis. The resources for automated data video generation using VLMs are available at `github.com/vis-nlp/DataReel`, and the resources for the study of robustness of VLMs to misleading charts are available at `github.com/vis-nlp/visDeception`. We hope these resources will encourage further research in this area.

1.4 Organization of the Thesis

In this section, we outline the structure of the remaining chapters of this proposal. ***Chapter 2*** reviews related work on data-driven storytelling, chart understanding tasks, data video generation, the use of VLMs as judges, misleading visualization designs, and their effects on VLM responses. ***Chapter 3*** describes the construction of our data video evaluation benchmark, along with a multi-agentic data video generation setup that outperforms single VLM approaches. ***Chapter 4*** investigates the effects of misleading chart designs on vision–language models, discusses a new metric to quantify the effect of these designs on VLMs and explores inference-time mitigation strategies. Finally, ***Chapter 5*** concludes the thesis and presents directions for future research.

2 Literature Review

In this section, we first review prior work on story generation. When exploring various chart related downstream tasks where language models have been utilized. We then look into recent studies that utilize VLM for story generation and data video generation. We then explore different works where VLMs were used as a Judge. We then summarize research on misleading visual designs in charts and their effects on human interpretations. We also highlight how the recent studies are insufficient to understand the effects of such designs on VLMs. We ultimately highlight the need for benchmarks to access the capabilities of VLMs in both generating and understanding robust and coherent data visualizations.

2.1 Story Generation Tasks

Automated story generation focuses on producing coherent sequences of events under open-ended constraints [59]. Prior work has explored textual [54], visual [61, 24], and multimodal [13] story generation, often focusing on narrative coherence and

character-centric progression.

In contrast, **data-driven storytelling** focuses on generating multimodal narratives in which visualizations communicate patterns, trends, and outliers in data, while accompanying text explains and contextualizes them [83, 85, 46]. Early work in this area emphasized extracting and ranking salient insights from data tables using statistical measures [29, 101]. Building on this foundation, systems such as DataShot [107] and Calliope [93] generate visualizations paired with textual captions to present key data facts, while later tools including Erato [98] and Socrates [113] incorporate user interaction to guide narrative structure and content selection.

2.2 Chart-related Downstream Tasks

Several downstream tasks associated with charts have been proposed in recent years. Chart question answering focuses on answering factual or reasoning-based questions about charts [68, 74], while open-ended chart QA generates explanatory responses [51]. Chart summarization aims to produce concise textual descriptions of charts [52, 100], whereas chart-to-table tasks extract underlying data from chart images [23, 70, 73]. Others focus on verifying claims about charts [3, 5]. Recent work has also focused on automated visualization generation from text queries [81]; however, they focused on static charts rather than animated charts.

2.3 VLMs for Story Generation

Recent work has begun exploring the use of LLMs for data story generation, particularly for tasks such as story planning and explanation generation. Broadly, prior research has examined LLMs from four perspectives [42]. First, LLMs show strong potential as tools for data exploration and analysis. Systems such as Urania [40] use LLMs to answer user queries through visualizations, while WaitGPT [117] visualizes data analysis code. Qu *et al.* [80] further demonstrate that LLMs can support novice users in data analysis tasks. Second, LLMs are increasingly used to generate narrations for data storytelling. Systems like Data Narrative [49] automatically produce static data stories directly from raw data using LLMs. Third, visualization generation is a core component of storytelling frameworks, and LLMs have proven effective in this area. Works such as VisEval [17] and Text2Vis [81] study their capabilities and limitations in visualization generation. Fourth, LLMs are also being used to build interactive storytelling systems, as seen in Jupybara [105]. However, most existing work focuses on stories with static visualizations. Systems that support dynamic or animated, chart-centric storytelling with LLMs remain relatively underexplored.

2.4 VLMs for Data Video Generation

In contrast to static story generation settings, data videos introduce fundamentally different challenges, requiring the coordinated creation of visual encodings, animation, narrative structure, and audio. End-to-end approaches for automated data video generation remain rare [90]. Early work relied on predefined templates and heuristic rules [8, 92], which limit diversity and scalability. More recent studies explore interactive authoring tools and human–LLM collaboration [108], as well as annotation-guided generation [87, 88], where users explicitly specify animation behavior. While effective as authoring aids, these approaches require substantial human input and domain expertise. Some work has begun to explore LLM-based multi-agent systems for data video generation [89]. However, the absence of systematic, model-centric evaluation makes it difficult to accurately assess model capabilities. More recently, Li et al. [60] introduced a benchmark for animated charts, but it focuses only on generating animations from straightforward inputs with explicit descriptions. In practice, users creating data stories may not already know the key insights or the appropriate animations to highlight them. This limitation highlights the need for a standardized benchmark and evaluation frameworks for fully automated data video generation.

2.5 Using VLM as a Judge

The use of language models as judges in place of a static evaluation metric has become increasingly common in natural language processing. Early approaches such as BERTScore [122] leveraged contextual embeddings from BERT [28] to measure semantic similarity between generated and reference text. With the emergence of more advanced large language models such as GPT-3 [14], these models began to be used more broadly as evaluators across a variety of tasks. For example, GPTScore [33] utilizes the generative capabilities of LLMs to assess outputs in different evaluation contexts. Recent work in visualization has similarly adopted LLM-based evaluation frameworks. In general, LLM-as-a-judge approaches can be categorized into two types: pairwise and pointwise. In pairwise settings, the LLM selects the better response from multiple candidates [34], while in pointwise settings, the LLM assigns scores to outputs based on predefined criteria [123], which can also be implemented by prompting the model to answer structured questions about the output. Benchmarks such as VisEval [17] and Text2Vis [81] use VLMs to evaluate the quality of generated visualizations. VisEval [17] employs a VLM-based question answering framework to assess aspects such as readability and layout correctness, whereas Text2Vis [81] adopts a pointwise scoring approach in which generated visualizations are evaluated across multiple criteria including readability, quality,

and chart correctness. DataNarrative [49] also utilizes VLMs as judges in a pairwise setting to compare generated data stories from different models, and demonstrates through human evaluation that VLM-based judgments exhibit high agreement with human evaluators.

2.6 Misleading Visual Designs

The potential for charts to mislead has been recognized since the 1950s, with classic works like *"How to Lie with Statistics"* [44] illustrating how visual and statistical representations can distort information and contribute to misinformation. Researchers have examined this issue from multiple perspectives. Some studies have investigated deception strategies in common chart types, i.e., graphs, lines, and pie, through user studies assessing their impact on interpretation and decision-making [79, 57]. Lan *et al.* conducted five focus group studies to investigate the underlying reasons behind visualization design flaws [55]. Ge *et al.* proposed a definition of visualization literacy that includes the ability to recognize visual deception, arguing that this skill is a vital component of overall visual literacy [35]. Lisnic *et al.* [63, 64] analyzed various tactics used to spread misleading narratives using deceptive visualization in social media. Some studies have focused on very specific aspects of visual deception tactics, such as axis truncation [26, 67]. On the other hand, Fan *et al.* [30] explored techniques to mitigate such misleading designs

in line charts. While the effects of misleading charts on human interpretation have been widely studied, their impact on VLMs remains largely unexplored.

2.7 VLMs for Visualization Understanding

Significant improvements have been made in chart reasoning and understanding due to the advances in large vision-language models [49]. These models can be categorized into: general-purpose multimodal models and chart-specific models. The general-purpose multimodal models can be closed-source or open-source. The closed-source models [2, 37] achieve the state-of-the-art performance on chart understanding benchmarks [69, 109] and often outperform their open-source counterparts [118, 58, 19, 115, 1]. However, the open-source ones are also making rapid progress to close the gap. For instance, the possibility of fine-tuning open-source models has led to the development of various chart-specific models [76, 72, 121, 71], which demonstrate strong performance on various benchmarks [69, 4, 53]. While existing VLMs perform very well on chart understanding tasks like ChartQA [69]; their ability to detect and accurately respond to questions about various types of misleading charts remains largely under-explored.

2.8 Effects of Misleading Visual Designs on VLMs

For interpreting misleading charts, while recent studies have examined the ability of language models, the evaluation contexts often diverge from the conditions under which such visuals are encountered in real-world scenarios. For example, Lo et al. [65] was the first to investigate the detection of misleading charts using real-world examples; however, the prompts provided to the models often implicitly signaled the presence of misleading elements. Alexander et al. [6] conducted a similar study on a different corpus consisting of misleading tweets. They use both the tweet text and the attached chart with the tweet to determine if the tweet is misleading or not. These setups contrast with real-world scenarios, where users typically remain unaware that a chart may be deceptive and simply pose questions about the chart without any suspicion of misinformation. Our study addresses the effects of misleading designs on VLMs in this more naturalistic setting.

2.9 Mitigating Effects of Misleading Visualization

Alongside works examining the impact of misleading visualizations on models, recent studies have also explored methods to mitigate the challenges posed by such designs. While these efforts represent important progress, several limitations remain. For instance, Misleading ChartQA [20] introduces a large-scale benchmark with 3,026

misleading chart question–answer pairs, but it does not include corresponding non-misleading versions of the same charts. This makes it difficult to isolate the effect of the misleading design itself. In addition, its mitigation strategy relies on explicitly specifying known types of misleading designs, which limits adaptability to unseen or novel distortions without modifying prompts. Similarly, MisVisFix [27] proposes a VLM-based inference-time approach for detecting and correcting misleading visualizations. However, its effectiveness depends on prompting the model to look for specific categories of misleading designs and often requires multiple prompt stages. As a result, there is a need for a simpler and more flexible inference-time solution that can dynamically handle diverse and previously unseen misleading visualizations.

2.10 Summary

In the context of robust and coherent data visualization generation, we observe that the use of VLMs for fully automated multimodal data video generation remains underexplored. At the same time, although recent work has examined the impact of misleading visualization designs on VLMs in naturalistic settings, there is still no benchmark that enables such a controlled study. This thesis addresses both of these gaps.

3 Data Video Generation

In this chapter, we address our first problem statement, which focuses on the limited exploration of VLMs for multi-modal data video generation, primarily due to the absence of standardized benchmarks and evaluation frameworks for systematic development and comparison. To this end, we first formulate the task of automated data video generation using VLMs and curate our benchmark, DATAREEL, along with an evaluation framework. We then evaluate state-of-the-art VLMs on this benchmark, assessing their capabilities on the data video generation task **(RQ1)**. Next, we propose a multi-agentic data video generation setup and show that it produces better data videos than single-model generation using our evaluation framework **(RQ2)**. Finally, we analyze data reels generated by the single-VLM setup to identify common errors and recurring issues that need to be addressed to advance automated data reel generation **(RQ3)**. Figure 3.1 an overview of our benchmarking process, and Figure 3.2 shows a sample data reel generated by our multi-agentic system.

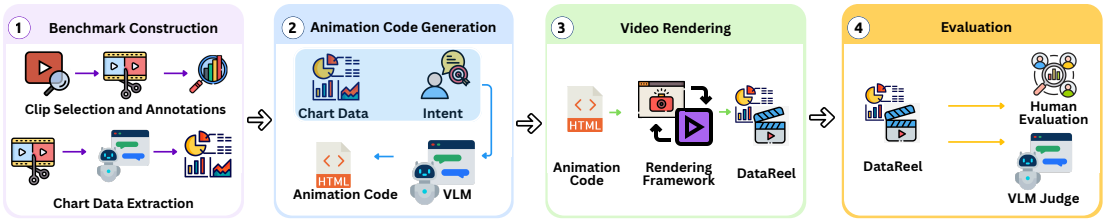


Figure 3.1: Overview of our benchmarking process: (1) We construct DATAREEL, our data video generation benchmark by collecting real-world *datareels*, annotating them, and extracting their underlying chart data. (2) Given chart data and user intent, one or multiple agentic VLMs generate the animation code. (3) This code is executed within a rendering framework to produce the final data video. (4) The resulting videos are evaluated through both human assessment and a VLM-based judge.

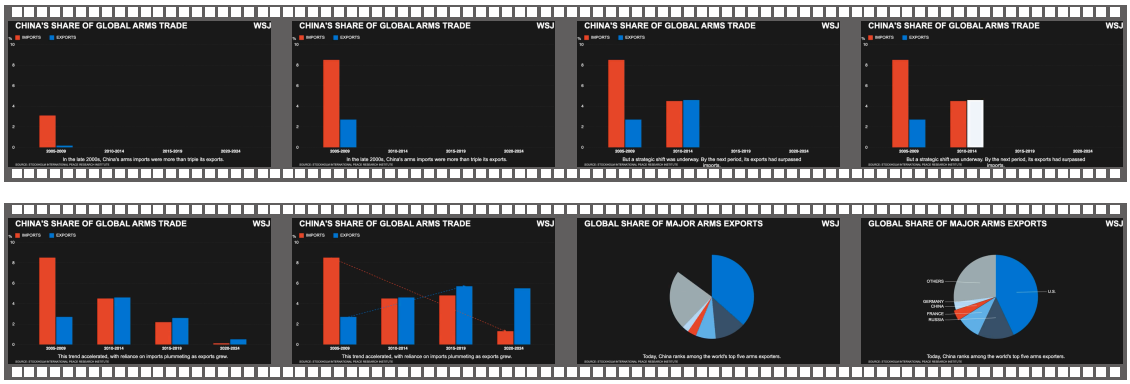


Figure 3.2: A datareel generated by the Multi Agentic Approch utilizing Gemini 2.5 Pro.

3.1 Task Formulation

In our benchmark, we refer to short data video clips referred to as *datareels*. We define the data video clip or *datareel* generation task as generating a chart animation based video clip from a given data table, user intent, and target animation duration. The output will be in HTML format, consisting of a sequence of chart animations, paired with narrative subtitles, that together form a coherent data-driven story consistent with the user's stated intent. Using headless browser, the generated animation is rendered and screenshots of the browser are captured at regular intervals, ultimately assembled into a video.

In real-world *datareels* are typically delivered through spoken voiceovers. However, to simplify the task for current models and to avoid challenges related to audio–video synchronization [31], we represent narration exclusively through on-screen subtitles in our task setup.

3.2 The DATAREEL Benchmark

We first develop a novel benchmark for data video generation. Our benchmark for the data video generation task comprises *datareels*—short, animated clips that communicate insights through chart-based visualizations synchronized with narration. Unlike conventional news videos, which primarily rely on real-world footage

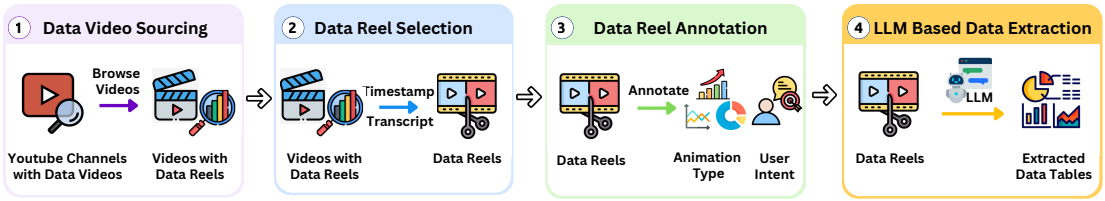


Figure 3.3: Overview of our benchmark construction pipeline: (1) We first identify videos that contain *datareels* from YouTube channels that regularly publish data videos. (2) We identify the timestamp of the *datareel* from each of the selected videos along with the transcript. (3) Annotators then annotate animation related metadata such as animation type and user intent. (4) LLMs are used to extract chart data from the *datareels* in tabular format.

accompanied by spoken commentary, *datareels* place animated data visualizations at the core of the narrative, making them a distinct and challenging medium for automated generation.

We constructed DATAREEL through a four-stage pipeline: *(i)* data video sourcing, *(ii) datareel* selection, *(iii) datareel* annotation, and *(iv)* LLM based data extraction. The overview of the benchmark construction has been shown in Fig. 3.3.

3.2.1 Data Video Sourcing

We focused on collecting *datareels* from real-world, high-impact sources where data-driven storytelling is routinely practiced. In contrast to standalone data videos,

datareels often appear sparsely embedded within longer news or documentary videos, making their identification non-trivial. To guide sourcing, we reviewed prior analyses of data videos in journalism [7, 21] and examined the channels used in those studies. We then augmented this list through a systematic exploration of news-centric YouTube channels that regularly publish videos containing animated data visualizations. This process resulted in a curated set of 14 channels that consistently produce data-driven video content.

3.2.2 *Datareel* Selection

Each annotator was assigned a subset of channels identified in the sourcing stage. Annotators skimmed through videos to identify segments containing *datareels* and recorded precise start and end timestamps for each clip. For every extracted clip, annotators also collected the corresponding narration transcript. They examined Channels in reverse chronological order to prioritize recent content, as animation quality and production practices have improved over time. A single video could yield multiple *datareels* when distinct animated visualization segments were present.

3.2.3 *Datareel* Annotation

For each selected clip, we recorded the types of animations present in the chart. To ensure consistent and systematic annotation, we adopted the animation taxonomy

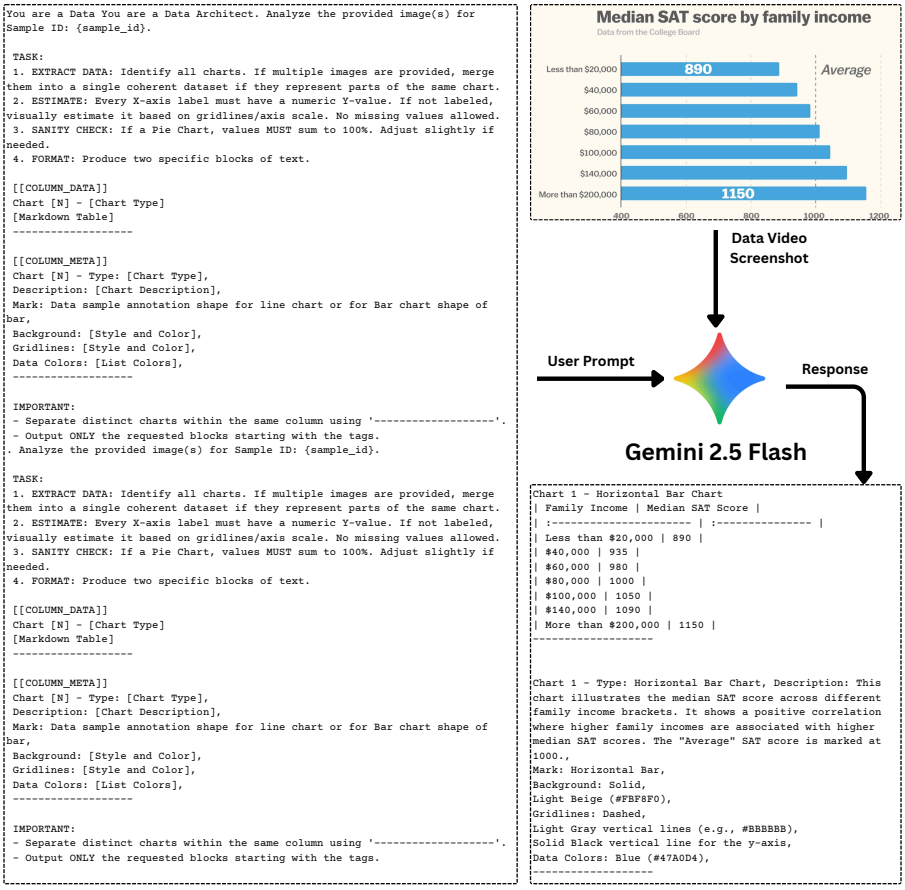


Figure 3.4: The figure presents an overview of the chart data extraction process using the Gemini-2.5-flash [102] model.

proposed by Shi et al., which is specifically designed for real-world data videos [94]. A comprehensive list of the animation types we considered and observed is provided in Table 3.2. In addition to animation types, we also annotated the intent of each clip. Here, intent refers to the narrative conveyed by the datareel, as expressed through the combined effect of both the animation and the accompanying narration.

3.2.4 LLM Based Data Extraction

Evaluating models on *datareel* generation requires access to the underlying chart data; however, such data are rarely released with data videos. While some videos cite data sources in their descriptions, the corresponding tables are often unavailable or difficult to recover. To address this limitation, annotators captured screenshots of each chart and employed large language models to extract tabular data directly from these images, following an approach similar to [75]. A subset of extracted tables was manually verified to assess feasibility and quality. As our benchmark emphasizes animation and narrative alignment rather than precise numerical reconstruction, minor extraction errors are acceptable and do not affect the validity of the evaluation.

We utilize Gemini-2.5-Flash [102] to extract data from chart images. The model was given a screenshot of the chart that we collected during our data collection stage. The annotators were instructed to collect the screenshot when the entire chart was visible. In case the animations occluded particular portions of the chart or if the same clip consisted of multiple animated segments, multiple images were collected and they were all sent to the model together. The models identified both the data present in the chart and associated metadata like colors of different portions, annotation, etc. The detailed overview of the process is shown in Fig. 3.4

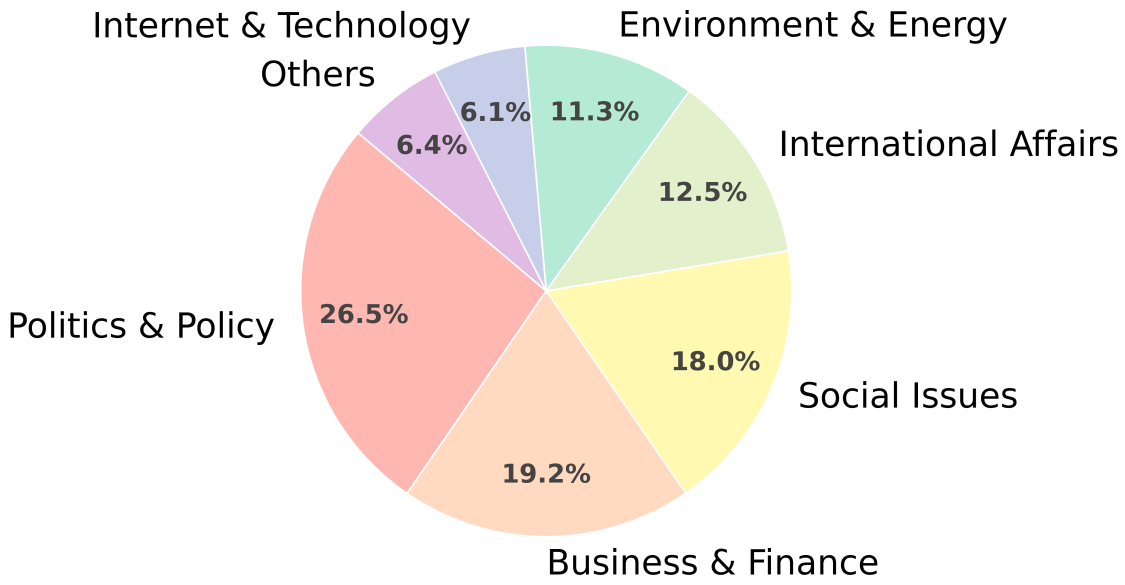


Figure 3.5: Topic distribution for the samples in DATAREEL.

3.2.5 Features of DATAREEL

We analyze the corpus statistics of DATAREEL to characterize its key properties and highlight the challenges it poses for automated data video generation.

Topical and Source Diversity. DATAREEL spans a wide range of real-world topics, including *Politics & Policy*, *Business & Finance*, *Social Issues*, *International Affairs*, and *Environment & Energy* (Figure 3.5), reflecting the broad use of data-driven video storytelling across journalism, education, and public communication.

Channel	**# Reels (%)**	**Channel**	**# Reels (%)**
Vox	112 (34.1%)	The Guardian	6 (1.8%)
The Wall Street Journal	82 (25.0%)	Kurzgesagt – In a Nutshell	4 (1.2%)
TLDR News Global	41 (12.5%)	The Gravel Institute	3 (0.9%)
Economics Explained	30 (9.1%)	Our Changing Climate	2 (0.6%)
Polymatter	23 (7.0%)	CBC News	2 (0.6%)
TLDR News	19 (5.8%)	The Infographics Show	2 (0.6%)
Politizane	1 (0.3%)	TLDR News EU	1 (0.3%)

Table 3.1: Distribution of Data Reels by Source Channel. Counts are followed by percentages in parentheses.

The benchmark aggregates content from 14 distinct YouTube channels (Table 3.1), providing diverse coverage without reliance on a single dominant source.

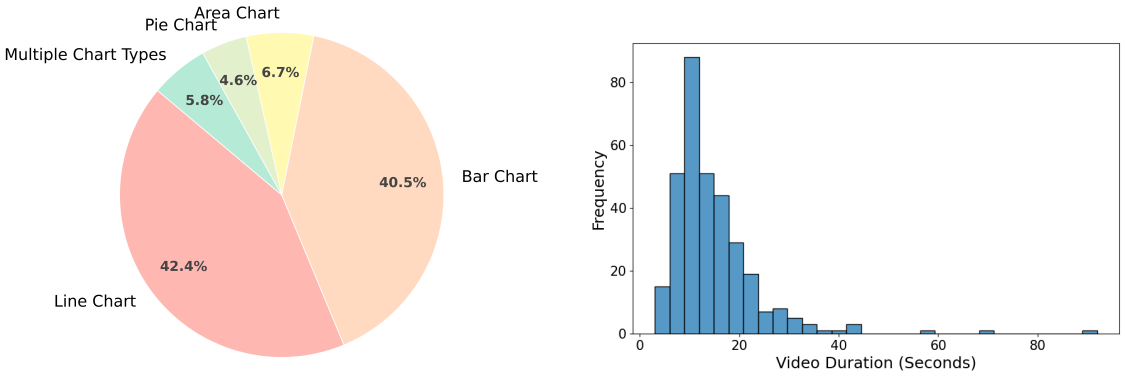
Variety of Chart Types and Visual Forms. As shown in Figure 3.6a, DATA-REEL includes a diverse set of chart types. Bar and line charts are most prevalent, followed by pie charts, area charts, and reels containing multiple chart types. This distribution mirrors real-world data videos, where authors favor chart forms that support animated emphasis, comparison, and temporal progression, while also increasing the complexity of animation planning and coordination.

Animation Category	**# Reels (%)**
Emphasis	297 (44.3%)
Suspense	244 (36.4%)
Comparison	105 (15.7%)
Ellipsis	17 (2.5%)
Cohering	4 (0.6%)
Focalization	3 (0.4%)

Table 3.2: **Distribution of Main Animation Categories**. Counts are followed by percentages in parentheses.

Rich and Diverse Animation Usage. Table 3.2 summarizes the distribution of high-level animation types in the benchmark. Emphasis-driven animations (44.3%) and suspense-oriented animations (36.4%) are most common, reflecting the narrative role of animation in guiding viewer attention and pacing information disclosure. Comparison-based animations account for 15.7%, while more subtle narrative devices such as ellipsis, cohering, and focalization appear less frequently.

Long and Multimodal Narratives. Unlike existing chart-centric benchmarks that focus on short textual outputs such as answers or summaries, DATAREEL contains longer narration segments aligned with animated visualizations. Figures 3.7a



(a) Chart type distribution.

(b) Video duration distribution.

Figure 3.6: Dataset characteristics of . Distribution of (a) chart types and (b) *datareel* durations.

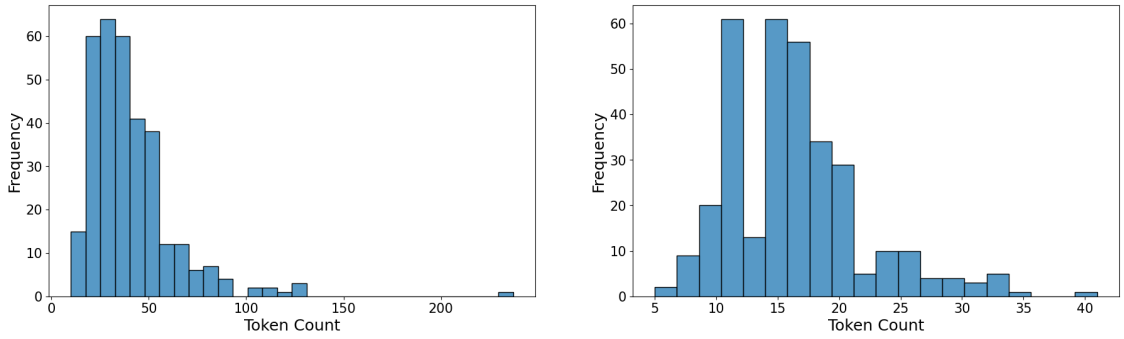
and 3.7b show wide, heavy-tailed distributions in narration and intent length. *datareel* durations also vary substantially (Figure 3.6b), with most clips under 20 seconds and a long tail of longer segments.

3.3 Methodology

In this section, we first describe the two different *datareel* generation approaches we have used. Then we also describe our evaluation framework and experimental setup.

3.3.1 *Datareel* Generation Approches

We evaluate LLMs under two distinct experimental settings:



(a) Transcript length (tokens).

(b) Intent description length (tokens).

Figure 3.7: **Token distribution statistics.** Distribution of token counts for (a) narration transcripts and (b) intent descriptions.

***(i)* Single Agent Prompting:** In this approach, we benchmark the models using a one-shot generation paradigm. We begin with a minimal initial prompt and iteratively refine it based on the quality of the outputs produced by the models. The final prompt incorporates strict requirements for HTML output formatting, provides guidance on the types of animations to include, and specifies the essential components necessary for correct rendering. Through multiple rounds of prompt refinement, the models progressively generate higher-quality data videos. The prompt for this approach has been shown in Fig. A.9 and Fig. A.10.

***(ii)* Multi Agentic Framework:** Recent works have shown that LLMs excel at multi-agentic setups [32, 125], where a number of LLMs are used, each given a certain role to perform a sub-task of an overall complex tasks. Inspired by the

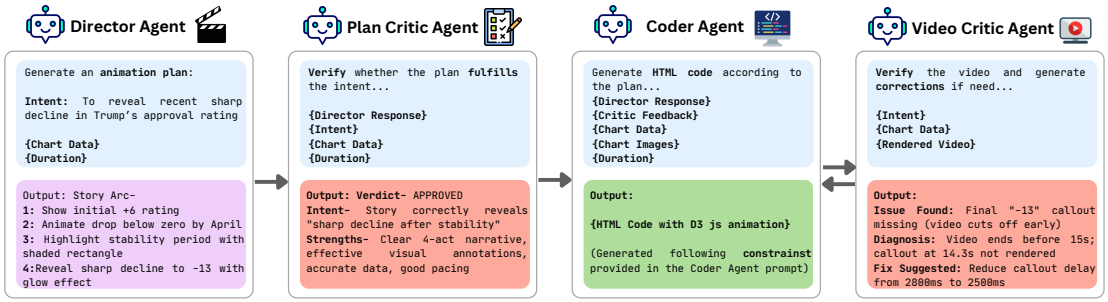


Figure 3.8: Overview of the multi-agent framework: (1) The Director Agent converts the input intent, chart data, and duration into a structured animation plan. (2) The Plan Critic Agent reviews the plan to verify that it aligns with the intent. (3) The Coder Agent translates the approved plan into executable HTML and D3.js animation code. (4) The Video Critic Agent evaluates the rendered video to detect issues in timing, narration–animation alignment, or visualization quality, and sends corrections to the coder agent for regeneration if needed.

applications of such agentic systems in the domain of static story generation [49], we propose a multi agentic system for the automated generation of data videos. As shown in figure 3.8, the framework consists of 4 different LLM agents with different roles. The agents are described below:

1. Director Agent: Given a data table, video intent, duration and a chart screenshot as input, the director agent is tasked to generate a scene by scene plan with subtitles for the animation.

2. Plan Critic Agent: The plan critic agent takes the complete plan and verifies

whether the intent is fulfilled, and makes corrections to the plan.

3. Coder Agent: Based on the corrected plan, this agent generates HTML code that plays an animation regarding the data.

4. Video Critic Agent: This agent evaluates the rendered video to identify and correct visual or temporal issues, such as still frames, misalignment between animations and subtitles, text occlusion, and unintended visual overlaps.

An example produced using this approach is shown in Figure 3.2. The prompts for each of the agents are discussed in Appendix A.

3.3.2 Evaluation Framework

We evaluate the generated *datareels* using two complementary judgment setups. First, we assess the regenerated HTML files using LLM-as-a-judge evaluation to analyze and critique the code. This setup is applied to outputs from all models using only the generated HTML code. Based on these results, we select the best-performing model, use it to implement the multi-agent framework, and render the HTML outputs generated by both the single-model and its multi-agent versions. We then use a VLM-as-a-judge to evaluate the rendered videos in a pair-wise manner, and validate these results through human evaluation.

HTML based Evaluation Framework: Following [49] and [81], our evaluation

suite employs an LLM-as-a-judge framework to assess the generated HTML, leveraging the strong code understanding and critique capabilities of large language models. Each output is scored on a 0–5 scale across four criteria that can be directly inferred from the HTML source without rendering the video. *Narrative Quality* evaluates whether the intended message is conveyed as a coherent and well-structured visual story. *Informativeness* measures the depth of insight provided by the subtitles and animations, rewarding interpretations that go beyond merely restating data values. *Subtitle–Transcript Similarity* assesses how faithfully the generated subtitles preserve the meaning and intent of the original narration transcript. Finally, *Code Correctness* examines whether the HTML structure and logic properly implement the intended animation plan.

Video centric Evaluation Framework: We use VLM-as-a-judge to conduct pairwise comparisons between two generated videos at a time. This setup enables us to evaluate the multi-agent system against the single-model approach using the same base model. Our preliminary experiments suggest that VLMs are more reliable in comparative (pairwise) judgments for video data. Based on this insight, we design a blind evaluation protocol in which the VLM is shown two videos without revealing their associated models and is asked to determine which video performs better across three criteria: *Visualization Quality*, *Subtitle–Animation Coherence*, and *Style Consistency*. For each criterion, the judge may select a winner or declare a tie, and

the overall winner is determined based on performance across the three categories. Visualization Quality evaluates whether the charts are rendered correctly, remain visually readable, and are free from clipping, overlapping, or rendering artifacts. Subtitle–Animation Coherence assesses whether subtitles are accurately aligned with the visual content, including timing synchronization and semantic consistency between text and animation. Style Consistency measures how closely each video adheres to the reference visual style in terms of color palette, layout, chart type, and overall design coherence.

Human Evaluation Framework: For the human evaluation, we follow the same pairwise comparison protocol used by the VLM judge. The Human annotators were presented with pairs of videos and asked to determine which one performs better across three criteria: *Visualization Quality*, *Subtitle–Animation Coherence*, and *Style Consistency*. Each video pair was evaluated with 2 annotator in 2 seperate rounds of evaluation. The overall winner is the model that outperforms the other in more categories, although ties are also allowed, consistent with the VLM evaluation scheme. Finally, we compute the agreement between human judgments and VLM selections to quantify how closely the VLM judge aligns with human preferences and also how well the annotators agree with eachother.

3.3.3 Model Selection and Experiment Setup

We evaluate four open-source and two closed-source mainstream Vision-Language Models (VLMs) on DATAREEL, covering diverse dataset sizes and domains. The evaluated models include Gemini 2.5 Pro [25], GPT-4.1 Mini [2], GPT-5.4 Mini [96], Claude 4.5 Opus [11], InternVL 3.5–8B [106], and Qwen 2.5 VL–7B [12].

Once the code for the *datareel* is generated, we use a lightweight Python framework with a headless browser and repeated screenshot capturing to convert the code into a *datareel*. All six models are evaluated under a single-model prompting setup. The best-performing model is then selected for the multi-agent framework. In both settings, the judge model is Gemini 3 Pro [38].

3.4 Result Analysis

In this section, we first present the findings from our experiments. We then conduct an ablation study on the different components of our multi-agentic setup, followed by a qualitative analysis to identify common issues and errors in the data videos generated by the models.

3.4.1 HTML based Evaluation

Table 3.3 gives us a look at the performance of the different models. For the 328 samples in the dataset, the metrics represent the average score achieved by different models. We observe that the closed source models are doing quite well across the four different metrics. Gemini 2.5 pro is the best performer here, followed by GPT 4.1 mini. GPT-5.4 Mini performs similarly to GPT-4.1 Mini overall, except on Subtitle–Transcript Similarity, where GPT-4.1 Mini achieves the best results. Claude Opus despite being a coding model falls short on the task. On the other hand, open-source models perform significantly worse. One big reason for this bad performance is size of the HTML files required to properly express the narrative and intent, which often is very near the context window limits of the smaller models. These results demonstrate that closed source models have good potential in data video generation, but open source models, on the other hand, still has a long way to go. This addresses our **RQ1**: *How effective are current state-of-the-art VLMs at generating multi-modal, non-static visualizations, such as data videos?*. While state of the art models can generate data video, their capabilities vary greatly.

Model	Narrative Quality	Informativeness	Subtitle–Transcript Similarity	Code Correctness	**Overall**
Gemini 2.5 Pro	**4.66**	**4.63**	3.33	**4.50**	**4.28**
GPT 4.1 mini	4.33	4.22	**3.37**	3.52	3.86
GPT 5.4 mini	4.49	4.39	2.77	3.51	3.79
Claude Opus 4.5	4.11	4.20	3.24	3.55	3.78
InternVL 3.5–8B	1.26	1.30	1.77	0.80	1.28
Qwen 2.5 VL–7B	0.92	0.92	1.37	0.67	0.97

Table 3.3: HTML-based evaluation of generated code quality across models.

3.4.2 Video Centric Evaluation

The next step of our evaluation suite involved a video-based comparison between the single-agent and multi-agent systems, both implemented using Gemini 2.5 Pro. We selecte 150 videos, and then send 2 videos side by side to the VLM Judge, which was Gemini 3 pro version, without mentioning which is the reel generated by the single prompt approach and which one is generated by the agentic approach. The results is disclosed in Table 3.4, where we can see that the VLM judge preferred the *datareel* generated by the multi-agentic framework 138 times out of the 150 evaluation samples.

Component	Human Judgment			VLM Judgment		
	Agentic Wins	**Agentic Loss**	**Ties**	**Agentic Wins**	**Agentic Loss**	**Ties**
Visualization Quality	61	51	38	103	21	26
Subtitle–Animation Coherence	56	36	58	70	14	66
Style Consistency	124	19	7	148	2	0
Overall	101	34	15	138	12	0

Table 3.4: Comparison of human and VLM judgments in pairwise evaluation of multi-agentic vs single-model outputs across evaluation components. Total samples used in this evalaution was 150.

3.4.3 Human Evaluation

We also conducted two rounds of human evaluation in order to validate the results obtained by the VLM judge during the Video Centric Evaluation. Table 3.4 presents the results for both the VLM judge and the human judges in one of the 2 rounds of human evaluation for the datareels. Firstly, we observe that the VLM judge has a higher tendancy of preferring reels generated by the multi-agentic approach. Humans preferred the agentic verison 101 times out of the 150 samples, which is not as many as the VLM judge. Both the Human and VLM judges are seen to prefer the videos generated by the multi-agentic setup.

Table 3.5 presents the agreement levels between human annotators across the two separate rounds, their agreement with the VLM judge, and the comparison

Comparison	Visualization Quality	Subtitle–Animation Coherence	Style Consistency	Overall
Human Evaluation 1 vs LLM	40.7%	41.3%	82.7%	68.7%
Human Evaluation 2 vs LLM	50.0%	46.0%	75.3%	63.3%
Human Evaluation 1 vs Human Evaluation 2	48.7%	48.7%	66.7%	63.3%
Human Consensus vs LLM	53.4% (n=73)	49.3% (n=73)	94.0% (n=100)	80.0% (n=95)

Table 3.5: Agreement between human evaluators and VLM judgments. Here, Human Consensus signifies cases where both annotators independently produced the same judgement.

between consensus human annotations and VLM annotations. We can observe that overall the 2 humans evaluation rounds have agreements scores of 68.7% and 63.3%. The two rounds of human evaluation show an agreement score of 63.3% between annotators, indicating that evaluating data reels is inherently subjective and can be challenging even for humans. We then restrict our analysis to samples where both human annotators agree, and find that the agreement with the LLM rises to **80%** when considering only these consensus cases. This further validates the results in Table 3.4, showing that both human evaluators and the VLM judge agree, to a certain extent, that videos generated by the multi-agent framework are better.

We observe that one of the big reasons behind the preference of the agentic version of the videos by the VLM judge is the style consistency factor. The videos generate by the single model approach are often at times very simple in terms of their designs choices and color palettes, while the actual references are much more

sophisticated with rich deisgn choices. The model is unable to properly replicate that if only a one shot generate approach is used. Table 3.4 shows that humans in 124 cases out of 150 believed that the agentic approach did a better job at following the reference judge. The VLM judge on the other hand chose the agentic version as better at replication 148 times, giving us an agreement of 82.7% in Table 3.5, and ultimately 94% when only cases of human consensus are considered. This shows that generally the multi-agentic setup generates better data videos, which answers our **RQ2**: *Can agentic capabilities be leveraged to improve the performance of VLMs in generating multi-modal visualizations, compared to single-model prompting?*

3.4.4 Ablation Study

In order to verify the importance of the two critic agents, we conducted an ablation study where we used only the Director agent and Coder agent for the purpose of *data reel* generation. We then compared the results with the ones generated by the single agent approach using the VLM judge. The results have been shown in Table 3.6. We observe that, compared to Table 3.4, the judge prefers the two-agent approach 117 times, which was previously 138 when all four agents were used. The addition of the critic agents improve the Visualization quality of the agentic workflow, as it increases from 67 wins with only director and coder agent to 103 wins out of 150 when the critic agents are used. Style consistency remains same for both cases,

Component	Wins	Loss	Ties
Visualization Quality	67	51	32
Subtitle–Animation Coherence	50	44	56
Style Consistency	147	0	3
Overall	117	33	0

Table 3.6: Ablation study comparing the agentic pipeline without critic components (No-Critic) against the Single-Shot baseline across evaluation components. Total samples for this evalaution was 150.

whereas Subtitle–Animation Coherence also sees a slight improvement.

3.4.5 Qualitative Analysis

In order to answer **RQ3**: *What are the common issues and challenges prevalent in data videos generated by VLMs?*, we investigate the common errors found in the data videos generated by the models. From each of the models we examined, we randomly sampled 25 *data reels* that were successfully renderable to get a better understanding of the challenges and issues that are persistent in the generated videos. We identified the following issues in the current state of the art models-

Reels with no Animation: A very common scenario with the models, especially

with the open source ones, where they produced HTML files that contained a chart only, without any visible animation or subtitle. We observed this issue in the closed source models like GPT 4.1 mini and GPT 5.4 mini too. The charts rendered by these models in Fig. 3.11(a), Fig. 3.10b and Fig. 3.12 were incomplete, with issues like clipping, text overlapping, irregular placement of chart objects like bars, axes, etc. This demonstrates the challenging nature of the task for the models involved.

Lack of sync with subtitle: Another common issue plaguing the generated reels was the lack of coherence between the animations and the subtitle. We often observe cases (Fig. 3.11(c) and Fig. 3.11(d)) where the animation starts much later than the appearance of the subtitle. There is also an issue of pacing, where it is seen that the LLMs struggle to keep the animation paced at a speed that is readable and enjoyable to the users, sometimes turning out to be too fast or too slow.

Overlapping and clipping texts and chart objects: Another prominent issue with the generated data reels was the overlapping and clipping of text and chart objects such as bars, axes, and legends, as seen in Fig. 3.9(a). This makes the subtitles of the reels difficult to understand and follow. Occlusion of subtitles due to background color was also noticed in a number of datareels.

External Component Errors: When asked to follow the reference, it was observed that some models Fig. 3.11(b), specially the multi-agentic approach tried utilizing external objects like pictures as background, icons etc. While this worked for most

files, some cases were observed where the models hallucinated image URL links, which ultimately did not render the desired background for the reels.

Lack of Adherence to the Reference: The single-prompt approach failed in most cases to follow the provided reference image. Table 3.4 makes this clear when examining the metric for style consistency. The human judges and the VLM judge are in agreement that the multi-agentic approach is much better at following the reference, compared to the single shot approach in case of Gemini 2.5 pro. Similar phenomenons were observed for the single prompt approach for the other models as well, like the rendered videos from Claude Opus, shown in Fig. 3.9, almost always features a purple background, ignoring the reference image that was provided with the prompt.

Instability in Animations and Charts: In some cases, the animations or charts appear jittery. For example, in Fig. 3.9(c) & Fig. 3.9(d) the visual elements appear to turn on and off intermittently, which disrupts the continuity of the display. Whenever the system attempts to remove focus from a specific part of the chart, the entire visualization tends to jitter and behave as if the whole graph is restarting, rather than smoothly unfocusing only the intended component.

Improper Chart Positioning: In several instances like Fig. 3.10a, the generated charts appear at distant or unexpected locations on the screen rather than near the center of the frame. This misplacement makes the visualization difficult to focus on

and reduces its overall clarity. Similarly, subtitles also appear in unintended positions, misaligned with the charts, further reducing the readability and effectiveness of the presentation.

3.5 Summary

Overall, in this chapter we address our problem statement on the use of VLMs for multi-modal data visualization generation. We show that current VLMs can generate data videos, although performance varies significantly across models **(RQ1)**. We demonstrate that a multi-agentic setup improves performance on this task **(RQ2)**. We also identify several recurring challenges, including lack of synchronization between animation components and overlapping elements among others **(RQ3)**. The next chapter makes an important investigation into the robustness of VLMs when dealing with misleading data visualizations.

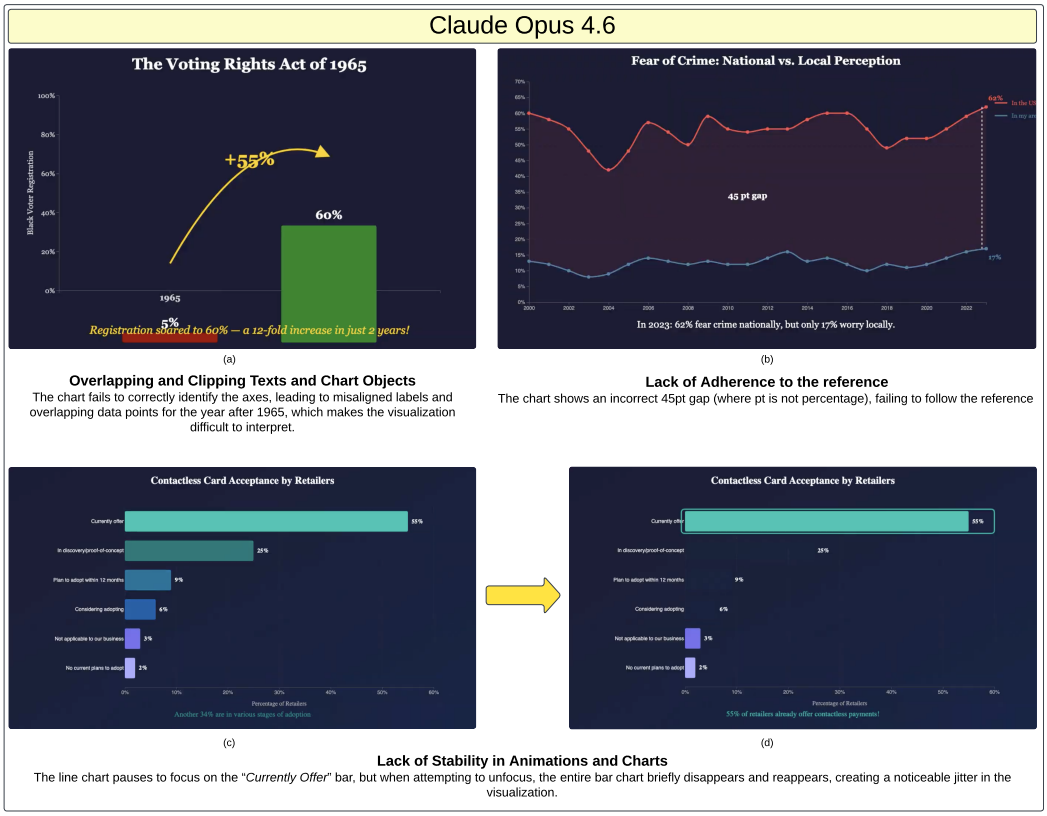
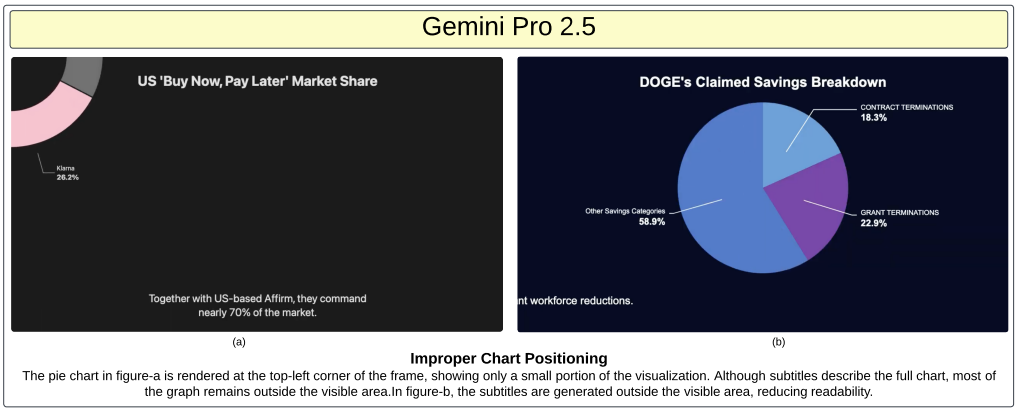
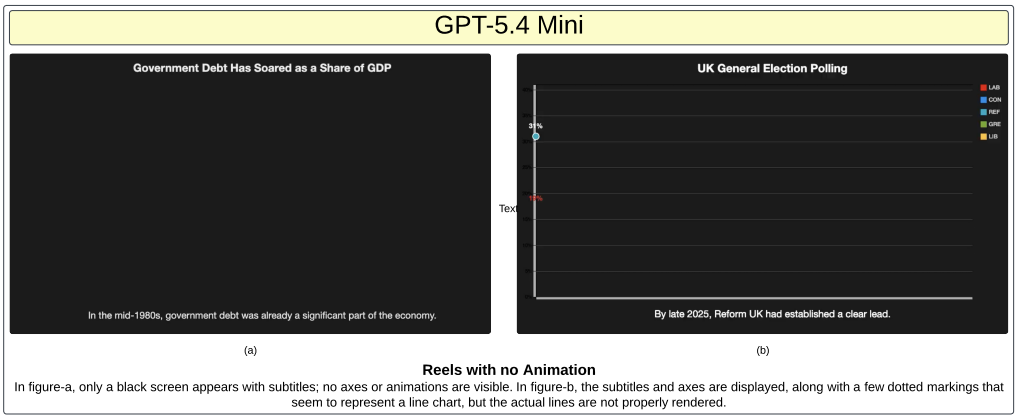


Figure 3.9: Claude Opus 4.6 exhibits overlapping chart elements, incorrect axis interpretation, an erroneous 45pt gap that does not correspond to the reference, and unstable animations where the chart briefly disappears and reappears when shifting focus



(a) Failure case from Gemini Pro 2.5.



(b) Failure cases from ChatGPT 5.4 Mini.

Figure 3.10: Examples of failure cases in chart visualization tasks. In (a), Gemini Pro 2.5 shows improper chart positioning, with the pie chart partially rendered and subtitles appearing outside the visible frame. In (b), ChatGPT 5.4 Mini produces a black screen with subtitles but no visible chart elements, or renders incomplete structures with scattered points instead of coherent lines.

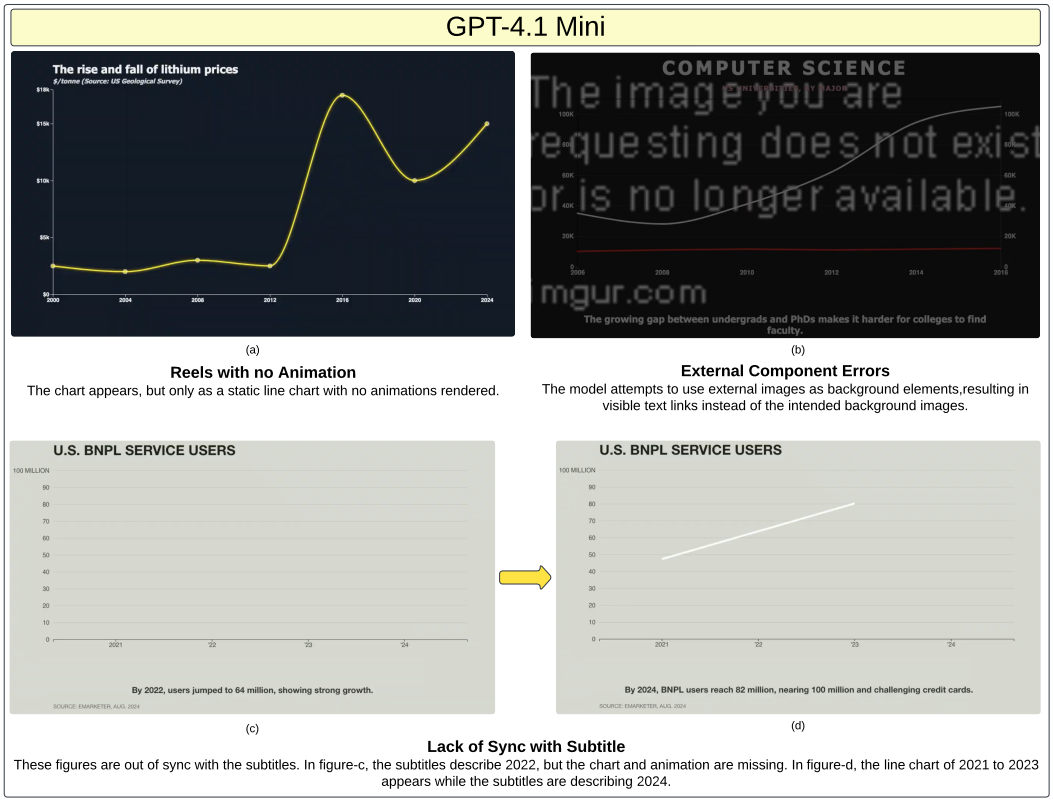


Figure 3.11: GPT-4.1 mini shows missing animations and synchronization issues, where only static charts are generated and the chart updates appear out of sync with the subtitles. External component errors are also observed, such as failed loading of background images due to invalid or hallucinated URLs.

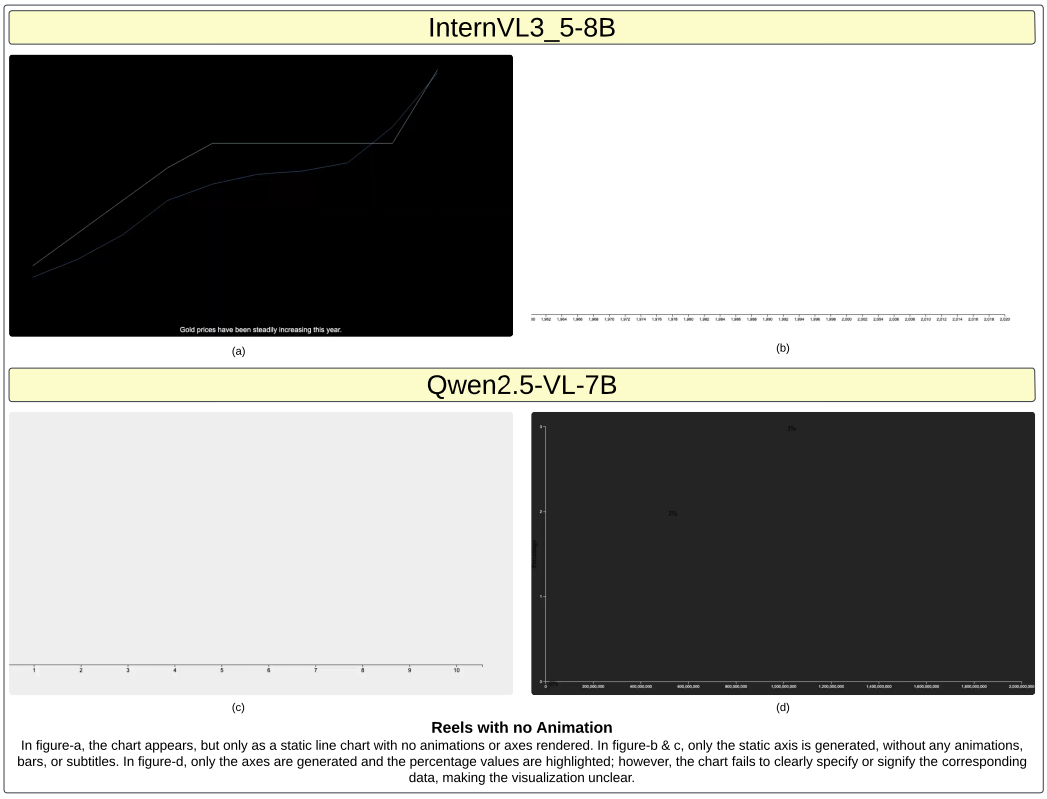


Figure 3.12: Failure cases from open-source vision–language models when generating chart-based reel content. (a–b) Outputs from InternVL3.5-8B and (c–d) outputs from Qwen2.5-VL-7B. Instead of producing a complete animated chart, the models generate incomplete visualizations such as static axes, partial lines, or highlighted percentages without proper data representation, resulting in unclear or blank diagrams.

4 Robustness of VLMs to Misleading Visualization

In this chapter, we address our problem statement related to the limitations of existing benchmarks in evaluating VLM robustness to misleading visual designs, particularly their lack of controlled experimental settings, large-scale datasets, and standardized evaluation metrics for rigorous analysis. We propose a standardized benchmark for this problem, grounded in realistic settings. Using this benchmark, we demonstrate that different misleading designs affect model behavior to varying degrees **(RQ4)**, and that different models show varying degrees of robustness **(RQ5)**. We further introduce a novel metric tailored to this task and propose a multi-agentic inference-time approach to mitigate the effects of misleading visualizations **(RQ6)**.

Recent studies have shown that misleading chart designs can influence real-world human decision-making [79, 57]. Charts are the building blocks of data stories. As we benchmark vision–language models (VLMs) for generating multimodal visualizations such as data videos, it is therefore important that these models are robust to misleading chart designs. Models that lack such robustness may produce biased

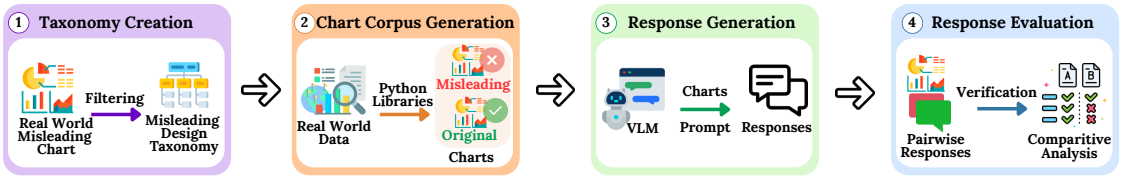


Figure 4.1: Overview of our methodology: **(1)** Taxonomy development, **(2)** Misleading chart corpus construction, **(3)** Prompt creation and VLM response generation, **(4)** Evaluating VLMs by comparing responses across original and misleading charts.

interpretations, which can subsequently propagate into biased data videos or data stories. For this investigation, we focus on static charts, allowing us to isolate and assess robustness to misleading designs.

4.1 Methodology

Our investigation follows four stages to identify whether and how misleading chart designs affect the performance of VLMs. First, we develop a taxonomy of common misleading chart designs. Second, we construct a chart corpus containing both controlled (accurate) and misleading versions of each chart. Third, we design prompts to systematically examine how different misleading designs impact model responses. Finally, we develop an evaluation framework to identify which VLMs are more vulnerable to these misleading designs. Below, we detail each stage.

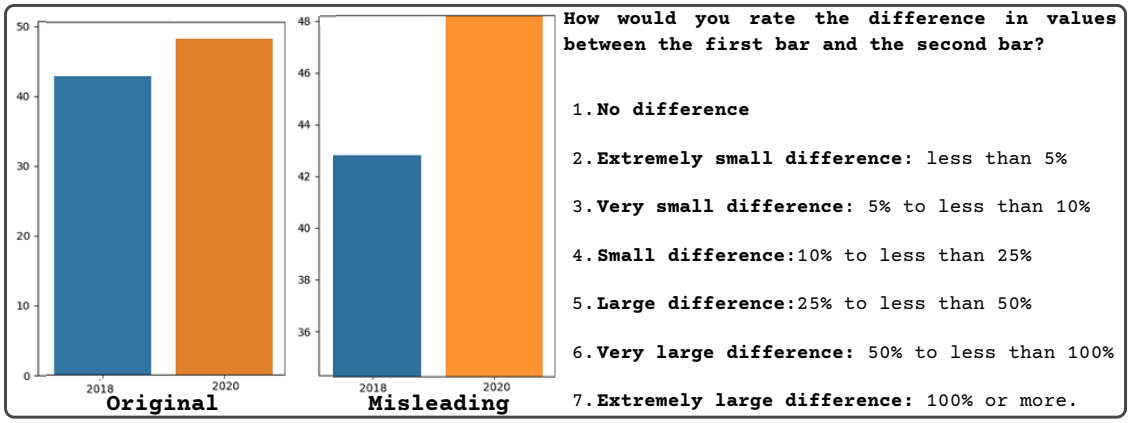


Figure 4.2: An example question prompt featuring a bar chart (*truncated axis*). Despite representing the same data, axis truncation on the misleading chart exaggerates the difference between the bars.

4.1.1 Taxonomy Creation

We began by reviewing prior work on real-world misleading visualizations, which often involve design-based deception alongside misleading captions or framing [66, 63]. Inspired by these examples, we identified recurring patterns of chart designs that, while statistically accurate, can still distort interpretation, such as inverting an axis to reverse the perceived trend. To ensure systematic evaluation, we refined our focus to include only those misleading designs compatible with survey-style, generalized questions. This choice enabled consistent and controlled testing of VLM susceptibility, following the methodology of Lauer *et al.* [57]. The selected chart types also allowed us to pair each visualization with generic, reusable questions—for

instance, in a bar chart, a simple prompt comparing the heights of two bars can reliably reveal the effect of a misleading design, without the need for chart-specific question engineering. To maintain this level of generality and control, we excluded forms of multimodal deception (e.g., misleading captions or annotations) and complex multi-view dashboards, where consistent survey-style prompts would be difficult to apply.

The final taxonomy is shown in Figure 4.3. Initially, we categorize misleading chart designs into two broad categories: *(a)* **Axis Manipulation**—alterations to chart axes that distort data implications, and *(b)* **Misleading Encoding**—visual encodings that deviate from best practices and misrepresent trends or magnitudes. Each category includes four representative types of visual distortions that can potentially mislead users. These designs have been widely recognized for their deceptive potential in prior studies [57, 79, 99].

The subcategories of Axis Manipulation include: *(i)* **Truncated Axis**: *y-axis* does not start at zero, exaggerating differences between bars (Figure 4.3(a)), *(ii)* **Aspect Ratio**: scales are distorted, misleading the viewer about the rate of change (Figure 4.3(b)), *(iii)* **Dual Axis**: two variables with different scales are plotted against separate *y-axes*, often suggesting a spurious correlation (Figure 4.3(c)), *(iv)* **Inverted Axis**: *Y-axis* is flipped, which can reverse the perceived trend direction (Figure 4.3(d)).

Figure 4.3: Examples of misleading chart designs.

The subcategories of misleading encoding include: *(i)* **Distorted Projection**: perspective effects make some segments of a 3D chart appear larger than they are (Figure 4.3(e)), *(ii)* **Data-Visual Disproportion**: bubble size is mapped to *radius* instead of *area*, visually inflating differences (Figure 4.3(f)), *(iii)* **Inappropriate Continuous Encoding**: a line chart is used to plot categorical variables, falsely implying continuity or trend (Figure 4.3(g)), *(iv)* **Inappropriate Categorical**

Category	Misleading Design	Chart Type
Axis Manipulation	`Truncated Axis` [Fig. 4.3 (a)]	Bar
	`Aspect Ratio` [Fig. 4.3 (b)]	Line
	`Dual Axis` [Fig. 4.3 (c)]	Line
	`Inverted Axis` [Fig. 4.3 (d)]	Line
Misleading Encoding	`Distorted Projection` [Fig. 4.3 (e)]	Pie
	`Data Visual Disproportion` [Fig. 4.3 (f)]	Bubble
	`Inappropriate Continuous Encoding` [Fig. 4.3 (g)]	Bar
	`Inappropriate Categorical Encoding` [Fig. 4.3 (h)]	Line

Table 4.1: Taxonomy of misleading chart designs we evaluated.

Encoding: continuous data is displayed using bar charts, disrupting perception of gradual variation (Figure 4.3(h)).

4.1.2 Chart Corpus Construction

While previous studies [66] have explored misleading charts, a major limitation is the absence of chart benchmarks containing both misleading charts and their corresponding original versions. Such paired examples are critical for isolating the impact of misleading design choices on VLMs. By controlling all other variables and altering only the chart design, we can directly attribute any differences in model responses to the misleading elements. Therefore, we first construct a corpus consisting of 800 original charts (100 charts of each of the 8 misleading subcategories), each paired with a misleading variant using the data tables from the benchmark ChartQA [69] corpus. We utilize widely adopted Python libraries [48, 110] to

generate both the original and misleading chart images. We ensured that both chart images (original and misleading) maintained consistent visual styles (e.g., chart type, color scheme, layout, typography), isolating the impact of the misleading design alone.

4.1.3 Response Generation

The response generation phase consists of two main components: *(i) Prompt Construction* and *(ii) Model Selection*.

(i) Prompt Construction: To evaluate VLM responses across various misleading chart types, we design a prompt P comprising three elements: *(a)* a general instruction f, *(b)* a chart-specific question C, and *(c)* a chart image I (original or misleading). The instruction f defines the task scope and explicit formatting guidelines for the response. The question C is a multiple-choice item about the chart, requiring the model to select an answer on a 7-point Likert scale [62] alongside a justification for its choice. To capture differences across misleading designs, we define a distinct chart-specific question $c_{\text{type}} \in C$ for each misleading chart type.

Figure 4.2 illustrates an example prompt used in our study, which includes a general instruction, a misleading bar chart exhibiting a *truncated axis*, and a multiple-choice question. The instructions direct the model to select the most appropriate answer from the listed options and justify its selection in a structured manner.

Our prompt design aims to assess whether VLMs are susceptible to misleading visualizations. Ideally, a robust model should recognize the misleading element in the chart, provide consistent answers across both the original and misleading versions, and articulate a clear rationale explaining why the chart is misleading, if applicable.

Each question in our benchmark is designed to reveal the specific effect of the corresponding misleading chart design. We adapt and extend the question formulations from Lauer et al. [57] by incorporating a neutral response option to allow nuanced interpretations. The following guidelines inform our question generation strategy:

• **Truncated axis (bar charts):** As this design amplifies visual differences of values, questions focus on comparing bar magnitudes.

• **Trend distortions (line charts; e.g., aspect ratio, inverted axes):** Questions evaluate perceived trend direction and strength.

• **Distorted projections and visual disproportion (e.g., pie charts, area charts):** Questions prompt models to compare relative magnitudes of chart segments.

• **Inappropriate data encoding (e.g., continuous vs. categorical mismatches):** Questions examine whether the model erroneously infers trends in discretized continuous data or fails to recognize categorical data where no trend

should be inferred.

(ii) Model Selection: We evaluate a diverse set of VLMs, including both proprietary and open-source models of varying sizes. Proprietary models such as GPT-4o [2], Gemini-2.5-Pro [37], and Claude-3.7-Sonnet [10] are included due to their state-of-the-art performance on chart understanding benchmarks like ChartQA [69]. Lightweight variants (e.g., GPT-4o-mini, Claude-3.5-Haiku) are also considered to assess performance under limited compute. For open-source models, to ensure a good balance between performance and computational efficiency, we select models below 10B parameters: Phi-4-5.57B [1], Qwen2.5-VL-7B [118], LLaVA-v1.6-Mistral-7B [58], MiniCPM-V-2.6-8B [120] and InternLM-7B [19]. For response generation, we use the default decoding parameters of each model: respective API endpoints for closed-source models, and HuggingFace [112] for open-source models.

4.1.4 Response Evaluation

The focus of our study was on how misleading visualizations can distort and influence model responses, similar to their effects on human judgment. To evaluate this, we examined whether VLMs' responses change when the same prompt is applied to both the original and a misleading version of a chart, assessing their sensitivity to visual distortions. A model is considered misled if its answer changes between the original and the misleading chart, with the effect quantified by the absolute difference in

Models	Aspect Ratio		Distorted Projection		Dual Axis		Inappropriate Categorical Enc.		Inappropriate Continuous Enc.		Inverted Axis		Data Visual Disproportion		Truncated Axis	
	z value	p	z value	p	z value	p	z value	p	z value	p	z value	p	z value	p	z value	p
Closed-Source Models																
GPT-4o	-3.25	$\mathbf{9.6e^{-4}}$	-0.80	0.41	-3.10	$\mathbf{1.6e^{-3}}$	-1.26	0.19	-8.14	$\mathbf{9.1e^{-17}}$	-7.78	$\mathbf{1.3e^{-15}}$	-2.27	$\mathbf{1.9e^{-2}}$	-7.61	$\mathbf{6.4e^{-15}}$
GPT-4o-mini	-7.41	$\mathbf{7.3e^{-14}}$	-0.44	0.63	-3.03	$\mathbf{2.0e^{-3}}$	-0.29	0.75	-6.96	$\mathbf{3.8e^{-13}}$	-7.97	$\mathbf{2.8e^{-16}}$	-2.18	$\mathbf{1.7e^{-4}}$	-6.45	$\mathbf{1.3e^{-12}}$
Gemini-2.5-Pro	-2.97	$\mathbf{6.6e^{-4}}$	-0.83	0.39	-1.53	0.12	-0.18	0.55	-1.5	0.10	-2.8	$\mathbf{1.9e^{-3}}$	-0.15	0.87	-5.6	$\mathbf{6.8e^{-9}}$
Claude-3.7-Sonnet	-0.59	0.54	-2.00	$\mathbf{3.3e^{-2}}$	-4.34	$\mathbf{9.3e^{-6}}$	-1.27	0.17	-4.45	$\mathbf{6.7e^{-6}}$	-6.99	$\mathbf{1.2e^{-12}}$	-2.44	$\mathbf{1.2e^{-2}}$	-3.33	$\mathbf{2.2e^{-4}}$
Claude-3.5-Haiku	-6.21	$\mathbf{2.3e^{-11}}$	-1.30	0.17	-4.31	$\mathbf{3.7e^{-6}}$	-0.15	0.87	-7.32	$\mathbf{1.5e^{-13}}$	-7.58	$\mathbf{1.3e^{-14}}$	-2.00	$\mathbf{3.0e^{-2}}$	-3.54	$\mathbf{9.3e^{-5}}$
Open-Source Models																
LLaVA-v1.6-Mistral-7B	-2.02	$\mathbf{2.5e^{-2}}$	-2.63	$\mathbf{4.9e^{-3}}$	-1.00	0.31	-5.09	$\mathbf{2.8e^{-8}}$	-8.59	$\mathbf{4.2e^{-23}}$	-5.44	$\mathbf{3.8e^{-10}}$	-6.90	$\mathbf{3.7e^{-13}}$	-1.00	0.32
InternLM-7B	-6.27	$\mathbf{4.2e^{-11}}$	-2.02	$\mathbf{3.9e^{-2}}$	-1.57	0.11	-0.03	0.97	-1.57	$\mathbf{9.7e^{-2}}$	-6.45	$\mathbf{1.2e^{-13}}$	-1.00	0.32	-1.00	0.32
Phi-4-5.57B	-4.33	$\mathbf{1.3e^{-5}}$	-1.34	0.15	-0.73	0.46	-4.07	$\mathbf{2.8e^{-5}}$	-2.06	$\mathbf{3.4e^{-2}}$	-7.18	$\mathbf{1.8e^{-15}}$	-4.38	$\mathbf{5.3e^{-7}}$	-5.5	$\mathbf{3.9e^{-10}}$
Qwen2.5-VL-7B	-6.95	$\mathbf{2.43e^{-13}}$	-1.33	0.18	-3.48	$\mathbf{1.9e^{-4}}$	-5.05	$\mathbf{3.8e^{-7}}$	-2.52	$\mathbf{1.03e^{-52}}$	-7.29	$\mathbf{2.8e^{-15}}$	-1.22	0.16	-5.96	$\mathbf{1.8e^{-9}}$
MiniCPM-V2.6-8B	-4.56	$\mathbf{4.3e^{-6}}$	-0.14	0.89	-0.86	0.38	-1.90	$5.5e^{-2}$	-3.83	$\mathbf{1.02e^{-4}}$	-7.35	$\mathbf{1.3e^{-14}}$	-0.69	0.44	-1.91	$5.2e^{-2}$

Table 4.2: Results of the Wilcoxon signed-rank test comparing original and misleading chart pairs. Here, $p < 0.05$ (in **bold** font) demonstrates that the model is significantly influenced by the misleading chart design, and the corresponding z value indicates the magnitude of this effect.

Likert scores between the two versions. In order to assess the statistical significance of the difference in responses caused by misleading chart designs, we conduct the Wilcoxon Signed Rank test [111], as the Shapiro-Wilk test [86] confirmed that the responses do not follow a normal distribution. We evaluate each model's responses to original and misleading charts as paired data and consider a misleading design to have a statistically significant effect if its p-value <0.05.

4.2 Results and Discussion

We present both qualitative and quantitative results addressing our two research questions. First, we analyze how misleading chart designs affect VLM responses (**RQ5**), followed by an evaluation of model robustness across different deception types (**RQ6**). Table 4.2 shows VLM performance on chart pairs using the Wilcoxon signed-rank test; a p-value <0.05 indicates a statistically significant difference, suggesting the model was misled.

Susceptibility to Types of Misleading Chart Designs (RQ4): Our findings confirm that all VLMs are vulnerable to misleading designs, with varying levels of susceptibility (Table 4.2). The misleading designs related to *axis manipulation*, particularly inverted axes, affect all 10 models, followed by aspect ratio and truncated axis distortions, which impact 9 and 7 models respectively. In contrast, misleading encoding designs like inappropriate categorical encoding or distorted projection had minimal impact (affecting only 3 models), although inappropriate continuous encoding showed moderate effects. These suggest that current VLMs are more prone to errors in interpreting spatial scale and structure than in detecting improper data encodings.

Robustness of VLMs to Misleading Chart Designs (RQ5): From Table 4.2, we find that except Gemini-2.5-Pro, other closed-source models exhibit a strong

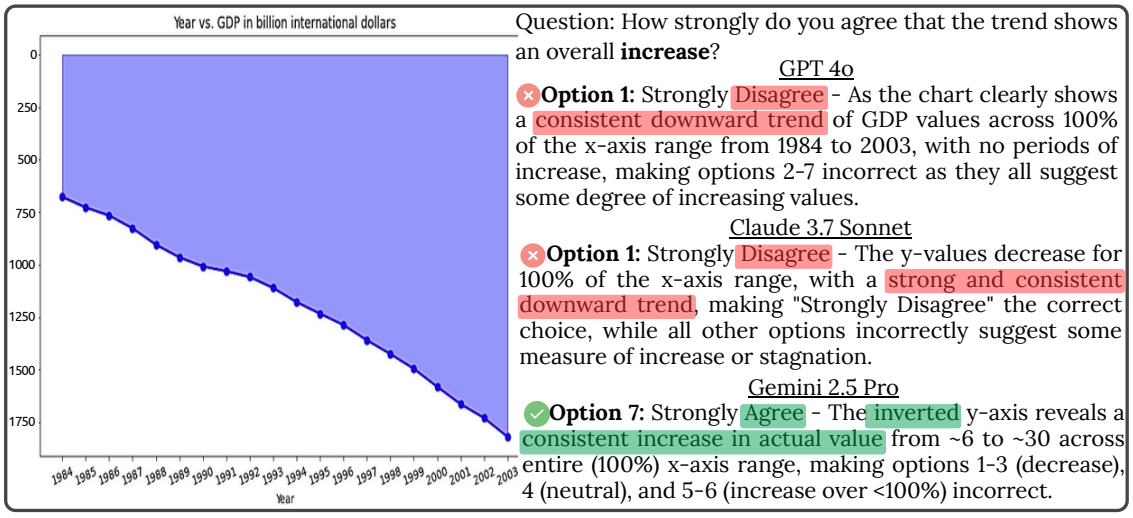


Figure 4.4: VLM responses to an inverted y-axis design type. Only Gemini correctly detects the axis inversion and adjusts its answer.

tendency to be deceived by misleading designs demonstrating vulnerability across six different misleading categories. In contrast, Gemini-2.5-Pro demonstrated stronger robustness, affected by only three.

In terms of open-source models, LLaVA, Phi-4, and Qwen were misled by six misleading types, while the MiniCPM-V2.6 demonstrated the highest robustness, misled by only three. Although both open and closed models struggled with axis manipulations like inverted axes and aspect ratios, key differences emerged: open-source models were uniquely affected by inappropriate categorical encoding and showed less sensitivity to truncated axes, an area where all closed models failed significantly.

Case Study for inverted axis: Given Gemini-2.5-Pro's top performance, we closely examined its behavior on inverted-axis charts—a category where other models performed poorly. Although statistical analysis still indicates a significant effect of this misleading design on Gemini (with moderately elevated z and p values), its performance stands out compared to other models. Upon close examination of the 100 samples of this type, we observed that Gemini correctly identified the inversion in the y-axis in 79% of the cases. Figure 4.4 illustrates an example of such a case where Gemini explicitly references the axis inversion in its explanation and chooses the correct answer, demonstrating greater awareness of the misleading design. In contrast, none of the open-source models could identify the axis inversion in their responses, often leading to incorrect predictions. These findings suggest that Gemini-2.5-Pro's advanced reasoning capabilities could be attributed to its greater performance.

4.3 Quantifying Deception

In our previous experiments, we established that models are indeed susceptible to misleading visualization designs. However, existing metrics fail to capture the extent of this deception. To address this, we introduce Deception Score, a distance-based metric that isolates the impact of visual manipulation from inherent model error, enabling more meaningful comparisons across different VLMs and categories of

deception. Using the Deception Score, we extend our evaluation to additional VLMs, further consolidating our findings for **RQ4** and **RQ5**, and reinforcing that different misleading designs affect VLM responses to varying degrees.

4.3.1 Motivation

When a VLM interprets a chart and selects an answer from our ordinal scale, its response may deviate from the gold answer for two distinct reasons:

1. **Baseline model error:** Even on correctly designed (original) charts, models may misinterpret the visualization due to limitations in visual reasoning, numerical estimation, or chart comprehension. This error reflects the model's inherent capability and is independent of any deceptive design.

2. **Deception-induced error:** When viewing a misleading chart, the model may be further led astray by the deceptive visual design, resulting in responses that deviate more substantially from the gold answer.

To measure the true effect of deception, we must disentangle these two sources of error.

4.3.2 Deception Metric Definition

We define **deception score** in a Likert-scale based question answering setup as the absolute numeric difference between the options chosen by the VLM in presence and the option chosen in absence of misleading designs. In our setup, the Likert-scale consists of seven points (Figure 4.2). The equation for the deception score is as follows:

$$\text{Deception Score} = \frac{1}{N} \sum_{i=1}^{N} [D(M_i, G_i) - D(O_i, G_i)] \tag{4.1}$$

where for each chart pair i:

- G_i is the Gold answer
- O_i is the answer the model produced for the original chart
- M_i is the answer the model produced for the misleading chart
- $D(M_i, G_i)$ is the distance between the model's response on the misleading chart and the gold answer
- $D(O_i, G_i)$ is the distance between the model's response on the original (non-misleading) chart and the gold answer
- N is the total number of paired chart samples

For our ordinal response scales, we compute D as the absolute difference between the selected option number and the gold option number. For instance, if the gold answer is option 3 and the model selects option 5, the distance is $|5 - 3| = 2$.

4.3.3 Interpretation

The deception score has an intuitive interpretation:

- **Positive values** indicate that the misleading design successfully deceived the model, causing it to deviate further from the correct answer than it would on an equivalent non-misleading chart.
- **Zero** indicates that the misleading design had no effect, the model's error on the misleading chart is the same as on the original chart.
- **Negative values** indicate that the model actually performed better on the misleading chart than the original.

4.3.4 Extending the Misleading Chart Evaluation Framework

To evaluate models using the deception score within our Likert-scale–based question answering setup, we require gold answers for each question. To support this, and to account for different misleading design subtypes, we extend our benchmark by programmatically deriving gold answers from the underlying data tables. This

Figure 4.5: Different Deception Types and Sub-types in the Extended Taxonomy.

extension transforms our evaluation from a purely comparative analysis into a proper accuracy-based assessment with deception score, allowing us to measure both susceptibility to deception and its magnitude. Additionally, we broaden the taxonomy by introducing new deception categories and increasing subtype diversity within existing ones, resulting in a more comprehensive evaluation resource.

Building on the taxonomy introduced in Table 4.1, we retain the core focus on

design-based deception—cases where charts remain visually plausible but use design choices that distort interpretation of the underlying data. In this extension, we also collect the corresponding gold answers. For example, in the bar chart shown in Figure 4.2, a question comparing relative bar heights has a mathematically determinable correct answer derived directly from the underlying data table, independent of any visual distortion in the chart.

The extended benchmark is organized around eight major forms of misleading chart design: *(i) aspect ratio distortion*, *(ii) data–visual disproportion*, *(iii) distorted projection*, *(iv) inappropriate encoding*, *(v) dual axis*, *(vi) inappropriate color coding*, *(vii) inverted axis*, and *(viii) truncated axis*. Compared to our earlier taxonomy, this represents both continuity and expansion:

- **Retained categories:** Truncated axis, aspect ratio, dual axis (now under inappropriate axis design), inverted axis, distorted projection, data–visual disproportion, inappropriate continuous encoding, and inappropriate categorical encoding are carried forward from the previous dataset.

- **Expanded subtypes:** Several categories now include additional subtypes. For example, *aspect ratio* includes both *compression* (from the earlier dataset) and newly added *expansion* examples. Similarly, *distorted projection* now covers both 3D pie charts (previously included) and newly added 3D donut

charts. The *truncated axis* category has been expanded to include multiple truncation styles: normal truncation, middle truncation, horizontal truncation, and two-bar comparisons.

- **Reorganized categories:** The previously separate encoding-based categories are now unified under *inappropriate encoding*, with two subtypes: inappropriate categorical encoding and inappropriate continuous encoding.

- **New category:** We introduce *inappropriate color coding* specifically for choropleth maps, covering misleading color choices such as rainbow palettes, disordered color scales, and overly similar colors. This category was not present in our earlier work and addresses an important class of real-world visualization deception.

Table 4.3 provides a detailed breakdown of the extended taxonomy. Figure 4.5 shows examples from the different types and subtypes of misleading designs.

To generate the chart images, we use multiple rendering frameworks, including **Matplotlib**, **Vega-Lite**, and **Highcharts.js**. We use this mixed generation pipeline because different chart families are more naturally supported by different tools. For example, several legacy chart types from the previous benchmark are preserved in Matplotlib, many of the newly added structured chart variants are generated in Vega-Lite, and the 3D donut-chart subset is rendered with Highcharts.js.

Major Category	Subcategories Summary	Chart Type
Truncated Axis	Middle Truncation (25); Vertical Truncation (50); Horizontal Truncation (25)	Bar
Aspect Ratio	Expansion (50); Compression (50)	Line
Dual Axis	Dual-axis (100)	Line
Inverted Axis	Inverted Axis (100)	Line
Distorted Projection	Pie Charts (75); Donut Charts (25)	Pie, Donut
Data–Visual Disproportion	Data–Visual Disproportion (100)	Bubble, Scatter
Inappropriate Encoding	Inappropriate Categorical Encoding (50); Inappropriate Continuous Encoding (50)	Line, Bar
Inappropriate Color Coding	Close Colors (38); Rainbow Colors (25); Disordered Colors (37)	Choropleth Map

Table 4.3: Breakdown of the extended benchmark by major deception category, subcategory, and chart type. Each major category contains 100 paired original–misleading chart instances, yielding 800 pairs (1,600 chart images) in total.

4.3.5 Evaluating Models Using Deception Score

We evaluated 10 models on the extended benchmark using the proposed deception score metric, with results presented in Table 4.4. Among the evaluated systems, Qwen3-VL-8B exhibits the highest overall vulnerability, while Gemini-2.5-Flash and Gemini-2.5-Pro achieve the lowest aggregate Deception Scores. More broadly, proprietary models tend to exhibit lower overall deception scores than most open-source systems, although no model is consistently robust across all deception categories.

At the category level, *Inverted Axis* produces the largest Deception Scores across nearly all models, indicating that reversing axis orientation fundamentally disrupts VLM interpretation of visual trends. *Inappropriate Encoding* also yields consistently

Models	Aspect Ratio	Distorted Projection	Dual Axis	Inappropriate Encoding	Inappropriate Color Coding	Inverted Axis	Data Visual Disproportion	Truncated Axis	Overall
Closed-Source Models									
GPT-4o	0.02	0.03	**0.25**	0.50	**0.41**	1.97	0.46	0.38	0.502
Claude-4.5-Haiku	-0.06	0.00	0.22	0.32	0.06	2.19	0.30	0.34	0.422
Gemini-2.5-Pro	-0.07	-0.05	-0.05	0.23	0.37	0.18	0.19	-0.01	0.099
Gemini-2.5-Flash	0.02	-0.06	0.07	0.02	0.26	0.39	0.16	0.01	0.109
Open-Source Models									
Gemma3-12B-IT	0.00	**0.09**	0.01	0.68	-0.15	1.99	0.05	-0.09	0.323
Pixtral-12B	**0.30**	0.06	-0.17	0.35	0.12	2.43	**0.62**	0.53	0.530
Qwen3-VL-2B	-0.05	0.07	0.00	-0.17	-0.15	2.57	-0.26	0.00	0.251
Qwen3-VL-4B	0.29	0.08	-0.04	0.34	0.22	**2.61**	0.36	0.63	0.561
Qwen3-VL-8B	0.07	**0.09**	0.15	1.58	0.17	2.09	0.29	**0.66**	**0.638**
Qwen3-VL-32B	0.17	-0.01	0.08	**1.80**	0.16	2.11	0.51	0.15	0.621

Table 4.4: Deception scores (DS) for each misleading chart category. DS measures the average increase in error when viewing misleading charts compared to original charts. The highest DS in each category is shown in **bold**.

high scores, suggesting that VLMs strongly depend on learned associations between chart types and semantic interpretation. Similarly, *Truncated Axis* and *Data–Visual Disproportion* generate elevated deception scores for several models, particularly among open-source systems. In contrast, *Aspect Ratio Distortion* and *Distorted Projection* generally produce comparatively small deception scores.

These findings reinforce our answer to both **RQ4** and **RQ5**, indicating models are affected by misleading design and their answer changes solely for this aspect. The models do vary significantly in their susceptibility to visual deception, but all

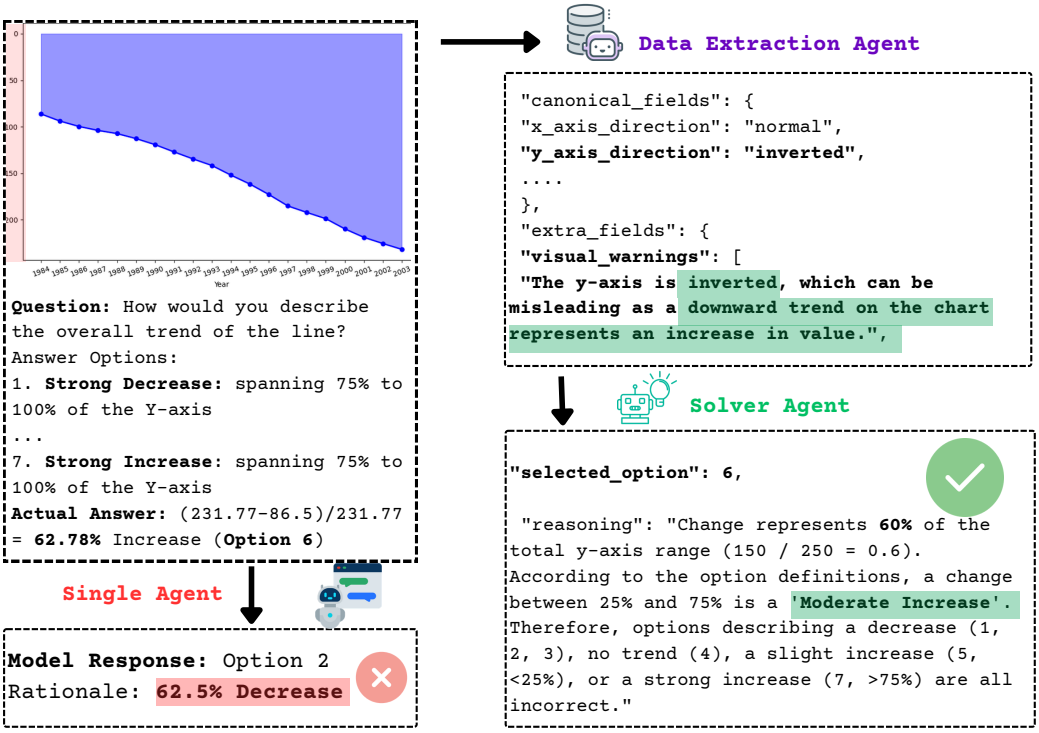


Figure 4.6: An example showing how using a Multi-Agentic approach as opposed to a single LLM helps in mitigating deception in case of inverted axis

models exhibit positive deception scores.

4.4 Mitigating Deception with a Multi-Agentic Approach

4.4.1 Methodology

To address **RQ6:** *Can inference-time mitigation approaches improve the robustness of these VLMs to misleading chart designs?*, we propose a two-agent pipeline that introduces an intermediate data extraction step prior to question answering. Our

hypothesis is that models are less likely to be misled when reasoning is grounded in the underlying data table rather than the visual chart. Given that VLMs are effective at extracting structured data from charts, providing this extracted data for downstream reasoning can help reduce the influence of misleading visual cues, potentially allowing the model to recognize inconsistencies without explicit guidance.

Our approach is illustrated in Figure 4.6. In the standard single-VLM setting, models often rely directly on visual appearance and therefore become susceptible to deceptive design choices such as inverted axes or truncated baselines. To mitigate this issue, we introduce a two-agent framework consisting of a *Data Extraction Agent* and a *Solver Agent*.

Stage 1: Structured Chart Extraction. The *Data Extraction Agent* analyzes the chart image and generates a structured Vega-Lite-style JSON representation describing the chart structure and underlying data. The extracted representation includes axis properties (e.g., scale direction, minimum and maximum values, scale type, and baseline configuration), visual encodings, legend mappings, and reconstructed data points. Importantly, during this structured extraction process, stronger models frequently generate additional metadata describing potentially misleading visual properties, even without explicit prompting to detect deception. For example, models often identify truncated baselines, inverted axes, or abnormal scaling behaviors while reconstructing chart structure. These observations are typically recorded in an

auxiliary `visual_warnings` field generated by the model itself rather than enforced by the prompting format.

Stage 2: Grounded Reasoning. The *Solver Agent* answers the chart question using both the original chart image and the structured metadata produced by the extraction stage. Providing explicit structural representations encourages the model to reason from reconstructed chart semantics and numerical relationships instead of relying primarily on potentially misleading visual appearance. As illustrated in Figure 4.6, this additional grounding often corrects failures observed in the single-VLM setting, particularly for deception types involving axis manipulation and misleading visual scaling.

4.4.2 Result Analysis

Table 4.5 compares the overall deception scores of single-agent and multi-agent settings across closed-source and open-source VLMs. Overall, the multi-agent setup reduces deception scores for most models, suggesting that explicitly extracting chart metadata before answering helps models become more robust to misleading visual designs. Among the closed-source models, GPT-4o, Claude Haiku, and Gemini 2.5 Pro all show clear reductions in deception score, with Gemini 2.5 Pro achieving the strongest improvement, dropping from 0.099 to 0.024. This indicates that the model's responses became more consistent between original and misleading

Model	Single Agent	Multi-Agent	Change
Closed-Source Models			
GPT-4o	0.502	0.255	0.247 ↓
Claude Haiku	0.422	0.228	0.194 ↓
Gemini 2.5 Pro	0.099	0.024	0.075 ↓
Gemini 2.5 Flash	0.109	0.135	0.026 ↑
Open-Source Models			
Gemma3 12B IT	0.323	0.211	0.112 ↓
Pixtral 12B	0.530	0.395	0.135 ↓
Qwen3-VL 2B	0.251	0.275	0.024 ↑
Qwen3-VL 4B	0.561	0.339	0.222 ↓
Qwen3-VL 8B	0.638	0.513	0.125 ↓
Qwen3-VL 32B	0.621	0.365	0.256 ↓

Table 4.5: Deception Scores: Single LLM vs Multi-Agent Setup (Overall). The Change column indicates whether the deception score decreases or increases under the multi-agent setup.

charts when grounded with structured chart metadata. However, Gemini 2.5 Flash shows the opposite trend, with its deception score increasing from 0.109 to 0.135, suggesting that the additional metadata does not consistently improve robustness for all models.

A similar pattern is observed among open-source models. Most open-source VLMs benefit from the multi-agent setup, including Gemma3 12B IT, Pixtral 12B,

Qwen3-VL 4B, and Qwen3-VL 8B. Qwen3-VL 32B shows a particularly strong improvement, decreasing from 0.621 to 0.365, while Qwen3-VL 8B remains the most susceptible overall despite improvement, with the highest deception score in both settings. However, not all open-source models improve. Qwen3-VL 2B show increased deception scores under the multi-agent setup. These results suggest that while the proposed multi-agent approach is generally effective, its benefits are not uniform across model families or sizes.

In relation to **RQ6:** *Can inference-time mitigation approaches improve the robustness of these VLMs to misleading chart designs?*, these findings suggest that inference-time multi-step approaches can meaningfully reduce the impact of misleading visualizations for certain models in majority of cases, but they do not universally improve robustness.

4.4.3 Qualitative Analysis

To better understand why the multi-agent setup succeeds for most models and fails in some, we qualitatively compare Gemini 2.5 Pro and Gemini 2.5 Flash across a number of representative misleading chart examples from the different chart types. These cases show that detecting a misleading design is not always sufficient; the model must also use that detection to override the misleading visual impression and ground its answer in the underlying values. Figure 4.7 illustrates this distinction

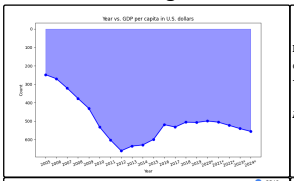
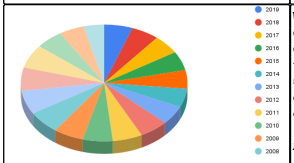
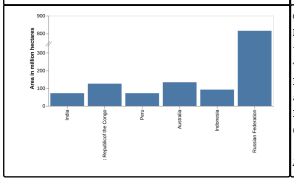
Misleading Chart	QA	Gemini 2.5 Pro	Gemini 2.5 Flash
	How would you describe the overall trend of the line? Answer: Option 6	**Data Extraction Agent** visual_warnings: The y-axis is inverted, which can be misleading as downward slopes represent increases in value. **Solver Agent** selected_option: 6 - Modereate Increase, reasoning: The chart has an inverted Y-axis, as noted in the metadata, meaning a downward visual trend represents an increase	**Data Extraction Agent** "The y-axis is inverted, with 0 at the top and increasing values downwards, which can be counter-intuitive for typical quantitative data representation." **Solver Agent** selected_option: 2 - Moderate Decrease, reasoning: Signifies an overall decrease in the actual GDP per capita
	Which statement best describes the disparity between the highest and the second highest quantity in the pie chart? Answer: Option 2	**Data Extraction Agent** visual_warnings: Pie chart with many slices of similar size ... Pseudo-3D effect can distort the perceived size of the slice **Solver Agent** selected_option: 1 - No disparity, reasoning: all segments appear to be of equal size. The metadata confirms this observation	**Data Extraction Agent** visual_warnings: The 3D effect can distort the perceived size of the pie slices. Precise values for each slice are not provided **Solver Agent** selected_option: 7 - Very large disparity, reasoning: ratio is 'more than 5x' the quantity of the second-largest segment
	Question: How different is the value of China compared to Slovak Republic in the chart? Answer: Options 7	**Data Extraction Agent** Visually, the most populous countries like China and India appear to have the highest values **Solver Agent** selected_option: 7 - Highest Difference, reasoning: China is colored with the darkest shade of blue, the Slovak Republic is colored with the lightest shade of blue	**Data Extraction Agent** visual_warnings: Most countries appear to be in the lowest value range of the color scale, making differentiation between countries with low values difficult **Solver Agent** selected_option: 1 - No Difference, reasoning: Both China and the Slovak Republic are colored in the lightest blue shade on the map
	Question: How would you rate the difference in values between Russian Federation and Democratic Republicof the Congo? Answer: Options 6	**Data Extraction Agent** The y-axis has a scale break between 300 and 800, which visually compresses the difference between the smaller bars and the largest bar (Russian Federation). **Solver Agent** selected_option: 6 - Very Large Difference, reasoning: the value for the Russian Federation is 820 million hectares, and for the Congo, it is 125 million hectares	**Data Extraction Agent** visual_warnings: The y-axis features a discontinuity symbol between 300 and 800, Russian Federation falls within this visually compressed or skipped range **Solver Agent** selected_option: 5 - Large Difference, reasoning: Value for the Russian Federation is approximately 325 million hectares, and for Congo, it is 125 million hectares

Figure 4.7: A qualitative analysis showing cases where mitigation approach has been successful for the best model, Gemini 2.5 pro, but failed in case of its light-weight counterpart, Gemini 2.5 flash.

across inverted-axis, disotrted projection, inappropriate color coding, and truncated axis examples.

In the top-most example of figure 4.7, the misleading design in inverted axis, and both models identify the misleading axis configuration, but they use this information differently. Gemini 2.5 Pro's Data Extraction Agent warns that "The y-axis is inverted," and adds that this "can be misleading as downward slopes represent increases in value.," and its Solver Agent selects *Option 6: Moderate Increase*. Its

reasoning explicitly states that although "The chart has an inverted Y-axis, as noted in the metadata". In contrast, Gemini 2.5 Flash also detects that "The y-axis is inverted" and that lower values are plotted higher, creating a misleading downward trend. However, its Solver Agent still selects *Option 2: Moderate Decrease*, reasoning only from the apparent visual trend. Thus, Gemini 2.5 Flash detected the inversion but still answered based on the visual appearance, whereas Gemini 2.5 Pro explicitly corrected for the inversion and used the actual numerical data to determine the true trend: increase, not decrease.

A similar pattern appears in the distorted projection pie chart example. Both models detect that the 3D effect can distort perceived slice sizes. Gemini 2.5 Pro notes that "the use of a pseudo-3D effect can distort the perception of slice sizes" and that "the slices appear to be of very similar size." Its Solver Agent selects *Option 1: No disparity*, reasoning that the pie chart displays many slices that are visually almost identical. Gemini 2.5 Flash also recognizes that "the 3D effect can distort the perceived size of the pie slices" and notes that precise values are not provided. However, its Solver Agent selects *Option 7: Very large disparity*. Both models are incorrect in this case, but Gemini 2.5 Pro is closer to the correct answer because it notes the actual data pattern, while Gemini 2.5 Flash estimates proportions.

The inappropriate color-coding map further shows how errors can originate from the Data Extraction Agent can affect the Solver Agent. Gemini 2.5 Pro

correctly identifies that visually, highly populated countries such as China and India appear to have the highest values. Its Solver Agent selects *Option 7: Highest Difference*, reasoning that China is shown with the darkest blue shade, while the Slovak Republic is shown with the lightest blue shade. By contrast, Gemini 2.5 Flash's Data Extraction Agent states that most countries appear to be in the lowest value range, falling victim to the misleading design. Its Solver Agent then incorrectly selects *Option 1: No Difference*, reasoning that both China and the Slovak Republic are colored in the lightest blue shade.

Finally, in the truncated axis examples, Gemini 2.5 Pro's Data Extraction Agent identifies that "the y-axis has a scale break between 300 and 800," which visually compresses the difference between smaller bars and the largest bar, Russian Federation. Its Solver Agent selects *Option 6: Very Large Difference*, using the values of approximately 820 million hectares for Russian Federation and 125 million hectares for Congo. Gemini 2.5 Flash also detects the discontinuity symbol and notes that Russian Federation falls within a visually compressed or skipped range. However, its Solver Agent selects *Option 5: Large Difference*, estimating Russian Federation as 325 million hectares, again falling victim to the misleading design.

Overall, these qualitative examples suggest that the multi-agent approach is most effective when the Metadata Agent accurately extracts the relevant chart information and the Answer Agent appropriately uses both the extracted metadata

and the chart image. This also indicates that stronger models are likely to benefit more from the mitigation pipeline, as observed in the cases of Gemini 2.5 Pro.

In relation to **RQ6:** *Can inference-time mitigation approaches improve the robustness of these VLMs to misleading chart designs?*, in light of this qualitative analysis, we can say that the for the mitigation approach to work, the model has to be capable of not only detection misleading designs, but also taking that detection into account when answering question about affected charts.

4.4.4 Summary

In this chapter, we address the limitations of existing benchmarks in evaluating VLM robustness, particularly the lack of metrics to quantify model susceptibility to misleading designs. We construct a dataset of misleading visualization designs and systematically analyze which designs most influence model behavior (**RQ4**) and which models are most affected (**RQ5**). To support this, we introduce a metric that isolates the impact of visual deception from inherent model error, demonstrating that different models are affected at varying degrees. We further propose an inference-time mitigation approach to improve robustness against misleading design choices (**RQ6**). In the next Chapter we conclude this thesis with some concluding remarks and directions for future research.

5 Conclusions

In this chapter, we conclude the thesis and outline potential directions for future extensions of the work.

5.1 Concluding Remarks

This thesis explored the capabilities and limitations of vision–language models (VLMs) in the context of data-driven visual storytelling, addressing both the generation of coherent data videos and the robustness of models to misleading visual designs.

Firstly, we introduced the novel task of automated *datareel* generation and proposed DATAREEL, a benchmark of real-world data reels that captures the complexity of multi-modal storytelling involving chart animations and aligned narration. Our experiments demonstrated that even VLM models struggle with this task, often producing outputs with issues in visual encoding, animation stability, and narrative coherence. We proposed a multi-agent framework that decomposes the

generation process into planning, coding, and critique stages, leading to improved performance over single-model prompting.

Secondly, we investigated the susceptibility of VLMs to misleading visual designs through a taxonomy and large-scale benchmark of paired original and misleading charts, revealing that models are highly sensitive to subtle visual distortions despite unchanged underlying data. We further introduced a distance-based deception metric to quantify this behavior and showed that multi-agent inference-time mitigation strategies can reduce, but not eliminate, such vulnerabilities.

Together, this thesis provides a unified perspective on advancing VLMs for both generating coherent data narratives and ensuring robustness in the presence of misleading visual inputs, laying the foundation for more reliable and interpretable AI systems for real-world data communication.

5.2 Future Works

There are several promising directions for future research building on this thesis, both in terms of advancing multi-model visualization generation and ensuring robustness of VLM when dealing with them.

Firstly, for data video clip or *datareel* generation, future work can focus on training models specifically tailored for this task, as well as addressing the common errors and limitations identified in our analysis. Inclusion of voice-overs in place

of on-screen subtitle, and the syncing of audio and video can be the next major contribution.

Secondly, the current setup for generating *datareels* can be extended to full-length data videos, which are widely used by journalism outlets. This would involve combining multiple *datareels* with coherent real-world footage to create more engaging and comprehensive data stories. Capabilities of VLMs, such as image and video retrieval from web sources, can be leveraged for this purpose, along with training VLMs to handle higher-level tasks such as directing and editing full-length data videos in a manner similar to professional production workflows.

Thirdly, the study of VLM robustness can be extended to more complex multi-modal visualization formats, such as data videos and dashboards. As efforts are made to improve model robustness against misleading inputs, it is equally important to investigate whether VLMs themselves generate misleading designs in visualization tasks.

Overall, we believe this thesis provides a foundational step toward advancing research on VLMs for data visualization.

Bibliography

[1] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[3] Mubashara Akhtar, Oana Cocarascu, and Elena Simperl. Reading and reasoning over chart images for evidence-based automated fact-checking. In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 399–414, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.30. URL `https://aclanthology.org/2023.findings-eacl.30`.

[4] Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. Chartcheck: An evidence-based fact-checking dataset over real-world chart images, 2023.

[5] Mubashara Akhtar, Nikesh Subedi, Vivek Gupta, Sahar Tahmasebi, Oana Cocarascu, and Elena Simperl. Chartcheck: Explainable fact-checking over real-world chart images, 2024.

[6] Jason Alexander, Priyal Nanda, Kai-Cheng Yang, and Ali Sarvghad. Can gpt-4 models detect misleading visualizations? In *2024 IEEE VIS*, pages 106–110, 2024.

[7] Fereshteh Amini, Nathalie Henry Riche, Bongshin Lee, Christophe Hurter, and Pourang Irani. Understanding data videos: Looking at narrative visualization

through the cinematography lens. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*, pages 1459–1468, 2015.

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

[9] 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 *Proceedings of the 2018 international conference on advanced visual interfaces*, pages 1–9, 2018.

[10] Anthropic. Claude 3.7 sonnet and claude code, February 2025. URL `https://www.anthropic.com/news/claude-3-7-sonnet`. Accessed: 2025-07-01.

[11] Anthropic. System card: Claude opus 4.5, 2025. URL `https://assets.anthropic.com/m/64823ba7485345a7/Claude-Opus-4-5-System-Card.pdf`.

[12] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun 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, 2025. URL `https://arxiv.org/abs/2502.13923`.

[13] Eden Bensaid, Mauro Martino, Benjamin Hoover, and Hendrik Strobelt. Fairytailor: A multimodal generative framework for storytelling. *arXiv preprint arXiv:2108.04324*, 2021.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[15] Alberto Cairo. *How charts lie: Getting smarter about visual information*. WW Norton & Company, 2019.

[16] Guangyao Chen, Siwei Dong, Yu Shu, Ge Zhang, Jaward Sesay, Börje F. Karlsson, Jie Fu, and Yemin Shi. Autoagents: A framework for automatic agent generation, 2024.

[17] Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. Viseval: A benchmark for data visualization in the era of large language models. *IEEE TVCG*, 2024.

[18] Siming Chen et al. Social media visual analytics. In *Computer Graphics Forum*, volume 36, pages 563–587. Wiley Online Library, 2017.

[19] Zhe Chen, Weiyun Wang, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024.

[20] Zixin Chen, Sicheng Song, Kashun Shum, Yanna Lin, Rui Sheng, Weiqi Wang, and Huamin Qu. Unmasking deceptive visuals: Benchmarking multimodal large language models on misleading chart question answering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 13767–13800, 2025.

[21] 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. In *Computer Graphics Forum*, volume 41, pages 527–539. Wiley Online Library, 2022.

[22] Kiroong Choe, Chaerin Lee, et al. Enhancing data literacy on-demand: Llms as guides for novices in chart interpretation. *IEEE TVCG*, 2024.

[23] Jinho Choi, Sanghun Jung, Deok Gun Park, Jaegul Choo, and Niklas Elmqvist. Visualizing for the non-visual: Enabling the visually impaired to use visualization. In *Computer Graphics Forum*, volume 38, pages 249–260. Wiley Online Library, 2019.

[24] Neil Cohn. Visual narrative comprehension: Universal or not? *Psychonomic Bulletin & Review*, 27(2):266–285, 2020.

[25] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

[26] Michael Correll, Enrico Bertini, and Steven Franconeri. Truncating the y-axis: Threat or menace? In *Proc. of the 2020 CHI*, pages 1–12, 2020.

[27] Amit Kumar Das and Klaus Mueller. Misvisfix: An interactive dashboard for detecting, explaining, and correcting misleading visualizations using large language models. *IEEE Transactions on Visualization and Computer Graphics*, 2025.

[28] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[29] Rui Ding, Shi Han, Yong Xu, Haidong Zhang, and Dongmei Zhang. Quicklinsights: Quick and automatic discovery of insights from multi-dimensional data. In *Proceedings of the 2019 International Conference on Management of Data*, SIGMOD '19, page 317–332, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450356435. doi: 10.1145/3299869.3314037. URL `https://doi.org/10.1145/3299869.3314037`.

[30] Arlen Fan, Yuxin Ma, Michelle Mancenido, and Ross Maciejewski. Annotating line charts for addressing deception. In *Proc. of the 2022 CHI*, pages 1–12, 2022.

[31] Clara Fernandez-Labrador, Mertcan Akçay, Eitan Abecassis, Joan Massich, and Christopher Schroers. Divas: Video and audio synchronization with dynamic frame rates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26846–26854, 2024.

[32] Adam Fourney, Gagan Bansal, Hussein Mozannar, Cheng Tan, Eduardo Salinas, Friederike Niedtner, Grace Proebsting, Griffin Bassman, Jack Gerrits, Jacob Alber, et al. Magentic-one: A generalist multi-agent system for solving complex tasks. *arXiv preprint arXiv:2411.04468*, 2024.

[33] Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, 2024.

[34] Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*, 2023.

[35] Lily W Ge, Yuan Cui, and Matthew Kay. Calvi: Critical thinking assessment for literacy in visualizations. In *Proc. of the 2023 CHI*, pages 1–18, 2023.

[36] Yingqiang Ge, Wenyue Hua, Kai Mei, jianchao ji, Juntao Tan, Shuyuan Xu, Zelong Li, and Yongfeng Zhang. Openagi: When llm meets domain experts. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 5539–5568. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1190733f217404edc8a7f4e15a57f301-Paper-Datasets_and_Benchmarks.pdf.

[37] Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, and et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

[38] Google. Gemini 3.1 pro preview, 2026. URL https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview.

[39] Joshua Gorniak, Yoon Kim, Donglai Wei, and Nam Wook Kim. Vizability: Enhancing chart accessibility with llm-based conversational interaction. In *Proc. of the 37th Annual ACM Symposium on UIST*, pages 1–19, 2024.

[40] Yi Guo, Xiaoyu Qi, Haoyang Li, Jing Zhang, Danqing Shi, Qing Chen, Daniel Weiskopf, and Nan Cao. Urania: Visualizing data analysis pipelines for natural language-based data exploration. *ACM Transactions on Interactive Intelligent Systems*, 16(1):1–26, 2026.

[41] Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. Chartllama: A multimodal llm for chart understanding and generation, 2023.

[42] Yi He, Shixiong Cao, Yang Shi, Qing Chen, Ke Xu, and Nan Cao. Leveraging large models for crafting narrative visualization: A survey, 2024. URL https://arxiv.org/abs/2401.14010.

[43] Enamul Hoque, Parsa Kavehzadeh, and Ahmed Masry. Chart question answering: State of the art and future directions, 2022.

[44] Darrell Huff. *How to lie with statistics*. Penguin UK, 2023.

[45] Jessica Hullman and Nick Diakopoulos. Visualization rhetoric: Framing effects in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2231–2240, 2011. doi: 10.1109/TVCG.2011.255.

[46] Jessica Hullman, Nicholas Diakopoulos, and Eytan Adar. Contextifier: Automatic generation of annotated stock visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 2707–2716, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318990. doi: 10.1145/2470654.2481374. URL https://doi.org/10.1145/2470654.2481374.

[47] Jessica Hullman, Steven Drucker, Nathalie Henry Riche, Bongshin Lee, Danyel Fisher, and Eytan Adar. A deeper understanding of sequence in narrative visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2406–2415, 2013. doi: 10.1109/TVCG.2013.119.

[48] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

[49] Mohammed Saidul Islam, Md Tahmid Rahman Laskar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. DataNarrative: Automated data-driven storytelling with visualizations and texts. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19253–19286, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.emnlp-main.1073.

[50] Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*, 2024.

[51] Shankar Kantharaj, Xuan Long Do, Rixie Tiffany Leong, Jia Qing Tan, Enamul Hoque, and Shafiq Joty. OpenCQA: Open-ended question answering with charts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11817–11837, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.811. URL https://aclanthology.org/2022.emnlp-main.811.

[52] Shankar Kantharaj, Rixie Tiffany Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4005–4023, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.277. URL https://aclanthology.org/2022.acl-long.277.

[53] Shankar Kantharaj, Rixie Tiffany Ko Leong, Xiang Lin, Ahmed Masry, Megh Thakkar, Enamul Hoque, and Shafiq Joty. Chart-to-text: A large-scale benchmark for chart summarization. *arXiv preprint arXiv:2203.06486*, 2022.

[54] Deept Kumar, Naren Ramakrishnan, Richard F Helm, and Malcolm Potts. Algorithms for storytelling. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 604–610, 2006.

[55] Xingyu Lan and Yu Liu. "i came across a junk": Understanding design flaws of data visualization from the public's perspective. *IEEE TVCG*, 2024.

[56] Xingyu Lan, Yang Shi, Yanqiu Wu, Xiaohan Jiao, and Nan Cao. Kineticharts: Augmenting affective expressiveness of charts in data stories with animation design. *IEEE Transactions on Visualization and Computer Graphics*, 28(1): 933–943, 2022. doi: 10.1109/TVCG.2021.3114775.

[57] Claire Lauer and Shaun O'Brien. The deceptive potential of common design tactics used in data visualizations. In *Proc. of SIGDOC 2020*, pages 1–9, 2020.

[58] Bo Li, Yuanhan Zhang, Guo, et al. Llava-onevision: Easy visual task transfer. *arXiv:2408.03326*, 2024.

[59] Boyang Li, Stephen Lee-Urban, George Johnston, and Mark Riedl. Story generation with crowdsourced plot graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 27, pages 598–604, 2013.

[60] Bozheng Li, Miao Yang, Zhenhan Chen, Jiawang Cao, Mushui Liu, Yi Lu, Yongliang Wu, Bin Zhang, Yangguang Ji, Licheng Tang, et al. Opusanimation: Code-based dynamic chart generation. *arXiv preprint arXiv:2510.03341*, 2025.

[61] Yitong Li, Zhe Gan, Yelong Shen, Jingjing Liu, Yu Cheng, Yuexin Wu, Lawrence Carin, David Carlson, and Jianfeng Gao. Storygan: A sequential conditional gan for story visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6329–6338, 2019.

[62] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[63] Maxim Lisnic, Cole Polychronis, Alexander Lex, and Marina Kogan. Misleading beyond visual tricks: How people actually lie with charts. In *Proc. of the 2023 CHI*, pages 1–21, 2023.

[64] Maxim Lisnic, Alexander Lex, and Marina Kogan. "yeah, this graph doesn't show that": Analysis of online engagement with misleading data visualizations. In *Proceedings of the 2024 CHI*, 2024. ISBN 9798400703300. doi: 10.1145/3613904.3642448. URL https://doi.org/10.1145/3613904.3642448.

[65] Leo Yu-Ho Lo and Huamin Qu. How good (or bad) are llms at detecting misleading visualizations? *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[66] Leo Yu-Ho Lo, Ayush Gupta, Kento Shigyo, Aoyu Wu, Enrico Bertini, and Huamin Qu. Misinformed by visualization: What do we learn from misinformative visualizations? In *Computer Graphics Forum*, volume 41, pages 515–525. Wiley Online Library, 2022.

[67] Sheng Long and Matthew Kay. To cut or not to cut? a systematic exploration of y-axis truncation. In *Proc. of the 2024 CHI*, pages 1–12, 2024.

[68] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177.

[69] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of ACL 2022*, pages 2263–2279, 2022.

[70] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (to appear)*. Association for Computational Linguistics, December 2023.

[71] Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. Unichart: A universal vision-language pretrained model for chart comprehension and reasoning. In *Proc. of EMNLP 2023*, pages 14662–14684, 2023.

[72] Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartInstruct: Instruction tuning for chart comprehension and reasoning. In *Findings of ACL 2024*, pages 10387–10409, August 2024. URL https://aclanthology.org/2024.findings-acl.619/.

[73] Ahmed Masry, Mehrad Shahmohammadi, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. Chartinstruct: Instruction tuning for chart comprehension and reasoning, 2024.

[74] 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, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19123–19151, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.978. URL https://aclanthology.org/2025.findings-acl.978/.

[75] Ahmed Masry, Abhay Puri, Masoud Hashemi, Juan A Rodriguez, Megh Thakkar, Khyati Mahajan, Vikas Yadav, Sathwik Tejaswi Madhusudhan, Alexandre Piché, Dzmitry Bahdanau, et al. Bigcharts-r1: Enhanced chart reasoning with visual reinforcement finetuning. *arXiv preprint arXiv:2508.09804*, 2025.

[76] Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. ChartGemma: Visual instruction-tuning for chart reasoning in the wild. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 625–643, January 2025. URL https://aclanthology.org/2025.coling-industry.54/.

[77] S. McKenna, N. Henry Riche, B. Lee, J. Boy, and M. Meyer. Visual narrative flow: Exploring factors shaping data visualization story reading experiences. *Computer Graphics Forum*, 36(3):377–387, 2017. doi: https://doi.org/10.1111/cgf.13195. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.13195.

[78] Ali Modarressi, Ayyoob Imani, Mohsen Fayyaz, and Hinrich Schütze. Ret-llm: Towards a general read-write memory for large language models, 2023.

[79] Anshul Vikram Pandey et al. How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques. In *Proc. of CHI 2015*, pages 1469–1478, 2015.

[80] Feiyuan Qu, Tan Tang, Zeyang Fu, Yan Chen, Hanze Jia, Junming Gao, Songela Nurdawulieti, and Yingcai Wu. Visualizing tree-of-analysis: Facili-

tating conversational visual analytics for novices. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pages 1–20, 2026.

[81] Mizanur Rahman, Md Tahmid Rahman Laskar, Shafiq Joty, and Enamul Hoque. Text2vis: A challenging and diverse benchmark for generating multimodal visualizations from text. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31837–31862, 2025.

[82] Raian Rahman, Rizvi Hasan, Abdullah Al Farhad, Md Tahmid Rahman Laskar, Md Hamjajul Ashmafee, and Abu Raihan Mostofa Kamal. Chartsumm: A comprehensive benchmark for automatic chart summarization of long and short summaries. *arXiv preprint arXiv:2304.13620*, 2023.

[83] Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos, and Sheelagh Carpendale. *Data-driven storytelling*. CRC Press, 2018.

[84] James A Roberts and Meredith E David. Technology affordances, social media engagement, and social media addiction: An investigation of tiktok, instagram reels, and youtube shorts. *Cyberpsychology, Behavior, and Social Networking*, 28(5):318–325, 2025.

[85] Edward Segel and Jeffrey Heer. Narrative visualization: Telling stories with data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6): 1139–1148, 2010. doi: 10.1109/TVCG.2010.179.

[86] Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611, 1965.

[87] Leixian Shen, Yizhi Zhang, Haidong Zhang, and Yun Wang. Data player: Automatic generation of data videos with narration-animation interplay. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):109–119, 2023.

[88] Leixian Shen, Haotian Li, Yun Wang, Tianqi Luo, Yuyu Luo, and Huamin Qu. Data playwright: Authoring data videos with annotated narration. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[89] Leixian Shen, Haotian Li, Yun Wang, and Huamin Qu. From data to story: Towards automatic animated data video creation with llm-based multi-agent systems, 2024. URL `https://arxiv.org/abs/2408.03876`.

[90] Leixian Shen, Haotian Li, Yun Wang, and Huamin Qu. Reflecting on design paradigms of animated data video tools. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–21, 2025.

[91] D. Shi, Y. Shi, X. Xu, N. Chen, S. Fu, H. Wu, and N. Cao. Task-oriented optimal sequencing of visualization charts. In *2019 IEEE Visualization in Data Science (VDS)*, pages 58–66, Los Alamitos, CA, USA, oct 2019. IEEE Computer Society. doi: 10.1109/VDS48975.2019.8973383. URL https://doi.ieeecomputersociety.org/10.1109/VDS48975.2019.8973383.

[92] Danqing Shi, Fuling Sun, Xinyue Xu, Xingyu Lan, David Gotz, and Nan Cao. Autoclips: An automatic approach to video generation from data facts. In *Computer Graphics Forum*, volume 40, pages 495–505. Wiley Online Library, 2021.

[93] Danqing Shi, Xinyue Xu, Fuling Sun, Yang Shi, and Nan Cao. Calliope: Automatic visual data story generation from a spreadsheet. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):453–463, 2021. doi: 10.1109/TVCG.2020.3030403.

[94] 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 *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380966. doi: 10.1145/3411764.3445337. URL https://doi.org/10.1145/3411764.3445337.

[95] Yang Shi, Zhaorui Li, Lingfei Xu, and Nan Cao. Understanding the design space for animated narratives applied to illustrations. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380959. doi: 10.1145/3411763.3451840. URL https://doi.org/10.1145/3411763.3451840.

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

[97] Jennifer G Stadler, Kipp Donlon, Jordan D Siewert, Tessa Franken, and Nathaniel E Lewis. Improving the efficiency and ease of healthcare analysis through use of data visualization dashboards. *Big data*, 4(2):129–135, 2016.

[98] Mengdi Sun, Ligan Cai, Weiwei Cui, Yanqiu Wu, Yang Shi, and Nan Cao. Erato: Cooperative data story editing via fact interpolation. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):983–993, 2023. doi: 10.1109/TVCG.2022.3209428.

[99] Danielle Albers Szafir. The good, the bad, and the biased: Five ways visualizations can mislead (and how to fix them). *interactions*, 25(4):26–33, 2018.

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

[101] Bo Tang, Shi Han, Man Lung Yiu, Rui Ding, and Dongmei Zhang. Extracting top-k insights from multi-dimensional data. In *Proceedings of the 2017 ACM International Conference on Management of Data*, SIGMOD '17, page 1509–1524, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450341974. doi: 10.1145/3035918.3035922. URL https://doi.org/10.1145/3035918.3035922.

[102] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, and Jean-Baptiste Alayrac et al. Gemini: A family of highly capable multimodal models, 2023.

[103] Mohammed Majbah Uddin, Rahmat Ullah, and Mohammad Moniruzzaman. Data visualization in annual reports–impacting investment decisions. *International Journal for Multidisciplinary Research*, 6(5), 2024.

[104] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023.

[105] Huichen Will Wang, Larry Birnbaum, and Vidya Setlur. Jupybara: Operationalizing a design space for actionable data analysis and storytelling with llms. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–24, 2025.

[106] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

[107] Yun Wang, Zhida Sun, Haidong Zhang, Weiwei Cui, Ke Xu, Xiaojuan Ma, and Dongmei Zhang. Datashot: Automatic generation of fact sheets from tabular data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):895–905, 2020. doi: 10.1109/TVCG.2019.2934398.

[108] Yun Wang, Leixian Shen, Zhengxin You, Xinhuan Shu, Bongshin Lee, John Thompson, Haidong Zhang, and Dongmei Zhang. Wonderflow: Narration-centric design of animated data videos. *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[109] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

[110] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[111] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.

[112] Thomas Wolf, Lysandre Debut, Victor Sanh, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*, pages 38–45, October 2020. URL `https://www.aclweb.org/anthology/2020.emnlp-demos.6`.

[113] G. Wu, S. Guo, J. Hoffswell, G. Chan, R. A. Rossi, and E. Koh. Socrates: Data story generation via adaptive machine-guided elicitation of user feedback. *IEEE Transactions on Visualization & Computer Graphics*, 30(01):131–141, jan 2024. ISSN 1941-0506. doi: 10.1109/TVCG.2023.3327363.

[114] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. Autogen: Enabling next-gen llm applications via multi-agent conversation, 2023.

[115] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

[116] Kaige Xie and Mark Riedl. Creating suspenseful stories: Iterative planning with large language models. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2391–2407, St. Julian's, Malta, March 2024. Association for Computational Linguistics. URL `https://aclanthology.org/2024.eacl-long.147`.

[117] Liwenhan Xie, Chengbo Zheng, Haijun Xia, Huamin Qu, and Chen Zhu-Tian. Waitgpt: Monitoring and steering conversational llm agent in data analysis with on-the-fly code visualization. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, pages 1–14, 2024.

[118] An Yang, Baosong Yang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[119] Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions, 2023.

[120] Yuan Yao, Tianyu Yu, Ao Zhang, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

[121] Liang Zhang, Anwen Hu, Haiyang Xu, Ming Yan, et al. Tinychart: Efficient chart understanding with visual token merging and program-of-thoughts learning. *arXiv preprint arXiv:2404.16635*, 2024.

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

[123] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

[124] Wangchunshu Zhou, Yuchen Eleanor Jiang, Peng Cui, Tiannan Wang, Zhenxin Xiao, Yifan Hou, Ryan Cotterell, and Mrinmaya Sachan. Recurrentgpt: Interactive generation of (arbitrarily) long text, 2023.

[125] Yuxuan Zhu, Antony Kellermann, Akul Gupta, Philip Li, Richard Fang, Rohan Bindu, and Daniel Kang. Teams of llm agents can exploit zero-day vulnerabilities. *arXiv preprint arXiv:2406.01637*, 2024.

A Appendix

This appendix contains the full system prompts used in our data video generation pipeline and misleading chart evaluation framework. Each prompt is presented across multiple figures for readability.

Data Video Generation Prompts

- **Data Video Generator System Prompt** (Figures A.9–A.11): The baseline single-agent prompt for generating animated data videos.
- **Director Prompt** (Figures A.12–A.13): Used by the Director agent to create structured animation plans.
- **Plan Critic Prompt** (Figures A.14–A.15): Used to validate and refine the proposed animation plan.
- **Coder Prompt** (Figures A.16–A.18): Used by the Coder agent to generate D3.js HTML animations.

- **Video Critic Prompt** (Figures A.19–A.20): Used to evaluate rendered video output and provide feedback.
- **HTML Judge Prompt** (Figures A.21–A.23): Evaluation prompt for scoring generated videos on narrative quality, informativeness, transcript similarity, and code correctness.
- **Pairwise VLM Judge Prompt** (Figures A.24–A.25): Evaluation prompt for pairwise comparison of generated videos.

Misleading Chart Evaluation Questions

Each evaluation prompt instructs the VLM to analyze the chart, provide a brief explanation of its reasoning, and select an answer. The unique question and answer options for each misleading chart type are presented in Figures A.1–A.4. Generalized versions of these prompts, where specific entity mentions (e.g., state names, country names) are replaced with generic placeholders for reproducibility, are presented in Figures A.5–A.8.

Misleading Chart Evaluation Prompts in Extended Dataset

The following figures present Misleading Chart Evaluation Prompts that were used in the extended version of the Dataset. These prompts are shown Figures A.5–A.8.

Misleading Chart Evaluation Prompts (1/4)

Truncated Axis

How would you rate the difference in values between the first bar and the second bar?

1. No difference: The values of the two bars are exactly equal.
2. Extremely small difference: The difference is less than 5% of the larger value.
3. Very small difference: The difference is 5% to less than 10% of the larger value.
4. Small difference: The difference is 10% to less than 25% of the larger value.
5. Large difference: The difference is 25% to less than 50% of the larger value.
6. Very large difference: The difference is 50% to less than 100% of the larger value.
7. Extremely large difference: The difference is 100% or more (i.e., one bar is at least double the other).

Inverted Axis

How strongly do you agree that the trend shows an overall increase?

1. Strongly Disagree: The y-values decrease for 100% of the x-axis range, with a strong and consistent downward trend.
2. Disagree: The y-values decrease for 50% to less than 100% of the x-axis range.
3. Somewhat Disagree: The y-values decrease for 1% to less than 50% of the x-axis range.
4. Neutral: The y-values neither show a clear increase nor a decrease across the x-axis.
5. Somewhat Agree: The y-values increase for 1% to less than 50% of the x-axis range.
6. Agree: The y-values increase for 50% to less than 100% of the x-axis range.
7. Strongly Agree: The y-values increase for 100% of the x-axis range, with a strong and consistent upward trend.

Figure A.1: Evaluation prompts for truncated axis and inverted axis.

Misleading Chart Evaluation Prompts (2/4)

Aspect Ratio

How strong is the trend depicted in the chart?

1. Neutral: The trend is neither clearly increasing nor decreasing, with no consistent direction.

2. Very Weak: The trend is nearly flat, with barely any noticeable change, and occupies less than 10% of the x-axis.

3. Weak: The trend is shallow and difficult to distinguish, with small changes in y-values, and occupies around 10% to 30% of the x-axis.

4. Slightly Weak: The trend is present but not very strong, showing fluctuations, and occupies around 30% to 50% of the x-axis.

5. Slightly Strong: The trend is noticeable and somewhat steady, with a clear direction, and occupies around 50% to 70% of the x-axis.

6. Strong: The trend is distinct and sustained, with a well-defined direction, and occupies around 70% to 90% of the x-axis.

7. Very Strong: The trend is steep and consistent, dominating the entire chart, and occupies more than 90% of the x-axis.

Data-Visual Disproportion

Which statement best describes the disparity between the quantities represented by the circles?

1. There is no disparity: The circles represent equal quantities.

2. The disparity is very small: The larger circle represents 1x to 1.5x the quantity of the smaller circle.

3. The disparity is small: The larger circle represents 1.5x to 2x the quantity of the smaller circle.

4. The disparity is moderate: The larger circle represents 2x to 3x the quantity of the smaller circle.

5. The disparity is large: The larger circle represents 3x to 5x the quantity of the smaller circle.

6. The disparity is very large: The larger circle represents 5x to 10x the quantity of the smaller circle.

7. The disparity is extremely large: The larger circle represents more than 10x the quantity of the smaller circle.

Figure A.2: Evaluation prompts for aspect ratio and data-visual disproportion.

Misleading Chart Evaluation Prompts (3/4)

Dual Axis

How would you rate the difference between the trends of the two lines in the chart?

1. No Similarity: The trend is similar for 0% of the chart, showing completely different movements.

2. Extremely Small Similarity: The trend is similar for less than 5% of the chart, with only brief matching segments.

3. Very Small Similarity: The trend is similar for 5% to less than 10% of the chart, with minimal overlap.

4. Small Similarity: The trend is similar for 10% to less than 25% of the chart, but differences dominate.

5. Large Similarity: The trend is similar for 25% to less than 50% of the chart, showing a mix of similarities and differences.

6. Very Large Similarity: The trend is similar for 50% to less than 100% of the chart, with only minor differences.

7. Extremely Large Similarity: The trend is similar for 100% of the chart, following the same direction, slope, and fluctuations.

Distorted Projection

Which statement best describes the disparity between the highest and the second highest quantity in the pie chart?

1. No disparity: The two largest segments represent equal quantities.

2. Extremely small disparity: The largest segment represents 1x to more than 1.25x the quantity of the second-largest segment.

3. Very small disparity: The largest segment represents 1.25x to more than 1.5x the quantity of the second-largest segment.

4. Small disparity: The largest segment represents 1.5x to more than 2x the quantity of the second-largest segment.

5. Moderate disparity: The largest segment represents 2x to more than 3x the quantity of the second-largest segment.

6. Large disparity: The largest segment represents 3x to more than 5x the quantity of the second-largest segment.

7. Very large disparity: The largest segment represents more than 5x the quantity of the second-largest segment.

Figure A.3: Evaluation prompts for dual axis and distorted projection.

Misleading Chart Evaluation Prompts (4/4)

Inappropriate Continuous Encoding

How would you describe the overall trend depicted in the chart?

1. No Trend: The data primarily represents categorical values.
2. Extremely Weak: The y-values change for 1% to less than 5% of the x-axis range.
3. Very Weak: The y-values change for 5% to less than 10% of the x-axis range.
4. Weak: The y-values change for 10% to less than 20% of the x-axis range.
5. Strong: The y-values change for 20% to less than 50% of the x-axis range.
6. Very Strong: The y-values change for 50% to less than 75% of the x-axis range.
7. Extremely Strong: The y-values change for 75% to 100% of the x-axis range.

Inappropriate Categorical Encoding

How would you describe the overall trend depicted in the chart?

1. Neutral: There is no significant upward or downward trend in the chart.
2. Extremely Weak: The y-values change for 1% to less than 5% of the x-axis range.
3. Very Weak: The y-values change for 5% to less than 10% of the x-axis range.
4. Weak: The y-values change for 10% to less than 20% of the x-axis range.
5. Strong: The y-values change for 20% to less than 50% of the x-axis range.
6. Very Strong: The y-values change for 50% to less than 75% of the x-axis range.
7. Extremely Strong: The y-values change for 75% to 100% of the x-axis range.

Figure A.4: Evaluation prompts for inappropriate continuous and categorical encoding.

Misleading Chart Evaluation Prompts for Extended Dataset (1/4)

Truncated Axis

Question: How would you rate the difference in values between the first bar and the second bar?

Answer Options:

1. No difference: The values of the two bars are exactly equal.
2. Extremely small difference: The difference is less than 10% of the larger value.
3. Very small difference: The difference is 10% to less than 25% of the larger value.
4. Small difference: The difference is 25% to less than 50% of the larger value.
5. Large difference: The difference is 50% to less than 75% of the larger value.
6. Very large difference: The difference is 75% to less than 100% of the larger value.
7. Extremely large difference: The difference is 100% or more (i.e., one bar is at least double the other).

Inverted Axis

Question: How would you describe the overall trend of the line?

Answer Options:

1. Strong Decrease: A large and steep overall decline, spanning 75% to 100% of the Y-axis range.
2. Moderate Decrease: A noticeable downward trend, covering from 25% to less than 75% of the Y-axis range.
3. Slight Decrease: A gentle overall decline, changing by less than 25% of the Y-axis range.
4. No Clear Trend: Little to no net change across the chart.
5. Slight Increase: A gentle overall rise, changing by less than 25% of the Y-axis range.
6. Moderate Increase: A noticeable upward trend, covering from 25% to less than 75% of the Y-axis range.
7. Strong Increase: A large and steep overall rise, spanning 75% to 100% of the Y-axis range.

Figure A.5: Misleading Chart Evaluation Prompts in Extended Dataset for truncated axis and inverted axis.

Misleading Chart Evaluation Prompts for Extended Dataset (2/4)

Aspect Ratio

Question: How large is the biggest fluctuation in the chart compared to the overall change from the start to the end of the line?

Answer Options:

1. No fluctuation: The largest fluctuation is effectively zero.
2. Extremely small fluctuation: Largest fluctuation is $<$10% of the total start-to-end change.
3. Very small fluctuation: Largest fluctuation is 10%–25% of the total change.
4. Small fluctuation: Largest fluctuation is 25%–50% of the total change.
5. Large fluctuation: Largest fluctuation is 50%–75% of the total change.
6. Very large fluctuation: Largest fluctuation is 75%–100% of the total change.
7. Extremely large fluctuation: Largest fluctuation $\geq$100% of the total change.

Data-Visual Disproportion

Question: Which statement best describes the disparity between the quantities they represent?

*(For country-specific variants: Consider **Country A** and **Country B**. Which option best describes how many times larger the value of Country A is compared to Country B?)*

Answer Options:

1. Circle A represents a much larger quantity than Circle B (5× or more larger).
2. Circle A represents a moderately larger quantity than Circle B (between 2× and less than 5× larger).
3. Circle A represents a slightly larger quantity than Circle B (between 1× and less than 2× larger).
4. The circles represent approximately equal quantities (their areas are roughly the same).
5. Circle B represents a slightly larger quantity than Circle A (between 1× and less than 2× larger).
6. Circle B represents a moderately larger quantity than Circle A (between 2× and less than 5× larger).
7. Circle B represents a much larger quantity than Circle A (5× or more larger).

Figure A.6: Misleading Chart Evaluation Prompts in Extended Dataset for aspect ratio and data-visual disproportion.

Misleading Chart Evaluation Prompts for Extended Dataset (3/4)

Dual Axis (Inappropriate Axis Design)

Question: How similar are the trends of the two lines in the chart?

Answer Options:

1. No Similarity: Trends match for 0% of the chart.

2. Extremely Small Similarity: Trends match for <5% of the chart.

3. Very Small Similarity: Trends match for 5%–10%.

4. Small Similarity: Trends match for 10%–25%.

5. Large Similarity: Trends match for 25%–50%.

6. Very Large Similarity: Trends match for 50%–99%.

7. Extremely Large Similarity: Trends match for 100%.

Distorted Projection

Question: Which statement best describes the disparity between the highest and the second highest quantity in the pie chart?

Answer Options:

1. No disparity: The two largest segments represent equal quantities.

2. Extremely small disparity: The largest segment represents 1× to less 1.25× the quantity of the second-largest segment.

3. Very small disparity: The largest segment represents 1.25× to less 1.5× the quantity of the second-largest segment.

4. Small disparity: The largest segment represents 1.5× to less than 2× the quantity of the second-largest segment.

5. Moderate disparity: The largest segment represents 2× to less than 3× the quantity of the second-largest segment.

6. Large disparity: The largest segment represents 3× to less than 5× the quantity of the second-largest segment.

7. Very large disparity: The largest segment represents more than 5× the quantity of the second-largest segment.

Figure A.7: Misleading Chart Evaluation Prompts in Extended Dataset for dual axis and distorted projection.

Misleading Chart Evaluation Prompts in Extended Dataset (4/4)

Inappropriate Encoding

Question: How would you describe the overall trend of the chart?

Answer Options:

1. Strong Decrease: A large and steep overall decline, spanning 75% to 100% of the Y-axis range.
2. Moderate Decrease: A noticeable downward trend, covering from 25% to less than 75% of the Y-axis range.
3. Slight Decrease: A gentle overall decline, changing by less than 25% of the Y-axis range.
4. No Clear Trend: Little to no net change across the chart.
5. Slight Increase: A gentle overall rise, changing by less than 25% of the Y-axis range.
6. Moderate Increase: A noticeable upward trend, covering from 25% to less than 75% of the Y-axis range.
7. Strong Increase: A large and steep overall rise, spanning 75% to 100% of the Y-axis range.

Inappropriate Color Coding

Question: How different is the value of **Region A** compared to **Region B** in the chart?

(Region A and Region B refer to specific states or countries in the choropleth map.)

Answer Options:

1. No Difference: Both regions fall in the same color group on the legend.
2. Extremely Small Difference: The regions differ by 1–2 color groups.
3. Very Small Difference: The regions differ by 3–4 color groups.
4. Small Difference: The regions differ by 5–6 color groups.
5. Large Difference: The regions differ by 7–8 color groups.
6. Very Large Difference: The regions differ by 9–10 color groups.
7. Extremely Large Difference: The regions differ by 11–12 color groups, appearing at opposite ends of the legend.

Figure A.8: Misleading Chart Evaluation Prompts in Extended Dataset for inappropriate encoding and inappropriate color coding.

Data Video Generator System Prompt (Part 1/3)

```
You are generating a **self-contained HTML file** that animates one or more charts
for a short **data video scene** using D3 js.

The goal is to create an **animated visual story** that fulfills the stated
**intent of the scene**, using ONLY the provided data.

The input may contain MULTIPLE charts.
All charts belong to the SAME scene and must be animated within ONE SVG.

================================================
CORE REQUIREMENT (VERY IMPORTANT)
================================================

- Author a cohesive STORY that conveys the intent using animation.
- The story must be constructed from the provided data ONLY.
- The narrative must be expressed through **on-screen subtitles** and **visual animation**.
- The animation + subtitles together must clearly fulfill the stated **intent**.
- Subtitles should explain, highlight, compare, or summarize the data as needed.
- The story should unfold progressively over time, not all at once.
```

Figure A.9: Data Video Generator system prompt (Part 1) — overview and core requirements.

Data Video Generator System Prompt (Part 2/3)

```
================================================

STRICT RULES

================================================

- Output ONLY valid HTML

- Use exactly ONE <svg id="chart">

- The SVG represents the FULL video frame

- SVG MUST use a fixed resolution of 1280x720 pixels

- <svg id="chart" width="1280" height="720">

- Do NOT resize the SVG dynamically


- Define:  window.__VIDEO_DURATION__ = {duration}

- THIS TIMING IS STRICTLY ENFORCED DURING RENDERING

- Use the entire duration meaningfully

- Allow the story to unfold progressively

- Avoid finishing too early or rushing key moments


- Define functions:  resetChart() and scheduleNow()

- scheduleNow() MUST produce a complete animation every time it is called

- resetChart() MUST restore the SVG to a valid initial state

- Use setTimeout for ALL animations

- NEVER reference requestAnimationFrame (directly or indirectly)

- Do NOT load external assets except the D3 CDN

- Do NOT invent, interpolate, or modify data values

- Deterministic animation only (no randomness, no frame-based callbacks)
```

Figure A.10: Data Video Generator system prompt (Part 2) — strict execution rules.

Data Video Generator System Prompt (Part 3/3)

```
==================================================

SUBTITLE REQUIREMENTS (MANDATORY)

==================================================

- You MUST generate subtitles to narrate the story

- Subtitles for a scene must be long enough to be readable by the viewer

- Subtitles must appear within the SVG (not HTML overlays)

- Subtitles must be synchronized with animation events

- Subtitles must explain the story implied by the intent

- Subtitles must be readable, non-overlapping, and stay within SVG bounds

- Subtitles MUST fit entirely inside the 1280x720 frame at all times

- Subtitle text must wrap or line-break to avoid overflow

- Subtitles must never be clipped or cropped

- Use a consistent subtitle position (e.g., bottom-center)


==================================================

LAYOUT & LEGIBILITY CONSTRAINTS

==================================================

- All charts, subtitles, axes, labels, and annotations MUST fit within 1280x720

- Use explicit inner margins; do NOT place marks on SVG edges

- Reserve space for subtitles and annotations

- Ensure no overlap between visual elements

- Axis labels, tick labels, annotations, and subtitles must not collide

- Prioritize clarity and readability over visual flair
```

Figure A.11: Data Video Generator system prompt (Part 3) — subtitle requirements and layout constraints.

Director Prompt (Part 1/2)

```
You are a chart animation director.


The goal is to create an **animated visual story** that fulfills the stated

**intent of the scene**, using ONLY the provided data.


IMPORTANT: An image (ss.png) has been provided representing the required visual style.

Analyze the image to plan the animation layout.

================================================

INPUT INFORMATION

================================================

Intent of the scene:  {intent}

Chart Data (Ground Truth):  {data}

Time limit:  {duration} seconds

================================================

CORE REQUIREMENT (VERY IMPORTANT)

================================================

- The animation MUST focus on expressing the **intent** through visual storytelling.

- Author a cohesive STORY that conveys the intent using animation.

- The story must be constructed from the provided data ONLY.

- The narrative must be expressed through **on-screen subtitles** and **visual animation**.

- The animation + subtitles together must clearly fulfill the stated **intent**.

- Subtitles should explain, highlight, compare, or summarize the data as needed.

- The story should unfold progressively over time, not all at once.
```

Figure A.12: Director prompt (Part 1) — role definition, input information, and core requirements.

Director Prompt (Part 2/2)

```
==================================================

NARRATIVE ANIMATION STRATEGIES (GUIDANCE)

==================================================


Choose animation strategies that best support the **intent** and the story you

are constructing from the data:


- **Emphasis**:  highlight key values, peaks, outliers, or categories

- **Suspense**:  gradual reveal, delayed comparison, count-up, staged disclosure

- **Comparison**:  contrast groups, time periods, categories, or benchmarks

- **Ellipsis**:  de-emphasize less important data to focus attention


These strategies should be applied deliberately and coherently.


==================================================

TASK

==================================================


1.  Produce a structured animation plan (JSON).

2.  The plan must specify how to replicate the layout, chart type, and positioning seen in the

provided image.

3.  Create a "subtitles" array within the JSON. Each entry must have 'start', 'end', and 'text'.

4.  Ensure the visual animation stages align perfectly with these subtitle timestamps.

Return JSON only.
```

Figure A.13: Director prompt (Part 2) — narrative strategies and task specification.

Plan Critic Prompt (Part 1/2)

```
You are a senior animation consultant.  Review the plan against the source data, intent, and the

PROVIDED IMAGES for visual style.

Your primary task is to verify that the plan **fulfills the stated intent** of the scene.

==================================================

INPUT INFORMATION

==================================================

Intent of the scene:  {intent}

Chart Data (Ground Truth):  {data}

Time limit:  {duration} seconds

==================================================

PROPOSED PLAN

==================================================

{plan}

==================================================

CORE REQUIREMENT (VERY IMPORTANT)

==================================================

- The plan must author a cohesive STORY that conveys the intent using animation.

- The story must be constructed from the provided data ONLY.

- The narrative must be expressed through **on-screen subtitles** and **visual animation**.

- The animation + subtitles together must clearly fulfill the stated **intent**.

- Subtitles should explain, highlight, compare, or summarize the data as needed.

- The story should unfold progressively over time, not all at once.
```

Figure A.14: Plan Critic prompt (Part 1) — role definition, input information, proposed plan, and core requirements.

Plan Critic Prompt (Part 2/2)

```
==================================================

NARRATIVE ANIMATION STRATEGIES (GUIDANCE)

==================================================


Ensure the plan uses appropriate animation strategies:


- **Emphasis**:  highlight key values, peaks, outliers, or categories

- **Suspense**:  gradual reveal, delayed comparison, count-up, staged disclosure

- **Comparison**:  contrast groups, time periods, categories, or benchmarks

- **Ellipsis**:  de-emphasize less important data to focus attention

==================================================

CRITIQUE CHECKLIST

==================================================

- **intent fulfillment**:  Does the plan clearly express and fulfill the stated intent?  Is the

intent the central focus of the animation?

- visual style:  Does the plan's description align with the visual style provided in the images?

- accuracy:  Is the data representation correct?

- timing:  Is the flow realistic for {duration} seconds?

- narrative:  Does the plan tell a coherent story that supports the intent?

- strategies:  Are the animation strategies applied deliberately to reinforce the intent?

Return concise, actionable feedback.  If the intent is not fulfilled, explain what is missing

and how to fix it.
```

Figure A.15: Plan Critic prompt (Part 2) — narrative strategies and critique checklist.

Coder Prompt (Part 1/3)

```
You are a D3.js animation engineer generating a **self-contained HTML file** that animates

charts for a data video scene.

IMPORTANT: Images have been provided representing the required visual style.

You MUST analyze the attached images to extract and replicate:

1.  Exact Color Palette:  Hex codes for background, marks, and text.

2.  Typography:  Match font style and sizing.

3.  Layout:  Replicate padding and positioning of elements.

================================================

STRICT RULES

================================================

- Output ONLY valid HTML

- Use exactly ONE <svg id="chart">

- The SVG represents the FULL video frame

- SVG MUST use a fixed resolution of 1280x720 pixels

- <svg id="chart" width="1280" height="720">

- Do NOT resize the SVG dynamically


- Define:  window.__VIDEO_DURATION__ = {duration}

- THIS TIMING IS STRICTLY ENFORCED DURING RENDERING

- Use the entire duration meaningfully

- Allow the story to unfold progressively

- Avoid finishing too early or rushing key moments
```

Figure A.16: Coder prompt (Part 1) — role definition and initial strict rules.

Coder Prompt (Part 2/3)

```
- Define functions:  resetChart() and scheduleNow()

- scheduleNow() MUST produce a complete animation every time it is called

- resetChart() MUST restore the SVG to a valid initial state

- Use setTimeout for ALL animations

- NEVER reference requestAnimationFrame (directly or indirectly)

- Do NOT load external assets except the D3 CDN

- Do NOT invent, interpolate, or modify data values

- Deterministic animation only (no randomness, no frame-based callbacks)


================================================

SUBTITLE REQUIREMENTS (MANDATORY)

================================================


- You MUST generate subtitles to narrate the story

- Subtitles for a scene must be long enough to be readable by the viewer

- Subtitles must appear within the SVG (not HTML overlays)

- Subtitles must be synchronized with animation events

- Subtitles must explain the story implied by the intent

- Subtitles must be readable, non-overlapping, and stay within SVG bounds

- Subtitles MUST fit entirely inside the 1280x720 frame at all times

- Subtitle text must wrap or line-break to avoid overflow

- Subtitles must never be clipped or cropped

- Use a consistent subtitle position (e.g., bottom-center)
```

Figure A.17: Coder prompt (Part 2) — timing, animation functions, and subtitle requirements.

Coder Prompt (Part 3/3)

```
================================================

LAYOUT & LEGIBILITY CONSTRAINTS

================================================



- All charts, subtitles, axes, labels, and annotations MUST fit within 1280x720

- Use explicit inner margins; do NOT place marks on SVG edges

- Reserve space for subtitles and annotations

- Ensure no overlap between visual elements

- Axis labels, tick labels, annotations, and subtitles must not collide

- Prioritize clarity and readability over visual flair



================================================

PLAN:

{plan}

FEEDBACK TO ADDRESS:

{feedback}

Chart Data (Ground Truth):

{data}
```

Figure A.18: Coder prompt (Part 3) — layout constraints and input placeholders.

Video Critic Prompt (Part 1/2)

```
Evaluate the rendered video against the source data, the plan, and the PROVIDED IMAGES.

Your primary task is to verify that the video **expresses the stated intent** through its

animations and subtitles.

================================================

INPUT INFORMATION

================================================

Intent of the scene:

{intent}


Chart Data (Ground Truth):

{data}

================================================

CORE REQUIREMENT (VERY IMPORTANT)

================================================

- The video must tell a cohesive STORY that conveys the intent using animation.

- The story must be constructed from the provided data ONLY.

- The narrative must be expressed through **on-screen subtitles** and **visual animation**.

- The animation + subtitles together must clearly fulfill the stated **intent**.

- Subtitles should explain, highlight, compare, or summarize the data as needed.

- The story should unfold progressively over time, not all at once.
```

Figure A.19: Video Critic prompt (Part 1) — role definition, input information, and core requirements.

Video Critic Prompt (Part 2/2)

```
================================================

ANIMATION ASSESSMENT (CRITICAL)

================================================

Carefully evaluate the animations in the video:

1.  **Animation Correctness**:  Are the animations taking place accurately or are there issues
like overlapping, clipping, or misplacement of text or visual elements?

2.  **Time Utilization**:  Is the video using the entire allotted duration effectively to tell
the story?  Is it too fast or too slow?  Does it end too early or rush key moments?

3.  **Intent Expression**:  Do the animations effectively express and support the intent?

4.  **Animation-Subtitle Sync**:  Are animations properly synchronized with subtitles?  Do they
appear together at the right moments?

5.  **Animation Quality**:  Are animations smooth, visible, and purposeful?  Are there too many,
too few, or poorly timed animations?

6.  **Animation Effectiveness**:  Do the animations help the viewer understand the data story?


Based on your assessment, provide specific feedback to:

- **ADD** animations if key moments lack visual emphasis

- **REMOVE** animations if they are distracting or redundant

- **CHANGE** animations if timing, duration, or style needs adjustment

- **EDIT** animations with subtitles if they are misaligned

If perfect, start with 'PASS'. Otherwise, provide specific D3.js/CSS fixes with clear
instructions on what animations to add, remove, change, or re-sync.
```

Figure A.20: Video Critic prompt (Part 2) — animation assessment and feedback guidelines.

HTML Judge Prompt (Part 1/3)

```
You are a critical evaluator assessing an automatically generated **data video**.


SAMPLE ID: {sample_id}


You are given:

- The complete HTML source used to render the visualization

- The intent of the scene

- The original narration transcript

- The data used to generate the visualization


IMPORTANT:

- Base your judgment ONLY on the provided HTML and text

- Do NOT assume animation quality beyond what the HTML explicitly implies

- Be very strict and objective in your evaluation


================================================

INPUT DATA

================================================


Intent:  {row['Intent']}

Original Transcript:  {row['Transcript']}

Chart Data:  {row['Chart Data']}

HTML Source (FULL): {html}
```

Figure A.21: HTML Judge prompt (Part 1) — role definition and input data.

HTML Judge Prompt (Part 2/3)

```
================================================

EVALUATION CRITERIA

================================================

1.  **Narrative Quality (0-5)**

Evaluate whether the intended message of the scene is covered and conveyed as a coherent story

through the sequence of animation scenes and subtitles.

- 0:  The animation does not reflect the stated intent and lacks any coherent story.

- 1:  The intent is barely addressed and the story is largely unclear or incomplete.

- 2:  The intent is partially covered, but the story is fragmented or poorly connected.

- 3:  The intent is mostly covered, but the narrative has noticeable gaps or weak transitions.

- 4:  The intent is clearly conveyed as a coherent story, with minor lapses in flow.

- 5:  The full intent is clearly and consistently conveyed through a well-structured narrative

across all animation scenes and subtitles.

2.  **Informativeness (0-5)**

Evaluate how insightful the subtitles and accompanying animations are, beyond simply reading or

restating the data.

- 0:  Subtitles provide no meaningful information or insight.

- 1:  Subtitles primarily read out data values with little to no interpretation.

- 2:  Some interpretation is present, but subtitles remain mostly descriptive or obvious.

- 3:  Subtitles provide moderate insights, though some segments still rely on data restatement.

- 4:  Subtitles frequently highlight meaningful patterns, trends, or comparisons supported by

animation.

- 5:  Subtitles consistently deliver clear, non-obvious insights into the data, effectively

reinforced by animation.
```

Figure A.22: HTML Judge prompt (Part 2) — narrative quality and informativeness criteria.

HTML Judge Prompt (Part 3/3)

```
3.  **Subtitle-Transcript Similarity (0-5)**

Evaluate how well the generated subtitles align with the original narration transcript.


- 0:  Completely diverges from transcript intent.

- 1:  Largely contradicts transcript.

- 2:  Partial overlap with missing key ideas.

- 3:  General alignment with omissions.

- 4:  Strong alignment with minor differences.

- 5:  Preserves transcript meaning throughout.


4.  **Code Correctness (0-5)**

Evaluate whether the **HTML code structure and logic correctly support the narration's intended

animation plan**.


- 0:  HTML code is clearly incorrect, broken, or unrelated to the intended narration.

- 1:  Major structural or logical issues prevent the HTML from supporting the narration.

- 2:  Some parts align with the narration, but key elements are missing or incorrect.

- 3:  HTML generally follows the narration plan, with noticeable inconsistencies or gaps.

- 4:  HTML correctly implements the narration plan with minor issues.

- 5:  HTML code cleanly and correctly implements the intended narration and animation plan

throughout.
```

Figure A.23: HTML Judge prompt (Part 3) — subtitle-transcript similarity and code correctness criteria.

Pairwise VLM Judge Prompt (Part 1/2)

```
You are a strict expert evaluator comparing two automatically generated **data videos**.


SAMPLE ID: {sample_id}


You are given:

1.  **Video A** - A generated data video (MP4)

2.  **Video B** - Another generated data video (MP4)

3.  A **reference screenshot** showing the expected visual style


Your task is to compare Video A and Video B based on the criteria below and determine which

video is better overall.

================================================

EVALUATION CRITERIA

================================================

1.  **Visualization Quality**


Definition:

Evaluate whether the visualization is rendered correctly and remains visually readable.


Consider:

- Are there rendering issues (blank screens, missing charts, visual glitches)?

- Is the chart readable (no overlapping elements, clipping, or unreadable text)?

- Is the visualization clean, legible, and well-structured?
```

Figure A.24: Pairwise VLM Judge prompt (Part 1) — role definition, input description, and visualization quality.

Pairwise VLM Judge Prompt (Part 2/2)

```
-------------------------

2.  **Subtitle-Animation Coherence**


Definition:

Evaluate alignment between what subtitles say and what the animation shows.


Consider:

- Do subtitles match what is being shown in the animation?

- Is there proper timing and synchronization?

- Do subtitles and visuals reinforce each other?

-------------------------

3.  **Style Consistency**


Definition:

Evaluate how well each video maintains the visual style shown in the reference screenshot.


Consider:

- Does the video use similar colors, chart type, layout, and design elements?

- How closely does each video follow the reference style?
```

Figure A.25: Pairwise VLM Judge prompt (Part 2) — subtitle-animation coherence and style consistency.