

Disentangling Visual and Factual Correctness in LVLMs' Visualization Literacy

Soohyun Lee, Jaeyoung Kim, Seokhyeon Park, Sihyeon Lee, Jiwon Song, Bohyoung Kim, Hyunjo Song, and Jinwook Seo

Abstract—Large Vision-Language Models (LVLMs) have shown remarkable capabilities in visualization interpretation, yet it remains unclear whether their responses reflect genuine reasoning over visual evidence or the influence of factual priors learned during training. Current evaluation methods mix these two sources, obscuring when correct visual interpretation is overridden by memorized factual knowledge. We present a disentanglement framework that systematically isolates visual correctness from factual correctness, revealing fundamental validity limitations in existing visualization literacy assessments. Through three complementary experiments with 15 state-of-the-art LVLMs, we demonstrate that: (1) Although several models achieve human-level performance on standard tests (VLAT), such performance may reflect factual recall rather than visual understanding, whereas randomized-data tests (reVLAT) underestimate visualization literacy when visual interpretation is correct but superseded by conflicting factual priors. (2) Using our Counterfactual Visualization Literacy Assessment Test (CVLAT) alongside capability-normalized arbitration metrics, we classify models by the sign of their visual-factual reliance index (VFR). This classification reveals a visualization-oriented majority and a factual knowledge-oriented minority, although several near-zero cases warrant cautious interpretation. The factual knowledge-oriented minority tends to override the chart with prior knowledge. A human baseline ($N = 30$) on the same counterfactual items confirms that people overwhelmingly follow the chart under conflict, providing a human reference point for visual-factual arbitration. (3) Prompt-based intervention can shift this prioritization, but its effectiveness is highly model-dependent and often direction-asymmetric, with some models responding strongly to only one prompt direction. Furthermore, high chart-reading capability does not predict prompt-controllability, indicating that visual-factual arbitration is not uniformly steerable. Overall, our findings demonstrate that LVLMs' high visualization accuracy is not sufficient evidence of faithful visual reasoning. We argue that reliable LVLm integration into visual analytics requires evaluating not only visualization literacy, but also how models arbitrate between visual evidence and factual priors, particularly when the two sources diverge. The CVLAT benchmark and code are available at <https://github.com/JaeyoungKim-HCIL/CVLAT>.

Index Terms—Visualization literacy, Large vision language models, Multimodal large language models, Evaluation study

I. INTRODUCTION

Recent advances in Large Language Models (LLMs) and Large Vision-Language Models (LVLMs) have demonstrated

Soohyun Lee, Seokhyeon Park, Sihyeon Lee, Jiwon Song and Jinwook Seo are with Seoul National University (e-mail: {shlee, shpark, sihyeon, jwsong}@hcl.snu.ac.kr; jseo@snu.ac.kr). Jaeyoung Kim is with MADI (e-mail: jykim@madit.com). Bohyoung Kim is with Hankuk University of Foreign Studies (e-mail: bkim@hufs.ac.kr). Hyunjo Song is with Soongsil University (e-mail: hsong@ssu.ac.kr).

Soohyun Lee and Jaeyoung Kim contributed equally (co-first authors). Hyunjo Song and Jinwook Seo are the corresponding authors.

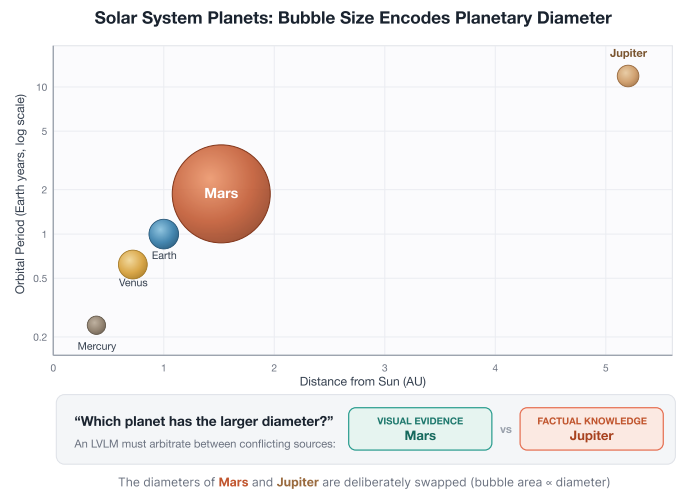


Fig. 1. A bubble chart of solar system planets where bubble size encodes planetary diameter, except Mars and Jupiter's sizes are deliberately swapped. When asked, "Which planet has the larger diameter?" an LVLm must choose between visual encoding (Mars appears larger) and its pre-trained factual knowledge (Jupiter is larger). This conflict illustrates the core challenge in evaluating visualization literacy: distinguishing genuine visual interpretation from factual recall.

strong capabilities across diverse visual analytics tasks, including data wrangling [1]–[3], visualization generation [4]–[7], and visualization interpretation [8]–[10]. Among these, visualization interpretation ability—commonly framed as visualization literacy—has received particular attention, as it plays a central role in determining whether LVLMs can reliably support analytical workflows, where accurate interpretation of charts and graphs is crucial for informed decision-making. A growing body of work has empirically evaluated this ability [11]–[16], establishing valuable performance baselines while highlighting the limits of current evaluation practices.

Despite these foundational efforts, current evaluations leave two key questions unresolved. First, existing studies have primarily evaluated a limited set of proprietary LVLMs, such as GPT-4 [17], Claude 3 Opus [18], and Gemini 1.5 Pro [19], while rapid model development has introduced numerous newer and open-source alternatives that remain underexamined. Second and more fundamentally, conventional accuracy-based assessment metrics fail to disentangle the basis of correctness—whether a correct answer reflects genuine visual interpretation or merely factual priors acquired during pre-training, and whether an incorrect answer stems from visual misreading or from overriding visual evidence with prior knowledge.

This second gap introduces a fundamental validity concern. Consider the bubble chart in Figure 1, where bubble size encodes planetary diameter, but the diameter values for Mars and Jupiter are intentionally reversed. When asked “Which planet has the larger diameter?”, an LVLM must choose between the visual evidence (Mars) and its pre-trained factual knowledge (Jupiter). If it responds “Jupiter,” existing accuracy metrics cannot determine whether it (1) correctly interpreted the chart but prioritized factual priors, or (2) failed to interpret the visualization in the first place. The inverse scenario is equally problematic: When visualizations align with real-world values, high accuracy may simply reflect factual recall rather than visual comprehension. Under current evaluation methods, these indistinguishable response sources make it impossible to assess what is genuinely being measured as visualization literacy.

To address these evaluation gaps and more precisely understand LVLMs’ visualization interpretation capabilities, we pose the following research questions:

- **RQ1 (Performance Assessment):** How do state-of-the-art proprietary and open-source LVLMs perform on visualization literacy tasks?
- **RQ2 (Conflict Resolution):** When visual information conflicts with factual knowledge, how do LVLMs prioritize between the two sources?
- **RQ3 (Preference Steering):** Can prompt engineering reliably shift LVLMs’ prioritization between visual evidence and factual priors under conflict conditions?

To investigate these questions, we propose a disentanglement framework that defines two key dimensions, visual correctness and factual correctness, in LVLM-based visualization interpretation (Sec. III). These two dimensions have typically been overlooked or treated as a single outcome in prior studies evaluating the performance of LVLMs. By making these sources of correctness explicit, our framework reveals how prevailing accuracy-based evaluation methods obscure whether reported performance reflects visual interpretation, factual recall, or a mixture of both.

Guided by this framework, we conduct three complementary experiments with 15 state-of-the-art LVLMs. First, we establish baseline performance through VLAT [20] and its randomized variant, reVLAT [12], which controls for factual priors while preserving visual encodings. Second, we introduce the Counterfactual Visualization Literacy Assessment Test (CVLAT), a diagnostic benchmark that systematically constructs visual-factual conflicts to measure how models arbitrate between visual evidence and stored knowledge when the two diverge. We position visual-factual arbitration not as an alternative to visualization literacy, but as one of its constituent facets—one that conventional accuracy-based tests entangle with perceptual decoding. By isolating and diagnosing this specific facet, CVLAT complements rather than replaces established assessments such as VLAT and reVLAT, which remain necessary for measuring perceptual decoding itself. Finally, we test whether prompt interventions can redirect this prioritization by comparing factual-priority and visual-priority prompts, assessing whether prioritization patterns are intrinsic

architectural tendencies or externally steerable preferences.

In summary, our main contributions are:

- A disentanglement framework that separates visual correctness from factual correctness in LVLM visualization literacy.
- A comprehensive empirical assessment of 15 state-of-the-art LVLMs, spanning both proprietary and open-source models.
- CVLAT: A novel diagnostic benchmark for measuring visual-factual arbitration within LVLM visualization literacy.
- An analysis of prompt engineering effectiveness in directing LVLMs’ prioritization between visual evidence and factual knowledge.

II. RELATED WORK

A. Visualization Literacy and Assessment

Visualization literacy has been defined in various ways in the literature [20]–[22]. The most well-known and concise definition of visualization literacy is *the ability and skill to read and interpret visually represented data in and to extract information from data visualizations* [20]. The concept of visualization literacy is connected with the utilization and adoption of visualization and visual analytics tools [23]–[25]. Consequently, there has been a growing interest in quantitatively assessing users’ visualization literacy [20], [21], [26], [27]. For example, Boy et al. [21] proposed an evaluation method based on item response theory (IRT) [28] for evaluating individuals’ visualization literacy in visualization types including *line graphs*, *bar charts*, and *scatterplots*. Börner et al. [22] assessed the visualization literacy of 273 science museum visitors through familiarity-based questions about various data visualizations. The most widely adopted assessment tool is the VLAT [20], which consists of 53 multiple-choice questions with 8 different types of tasks across 12 visualization types. Building upon this research, Pandey and Ottley proposed Mini-VLAT [26], a concise version of VLAT that reduces the number of questions to 12 while maintaining assessment validity. Ge et al. developed a precise definition of misleaders—decisions made in the construction of visualizations that can lead to conclusions not supported by the data—and proposed CALVI [27] to assess critical thinking about misleading visualizations.

While effective for humans, these tests may not transfer to LVLMs: trained on massive datasets that may include the test items themselves, LVLMs blend visual interpretation with pre-trained knowledge in ways that standard assessments were not designed to disentangle. Our work addresses this gap (Sec. III).

B. Visualization Literacy in Large Vision Language Models

In recent years, with advances in the visual understanding of AI models [29]–[31], the concept of visualization literacy has been extended to LVLMs. As these models continue to evolve, assessing LVLMs’ visualization literacy has emerged as a significant new research direction [11]–[15]. Across this body of research, the VLAT has emerged as the predominant

methodology, providing researchers with a standardized approach to measure and compare visualization literacy across different LVLM architectures. For instance, Bendeck and Stasko [11] evaluated GPT-4V's visualization literacy through a series of tests, including the VLAT. Their findings revealed that GPT-4V demonstrated strong capabilities in *identifying trends* and *extreme values*, while showing notable limitations in accurately retrieving specific values from visualizations. Similarly, Li et al. [14] assessed visualization literacy in GPT-4o, Claude 3 Opus, and Gemini 1.5 Pro with VLAT [20] and Mini-VLAT [26]. Their study found that these models outperformed the human baseline in *identifying correlations*, *clusters*, and *hierarchical structures*. In more recent research, Pandey and Ottley [13] evaluated the visualization literacy of GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, and Llama3.2-vision using both VLAT [20] and CALVI [27]. They reported that these LVLMs approached or exceeded human-level performance in tasks such as *trend identification* and *hierarchical structure detection*, but showed poor reliability in interpreting deceptive visualizations.

Notably, Hong et al. [12] assessed the visualization literacy of GPT-4V and Gemini using reVLAT—a modified version with randomized data while maintaining the same chart types and task types. Their study found that GPT-4V performed relatively well in *Finding Correlation Trends* and *Making Comparisons* when using *scatterplots*, achieving performance comparable to humans. However, unlike prior research [11], Hong et al. reported that GPT-4V showed weak performance in tasks such as *Finding Extremum*. Importantly, they observed that LVLMs exhibited a strong tendency to rely on their pre-existing knowledge rather than the visualization content when answering questions. These conflicting findings highlight the need for a more nuanced framework to assess whether LVLMs truly understand visualizations or merely leverage their pre-trained knowledge—a gap our research aims to address through empirical investigation (Sec. V).

Our work connects three strands of prior research. Relative to benchmarking studies such as Hong et al. [12], we systematically construct visual-factual conflicts grounded in shared factual priors (instead of randomizing data), introduce per-model arbitration metrics with capability controls (instead of reporting a population-level accuracy gap), and test whether prompt-based intervention can steer this reliance. Relative to a complementary line of work that pursues *structural* improvements to LVLM chart understanding via self-training [32], reasoning chains [33], [34], mixture-of-experts architectures [35], and chart-focused instruction tuning [36], our work shifts the focus toward diagnosing base-model arbitration, providing a benchmark against which such structurally improved models can be evaluated. We further connect to work that treats human-model correspondence as an empirical question in its own right, comparing human and LVLM literacy [37], contrasting human chart takeaways with LLM-generated ones [38], and probing how far LLMs can stand in for human participants in visualization studies [39]. Our $N = 30$ human baseline on CVLAT, recruited via Prolific (Sec. V-A2), extends this line of work by providing a direct human-LVLM comparison on the same arbitration items.

C. Cognitive Bias in Visualization Interpretation

The influence of human cognitive biases on visualization tasks has been actively researched [40], [41]. Human cognitive biases are psychological mechanisms that systematically distort information during decision-making processes [42], [43]. For example, Xiong et al. [44] demonstrated the “curse of knowledge” bias in data visualization, showing that people with prior knowledge of specific data patterns incorrectly assume others will find the same patterns visually salient, hindering effective communication of insights. These studies underscore that prior knowledge and biases can overshadow raw visual information in humans.

We draw on this body of literature as motivational context for examining whether LVLMs’ pretrained knowledge analogously interacts with visual interpretation, treating the human-LVLM analogy as an empirical question rather than a stipulated equivalence. While related phenomena in the context of LLMs and LVLMs are often discussed under various terms such as hallucination, model bias, or reasoning errors, their specific impact on visualization literacy assessment remains largely unexplored.

Recent work on hallucination in LLMs [45] has proposed taxonomies to classify hallucinations, but these frameworks focus primarily on text-based responses and do not address the unique challenges of visual interpretation tasks. Similarly, Guan et al. [46] introduced a diagnostic benchmark distinguishing between “language hallucination” and “visual illusion” in LVLMs, examining how models respond to manipulated images that contradict factual knowledge. However, their work focuses on general image understanding rather than the specific domain of visualization literacy assessment.

Building on these foundations, our work specifically investigates how LVLMs interpret visualizations, examining whether models rely more on visual information or factual knowledge when these two sources align or conflict. By establishing visual correctness and factual correctness as orthogonal dimensions, our work makes the visual-factual arbitration explicit and measurable. We use the cognitive-bias literature solely as motivational context for studying LVLM visualization literacy, without assuming shared cognitive mechanisms between humans and LVLMs.

III. PROBLEM STATEMENT

Prior studies that empirically evaluate the visualization literacy of LVLMs [11]–[15] predominantly utilize visualization literacy tests designed for humans [20], [26]. While these tests enable straightforward accuracy-based analysis for specific visualization and task types, they yield conflicting findings about LVLMs’ visualization capabilities. For example, Bendeck and Stasko [11] concluded through their replication study that the probability of GPT-4V solving problems by relying on prior knowledge rather than visualization interpretation was low. In contrast, Hong et al. [12] demonstrated that LVLMs heavily relied on their pre-existing knowledge to answer questions instead of utilizing information from the visualizations.

Given these conflicting findings, we propose that a thorough evaluation of LVLMs’ visualization literacy should extend

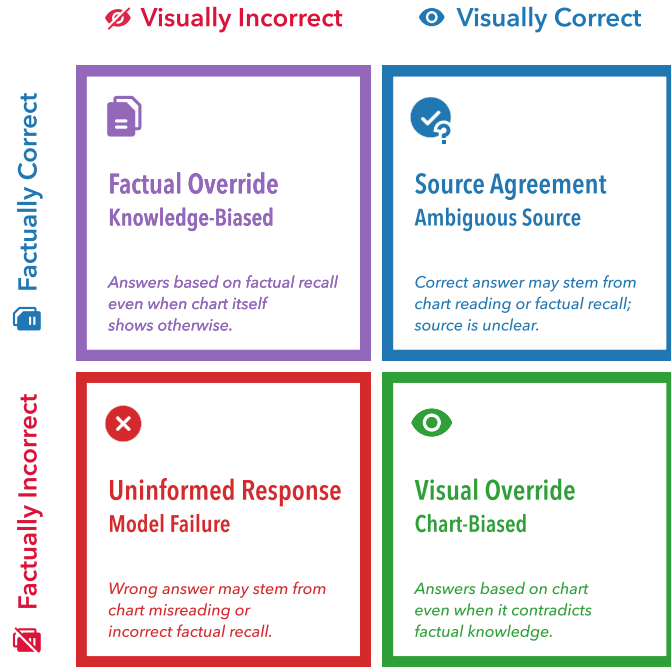


Fig. 2. A quadrant framework for evaluating how LVLMs balance visual information and pre-trained knowledge. The framework uses two dimensions: Visual Correctness measuring adherence to visual information, and Factual Correctness measuring alignment with factual knowledge. The four quadrants represent distinct cases of LVLM behavior when interpreting visualizations.

beyond simple accuracy metrics. Relying solely on these metrics may mask whether correct answers result from true visual understanding or from leveraging pre-trained knowledge. Additionally, it may not clarify whether incorrect answers indicate actual failures in interpretation or merely reflect a preference for factual knowledge over visual information.

We therefore propose a two-dimensional framework that disentangles visual interpretation from factual knowledge (Figure 2). Rather than evaluating responses along a single correct/incorrect axis, we assess them along two independent dimensions: visual correctness (whether responses align with the presented visualization) and factual correctness (whether responses align with real-world facts). Throughout our discussion, we use the terms “factual knowledge” and “prior knowledge” interchangeably to refer to the information that LVLMs have acquired during pre-training, which exists independently of the visual information.

This framework reveals critical patterns in LVLM behavior that single-axis accuracy metrics cannot capture. By analyzing responses across both dimensions, we identify two primary cases:

Aligned cases (Figure 3a) occur when visualization and factual knowledge agree:

- **Source Ambiguity** — Visually Correct (VC) and Factually Correct (FC): High accuracy may stem from either visual interpretation or knowledge recall, potentially inflating visualization literacy scores and making it unclear how much of the performance reflects actual visualization literacy.
- **Model Failure** — Visually Incorrect (VI) and Factually Incorrect (FI): This clearly indicates failure, but we cannot determine whether the failure stems from misinterpreting the visualization, incorrect knowledge, or both.

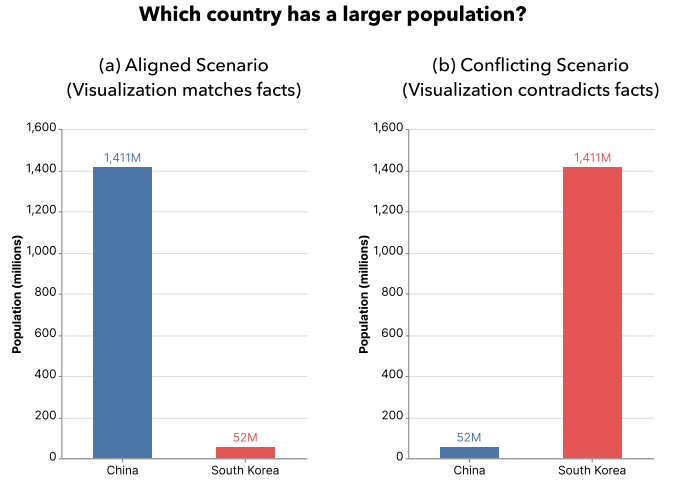


Fig. 3. Population comparison bar charts demonstrating aligned versus conflicting scenarios when asked “Which country has a larger population?”. (a) Aligned scenario where visualization matches factual knowledge (China > South Korea): answering “China” falls into the Source Ambiguity (VC/FC) quadrant, as we cannot determine if the model read the chart or recalled facts. (b) Conflicting scenario presenting counterfactual data where the visualization shows South Korea > China: answering “China” falls into the Factual Override (VI/FC) quadrant (prioritizing factual knowledge over visual information), while answering “South Korea” falls into the Visual Override (VC/FI) quadrant (correctly interpreting the visualization despite contradicting facts).

Conflicting cases (Figure 3b) emerge when visualization contradicts facts:

- **Visual Override** — Visually Correct (VC) and Factually Incorrect (FI): Following counterfactual visualizations suggests that the model relied on visual information rather than factual knowledge, prioritizing visual information over factuality.
- **Factual Override** — Visually Incorrect (VI) and Factually Correct (FC): Choosing facts over visualization may indicate either interpretation failure or knowledge prioritization, potentially underestimating visualization literacy.

This framework clarifies why existing evaluations generate inconsistent findings: they cannot distinguish between these fundamentally different response mechanisms in our two-dimensional space. We therefore designed three complementary experiments: Section IV establishes baseline performance using VLAT and reVLAT; Section V introduces CVLAT to explicitly test conflicting visual-factual cases; Section VI explores whether prompt engineering can shift models’ prioritization between visual evidence and factual knowledge.

IV. EXPERIMENT ONE: EVALUATING LVLMs’ VISUALIZATION LITERACY

In our first experiment, we address **RQ1**: What is the visualization literacy performance of state-of-the-art LVLMs, including both proprietary and open-source models?

TABLE I
SELECTED LVLMs FOR VISUALIZATION LITERACY ASSESSMENT

Model	Developer	Parameter Size
<i>Proprietary Models</i>		
GPT-5.5	OpenAI	-†
Claude-Opus-4.7	Anthropic	-†
Claude-Sonnet-4.6	Anthropic	-†
Claude-Haiku-4.5	Anthropic	-†
Gemini-3.1-Pro	Google	-†
Gemini-3.1-Flash-Lite	Google	-†
Grok-4.3	xAI	-†
Grok-4.20	xAI	-†
<i>Open-source Models</i>		
Llama4-Maverick	Meta	400B (17B active)*
Llama4-Scout	Meta	109B (17B active)*
Gemma-4-31B	Google	31B
Gemma-4-26B-A4B	Google	26B (4B active)*
Qwen3-VL-235B	Alibaba	235B (22B active)*
Qwen3-VL-32B	Alibaba	32B
Qwen3-VL-8B	Alibaba	8B

† Parameter sizes for proprietary models are not disclosed due to their proprietary architectures.

* Active parameters for Mixture-of-Experts (MoE) models.

Following prior methodologies [11]–[13], we assess a state-of-the-art LVLM suite using VLAT [20] and reVLAT [12]. Our evaluation covers the latest proprietary releases and recent open-source families (Table I), many of which have not yet been systematically evaluated for visualization literacy.

A. Model Selection

Our study aims to balance reproducibility with comprehensive coverage of current models. For proprietary models, we establish minimum reproducibility criteria by selecting only those that support temperature adjustment, which allows consistent result generation across runs. For models that additionally expose a reasoning-effort or extended-thinking control, we fix this control to its lowest available setting. Because extended internal reasoning is a known source of run-to-run variability even at temperature = 0 [47], the lowest-effort setting provides the more reproducible configuration.

Table I presents the 15 LVLMs selected for our evaluation. The exact API snapshots and local Hugging Face checkpoints used are documented in Appendix D. We choose to evaluate base models rather than fine-tuned variants for several reasons. First, using base models provides a clearer picture of the foundational visual literacy capabilities inherent to each architecture without task-specific optimizations that might obscure underlying limitations. Second, this approach ensures fair comparison across model families, as fine-tuning techniques and datasets vary widely between research teams and commercial providers, potentially introducing confounding factors into our analysis. Third, base models represent the most accessible versions for broader research communities, making our findings more generalizable and applicable across various downstream applications where custom fine-tuning may not be feasible. Finally, we deliberately include a range of open-source models such as Llama4, Gemma-4, and Qwen3-

VL to provide a more holistic perspective on visualization literacy capabilities across the LVLM landscape.

B. Prompt Design

We design two prompt conditions for our evaluation (see Appendix A for full prompt texts). The **Normal prompt**, based on Hong et al.'s [12] approach, instructs models to provide direct answers and select 'Omit' when uncertain, without requiring reasoning. We intentionally keep this prompt minimal to avoid confounding the measurement of baseline visualization literacy capabilities.

The **Explain prompt** incorporates chain-of-thought prompting [48] informed by semantic content frameworks from visualization research [49], [50]. It guides models through three stages aligned with Lundgard and Satyanarayan's framework [49]: (1) describing visual attention patterns, (2) extracting data values corresponding to elemental encoding recognition, and (3) explaining calculations and interpretations encompassing statistical relationships and perceptual patterns.

C. Experimental Design and Validity

1) *Test Set Selection and Rationale*: Using the quadrant framework introduced in Section III, we first examine VLAT, which constructs visualizations using data that generally aligns with real-world knowledge. Under this setting, incorrect responses fall into the Model Failure (**VI**∧**FI**) case: The model clearly fails, but we cannot determine whether the error stems from visual misinterpretation or incorrect factual recall. Similarly, correct responses fall into the Source Ambiguity (**VC**∧**FC**) case, where we cannot distinguish genuine visual interpretation from pretrained knowledge recall, potentially inflating visualization literacy scores.

To complement this assessment, we adapt reVLAT as our second test. reVLAT preserves VLAT's chart and task types but replaces underlying data with randomized values, which break the alignment between visualizations and the original real-world data. Under this setting, correct responses typically fall into the Visual Override (**VC**∧**FI**) case, indicating genuine visualization interpretation. However, incorrect responses remain ambiguous—they may indicate true failures to interpret the visualization or correspond to the Factual Override (**VI**∧**FC**) case, where factual priors outweigh visual evidence. Consequently, reVLAT may underestimate visualization literacy by not distinguishing between these two error sources.

2) *Experimental Protocol and Evaluation Methodology*: VLAT and reVLAT each contain 53 multiple-choice questions, with varying numbers of answer options (3, 4, or 5), including an 'Omit' option. To control for potential ordering effects reported in prior research [12], [13], we generate 120 answer-ordering variants per question. For 5-option questions, this includes all unique permutations. For questions with fewer options (3 or 4), we proportionally repeat each unique permutation to reach 120 variants. Our experimental design yields 6,360 question variations per model for each prompt-test combination (120 permutations × 53 questions). With 2 test types and 2 prompt conditions, each LVLM processes 25,440 distinct instances. Across the 15 LVLMs, this produces

TABLE II

PERFORMANCE COMPARISON OF LVLMS ON VISUALIZATION LITERACY TESTS. BEST PERFORMANCES AMONG PROPRIETARY MODELS AND BEST AMONG OPEN-SOURCE MODELS ARE HIGHLIGHTED IN **BOLD AND UNDERLINED**. THE Δ COLUMNS SHOW THE PERFORMANCE DROP FROM VLAT TO reVLAT.

Model	Provider	VLAT		reVLAT		Δ (VLAT–reVLAT)	
		Normal	Explain	Normal	Explain	Normal	Explain
<i>Human Baseline</i> [20]	–	65.50 (13.30)		–		–	
GPT-5.5	OpenAI	80.09 (26.93)	93.85 (20.69)	66.19 (34.81)	87.97 (28.52)	13.90	5.88
Claude-Opus-4.7	Anthropic	94.87 (15.38)	94.70 (17.45)	88.52 (26.29)	90.63 (25.59)	6.35	4.07
Claude-Sonnet-4.6	Anthropic	89.43 (26.43)	83.40 (24.71)	79.62 (36.88)	73.24 (37.17)	9.81	10.16
Claude-Haiku-4.5	Anthropic	77.09 (32.30)	72.08 (29.57)	67.61 (38.57)	64.65 (35.40)	9.48	7.43
Gemini-3.1-Pro	Google	99.73 (1.73)	98.79 (4.77)	93.29 (23.79)	93.21 (24.19)	6.44	5.58
Gemini-3.1-Flash-Lite	Google	77.33 (32.52)	90.19 (23.55)	69.70 (36.92)	88.00 (24.63)	7.63	2.19
Grok-4.3	xAI	44.36 (28.00)	77.72 (27.26)	33.52 (29.92)	72.09 (29.50)	10.84	5.63
Grok-4.20	xAI	37.89 (29.27)	74.95 (26.46)	37.66 (32.37)	70.77 (29.73)	0.23	4.18
<i>Proprietary Average</i>	–	75.10	85.71	67.01	80.07	8.09 [†]	5.64 [†]
Llama4-Maverick	Meta (Local)	60.58 (34.77)	67.75 (37.48)	49.58 (35.44)	69.07 (37.43)	11.00	-1.32
Llama4-Scout	Meta (Local)	48.51 (34.87)	61.37 (36.51)	40.25 (32.88)	55.72 (36.81)	8.26	5.65
Gemma-4-31B	Google (Local)	67.77 (35.80)	82.23 (33.84)	63.60 (39.75)	83.95 (32.70)	4.17	-1.72
Gemma-4-26B-A4B	Google (Local)	58.38 (39.83)	86.24 (28.67)	63.84 (35.93)	82.16 (31.67)	-5.46	4.08
Qwen3-VL-235B	Alibaba	60.03 (35.47)	78.95 (31.54)	51.37 (38.28)	78.43 (33.10)	8.66	0.52
Qwen3-VL-32B	Alibaba (Local)	67.01 (36.15)	81.97 (29.10)	54.48 (38.16)	79.21 (34.43)	12.53	2.76
Qwen3-VL-8B	Alibaba (Local)	53.22 (31.50)	67.41 (32.62)	38.79 (35.74)	65.14 (36.30)	14.43	2.27
<i>Open-source Average</i>	–	59.36	75.13	51.70	73.38	9.22 [†]	2.62 [†]

* Values shown as: mean accuracy percentage (standard deviation) for VLAT and reVLAT columns; percentage point differences for Δ columns.

* Human Baseline represents average human performance on VLAT reported by Lee et al. [20].

* Negative Δ values indicate better performance on reVLAT than VLAT.

[†] Average of absolute values.

381,600 total trials, providing robust statistical power for our comparative analysis.

Given the scale of our experiment, we adopt an efficient and consistent evaluation pipeline. Following established practices [46], [51], we use GPT-5.4-nano to extract letter-based answers from model responses. The extraction process employs structured prompts to identify the selected option (i.e., letter-based answer) from each model's response text, accommodating varied response formats in which models provide explanations, qualifiers, or expressions of uncertainty alongside their answers. To ensure consistency and robustness in answer extraction, we conduct five independent extraction runs with GPT-5.4-nano (temperature=0) and apply majority voting to determine the final answer for each trial. The same extraction pipeline is reused in all subsequent experiments and capability-reference conditions; across all 813,600 resulting trials in the full study suite, the pipeline produced a structured letter selection for every response, with no unparseable extractions.

3) *Analysis Methodology*: For each model and experimental condition we report mean accuracy and standard deviation across questions, where per-question accuracy is computed as the mean over the 120 permutations to control for ordering effects. We summarize Experiment 1 dispersion with standard deviations across the 53 items, whereas Experiments 2 and 3, which test directional hypotheses about arbitration shifts, additionally report bootstrap confidence intervals. We also analyze performance across visualization types (e.g., bar charts, line graphs) and task categories (e.g., trend identification, value retrieval) to identify model-specific strengths and limitations. Detailed breakdowns are provided in Appendix B.

D. Experimental Results

1) *Overall Performance Analysis*: Table II presents the performance results of 15 state-of-the-art LVLMS on both VLAT and reVLAT under Normal and Explain prompt conditions.

On VLAT, which contains visualizations generally aligned with real-world data, several models demonstrated strong performance. Under the Normal prompt condition, six proprietary LVLMS exceeded the human baseline of 65.50%: Gemini-3.1-Pro achieved the highest accuracy at 99.73%, followed by Claude-Opus-4.7 (94.87%), Claude-Sonnet-4.6 (89.43%), GPT-5.5 (80.09%), Gemini-3.1-Flash-Lite (77.33%), and Claude-Haiku-4.5 (77.09%). Among open-source models, Gemma-4-31B (67.77%) and Qwen3-VL-32B (67.01%) also approached or exceeded the human baseline.

The Explain prompt condition produced highly heterogeneous effects across models. The two Grok variants showed the most dramatic improvements (Grok-4.20: from 37.89% to 74.95%; Grok-4.3: from 44.36% to 77.72%), and GPT-5.5 also improved substantially from 80.09% to 93.85%. In contrast, the two non-Opus Claude variants exhibited *negative* Explain-vs.-Normal effects on VLAT (Claude-Sonnet-4.6: from 89.43% to 83.40%; Claude-Haiku-4.5: from 77.09% to 72.08%), indicating that the structured Explain prompt did not benefit these models' accuracy. Meanwhile, Gemini-3.1-Pro and Claude-Opus-4.7, both already near-ceiling under the Normal condition, showed essentially no Explain effect.

However, performance changed dramatically on reVLAT, which uses randomized data. Most models experienced substantial performance drops, with GPT-5.5 declining from 80.09% to 66.19% and Claude-Sonnet-4.6 dropping from 89.43% to 79.62%. Among the top-performing proprietary

models, Gemini-3.1-Pro and Claude-Opus-4.7 retain remarkably high reVLAT accuracy (93.29% and 88.52%, drops of only 6.44pp and 6.35pp, respectively). Grok-4.20 shows the smallest absolute drop in the suite (0.23pp from 37.89% to 37.66%), though from a low baseline. Among open-source models, Gemma-4-26B-A4B (63.84%) and Gemma-4-31B (63.60%) achieve the highest reVLAT Normal accuracies. This substantial performance gap between VLAT and reVLAT reflects the ambiguity in evaluation described in our quadrant framework (Sec. III). On VLAT, where visual and factual information align, high accuracy may stem from either genuine visual interpretation or factual recall (Source Ambiguity ($VC \wedge FC$)); on reVLAT, where factual priors no longer apply, the performance drop suggests that some models relied on factual knowledge rather than visual interpretation when answering VLAT questions.

When examining group-level performance, proprietary models achieved higher average accuracies than open-source models: 75.10% (Normal) and 85.71% (Explain) on VLAT, compared to 59.36% and 75.13% for open-source models. However, individual comparisons reveal a more nuanced picture: several open-source models (Gemma-4-31B, Qwen3-VL-32B, Llama4-Maverick) match or exceed the lower-performing proprietary models like Grok-4.20 and Grok-4.3.

The pattern of performance drops from VLAT to reVLAT depends on the prompt condition. Under the Normal prompt, the proprietary group exhibits a slightly smaller drop (8.09pp) than the open-source group (9.22pp). Under the Explain prompt, however, the open-source group exhibits a markedly smaller drop (2.62pp) than the proprietary group (5.64pp). Notably, several open-source models actually show an inverse pattern, performing better on reVLAT than VLAT (e.g., Gemma-4-31B: 82.23% on VLAT, 83.95% on reVLAT; and Llama4-Maverick: 67.75% on VLAT, 69.07% on reVLAT). Within the proprietary group, Gemini-3.1-Pro and Claude-Opus-4.7 stand out for their robustness to data randomization, showing minimal drops of 6.44pp and 6.35pp under the Normal-prompt condition, respectively. These divergent group-level patterns suggest that some recent models show smaller accuracy drops under data randomization, consistent with reduced dependence on aligned factual priors, though, as Experiment 2 will show, such robustness is conceptually distinct from visual-factual arbitration.

Prior research [12] attributed these performance differences to LVLMS' reliance on pre-existing knowledge rather than visual interpretation. Our quadrant framework provides a more precise account of this phenomenon. The observed performance patterns reveal two key insights:

First, on VLAT, where visual information aligns with real-world facts and thus responses fall into the Source Ambiguity ($VC \wedge FC$) and Model Failure ($VI \wedge FI$) cases, we cannot definitively determine whether high accuracy stems from genuine visualization interpretation or from factual recall. As a result, VLAT scores may overstate visualization literacy by masking the underlying basis of correctness.

Second, the dramatic performance drop on reVLAT raises an interpretive challenge. Although such declines might suggest weak visualization interpretation abilities, they may alterna-

tively indicate that models correctly interpreted the visualizations but prioritized their factual knowledge when generating responses, corresponding to the Factual Override ($VI \wedge FC$) case. This distinction is crucial: if models are indeed interpreting visualizations correctly but defaulting to factual knowledge under conflict, then current accuracy-based evaluation methods may systematically underestimate their true visualization literacy capabilities.

These findings motivate our subsequent experiments, which systematically examine how LVLMS resolve conflicts between visual information and factual knowledge under controlled conditions.

V. EXPERIMENT TWO: ASSESSING LVLMS' VISUALIZATION LITERACY WITH COUNTERFACTUAL VISUALIZATIONS

In our second experiment, we address **RQ2**: How do LVLMS respond when visual information conflicts with factual knowledge, and how do conflicts shape their information prioritization?

This connects to McNutt et al.'s 'Visualization Mirages' [52], where visual encodings interact with prior knowledge to cause misinterpretation. Whereas Mirages characterize human failures, we ask how LVLMS resolve visual-factual conflicts.

Previous studies [11], [13], [15] have evaluated LVLMS' visualization literacy using deceptive visualizations with different types of misleaders, including those arising from data curation [53], data wrangling [54], [55], and visualization design [56], [57]. While these approaches reveal important facets of visualization literacy, they do not isolate the specific challenge we address. Particularly, they did not explicitly examine how LVLMS arbitrate between visual evidence and factual knowledge when the two are in conflict during the visualization interpretation process.

This distinction is essential for accurately assessing visualization literacy, as it gives rise to two problematic assessment cases:

- 1) **Aligned visual-factual cases**: When visualizations align with factual knowledge, LVLMS may answer correctly by relying on prior knowledge rather than interpreting the visualization itself. Such responses fall into the Source Ambiguity ($VC \wedge FC$) case and can artificially inflate reported visualization literacy scores.
- 2) **Conflicting visual-factual cases**: When visualizations contradict factual knowledge, LVLMS may generate factually correct answers while ignoring visual information. These responses correspond to the Factual Override ($VI \wedge FC$) case and can lead to underestimation of true visualization literacy.

We focus on these two cases because they are the primary sources of systematic bias in accuracy-based evaluation; the remaining cases are either diagnostically unambiguous or non-informative.

To investigate these cases systematically, we introduce the Counterfactual Visualization Literacy Assessment Test (CVLAT), which uses counterfactual visualizations that deliberately conflict with widely-known facts. By inducing visual-

factual conflicts, CVLAT lets us estimate aggregate visual-factual arbitration tendencies under controlled conflict.

A. Experimental Design and Methodology

1) *Design of CVLAT and Rationale*: For a systematic assessment of how LVLMs prioritize between visual information and factual knowledge, we design the Counterfactual Visualization Literacy Assessment Test (CVLAT), adapting VLAT with deliberately counterfactual visualizations.

Although visualization literacy tests with randomized data such as reVLAT [12] provide useful controls, they are not well suited to our specific research goals. Data randomization does not guarantee that (1) models have prior knowledge about the topic domain of interest, (2) meaningful conflicts between visual and factual information occur, or (3) such conflicts can be used to analyze prioritization behavior. For example, reVLAT includes arbitrary height-weight scatterplots for which models lack any factual reference point. Without verifiable facts to contradict, models have no choice but to follow the visualization, preventing assessment of visual-factual prioritization. Unlike reVLAT, which randomizes chart values and thereby weakens the alignment with factual priors, CVLAT *preserves* shared factual priors and deliberately constructs visual-factual conflicts. This design allows us to operationalize visual-factual arbitration as a primary, controlled variable, rather than an incidental by-product of factual misalignment.

To address these limitations, CVLAT follows three key design principles:

First, we replace VLAT's original datasets with data drawn from domains strongly grounded in widely shared factual knowledge, including economic indicators, political demographics, and natural phenomena. This ensures that models are likely to possess relevant factual priors that can conflict with visual encodings.

Second, we exclude purely perceptual tasks such as *Find Anomalies* and *Find Clusters*. These tasks primarily involve pattern recognition rather than data interpretation or factual knowledge integration. This refinement reduces the original 53 VLAT questions to 48 questions that focus on data interpretation, where visual-factual conflicts can meaningfully arise.

Third, we structure answer options to explicitly detect information prioritization:

- a visually correct option derived from the counterfactual visualization
- a factually correct option based on real-world knowledge
- an 'Omit' option to allow uncertainty
- distractor options for 4- and 5-choice questions

The option structure together with the 120-permutation sweep instantiates the forced-choice paradigm of psychophysics [58], [59] and the syllogistic belief-bias paradigm of Trippas et al. [60]: contradictory signals are deliberately presented and the model must choose among labeled options including an explicit "Omit" escape, so that systematic per-model leans across all trials are interpretable as preferences.

2) *Human baseline study (design)*: To calibrate CVLAT difficulty against the established VLAT human baseline, we administered the same 48 CVLAT items to $N = 30$ Prolific participants using the same multiple-choice format and correction-for-guessing scoring as the original VLAT study [20]. Items were presented one per page with option order randomized per participant. Participants were not informed of the counterfactual construction. Results of this calibration are reported in Sec. V-B (alongside model results) and Appendix C.

3) *Experimental Protocol and Evaluation Methodology*: The Counterfactual Visualization Literacy Assessment Test (CVLAT) comprises 48 multiple-choice questions with varying option counts (3, 4, or 5 options), each including an 'Omit' choice. Following the methodology of Experiment 1, we address potential ordering effects by systematically generating all possible permutations based on the 5-option questions. This results in each model processing 5,760 distinct question variations (120 permutations $\times$ 48 questions) using the Normal prompt condition from Experiment 1, yielding 86,400 total trials across all 15 LVLMs.

4) *Capability references*: Alongside CVLAT, we administer two control conditions derived from the same item set: the *anonymized visual baseline* (V_{anon}) and the *Q-only* condition. The anonymized visual baseline presents the same counterfactual chart with domain-identifying cues replaced by neutral placeholders, removing the factual signal so that the resulting response distribution measures pure chart-reading capability. Symmetrically, the Q-only condition asks the same domain question without any accompanying chart, allowing its factual-correct rate (F_Q) to estimate the model's factual-prior availability. Because both conditions repurpose the exact CVLAT items rather than introducing an external control set, they disentangle chart-reading proficiency from factual-prior reliance without content-based confounds.

5) *Evaluation Metrics*: We define three complementary evaluation metrics to quantify how LVLMs prioritize visual information and factual knowledge under conflict:

- **Visualization Fidelity Score (VF Score)**: Measures adherence to visual information
- **Factual Alignment Score (FA Score)**: Measures reliance on factual knowledge
- **Visual-Factual Reliance Index (VFRI)**: Captures relative preference between visual and factual sources

In the definitions below, V_{CVLAT} and F_{CVLAT} denote visual-aligned and factual-aligned accuracy in CVLAT, while V_{anon} and F_Q serve as reference scores for capability normalization.

Correction for Guessing

To address the issue of guessing in multiple-choice questions, we apply a correction-for-guessing formula adapted from educational assessment literature [61]–[63]. For each question i , let S_i denote the rate of the target response, W_i the rate of distractor responses, C_i the number of answer options, and D_i the number of distractor options. The corrected score is:

$$\text{Score}_i = \max\left(0, S_i - \frac{W_i}{D_i}\right). \quad (1)$$

TABLE III

CVLAT EVALUATION RESULTS. **ANON V%** (VISUAL-SELECTION RATE ON THE ANONYMIZED VISUAL BASELINE) AND **Q-ONLY %** (TEXT-ONLY FACTUAL-KNOWLEDGE ACCURACY) ARE CAPABILITY REFERENCES USED TO NORMALIZE VF AND FA, RESPECTIVELY. THE **FALSE** COLUMN REPORTS THE SHARE OF RESPONSES THAT ARE NEITHER VISUALLY NOR FACTUALLY CORRECT (I.E., DISTRACTOR SELECTIONS OR OMIT), AND VISUAL, FACTUAL, AND FALSE SUM TO 100%. VFRI IS THE RELATIVE PREFERENCE BETWEEN VF AND FA (RANGE -1 PURELY FACTUAL TO $+1$ PURELY VISUAL). THE 95% CI COLUMN REPORTS THE PERCENTILE BOOTSTRAP INTERVAL ($B = 10,000$) OF THE PER-QUESTION VFRI. GROUP LABELS ARE DESCRIPTIVE, AND MODELS WHOSE 95% CI INCLUDES ZERO SHOULD BE INTERPRETED AS NEAR-BOUNDARY CASES. VF AND FA ARE CAPABILITY-NORMALIZED RATIOS AND ARE NOT BOUNDED ABOVE BY 1, WHICH IS WHY THE LOW-CAPABILITY GROK ROWS EXCEED 1. VFRI IS THE BOUNDED $[-1, +1]$ SUMMARY INDEX.

Model	Anon V	Q-only	Response Dist. (%)			Scores			
	(%)	(%)	Visual	Factual	False	VF	FA	VFRI	95% CI
Factual knowledge-oriented models (point-estimate VFRI < 0)									
Grok-4.20	35.83	57.64	17.45	57.76	24.79	0.543	1.068	-0.594	[−0.78, −0.37]
Qwen3-VL-32B	73.06	62.81	38.12	50.68	11.20	0.422	0.859	-0.496	[−0.72, −0.25]
Qwen3-VL-235B	59.60	62.90	27.57	53.85	18.58	0.352	0.707	-0.358	[−0.61, −0.10]
Llama4-Scout	60.69	76.65	25.30	55.05	19.65	0.380	0.904	-0.339	[−0.52, −0.15]
Grok-4.3	33.70	63.66	17.19	59.76	23.06	1.070	1.437	-0.293	[−0.48, −0.10]
Llama4-Maverick	72.03	84.83	33.25	54.18	12.57	0.386	0.541	-0.106	[−0.37, +0.15]
Visualization-oriented models (point-estimate VFRI ≥ 0)									
Qwen3-VL-8B	46.20	62.73	32.62	47.20	20.17	0.880	0.545	+0.048	[−0.27, +0.36]
Claude-Haiku-4.5	82.97	74.65	49.18	43.40	7.41	0.552	0.671	+0.120	[−0.15, +0.39]
Gemma-4-31B	70.94	73.80	45.97	38.18	15.85	0.585	0.413	+0.172	[−0.11, +0.45]
Gemma-4-26B-A4B	82.47	77.52	56.28	31.37	12.34	0.631	0.368	+0.208	[−0.07, +0.47]
GPT-5.5	75.26	84.70	49.01	38.33	12.66	0.644	0.394	+0.247	[−0.01, +0.50]
Claude-Sonnet-4.6	88.11	84.84	63.54	30.82	5.64	0.864	0.344	+0.310	[+0.05, +0.56]
Gemini-3.1-Flash-Lite	75.36	80.94	50.05	36.72	13.23	0.586	0.329	+0.325	[+0.04, +0.59]
Claude-Opus-4.7	91.70	83.21	79.44	13.65	6.91	0.792	0.235	+0.669	[+0.44, +0.87]
Gemini-3.1-Pro	94.25	88.87	90.80	6.11	3.09	0.987	0.043	+0.892	[+0.77, +0.98]

This formula penalizes random guessing by subtracting the expected contribution of distractor responses, ensuring that chance-level performance yields a score near zero. The number of distractors D_i depends on the specific condition. In CVLAT, both visual-correct and factual-correct options are treated as focal response categories by construction, yielding $D_i = C_i - 2$. In contrast, for the capability-reference conditions (V_{anon} , Q-only), only one focal target category exists, yielding $D_i = C_i - 1$.

Visualization Fidelity Score (VF Score)

The VF Score measures the extent to which LVLMS follow visual information when it contradicts factual knowledge, corresponding to the Visual Override (**VC**∧**FI**) case in our framework. We apply Equation 1 on CVLAT with S_i set to the visual-correct rate to obtain $V_{\text{CVLAT},i}$, and to the anonymized visual baseline under the same S_i setting to derive the chart-reading capability reference $V_{\text{anon},i}$. To prevent capability deficits (e.g., a model that simply cannot read the chart) from being mistaken for factual bias, we normalize $V_{\text{CVLAT},i}$ by the capability reference $V_{\text{anon},i}$:

$$VF_i = \frac{V_{\text{CVLAT},i}}{V_{\text{anon},i} + \varepsilon}, \quad (2)$$

where $\varepsilon = 10^{-6}$ prevents division by zero when the capability reference is near zero. The overall VF Score is the mean of VF_i across all questions.

Factual Alignment Score (FA Score)

The FA Score quantifies the extent to which LVLMS prioritize factual knowledge over contradictory visual information, corresponding to the Factual Override (**VI**∧**FC**) case in our framework. We apply Equation 1 on CVLAT with S_i set to

the factual-correct rate to obtain $F_{\text{CVLAT},i}$. Following the same formulation, we evaluate the Q-only condition to obtain the factual-prior availability reference $F_{Q,i}$. Symmetrically to the visual dimension (i.e., VF):

$$FA_i = \frac{F_{\text{CVLAT},i}}{F_{Q,i} + \varepsilon}. \quad (3)$$

The overall FA Score is the mean of FA_i across all questions. **Visual-Factual Reliance Index (VFRI)**

The VFRI combines the VF and FA scores into a single measure that reflects an LVLMS's relative preference between visual information and factual knowledge. For each question i , the index is defined as:

$$\text{VFRI}_i = \frac{VF_i - FA_i}{VF_i + FA_i + \varepsilon} \quad (4)$$

where ε is the same small constant defined for Eq. 2. The overall VFRI is the mean across all questions, ranging from -1 (strong factual preference) to $+1$ (strong visual preference).

B. Experimental Results

Human baseline. Following the original VLAT study's correction-for-guessing convention (the visually-correct option as the scoring target), mean human accuracy on CVLAT is 53.71% (SD 14.92), and raw uncorrected accuracy is 60.76% (SD 11.70). The corrected score is statistically equivalent to both the original VLAT human baseline (51.91%, SD 16.57, $N = 191$ [20]) and to a recent Prolific VLAT replication reported in the Mini-VLAT study (53.08%, SD 18.96, $N = 199$ [26]) within a ± 10 pp equivalence margin (see Appendix C for the scoring formula and formal equivalence

test). We report this as a calibration result establishing that CVLAT is not substantially harder for humans than VLAT, rather than as a full revalidation of VLAT-equivalent difficulty. Applying the same formula with the *Factual*-correct option as the target (i.e., counting how often participants chose the option matching real-world fact but contradicting the chart) yields a corrected score of -9.46% (SD 8.97, unclamped), which is substantially below chance. In the aggregate, participants did not override the chart with prior factual knowledge but predominantly followed the visual representation, a behavior shared by all 30 participants individually. For this factual-scored diagnostic, we intentionally use the unclamped correction-for-guessing score to effectively detect below-chance factual selection. Conversely, the LVLm VF/FA normalization continues to employ the clamped score in Eq. 1.

LVLm results. Table III presents the comprehensive evaluation results of 15 LVLms on CVLAT, revealing distinct prioritization patterns. Figure 4 provides a visual landscape of the corresponding VFRI-accuracy relationships across the suite. Based on the sign of their point-estimate VFRI, models are descriptively categorized into two distinct cohorts:

- 1) **Factual knowledge-oriented models:** Grok-4.20, Qwen3-VL-32B, Qwen3-VL-235B, Llama4-Scout, Grok-4.3, and Llama4-Maverick. These models tend to favor factual knowledge over visual information when faced with counterfactual visualizations, although the strength of this tendency varies across architectures.
- 2) **Visualization-oriented models:** Qwen3-VL-8B, Claude-Haiku-4.5, Gemma-4-31B, Gemma-4-26B-A4B, GPT-5.5, Claude-Sonnet-4.6, Gemini-3.1-Flash-Lite, Claude-Opus-4.7, and Gemini-3.1-Pro. These models tend to follow visual encodings over factual knowledge when faced with counterfactual visualizations, although the strength of this tendency varies across architectures.

Nine of fifteen models have positive point-estimate VFRI. The most recent proprietary releases, including Claude-Opus-4.7 (VFRI +0.669), Gemini-3.1-Pro (+0.892), Claude-Sonnet-4.6 (+0.310), Gemini-3.1-Flash-Lite (+0.325), and GPT-5.5 (+0.247), fall on the visualization-oriented side, while both Meta Llama4 checkpoints and both xAI Grok checkpoints remain factual knowledge-oriented. While prior work [12] reported that LVLms rely heavily on pre-existing knowledge when interpreting visualizations, our results show that this tendency is not uniform across architectures or generations. Individual LVLms exhibit distinct and measurable prioritization patterns rather than a single shared bias.

At the strong-visualization end, Gemini-3.1-Pro achieves the highest VFRI (+0.892) and a near-ceiling VF Score (0.987) while maintaining a remarkably low false-response rate (3.09%), indicating robust visual interpretation even when visualizations present counterfactual information. Claude-Opus-4.7 follows closely (VFRI +0.669, VF Score 0.792, false rate 6.91%). Both models also have near-ceiling chart-reading capability references (Anon V $\geq 91\%$), so their positive VFRI reflects an actual visual preference rather than a capability asymmetry. At the opposite end, Grok-4.20 has the strongest negative VFRI (-0.594). Even after capability normalization,

the model selects the factual option much more often than the visual option, consistent with aggregate factual-prior reliance relative to its measured chart-reading capability (Anon V 35.83%).

A notable contrast emerges within the Qwen3-VL family. The 8B checkpoint has a weak positive, near-boundary VFRI (+0.048), whereas the 32B and 235B checkpoints show clear factual orientation (VFRI -0.496 and -0.358 , respectively). Although recent research suggests that larger models can store substantially more factual knowledge, with capacity scaling linearly with model size [64], [65], our Q-only probe shows that factual-prior availability is nearly identical across the three sizes (62.73%, 62.81%, and 62.90% for 8B, 32B, and 235B). The within-family contrast therefore more directly reflects differences in how strongly each checkpoint arbitrates in favor of factual priors when those priors are equally available.

This pattern does not generalize across families. Both Llama4 checkpoints lean factual (Maverick -0.106 , Scout -0.339), both Gemma-4 checkpoints lean visual (positive VFRI), and all three Claude checkpoints lean visual. The Llama4 lean is notable because both checkpoints possess reasonable chart-reading capability (Anon V 72.03 and 60.69) yet exhibit Q-only factual rates (84.83 and 76.65) comparable to the strongest factual-prior models in the suite. Notably, recent proprietary frontier models, including Gemini-3.1-Pro (+0.892), Claude-Opus-4.7 (+0.669), and Claude-Sonnet-4.6 (+0.310), remain visualization-oriented despite their undisclosed but presumably larger parameter counts, suggesting that scale is only one factor among several—such as training-data composition, RLHF objectives, and instruction-tuning recipes—that shape arbitration behavior. Because vision-encoder, connector, and tokenizer details are not disclosed for the proprietary endpoints in our suite, systematically disentangling the specific impact of architecture versus model family remains a target for future work.

Relating these patterns to the human baseline reported above (gold diamond in Fig. 4) sharpens the human-model comparison. Humans are visualization-oriented without exception (all 30 participants have positive VFRI, with factual-scored accuracy well below chance), whereas the models split into oriented groups. The human-model divergence is therefore concentrated in the factual knowledge-oriented checkpoints (Grok-4.20, Grok-4.3, Qwen3-VL-32B, Qwen3-VL-235B, and Llama4-Scout), which select the factually-correct, chart-contradicting option far more often than any human participant did, whereas the visualization-oriented models resolve the conflict much as the human cohort does. Appendix E analyzes the per-question VFRI behind these aggregates for the human cohort and all 15 models, identifying which kinds of questions drive visually- or factually-oriented behavior.

VI. EXPERIMENT THREE: STEERING LVLms' INFORMATION PRIORITIZATION THROUGH PROMPT ENGINEERING

In our third experiment, we investigate **RQ3**: To what extent can prompt engineering shift LVLms' information prioritization between visual information and factual knowledge when the two sources conflict?

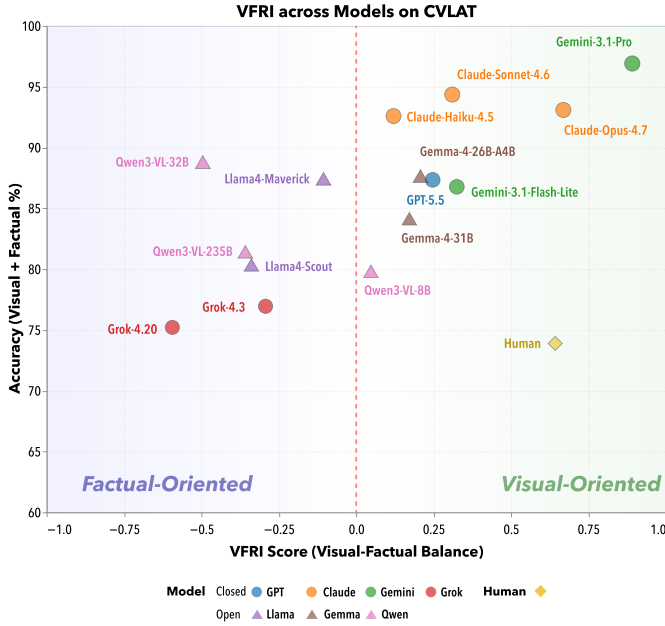


Fig. 4. Relationship between Visual-Factual Reliance Index (VFRI) and accuracy. The x-axis shows VFRI scores ranging from -1 (strong factual preference) to 1 (strong visual preference), while the y-axis represents accuracy (percentage of visual + factual correct responses). Marker shapes indicate model source type (circles for closed-source, triangles for open-source), and colors represent different model families. Per-model 95% bootstrap CIs are reported in Table III. The gold diamond marks the human cohort ($N = 30$) at its mean per-participant VFRI ($+0.64$) and accuracy (74.0%). The human VFRI is computed without capability normalization (Appendix C).

This investigation aims to determine whether the information prioritization patterns identified in Experiment 2 can be influenced through explicit prompt engineering, or whether they instead reflect more stable model-specific tendencies.

A. Prompt Design

We adopt an explicit prompting strategy to investigate whether prompt engineering can shift models' information prioritization preferences (see Appendix A for full prompt texts). Two contrasting prompts are designed: the **factual-priority prompt** instructs models to prioritize their factual knowledge over visual information when contradictions arise, while the **visual-priority prompt** instructs models to treat visual data as ground truth and respond based on visual information, even when it conflicts with their factual knowledge.

B. Experimental Protocol and Analysis Methodology

We follow the same experimental protocol as Experiment 2, using CVLAT with 48 questions and 120 permutations per question. Every model is evaluated under both the factual-priority and visual-priority prompts, yielding 11,520 trials per model and 172,800 trials across all 15 models.

For analysis, we compute the VF Score, FA Score, and VFRI under the prompt-engineering condition and compare them with the corresponding baseline scores from Experiment 2. Prompt effectiveness is quantified as the change (Δ) between intervention and baseline scores, where positive Δ VFRI values indicate shifts toward prioritizing visual information

and negative values indicate shifts toward prioritizing factual knowledge. Statistical significance is assessed using paired bootstrap testing (10,000 iterations, $\alpha = 0.05$), with 95% confidence intervals and significance labels reported in Table IV.

C. Experimental Results

Table IV and Figures 5–6 summarize the effects of prompt engineering on LVLMs' information prioritization. Models fall into four distinct controllability profiles, which serve as the organizational basis for the rows of Table IV.

Symmetric responders. Claude-Opus-4.7 exhibited one of the most pronounced responses to prompt engineering, shifting from a strongly visualization-oriented baseline (VFRI $+0.669$) to a clearly factual knowledge-oriented stance under the factual-priority prompt (VFRI -0.497 , $\Delta_F = -1.166$, $p < 0.001$), while extending further toward visual reliance under the visual-priority prompt (VFRI $+0.910$, $\Delta_V = +0.241$, $p < 0.01$). Gemini-3.1-Flash-Lite shows a comparably broad bidirectional range ($\Delta_F = -1.012$, $\Delta_V = +0.386$, both $p \leq 0.01$). Claude-Sonnet-4.6, Gemma-4-31B, and Gemma-4-26B-A4B also belong to this profile, shifting significantly in both directions as instructed. Claude-Haiku-4.5 is directionally consistent in both directions but neither shift reaches statistical significance, suggesting an underpowered variant of the same underlying pattern.

F-priority-collapsing models. Qwen3-VL-32B and Qwen3-VL-235B exhibited a counter-intuitive pattern: the factual-priority prompt triggered a shift *toward* visual processing instead of deepening further factual alignment (positive and significant Δ_F : $+0.759$ *** for 32B; $+0.322$ * for 235B), while their visual-priority shifts remained uniformly positive and substantial (Δ_V up to $+1.096$ for 32B). The pattern is accompanied by a sharp surge in generation length: median completion tokens expanded from 7 in the baseline to 212 under the factual-priority condition for Qwen3-VL-32B, and from 3 to 136 for Qwen3-VL-235B. Rather than providing a direct answer, factual-priority outputs typically begin with an explicit breakdown of the chart. On a stacked-area GDP item, for example, Qwen3-VL-32B transitioned from a single-token baseline response of “(a)” to an extended text starting, “*The chart provided shows GDP Growth Trends for the US, China, and Japan from 2009 to 2014 ... In 2009, the US GDP is around \$14,000 billion ...*”. We interpret this in Sec. VII as a deliberation-activation effect. By compelling the model to verify facts, the factual-priority prompt elicits longer, chart-referencing outputs that appear to surface visual evidence the terse baseline elided, eventually driving the model's final state away from (not toward) the prompted direction.

V-priority-insensitive models. GPT-5.5 exhibited a strong factual-priority response ($\Delta_F = -0.784$, $p < 0.001$) but no detectable visual-priority response ($\Delta_V = -0.062$, ns); uniquely within the evaluation suite, its visual-priority point estimate was mildly negative. This asymmetry persists despite the model possessing an intact chart-reading capability. For example, on the tech-company quarterly-revenue bar chart (evaluating the comparison “IBM Q2 revenue is larger than Facebook's”), GPT-5.5 achieves a visual-correct rate of 100%

TABLE IV

EFFECTS OF PROMPT ENGINEERING ON LVLM INFORMATION PRIORITIZATION, GROUPED BY CONTROLLABILITY PROFILE (SEC. VI NARRATIVE). ALL VALUES ARE VISUAL-FACTUAL RELIANCE INDEX (VFRI). THE TWO Δ VFRI COLUMNS FLANK THE BASELINE CONDITION (Δ = PROMPTED – BASELINE, WHERE POSITIVE SHIFTS INDICATE A MOVEMENT TOWARD VISUAL EVIDENCE AND NEGATIVE SHIFTS TOWARD FACTUAL PRIORS). BRACKETS BELOW EACH Δ VALUE REPORT THE 95% CONFIDENCE INTERVAL FROM A PAIRED BOOTSTRAP ($B = 10,000$, SHIFT METHOD). THE TRAJECTORY COLUMN SHOWS THE THREE VFRI POSITIONS ON A $[-1, +1]$ AXIS, COLOR-CODED FOR INTERPRETATION: RED FOR FACTUAL-PRIORITY, BLACK FOR BASELINE, AND BLUE FOR VISUAL-PRIORITY. SIGNIFICANCE LEVELS: *** $p \leq 0.001$, ** $p \leq 0.01$, * $p \leq 0.05$, AND *ns* FOR NOT SIGNIFICANT.

Model	Factual Prompted	Δ_F VFRI	Baseline	Δ_V VFRI	Visual Prompted	Trajectory
<i>Symmetric responders</i>						
Claude-Haiku-4.5 [†]	+0.055	−0.066 ns [−0.221, +0.080]	+0.120	+0.177 ns [−0.013, +0.381]	+0.297	
Gemma-4-31B	−0.268	−0.440 * [−0.794, −0.064]	+0.172	+0.612 *** [+0.377, +0.862]	+0.784	
Gemma-4-26B-A4B	−0.081	−0.289 * [−0.554, −0.015]	+0.208	+0.340 *** [+0.189, +0.513]	+0.548	
Claude-Sonnet-4.6	−0.354	−0.665 *** [−0.903, −0.433]	+0.310	+0.095 * [+0.012, +0.208]	+0.405	
Gemini-3.1-Flash-Lite	−0.687	−1.012 *** [−1.286, −0.735]	+0.325	+0.386 ** [+0.191, +0.612]	+0.711	
Claude-Opus-4.7	−0.497	−1.166 *** [−1.438, −0.892]	+0.669	+0.241 ** [+0.093, +0.415]	+0.910	
<i>F-priority-collapsing</i>						
Qwen3-VL-32B	+0.264	+0.759 *** [+0.507, +1.016]	−0.496	+1.096 *** [+0.819, +1.369]	+0.600	
Qwen3-VL-235B	−0.036	+0.322 * [+0.049, +0.614]	−0.358	+0.733 *** [+0.446, +1.020]	+0.375	
<i>V-priority-insensitive</i>						
GPT-5.5	−0.537	−0.784 *** [−1.011, −0.555]	+0.247	−0.062 ns [−0.184, +0.039]	+0.185	
Gemini-3.1-Pro [‡]	−0.839	−1.731 *** [−1.883, −1.548]	+0.892	+0.028 ns [−0.002, +0.078]	+0.920	
<i>F-priority-insensitive</i>						
Grok-4.20	−0.473	+0.121 ns [−0.080, +0.358]	−0.594	+0.332 *** [+0.149, +0.535]	−0.262	
Llama4-Scout	−0.524	−0.186 ns [−0.417, +0.038]	−0.339	+0.416 *** [+0.187, +0.651]	+0.077	
Grok-4.3	−0.263	+0.030 ns [−0.053, +0.119]	−0.293	+0.166 ** [+0.050, +0.288]	−0.127	
Llama4-Maverick	−0.107	−0.001 ns [−0.246, +0.251]	−0.106	+0.340 * [+0.062, +0.625]	+0.234	
Qwen3-VL-8B	+0.107	+0.060 ns [−0.235, +0.363]	+0.048	+0.278 * [+0.024, +0.553]	+0.326	

* All VFRI values are capability-normalized per Sec. V; baseline column matches Table III exactly.

[†] Both shifts are directionally consistent with the Symmetric profile but neither reaches statistical significance, indicating an underpowered observation rather than confirmed bidirectional control.

[‡] Near-ceiling baseline (VFRI = +0.892); the absence of a significant Δ_V reflects limited headroom for further visual gain rather than insensitivity to the visual-priority prompt.

on the anonymized chart and 97% at baseline, yet this rate drops sharply to 29% under the visual-priority prompt. This indicates that chart-reading capability does not translate into visual reliance under conflict for this model, which instead defaults to its factual prior. Gemini-3.1-Pro exhibits a comparable profile, acting as a near-ceiling variant. Because it operates with a high baseline VFRI of +0.892, there is essentially no remaining headroom for further visual gain, resulting in a stark shift under the factual prompt but stagnation under the visual one ($\Delta_F = -1.731$ ***, $\Delta_V = +0.028$ ns).

F-priority-insensitive models. The Grok and Llama4 fam-

ilies, along with Qwen3-VL-8B, exhibited the mirror-image pattern: they demonstrated significant visual-priority responses (Δ_V ranging from +0.166 to +0.416, all $p \leq 0.05$; see Table IV) but factual-priority effects that fell below significance. These checkpoints comply effectively with the visual-priority instruction, yet the factual-priority instruction produces no statistically detectable shift toward factual reliance.

Taken together, prompt engineering is a viable but non-universal strategy for steering information prioritization. The heterogeneity across models is substantial, including divergent responses between comparably capable proprietary models

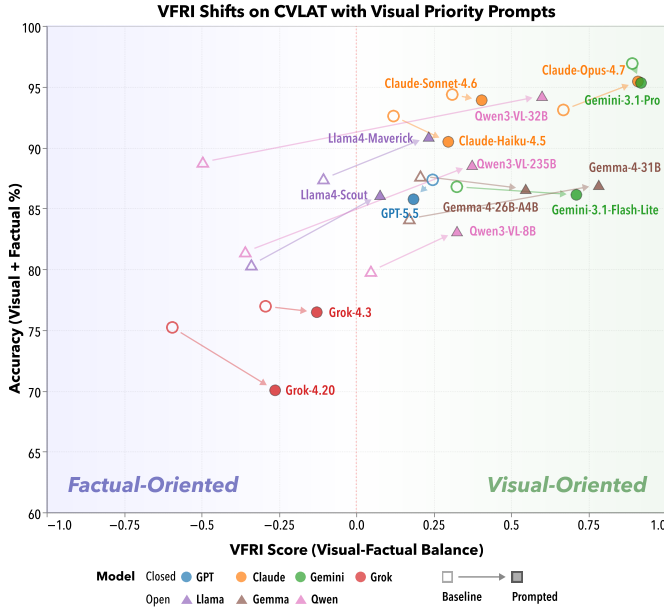


Fig. 5. Visual-priority prompt effect. Each model contributes a hollow baseline marker (start) and a filled visual-priority marker (end) connected by an arrow indicating the direction of change in VFRI and accuracy. Rightward movement along the VFRI axis indicates the prompt successfully pulled the model toward visual evidence. Per-model 95% paired-bootstrap CIs for Δ_V VFRI are reported in Table IV.

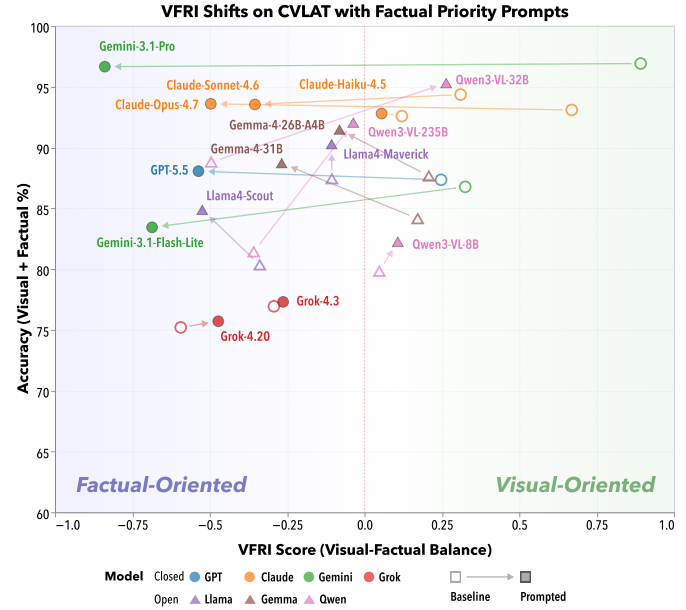


Fig. 6. Factual-priority prompt effect. Each model contributes a hollow baseline marker (start) and a filled factual-priority marker (end) connected by an arrow indicating the direction of change in VFRI and accuracy. Leftward movement indicates the prompt pulled the model toward factual priors, while rightward movement is the F-priority-collapsing anomaly discussed in Sec. VI and Sec. VII. Per-model 95% paired-bootstrap CIs for Δ_F VFRI are reported in Table IV.

(e.g., GPT-5.5 vs. Claude-Opus-4.7). This controllability is often non-reciprocal; single-direction insensitivity manifests in both visual and factual directions, and the F-priority-collapse profile reveals instances where factual-priority prompts inadvertently trigger the inverse behavior. Benchmark performance does not clearly align with these behavioral profiles. Higher-performing Gemini-3.1-Pro (top VLAT-Normal and VLAT-Explain in the suite) and GPT-5.5 (third-highest VLAT-Explain at 93.85%) are classified as V-priority-insensitive, while lower-performing Llama4 variants are F-priority-insensitive. Within-family controllability patterns are mixed. While the Claude, Gemma-4, Llama4, and Grok families consistently fall within a single profile, Qwen3-VL and Gemini diverge internally. The Qwen3-VL lineage shows a clear scale-associated split (8B in F-priority-insensitive, 32B and 235B in F-priority-collapsing) that mirrors the baseline VFRI gradient analyzed in Sec. V. Conversely, the Gemini divergence occurs across distinct scale tiers (Pro in V-priority-insensitive, Flash-Lite in Symmetric). The scale-dependent split observed in Qwen3-VL suggests that a model's intrinsic arbitration tendency and its susceptibility to prompt-based scaffolding covary with scale within the Qwen3-VL family.

VII. DISCUSSION

Recent studies have reported strong performance of LVLMs on visualization interpretation tasks, raising an important question: to what extent does such performance reflect genuine visualization literacy, as opposed to reliance on pre-trained factual knowledge? Our results show that accuracy alone is insufficient to answer this question. By disentangling visual correctness from factual correctness, we show that high

benchmark scores alone do not reveal how models arbitrate under conflict. Although nine of 15 models show positive point-estimate VFRI under the baseline CVLAT condition, Experiment 3 demonstrates that this visual orientation is by no means uniformly stable under prompt-based interventions. Rather, these intervention effects remain heavily constrained, direction-asymmetric, and highly model-dependent. Notably, this behavioral divergence has important implications for how visualization literacy should be evaluated and interpreted in LVLM-based analytical systems.

In line with prior empirical evaluations of LVLM visualization literacy, we assess a state-of-the-art LVLM suite on established benchmarks (Sec. IV, Table II). Several models approach or exceed the 65.5% human VLAT baseline yet generally perform worse on reVLAT. This divergence between VLAT and reVLAT performance underscores a core limitation of accuracy-based evaluation: high scores on aligned benchmarks do not necessarily indicate robust visualization interpretation. Instead, they may reflect reliance on factual priors that break down when visual and factual signals diverge. This observation directly motivates our disentanglement framework, which exposes systematically different response patterns, consistent with visual interpretation versus factual recall, that remain indistinguishable under conventional evaluation protocols.

Interpreted through our proposed quadrant framework, these results yield two key insights. First, on benchmarks such as VLAT where visual encodings align with real-world factual knowledge, high accuracy may be driven by recall of pre-trained knowledge rather than interpreting visual information, leading to inflated estimates of visualization literacy. Sec-

ond, on benchmarks with randomized data, such as reVLAT, models' tendency to prioritize factual knowledge over visual evidence can suppress correct visual interpretation, resulting in underestimation of their actual visualization literacy. These complementary biases help explain the inconsistent findings in prior studies [11], [12]. Differences in benchmark design cause evaluations to capture varying mixtures of visualization interpretation and factual recall, leading to inconsistent assessments of the same models even when similar performance metrics are used.

Building on this insight, we introduce CVLAT to explicitly investigate which information LVLMs prioritize when visual information and factual knowledge (i.e., pre-trained knowledge) conflict, and define metrics to quantify their tendencies. Our results reveal two distinct model groups (factual knowledge-oriented and visualization-oriented). Whereas Hong et al. [12] document that LVLMs rely on prior knowledge as a population-level pattern, our capability-normalized CVLAT metrics show that this reliance is not uniform: individual models exhibit distinct, measurable, and partially steerable prioritization patterns. CVLAT differs from reVLAT in that it preserves the factual signal and forces the two into conflict, surfacing arbitration behavior that is diagnostically relevant whenever visual evidence and prior knowledge disagree.

Finally, we examine whether explicit prompting can modulate LVLMs' information prioritization (Sec. VI). Models exhibit four prompt-controllability profiles. Symmetric responders move in the expected direction under both factual-priority and visual-priority prompts. F-priority-collapsing responders unexpectedly move toward visual evidence when instructed to prioritize factual knowledge. The remaining two profiles are one-sided: V-priority-insensitive models respond only to factual-priority prompts, whereas F-priority-insensitive models respond only to visual-priority prompts. Single-direction prompt failure is therefore a recurring outcome rather than an isolated curiosity, and prompt-controllability cannot be assumed to be reciprocal.

The output-length analysis reported in Sec. VI supports a deliberation-activation interpretation of the F-priority-collapse pattern (the factual-priority prompt activates chart engagement that the baseline behavior elided) rather than active override of visual evidence. We caution that this signature is suggestive rather than definitive. The factual-priority instruction is conditional (triggering only under recognized conflict) while the visual-priority instruction is unconditional, and sparse factual priors are an additional confound. Disentangling these mechanisms more precisely is left for future work.

Overall, the three evaluation instruments (VLAT, reVLAT, and CVLAT) play complementary roles. Together, they provide an assessment by isolating and operationalizing different capacities within LVLm visualization literacy. Specifically, VLAT measures perceptual decoding under aligned conditions, reVLAT measures perceptual decoding with factual priors largely neutralized via data randomization, and CVLAT captures arbitration dynamics when visual and factual signals conflict. The Q-only condition, administered alongside CVLAT as a capability reference, indexes factual grounding (i.e., factual-prior availability without visual evidence), and

the capability-normalization applied in VF/FA incorporates this factual-grounding layer directly into the arbitration metric. This multi-instrument approach addresses diagnostic gaps that no single benchmark resolves. Consequently, CVLAT directly reveals how models prioritize conflicting sources.

Consistent with this design, our human cohort ($N = 30$) under the same Normal prompt setting systematically followed the chart rather than oscillating between visual and factual options, with factual-scored accuracy well below chance (Sec. V-B, Appendix C). This robust human baseline indicates that the counterfactual construction within our benchmark does not collapse into an ambiguity-driven guessing game, even under minimal prompting.

Arbitration is central to deployed literacy because real-world workflows routinely encounter charts whose correct interpretation depends on which competing signal to trust. Therefore, CVLAT serves as a diagnostic designed to match a model's intrinsic orientation with specific use-case requirements, rather than functioning as a conventional, one-dimensional model leaderboard. Visual fidelity and factual prioritization represent descriptive orientations rather than normative judgments.

A. Implications for Visual Analytics System Design

Our findings have practical implications for the deployment of LVLMs in visual analytics systems, particularly in settings where faithful interpretation of visual evidence is critical. Such deployments face two mandates, reporting faithfully what a chart encodes while remaining factually correct about the world, and since neither is universally prior, the implications below center on making the arbitration between these mandates explicit and controllable rather than on prescribing a single priority for all deployments.

Model Selection: Under this framing, choosing a model also means choosing a de facto default arbitration policy. For applications requiring faithful visual interpretation (e.g., exploratory data analysis, anomaly detection, or visual validation), visualization-oriented models (indicated by a high positive VFRI in Table III) are preferable, as these models more reliably report what is visually encoded, even when the observed stimuli directly contradict their intrinsic factual expectations. Conversely, in deployments where misleading or adversarial charts are a concern, factual knowledge-oriented models may be preferable, as their tendency to favor factual priors over conflicting visual encodings could mitigate certain forms of visual misinformation, though we did not directly evaluate adversarial charts. For knowledge-intensive tasks where visualizations largely align with real-world facts, factual knowledge-oriented models may be more appropriate.

Prompt Engineering: Our results demonstrate that prompt engineering has limited and model-dependent effectiveness in redirecting information prioritization. While most evaluated models are responsive to at least one priority prompt, only approximately one-third exhibit bidirectional controllability, and some respond effectively in only a single direction (asymmetric sensitivity). Consequently, for systems requiring dynamic and adaptable information prioritization, developers should prefer models with empirically verified bidirectional

controllability rather than assuming prompt engineering alone suffices. Importantly, high visualization-literacy benchmark scores do not necessarily imply prompt-controllability. GPT-5.5 attains 93.85% on VLAT-Explain and a high anonymized-visual baseline (75.3%) on CVLAT, yet shows no significant visual-priority response (Sec. VI). Visual capability and prompt-controllability should therefore be verified separately rather than inferred from one another.

Beyond Single-Channel Responses: For high-stakes applications, robustness can be improved by pairing models with contrasting orientations, incorporating CVLAT-style uncertainty estimation, and adding human-in-the-loop review under strong visual-factual conflict or low model confidence. Beyond such safeguards, the single-answer format itself can be relaxed so that one response serves both mandates: a model can disclose a detected conflict explicitly (“the chart shows X , which contradicts the widely accepted value Y ”) or separate chart-extracted content from prior-knowledge context, echoing Lundgard and Satyanarayan’s semantic content framework [49]. The chart-referencing deliberation under factual-priority prompts (Sec. VI) suggests that instructions can elicit explicit chart engagement, although we did not test such reporting directly. Realizing either design also requires evaluation that scores conflict detection and disclosure separately from answer accuracy, a capability CVLAT does not yet provide.

B. Limitations and Future Work

While our study provides valuable insights, several limitations warrant future investigation.

Scope Limitations: Our evaluation covers common visualization types such as bar charts, line graphs, scatterplots, and pie charts. Generalizing our findings to more intricate chart types and formats (e.g., network diagrams, heatmaps with complex encoding schemes, 3D plots, domain-specific scientific charts, and interactive or animated formats) remains an open question that falls outside the scope of our current counterfactual framework. Additionally, our counterfactual design manipulates only numerical values. Non-numerical encoding dimensions, including color, spatial layout, and temporal ordering, remain largely unexplored and present a promising avenue that may reveal distinct prioritization mechanisms. Finally, we restrict our experiments to English prompts, which may overlook language-specific behaviors in multilingual or non-English LLMs.

Methodological Considerations: The CVLAT suite comprises 48 items, which is sufficient for identifying broad behavioral patterns but constrained for fine-grained or per-task analysis. In addition, the VFRI metric should be interpreted in conjunction with the false-response rate. A near-zero VFRI does not, by itself, indicate a balanced integration of sources; intermediate values can also arise from elevated false-response (neither visually nor factually correct) rates, signaling model instability or confusion rather than deliberate arbitration.

Future Directions: Promising directions for future work include the following. (1) Investigating whether structural and architectural interventions—such as targeted fine-tuning and

data-centric methods [32], [33], [35], [36], as well as controlled vision-encoder and connector ablations—can achieve cleaner decoupling or more predictable steering of visual-factual prioritization gradients. (2) Systematically varying conflict magnitude (ranging from subtle discrepancies to full reversals) and partial conflicts (where only a specific subset of data points is perturbed) to map precisely how model override rates depend on mismatch strength. (3) Exploring richer prompt designs beyond the binary priority framings evaluated in this work. In particular, this includes developing conflict-aware prompts that explicitly alert the model to potential visual-factual discrepancies, alongside incorporating few-shot demonstrations that contextually anchor visual override behaviors. (4) Resolving the underlying causality behind the F-priority-collapse pattern: while our output-length analysis is consistent with a hypothesis of heightened deliberation-activation, decoupling this from alternative explanations will require controlled, mechanism-level probes (e.g., matched conditional/unconditional instructions and direct access to internal model states). (5) Developing adaptive evaluation frameworks that algorithmically or systematically adjust task difficulty based on an individual model’s underlying baseline capabilities, mitigating both floor and ceiling effects across highly disparate model tiers.

VIII. CONCLUSION

This work provides a systematic examination of LLMs’ visualization literacy capabilities and exposes fundamental limitations in how it is currently evaluated. By introducing a quadrant framework that disentangles visual correctness from factual correctness, we show that existing accuracy-based evaluation methods often conflate genuine visualization interpretation with reliance on pre-trained knowledge. As a result, high performance does not necessarily imply faithful visualization understanding. Our experiments with 15 state-of-the-art LLMs, complemented by a human baseline study ($N = 30$, recruited via Prolific), demonstrate five main findings. First, several SOTA LLMs achieve visualization-literacy performance approaching or exceeding human baselines. Under the Normal prompt condition, Gemini-3.1-Pro reaches 99.7% on VLAT and 93.3% on reVLAT, while Claude-Opus-4.7 reaches 94.9% on VLAT and 88.5% on reVLAT. Second, the evaluated suite tilts visualization-oriented on average, but a factual knowledge-oriented minority shows statistically significant factual orientation that masks their underlying visualization literacy whenever visual and factual signals conflict. Third, our proposed CVLAT test reveals distinct model groups—factual knowledge-oriented and visualization-oriented—with measurable differences in information prioritization. Fourth, prompt engineering can alter information prioritization, but its effectiveness is highly model-dependent and often direction-asymmetric. Factual-priority instruction sometimes shifts models toward visual rather than factual evidence, and one-sided prompt failure occurs in either direction (V- or F-priority-insensitive), indicating that prompt-controllability is often not reciprocal and cannot be assumed from benchmark performance alone. Fifth, a human baseline shows that humans

systematically follow the chart rather than overriding it with factual priors (Sec. V-B, Appendix C). Overall, these findings have critical implications for deploying LVLMS in visual analytics systems, as practitioners must carefully consider not only overall accuracy, but also how models prioritize visual and factual information. More broadly, our work establishes a foundation for developing more nuanced evaluation frameworks and understanding the interplay between visual interpretation and factual knowledge in multimodal foundation models.

ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (No. NRF-2023R1A2C2005209; No. RS-2025-00520170; No. RS-2024-00456657), the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean government (MSIT) [No. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], the AI Seoul Tech Research Support Program of the Seoul Future Foundation, the SNU-Global Excellence Research Center establishment project, and the Hankuk University of Foreign Studies Research Fund. The ICT at Seoul National University provided research facilities for this study.

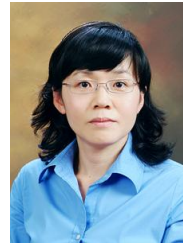
REFERENCES

- [1] J. Kim, S. Lee, H. Jeon, K.-J. Lee, H.-J. Bae, B. Kim, and J. Seo, "PhenoFlow: A human-LLM driven visual analytics system for exploring large and complex stroke datasets," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 1, pp. 470–480, 2025.
- [2] C. Wang, J. Thompson, and B. Lee, "Data formulator: AI-powered concept-driven visualization authoring," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 1, pp. 1128–1138, 2024.
- [3] C. Wang, B. Lee, S. M. Drucker, D. Marshall, and J. Gao, "Data formulator 2: Iterative creation of data visualizations, with AI transforming data along the way," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–17.
- [4] Y. Tian, W. Cui, D. Deng, X. Yi, Y. Yang, H. Zhang, and Y. Wu, "ChartGPT: Leveraging LLMs to generate charts from abstract natural language," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 3, pp. 1731–1745, 2025.
- [5] P. Maddigan and T. Susnjak, "Chat2VIS: Generating data visualizations via natural language using ChatGPT, Codex and GPT-3 large language models," *IEEE Access*, vol. 11, pp. 45 181–45 193, 2023.
- [6] V. Dibia, "LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2023, pp. 113–126.
- [7] J. Chen, J. Wu, J. Guo, V. Mohanty, X. Li, J. P. Ono, W. He, L. Ren, and D. Liu, "Interchat: Enhancing generative visual analytics using multimodal interactions," *Computer Graphics Forum*, vol. 44, no. 3, p. e70112, 2025.
- [8] S. Jung, H. Jeon, J. Rhee, and J. Seo, "Can VLMs assess similarity between graph visualizations?" *arXiv preprint arXiv:2504.09859*, 2025.
- [9] J. Wang, Q. Ye, L. Liu, N. L. Guo, and G. Hu, "Scientific figures interpreted by ChatGPT: strengths in plot recognition and limits in color perception," *NPJ Precision Oncology*, vol. 8, no. 1, p. 84, 2024.
- [10] K. Choe, C. Lee, S. Lee, J. Song, A. Cho, N. W. Kim, and J. Seo, "Enhancing data literacy on-demand: LLMs as guides for novices in chart interpretation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 9, pp. 4712–4727, 2025.
- [11] A. Bendeck and J. Stasko, "An empirical evaluation of the GPT-4 multimodal language model on visualization literacy tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 1, pp. 1105–1115, 2025.
- [12] J. Hong, C. Seto, A. Fan, and R. Maciejewski, "Do LLMs have visualization literacy? an evaluation on modified visualizations to test generalization in data interpretation," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 10, pp. 7004–7018, 2025.
- [13] S. Pandey and A. Ottley, "Benchmarking visual language models on standardized visualization literacy tests," *Computer Graphics Forum*, vol. 44, no. 3, p. e70137, 2025.
- [14] Z. Li, H. Miao, V. Pascucci, and S. Liu, "Visualization literacy of multimodal large language models: A comparative study," *arXiv preprint arXiv:2407.10996*, 2024.
- [15] L. Y.-H. Lo and H. Qu, "How good (or bad) are LLMs at detecting misleading visualizations?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 1, pp. 1116–1125, 2025.
- [16] K. Mukherjee, D. Ren, D. Moritz, and Y. Assogba, "EncQA: Benchmarking vision-language models on visual encodings for charts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 32, no. 1, pp. 648–658, 2026.
- [17] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "GPT-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [18] Anthropic, "The Claude 3 model family: Opus, Sonnet, Haiku," Anthropic, Tech. Rep., 2024. [Online]. Available: <https://www.anthropic.com/claude-3-model-card>
- [19] Gemini Team, P. Georgiev, V. I. Lei, R. Burnell, L. Bai, A. Gulati, G. Tanzer, D. Vincent, Z. Pan, S. Wang *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," *arXiv preprint arXiv:2403.05530*, 2024.
- [20] S. Lee, S.-H. Kim, and B. C. Kwon, "VLAT: Development of a visualization literacy assessment test," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 551–560, 2017.
- [21] J. Boy, R. A. Rensink, E. Bertini, and J.-D. Fekete, "A principled way of assessing visualization literacy," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 12, pp. 1963–1972, 2014.
- [22] K. Börner, A. Maltese, R. N. Balliet, and J. Heimlich, "Investigating aspects of data visualization literacy using 20 information visualizations and 273 science museum visitors," *Information Visualization*, vol. 15, no. 3, pp. 198–213, 2016.
- [23] A. V. Maltese, J. A. Harsh, and D. Svetina, "Data visualization literacy: Investigating data interpretation along the novice–expert continuum," *Journal of College Science Teaching*, vol. 45, no. 1, pp. 84–90, 2015.
- [24] M. Galesic and R. Garcia-Retamero, "Graph literacy: A cross-cultural comparison," *Medical Decision Making*, vol. 31, no. 3, pp. 444–457, 2011.
- [25] B. Bach, S. Huron, U. Hinrichs, J. C. Roberts, and S. Carpendale, "Special issue on visualization teaching and literacy," *IEEE Computer Graphics and Applications*, vol. 41, no. 06, pp. 13–14, 2021.
- [26] S. Pandey and A. Ottley, "Mini-VLAT: A short and effective measure of visualization literacy," in *Computer Graphics Forum*, vol. 42, no. 3. Wiley Online Library, 2023, pp. 1–11.
- [27] L. W. Ge, Y. Cui, and M. Kay, "CALVI: Critical thinking assessment for literacy in visualizations," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 2023, pp. 1–18.
- [28] F. B. Baker, *The basics of item response theory*. ERIC, 2001.
- [29] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34 892–34 916, 2023.
- [30] S. Park, Y. Song, S. Lee, J. Kim, and J. Seo, "Leveraging multimodal LLM for inspirational user interface search," in *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 2025, pp. 1–22.
- [31] S. Lee, M. Chang, S. Park, and J. Seo, "Assessing graphical perception of image embedding models using channel effectiveness," in *2024 IEEE visualization and visual analytics (VIS)*. IEEE, 2024, pp. 226–230.
- [32] M. Huang, H. Lai, X. Zhang, W. Wu, J. Ma, L. Zhang, and J. Liu, "EvoChart: A benchmark and a self-training approach towards real-world chart understanding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 4, 2025, pp. 3680–3688.
- [33] A. K. Das, M. Tarun, and K. Mueller, "Charts-of-thought: Enhancing LLM visualization literacy through structured data extraction," *IEEE Transactions on Visualization and Computer Graphics*, vol. 32, no. 1, pp. 1427–1437, 2026.
- [34] Y. Wu, L. Yan, L. Shen, Y. Wang, N. Tang, and Y. Luo, "ChartInsights: Evaluating multimodal large language models for low-level chart question answering," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 12 174–12 200.

- [35] Z. Xu, B. Qu, Y. Qi, S. Du, C. Xu, C. Yuan, and J. Guo, "ChartMoE: Mixture of diversely aligned expert connector for chart understanding," in *International Conference on Learning Representations (ICLR)*, 2025.
- [36] A. Masry, M. Thakkar, A. Bajaj, A. Kartha, E. Hoque, and S. Joty, "ChartGemma: Visual instruction-tuning for chart reasoning in the wild," in *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, 2025, pp. 625–643.
- [37] A. Verma, K. Mukherjee, C. Potts, E. Kreiss, and J. E. Fan, "CHART-6: Human-centered evaluation of data visualization understanding in vision-language models," *arXiv preprint arXiv:2505.17202*, 2025.
- [38] H. W. Wang, J. Hoffswell, S. M. T. Thane, V. S. Bursztyn, and C. X. Bearfield, "How aligned are human chart takeaways and LLM predictions? A case study on bar charts with varying layouts," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 1, pp. 536–546, 2025.
- [39] J. Satkunarajan, M. Abdelaal, S. Koch, K. Kurzhals, and D. Weiskopf, "Can LLMs simulate target users in visualization case studies?" *Computer Graphics Forum*, p. e70446, 2026.
- [40] E. Wall, L. M. Blaha, L. Franklin, and A. Endert, "Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics," in *2017 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 2017, pp. 104–115.
- [41] E. Dimara, S. Franconeri, C. Plaisant, A. Bezerianos, and P. Dragicevic, "A task-based taxonomy of cognitive biases for information visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 2, pp. 1413–1432, 2020.
- [42] D. Kahneman, *Thinking, fast and slow*. Macmillan, 2011.
- [43] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty," *Science*, vol. 185, no. 4157, pp. 1124–1131, 1974.
- [44] C. Xiong, L. Van Weelden, and S. Franconeri, "The curse of knowledge in visual data communication," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 10, pp. 3051–3062, 2020.
- [45] Y. Bang, Z. Ji, A. Schelten, A. Hartshorn, T. Fowler, C. Zhang, N. Cancedda, and P. Fung, "HalluLens: LLM hallucination benchmark," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2025, pp. 24 128–24 156.
- [46] T. Guan, F. Liu, X. Wu, R. Xian, Z. Li, X. Liu, X. Wang, L. Chen, F. Huang, Y. Yacoub *et al.*, "HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 14 375–14 385.
- [47] B. Atil, S. Aykent, A. Chittams, L. Fu, R. J. Passonneau, E. Radcliffe, G. R. Rajagopal, A. Sloan, T. Tudrej, F. Ture *et al.*, "Non-determinism of "deterministic" LLM system settings in hosted environments," in *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)*, 2025, pp. 135–148.
- [48] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 824–24 837, 2022.
- [49] A. Lundgard and A. Satyanarayan, "Accessible visualization via natural language descriptions: A four-level model of semantic content," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 1073–1083, 2022.
- [50] H.-K. Ko, H. Jeon, G. Park, D. H. Kim, N. W. Kim, J. Kim, and J. Seo, "Natural language dataset generation framework for visualizations powered by large language models," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 2024, pp. 1–22.
- [51] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu, "G-Eval: NLG evaluation using GPT-4 with better human alignment," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 2511–2522.
- [52] A. McNutt, G. Kindlmann, and M. Correll, "Surfacing visualization mirages," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–16.
- [53] W. Kim, B.-J. Choi, E.-K. Hong, S.-K. Kim, and D. Lee, "A taxonomy of dirty data," *Data Mining and Knowledge Discovery*, vol. 7, pp. 81–99, 2003.
- [54] A. Cockburn, C. Gutwin, and A. Dix, "HARK no more: On the preregistration of CHI experiments," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–12.
- [55] L. Y.-H. Lo, A. Gupta, K. Shigyo, A. Wu, E. Bertini, and H. Qu, "Misinformed by visualization: What do we learn from misinformative visualizations?" in *Computer Graphics Forum*, vol. 41, no. 3. Wiley Online Library, 2022, pp. 515–525.
- [56] M. Correll, M. Li, G. Kindlmann, and C. Scheidegger, "Looks good to me: Visualizations as sanity checks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 830–839, 2019.
- [57] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini, "How deceptive are deceptive visualizations? an empirical analysis of common distortion techniques," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1469–1478.
- [58] D. M. Green and J. A. Swets, *Signal detection theory and psychophysics*. Wiley New York, 1966.
- [59] N. A. Macmillan and C. D. Creelman, *Detection Theory: A User's Guide*, 2nd ed. Lawrence Erlbaum Associates, 2005.
- [60] D. Trippas, M. F. Verde, and S. J. Handley, "Using forced choice to test belief bias in syllogistic reasoning," *Cognition*, vol. 133, no. 3, pp. 586–600, 2014.
- [61] J. Diamond and W. Evans, "The correction for guessing," *Review of Educational Research*, vol. 43, no. 2, pp. 181–191, 1973.
- [62] R. B. Frary, "Formula scoring of multiple-choice tests (correction for guessing)," *Educational Measurement: Issues and Practice*, vol. 7, no. 2, pp. 33–38, 1988.
- [63] R. M. Thorndike, G. K. Cunningham, R. L. Thorndike, and E. P. Hagen, *Measurement and evaluation in psychology and education*. Macmillan Publishing Co, Inc, 1991.
- [64] Z. Allen-Zhu and Y. Li, "Physics of language models: Part 3.3, knowledge capacity scaling laws," *arXiv preprint arXiv:2404.05405*, 2024.
- [65] X. Lu, X. Li, Q. Cheng, K. Ding, X.-J. Huang, and X. Qiu, "Scaling laws for fact memorization of large language models," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 11 263–11 282.



Soohyun Lee is a Ph.D. student at the Human-Computer Interaction Laboratory under the Department of Computer Science and Engineering, Seoul National University, Korea. His research interests include Human-AI Interaction, Graphical Perception, and Data Analysis. Before starting his Ph.D. Program, he received a B.S. degree in Computer Science and Engineering and a B.A. degree in Statistics from Korea University.



Bohyoung Kim received the B.S. and M.S. degrees in computer science and the Ph.D. degree in computer science and engineering from Seoul National University, Seoul, South Korea, in 1995, 1997, and 2001, respectively. She is currently a professor in the Department of Biomedical Engineering, Hankuk University of Foreign Studies, South Korea. Her research interests include computer graphics, visualization, medical imaging, and bio-medical informatics.



Jaeyoung Kim is the Chief Technical Officer at MADi, where he leads the development of AI-driven clinical trial design optimization. His research interests include visual analytics, human-AI interaction, and machine learning for healthcare. He received the B.S. degree in Electrical and Computer Engineering in 2016 from Sungkyunkwan University and the Ph.D. degree in Computer Science and Engineering in 2025 from Seoul National University.



Hyunjo Song is currently an assistant professor at the School of Computer Science and Engineering, Soongsil University, South Korea. His research interests include human-computer interaction, information visualization, visual analytics, eye tracking and health informatics. He received the B.S. degree in computer science and engineering in 2009, the M.S. degree in electrical engineering in 2011, and the Ph.D. degree in electrical engineering and computer science in 2016, all from Seoul National University, Seoul, Korea.



Seokhyeon Park is a Postdoctoral Researcher at Seoul National University, Republic of Korea. His research interests include Human-AI Interaction, Visualization, and Interface Design. He received the B.S. degree in Computer Science and Engineering and the B.A. degree in Information Science and Culture in 2020, and the Ph.D. degree in Computer Science and Engineering in 2026, all from Seoul National University.



Jinwook Seo is a professor in the Department of Computer Science and Engineering, Seoul National University, where he is also the Director of the Human-Computer Interaction Laboratory. His research interests include Human-Computer Interaction, Information Visualization, and Biomedical Informatics. He received his PhD in Computer Science from the University of Maryland at College Park in 2005.



Sihyeon Lee is a Ph.D. Student at the Department of Computer Science and Engineering, Seoul National University, Korea. His research interests include Biomedical Informatics, Precision Health, and Medical LLM Agents. Before starting his Ph.D. program, he received a B.S. degree in AI Convergence from Soongsil University.



Jiwon Song is a Ph.D. student at the Human-Computer Interaction Laboratory under the Department of Computer Science and Engineering, Seoul National University, Korea. Her current research focuses on AI-integrated informal learning for adults. Before starting her Ph.D. program, she received a B.S. degree in Computer Science and Engineering and a B.A. degree in Information Science and Culture from Seoul National University.