# StructViz-Bench: A Controlled Study of Visualization-Format Dependence in Multimodal LLM Reasoning over Structured Data

**Anonymous authors**
Paper under double-blind review

## Abstract

Multimodal LLMs increasingly reason over structured data that reaches them *as a picture*—a bar chart, a heatmap, a node–link diagram—and the rendering is usually chosen incidentally by whatever tool drew the figure. We ask whether that arbitrary choice changes what the model concludes, and introduce StructViz-Bench, which treats visualization format as a *controlled* variable—fixing the underlying data and the question semantics and varying *only* how the data is drawn—and is, to our knowledge, the first to do so *across* structured modalities: tabular, time-series, and graph under one protocol. The benchmark comprises 3,795 base items rendered into 14 visualization types over four reasoning-depth categories. Evaluating several proprietary and open MLLMs, we establish that the choice matters, and by a wide margin: within-modality best–worst exact-match gaps reach 40.5pp, roughly half of all base questions (45–56% per model) change correctness under some format swap, and paired item-level tests are significant for all 12 model$\times$modality comparisons (Bonferroni-corrected $p < 0.01$). The effect is not an artifact of the scoring rule or of layout randomness, and it persists across model families, a $4.6\times$ parameter increase, and all four prompt styles we test—including chain-of-thought, which never helps (4.4–14.6pp below each model's best non-CoT prompt). An answerability audit of every (question, rendering) pair then shows what these gaps are made of: about half of all pairs do not encode the answer at all (a bar chart of column means has no rows; a drawing of a large graph has no node labels), and on those, accuracy sits at prior level for every model. Re-rendering the suite so that every format encodes the answer and re-evaluating an open model, the answer-less renderings recover 16–23pp, the graph listing loses 10pp once it stops printing node degrees, and the rendering effect that remains for the same information is 11pp on tables and 5pp on graphs; none is identified for time series, where no rendering but the complete listing beats a question-only prior, and no graph rendering beats it at all for this model. Printing the queried quantity matters more than the chart type. For practice, much of the accuracy cost is recoverable by choosing the rendering: a learned lookup that picks the best rendering per question type, with no weight update, recovers up to 19.5pp over a random format and 2–8pp over the best fixed format per modality (object-level CIs exclude zero for all seven models; a hand rule that only picks an encoding format captures 0.4–2.7pp of it); and multi-format fine-tuning raises cross-format agreement in every modality (Consistency Rate up 15–25pp) and narrows the accuracy gap in two of three, while a $\lambda$=0 control shows that a consistency regulariser adds agreement but no accuracy, partly by making errors consistent. We release the benchmark, the deterministic generation pipeline, and all evaluation and mitigation code.

## 1 Introduction

Multimodal large language models are increasingly asked to reason over structured data— spreadsheets, sensor traces, relational graphs—and that data rarely arrives as raw numbers. It reaches the model *as an image* of some rendering: a bar chart, a heatmap, a node–link diagram, or a screenshot



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

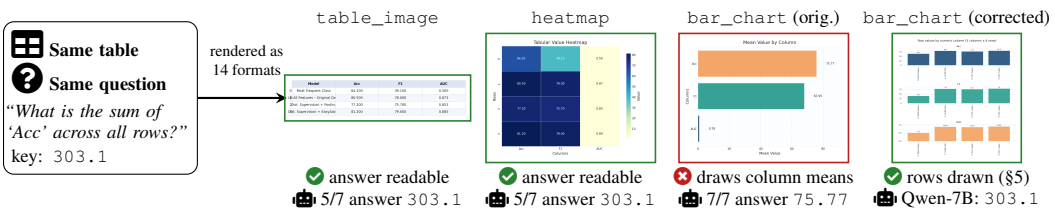


Figure 1: StructViz-Bench fixes the data and the question and varies only the rendering. Shown: one real item, four of its 14 renderings, and the seven evaluated models' answers. Accuracy on identical questions differs by up to 40.5pp across renderings and about half of all questions flip correctness under some format swap. Much of this is what the rendering encodes—the original bar chart draws column means, and every model reports the mean—and re-rendering the rows restores the answer.

of a table. The rendering is usually incidental, decided by whatever tool drew the figure, and it should not change what the model concludes. If it does, both a model's measured competence and its deployed behaviour hinge on a presentation choice that no one intended as a variable. We find that it does, and by a large margin (Figure 1).

We isolate the effect with a controlled design: the data object and the question are held fixed and *only* the rendering varies—not whether models can read charts, but how much the answer depends on a choice that leaves the data untouched. Across tabular, time-series, and graph data and 14 formats, exact-match accuracy on the same questions swings by up to 40.5 points, and roughly half of all base questions flip correctness under some format swap. The rendering axis has been probed one modality at a time (Mukhopadhyay et al., 2024; Deng et al., 2024; Wei et al., 2024; Daswani et al., 2024), never under a common protocol; textual format sensitivity is documented but stays within the text channel (Sclar et al., 2024; Ngweta et al., 2025; Ashury-Tahan et al., 2026); and chart benchmarks (Wang et al., 2024; Masry et al., 2025; 2022) draw each item once, so format dependence stays confounded with question difficulty and chart source. We close this gap and go further, asking *why* the effect arises and whether it can be reduced.

**Contributions.**

1. **A format-controlled benchmark across three structured modalities**: 3,795 base items, 14 formats (non-chart transforms such as GAF, recurrence, and adjacency alongside conventional charts), four reasoning-depth categories, and an audit of what each rendering shows, all released with the pipeline and code (Section 3).
2. **A large, robust, and explained effect.** Gaps of up to 40.5pp, flips on 45–56% of questions per model, and 12/12 model$\times$modality contrasts significant after correction survive layout randomness, model family, a $4.6\times$ parameter increase, and four prompt styles (chain-of-thought costs 4.4–14.6pp). An answerability audit, a question-only prior, a corrected rendering suite, and a precomputation $2\times2$ show that the gaps are dominated by what a rendering encodes and isolate the reading-side dependence that remains: 11pp on tables, 5pp on graphs, none identified for series (Sections 4–5).
3. **The accuracy cost is largely recoverable**: a learned format selector recovers up to 19.5pp over a random format (2–8pp over the best fixed format, object-level CIs excluding zero for all seven models) with no weight update, and multi-format LoRA fine-tuning raises cross-format agreement in every modality, with a $\lambda$=0 control showing that a consistency regulariser adds agreement, not accuracy (Section 6).

## 2 Related Work

**Foundation vision-language models.** Modern MLLMs pair a contrastive or generative vision encoder (Radford et al., 2021) with a large language model (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023). We evaluate a representative mix of open models—Qwen2.5-VL (Bai et al., 2025) and InternVL (Chen et al., 2024b;a)—and proprietary APIs (GPT-4o (OpenAI, 2024), the Gemini family (Gemini Team, Google, 2023; Google DeepMind, 2025; Gemini Team, Google, 2025), and Claude

(Anthropic, 2025)). General multimodal reasoning benchmarks such as MMMU (Yue et al., 2024) and MathVista (Lu et al., 2024) break results down by discipline and skill, but each item appears in exactly one rendering, so how it is drawn is never varied.

**Format sensitivity in the text channel.** LLMs are sensitive to non-semantic prompt formatting: FormatSpread finds accuracy spreads up to tens of points across semantically equivalent prompt formats (Sclar et al., 2024), prompt-style perturbations degrade few-shot accuracy unless the model sees a mixture of formats (Ngweta et al., 2025), and ToRR shows table reasoning swings under JSON/HTML/etc. serializations and structural perturbations (Ashury-Tahan et al., 2026). Every one of these perturbations lives in the *textual* input channel. We study the orthogonal and, for deployed MLLMs, more common axis: the same data reaching the model as a *rendered image*, where the manipulated variable is the picture itself rather than the text. Encoding graphs as text for LLMs (Fatemi et al., 2024) is a further text-channel analogue; we instead render graphs visually.

**Chart and structured-data understanding.** A long line of chart-QA benchmarks— FigureQA (Kahou et al., 2018), DVQA (Kafle et al., 2018), PlotQA (Methani et al., 2020), and ChartQA (Masry et al., 2022)—and more recent, larger or instruction-tuned efforts— ChartBench (Xu et al., 2023), ChartX (Xia et al., 2025), ChartLlama (Han et al., 2023), MMC (Liu et al., 2024), the human-validated CharXiv (Wang et al., 2024), and the more diverse ChartQAPro (Masry et al., 2025)—have advanced chart understanding substantially. Yet they share the design choice that makes format sensitivity impossible to isolate: the data and the question vary *together with* the rendering, so any format effect is confounded with question difficulty and chart source. They also stay within charts, leaving the other structured modalities untouched. None isolates the rendering as a controlled variable, and none spans tabular, time-series, and graph modalities under one protocol. Table 17 (Appendix H) summarises the comparison.

**Format-controlled studies, one modality at a time.** Several studies do hold data and question fixed while changing the rendering, each within one modality: Mukhopadhyay et al. (2024) re-render chart tables into many variants with identical questions; Deng et al. (2024) compare five textual against three image renderings of the same tables; Wei et al. (2024) vary the visual layout of a fixed graph; and Daswani et al. (2024) read the same series as a plot rather than as text; Das et al. (2024) compare text, motif, and image encodings of graphs, Liu et al. (2025) numeric against plotted series, Zhou et al. (2026) probe GAF and recurrence encodings for series captioning, Ho et al. (2026) present the same scientific evidence as a table or as a chart (their tables are our SciTabAlign source), and the contemporaneous Ahsan et al. (2026) aligns tables across text formats and rendered images. All report large swings, so they corroborate our central claim rather than compete with it. What is missing is a *common* protocol under which the three modalities are comparable, non-chart transforms sit alongside conventional charts in every modality, what each rendering actually shows is audited, every gap is checked against a question-only prior (Li et al., 2025), and the cause can be separated from the cure.

**Robustness of MLLMs.** Recent work diagnoses and mitigates modality interference and instruction sensitivity in MLLMs (Cai et al., 2025; Han et al., 2024). We connect to this line by treating visualization format as a controlled perturbation, quantifying its effect, and testing simple mitigations (a learned format selector and multi-format fine-tuning).

In short: (i) text-channel format sensitivity is well documented, and visual-rendering sensitivity has been shown one modality at a time; (ii) no prior study holds data and question fixed across tabular, time-series, and graph data under one protocol; (iii) we isolate the rendering as the manipulated variable across three structured modalities, separate what a rendering encodes from what a model reads, and show the accuracy cost is largely recoverable by choosing the rendering.

## 3 StructViz-Bench Construction

**Design principle.** For each base item we fix the underlying structured object and the question semantics and vary only the visualization format in which the object is rendered. This isolates visualization-format dependence from data difficulty and question phrasing. The pipeline is fully deterministic (global seed 42) with *no language model in the data or answer path*: every ground-truth

answer is computed programmatically from the same object (pandas/NumPy/NetworkX) that is later rendered, so answers are grounded by construction.

**Synthetic data.** We generate 20 datasets per modality from seeded factories: tabular (12 domains, 16–35 rows, mixed categorical/date/numeric columns with plausible clipping); time series (8 pattern families: linear trend, seasonal, trend+seasonal, random walk, step, spike-anomaly, gradual drift, multimodal; length 50–200 with structural annotations); graphs (8 topology families: Erdős–Rényi, Barabási–Albert, Watts–Strogatz, tree, grid, star, stochastic block model, bipartite; ≈24–96 nodes).

**Real-world data.** The same programmatic generators run over three real sources, changing only the data: SciTabAlign, the scientific-paper tables released with Ho et al. (2026); ETT (Zhou et al., 2021), electricity-transformer time series, windowed; and NetworkX graphs (Hagberg et al., 2008)—classic real-world graphs such as Karate and Les Misérables plus generator families—filtered to 15–100 nodes (Appendix M).

**Question–answer generation.** Each modality exposes a fixed inventory of task templates, each paired with a deterministic answer function over the source object (e.g., `df.iloc[0][col]`, `argmax(series)`, `nx.betweenness_centrality`). Guard clauses skip degenerate inputs (empty/short series, $<2$ numeric columns, singleton components). The full inventory is in Appendix A.

**Reasoning-depth categories.** Each template carries an authored (*reasoning_steps*, *requires_arithmetic*, *counterfactual*) triple mapped to 1-/2-/3-hop or counterfactual. We treat these as reasoning-*type* categories rather than a strictly ordinal difficulty ladder, and analyze their empirical calibration in Section 4.

**Rendering.** Data are rendered at $1024 \times 768$, 150 DPI into 14 formats —tabular: {bar chart, heatmap, table image, scatter plot, text-only}; time series: {line plot, GAF, recurrence plot, heatmap, text-only}; graph: {node-link, adjacency matrix, circular layout, text-only}. Layouts are deterministic (e.g., `spring_layout(seed=42)`), so all models receive pixel-identical inputs. Crucially, `text_only` is itself a *rendered image* of the data serialized as text; every condition—including `text_only`—is delivered through the vision channel, so only the visual format varies. Appendix C shows one object per modality in every format.

**Cross-modal pilot level.** Beyond single-object items, we include a small cross-modal pilot that places two objects of *different* modalities side by side in one image (table+time-series, table+graph, time-series+graph) and asks a question that cannot be answered from either panel alone (cross-modal aggregation, alignment, comparison, threshold, and counterfactual tasks). It is deliberately narrow: one table, one series, and one graph supply all three pairings, and 30 unique question–answer problems (10 per pairing) are each queried 15–25 times under distinct item ids (600 ids in total), so the level measures response variability on 30 problems rather than 600 independent questions. Each problem is rendered in five paired format combinations with the objects and the question held fixed. We report it in Section 4 as a pilot only.

**Scale.** The benchmark has 3,795 base items (1,765 tabular, 1,370 time-series, 660 graph); rendered across each modality's formats this yields 18,315 rendered instances per model.

**Quality control and honesty.** Ground truth is correct by construction; we additionally run a two-rater human audit of image-level answerability (Appendix F; it largely agrees with the code audit of Section 5), flag the ten (modality, task) cells whose answer is near-constant ($\geq$90% one answer; 556 questions; Appendix A), and report the format gaps with them excluded and against a question-only prior baseline (Section 4). All randomness derives from explicit seeds, so the corpus and metrics reproduce from code.

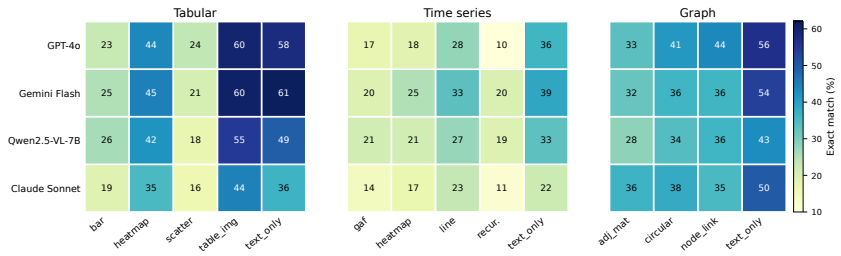


Figure 2: Exact match (%) by format and modality, four representative models (numbers in Table 1; Gemini-2.5, Qwen-32B, and InternVL appear in Tables 3 and 20). For the *same* questions, accuracy spans the low teens to the high fifties on the rendering alone; text-faithful formats and `line_plot` read best, `scatter_plot`, `recurrence_plot`, and `gaf` worst—largely because they do not encode what the questions ask (Section 5).

## 4 Format Sensitivity: Results

We evaluate GPT-4o, Gemini Flash, Gemini-2.5, Claude Sonnet, Qwen2.5-VL-7B/32B, and InternVL2.5-8B (18,315 rendered instances each; exact model identifiers and setup in Appendix G). Overall exact match ranges 20.3–36.2%. Within-modality best–worst gaps reach 40.5pp (tabular), and per-question flip rate is 45.2–55.9% for the four representative models of Table 1 (34.3–55.9% across all seven). Paired permutation tests (10,000 iterations) put every one of the 12 model×modality best–worst contrasts of those four models at the estimator's floor, raw $p = 1/10001$, i.e. Bonferroni-corrected $p = 0.0012 < 0.01$ (Cohen's $d$ 0.28–0.71). The extremes are selected on the tested sample, so these are post-hoc contrasts; selecting the pair on one random half of the *source objects* and testing it on the other keeps all 12 contrasts at corrected $p < 0.01$ in both directions (object-level sign-flip tests). Questions cluster within 299 source objects (12.7 per object on average), so all CIs and permutations in this paper resample or flip *objects*: the 12 gaps have 95% CI half-widths of 1.6–4.4pp and remain at the permutation floor under Holm correction.

**Against a question-only prior.** Following Li et al. (2025), a predictor that never sees the image and answers each (modality, task) with its most frequent training answer (objects split two-fold, 20 seeds) scores 24.0% on tables, 40.8% on series, and 40.9% on graphs (33.0% overall). Tables clear it by 20–37pp in the best format; graphs only through the text listing (+2–16pp; the drawings sit at or below it for three of four models); and *no model beats it on time series in any format* (best 38.7%), so the time-series rows of Table 1 measure how a rendering shifts a model's response bias against a skewed answer distribution, not how well it reads the series. (The same open model with the image withheld scores 19.1/19.9/27.7%, below this prior, so the prior is the stricter reference; Section 5(iv).) Ten (modality, task) cells (556 questions; Appendix A) have one answer covering ≥90% of questions; excluding them leaves every gap significant (tabular 31.6–41.6, series 14.0–22.0, graph 16.0–22.9pp) and flip rates at 44.8–57.0%.

**Where the effect lives.** The gap is largest for tabular data (best–worst 28–40.5pp across models) and smaller but significant for time series and graphs. It holds for synthetic and real-world sources alike, and is larger on the real-world objects in 11 of 12 model×modality cells: on SciTabAlign tables the four core models' gaps are 35.6–47.1pp against 10.2–29.0pp on the synthetic tables, and on the NetworkX graphs 21.2–35.6pp against 12.8–19.8pp (per-task and per-source tables in Appendix E). Excluding `text_only`, the gap among purely graphical formats is 28.4–39.7pp for tables and 8.1–17.4pp for series (all $p < 10^{-4}$) but only 3.0–10.3pp for graphs (significant for two models): the graph effect is carried largely by the text listing, which prints node degrees (Section 5). The difficulty "levels" are non-monotonic in accuracy (answer-space size dominates), so we treat them as reasoning types.

**Practical takeaway.** For a fixed model the rendering choice can be worth as much as a model upgrade (Claude Sonnet, `scatter_plot`→`table_image`: +28.4pp). The rule this supports is about *encoding*: prefer a rendering that shows the queried quantity and avoid one that cannot; Section 6 turns it into a per-task selector.

Table 1: Exact match (%) by visualization format, per modality (four representative models; the selector and ensemble tables report all seven). The within-row spread is the format-sensitivity gap; the best and worst format differ by model.

| *Tabular* Model | bar_chart | heatmap | scatter_plot | table_image | text_only |
|---|---|---|---|---|---|
| GPT-4o | 22.8 | 43.6 | 23.6 | 59.5 | 57.6 |
| Gemini Flash | 25.0 | 44.7 | 20.6 | 60.3 | 61.1 |
| Qwen2.5-VL-7B | 25.9 | 41.7 | 17.5 | 54.6 | 49.0 |
| Claude Sonnet | 18.9 | 34.8 | 16.1 | 44.5 | 36.5 |
| *Time series* Model | gaf | heatmap | line_plot | recurrence_plot | text_only |
| GPT-4o | 17.2 | 18.5 | 27.6 | 10.1 | 35.8 |
| Gemini Flash | 19.5 | 24.9 | 33.0 | 19.7 | 38.7 |
| Qwen2.5-VL-7B | 21.2 | 21.0 | 27.3 | 19.2 | 32.8 |
| Claude Sonnet | 14.1 | 17.0 | 23.3 | 11.3 | 22.0 |
| *Graph* (adjacency_matrix / circular_layout / node_link / text_only) | | | | | |
| GPT-4o | 33.3 | 41.2 | 43.6 | 56.4 | |
| Gemini Flash | 31.8 | 36.2 | 35.8 | 53.5 | |
| Qwen2.5-VL-7B | 28.2 | 33.9 | 36.2 | 43.0 | |
| Claude Sonnet | 36.5 | 37.9 | 34.8 | 49.5 | |

**Not a scoring artifact.** Gaps persist under a tolerance-based numeric metric (abs. 0.05 / rel. 0.01) and under categorical-answer normalisation (Appendix J); token-F1 (Appendix B) mirrors EM but normalises text, not numbers, so we do not lean on it.

**Not prompt-driven; not scale-solved.** Four prompt styles (*minimal*, *concise*, *detailed*, chain-of-thought; CoT, Wei et al., 2022) on a stratified 500-question subsample (Table 18) move overall accuracy by 4.4–16.1pp, yet the format gap survives all 16 model×style cells (tabular 15.5–36.9pp, for three models larger than the model's entire prompt spread; `scatter_plot` is last in 14 of the 16). Nor does scale: Qwen 7B→32B ($4.6\times$) raises overall EM by $+0.9$pp (object CI $[+0.1, +1.7]$) and moves the tabular gap 37.1→34.7pp (n.s.; CI $[-4.7, +0.1]$).

**Chain-of-thought does not help.** Under the single CoT template we tested it lands 4.4–14.6pp *below* each model's best non-CoT prompt (mean $-10.7$pp; object-cluster CIs exclude zero; Table 18, Appendix I), and where it narrows the format gap it levels *down* (Gemini's tabular gap 36.9→15.5pp because `table_image` collapses 55.8→34.8). Verbalised reasoning over a misread chart compounds the misreading, which is why Section 6 pursues input-side remedies.

**Cross-modal pilot (L2).** The benchmark also contains a cross-modal level in which *two* objects of different modalities are rendered side by side and the question requires both (e.g. "what is |table mean − series mean|?"). We report it as a pilot, not a result: it is built from a single (table, series, graph) triple and 30 unique question–answer problems, each queried 15–25 times under distinct item ids with an identical image (Section 3), so its 600 ids measure response variability on 30 problems and support no significance claim. Two observations are still worth recording. Models diverge sharply on it (overall exact match 1.3–55.8%; Claude Sonnet returns well-formed numbers read from one panel instead of combining the two), and the best paired rendering is not predictable from the single-object ranking: GPT-4o is read best with `text_only+text_only`, but Gemini's best table+series pairing is `scatter_plot+recurrence_plot`, the two formats that are *worst* in Table 1. Rebuilding this level from many independent object pairs, with a question-only baseline, is the obvious next step and we do not build on it here.

## 5 Why: What the Rendering Encodes vs. What the Model Reads

We separate two explanations with an audit of what each rendering actually encodes, not with intuition. Appendix D walks through three items format by format.

**Answerability audit.** For every (question, format) pair we classify, from the renderer code and the question, whether the answer is *visible* (V), *derivable* (D: all needed information is drawn but must be counted or computed), *absent* (A: the rendering does not encode it or truncates it away), or *constant* (C: the answer does not depend on the image). The rules, counts, and per-model table are in Appendix J; a released script applies them. The result reframes Table 1: 57% of tabular, 51% of time-series, and 47% of graph pairs are absent *by design*, because the tabular `bar_chart` draws only column means, `scatter_plot` only two columns, `table_image`/`text_only` only the first 12/14 rows, the graph drawings label nodes only below 30 nodes, and the time-series transforms discard values.

**(i) Absent answers explain the low end.** On absent pairs accuracy sits at prior level and model identity stops mattering: tabular `bar_chart` 5–9%, `heatmap` 3–8%, graph drawings 10–23% for every model (Table 19). (`value_extraction` drawn as a `bar_chart`: 1.7% for all four models, because a bar chart of column means contains no row.) The "worst" formats of Table 1 are worst because they do not encode what most questions ask.

**(ii) Present answers reorder the formats.** Where the answer is present, `bar_chart` becomes the *best* tabular rendering (82–85% for three models on the 401 pairs it can answer), `heatmap` reaches 61–76%, and the graph text listing 57–78% against 29–47% for the drawings; only `gaf` and `recurrence_plot` stay at 4–24% even where they could answer, below the 41% prior.

**(iii) On the original suite** a same-question contrast between two complete renderings exists only for tables: `table_image` vs. `text_only` on the 1,057 fully drawn tables differs by $+1.8$/$+1.9$pp for GPT-4o/Gemini (n.s.) and $+9.8$/$+12.9$pp for Qwen/Claude ($p < 10^{-4}$). Within one source (SciTabAlign) the tabular gap is 38–50pp on fully drawn tables and 24–39pp on truncated ones, so truncation lowers the best format rather than creating the gap; the graph listing's lead is concentrated in `degree_query` ($+31$–47pp there, $-4.9$ to $+2.2$pp elsewhere). (Attention mass on image tokens does not track accuracy, ruling out only the "not looking" story.)

**(iv) Corrected suite: what survives when every rendering encodes the answer.** We re-rendered the benchmark so that the absent category nearly disappears—row-level bars with labels and values, all rows in every table view, every node labelled, complete edge and point listings (all images $\leq 1024\times1024$; Appendix J)—and re-evaluated Qwen2.5-VL-7B on all 18,315 pairs plus a graph listing *without* degrees (Table 2). We now define *answerable* strictly, as readable to exact-match precision, so a labelled scatter plot without value annotations stays absent for value extraction, and we exclude the ten near-constant cells. Under this definition absent pairs fall from 57/51/47% to 23/42/0% (tables/series/graphs; the lossy series transforms are kept on purpose). First, the renderings that withheld the answer recover most: row-level bars $+22.9$pp, labelled scatter $+16.1$pp, the complete series listing $+10.3$pp (it was already the best series format), table and text views $+3.3$ to $+5.7$pp; the tabular gap falls from 37.1 to 24.2pp and its flip rate from 60.1 to 50.8%. Second, the graph listing loses 9.8pp once it stops printing degrees ($+10.6$pp on `degree_query` when they are added back, $+0.4$ elsewhere) and the graph gap falls from 14.8 to 7.9pp: the degree-annotation effect of (ii) is causal. Third, a rendering effect *remains* for the same model reading the same information, but it is smaller than the headline gaps: 11.3pp on the 186 tabular questions answerable in every format (ranking, outlier, correlation; `text_only` vs. `bar_chart`, CI $[2.7, 19.9]$, $p$=0.02; 12.8pp $[9.5, 16.2]$ over all non-constant tabular questions without value extraction), and 4.9pp on the 528 graph questions ($[1.7, 8.1]$, $p$=0.005). For series no such question exists: the only series questions answerable in every format are the two near-constant cells, where the 38pp spread is a yes/no response bias, and among the three non-lossy renderings on non-constant questions the gap is 21.4pp $[18.4, 24.5]$. Fourth, against the question-only prior of Section 4, every tabular rendering now clears it ($+9.7$ to $+33.9$pp), only the complete listing does for series ($+2.3$pp $[0.7, 4.0]$), and *no* graph rendering does (best `node_link` $-1.9$pp $[-4.4, +0.5]$): for this model, a fully labelled graph drawing is not worth more than the task's majority answer (with no image at all the same model scores 19.1/19.9/27.7%). Giving it the *raw data as text* instead—CSV, point list, edge list—scores 52.4/37.5/41.4%: within 5.5pp of the best image for tables and series (the rendered listing reads better than the token stream) and above every graph drawing, at 0.5–1.5 s per query against 2.8–4.4 s for an image; the deterministic tool route is 100% by construction. For a user who holds the data, an image is justified only where it is the cheaper channel or adds a precomputed quantity. Finally, a $2\times2$ on the 255 mean-comparison questions separates rendering from precomputation: with every row drawn in both, the full table and the row-level bar chart differ by $-5.5$pp ($[-10.8, 0.0]$, $p$=0.06); printing the column means raises both by $+14.9$ and $+12.2$pp, driven by the 31 mean-aggregation questions

Table 2: Corrected rendering suite (v2), Qwen2.5-VL-7B. "absent" is the share of (question, format) pairs whose answer the image does not encode to exact-match precision; "gap" the best–worst exact-match gap (pp) over all questions; "residual" the gap on the questions answerable in every format outside near-constant cells (n; object-cluster 95% CI); "vs. prior" the best v2 format minus the question-only prior (CI); "raw text" the same model given the data as text, no image.

| | absent pairs (%) | | gap, all questions | | residual (n; CI) | best vs. prior | raw text |
|---|---|---|---|---|---|---|---|
| Modality | v1 | v2 | v1 | v2 | v2 | v2 | EM (%) |
| Tabular | 56.8 | 23.2 | 37.1 | 24.2 | 11.3 (186; [2.7, 19.9]) | +33.9 [31.1, 36.8] | 52.4 |
| Time series | 51.4 | 41.7 | 13.6 | 24.0 | none identified (0) | +2.3 [0.7, 4.0] | 37.5 |
| Graph | 47.0 | 0.0 | 14.8 | 7.9 | 4.9 (528; [1.7, 8.1]) | −1.9 [−4.4, 0.5] | 41.4 |

where the printed mean *is* the answer (19→97%) and only +6.2/+3.6pp on the 224 comparisons; the interaction is +2.7pp $[-2.0, +7.6]$, so we cannot rule one out. Once both renderings encode the rows, chart type matters little; the v1 chart-type gaps were encoding, and precomputation matters where it prints the queried quantity. A second open model at its released 448-px single-tile input (InternVL2.5-8B, Appendix J) cannot read the denser v2 images—its table image falls from 33.9 to 7.3% on the same questions—yet replicates the degree effect (−9.3pp without degrees) and shows a tabular residual of 6.5pp $[2.5, 10.9]$ and none on graphs (+1.7pp $[-1.6, 4.9]$): the corrected suite also exposes input resolution as a first-order variable.

## 6 MITIGATION

**Format ensemble: headroom, not a remedy.** A majority vote over a question's renderings (Table 20, Appendix K) beats an arbitrary single format (+0.2 to +7.0pp) but *not* the best fixed format per modality (−3.4 to −13.8pp, object-cluster CIs below zero) at $K\times$ the cost; we keep it only for its oracle bound (correct if any format is, 40–67%), the headroom a *selector* can target (cf. Mixture-of-Formats, Ngweta et al., 2025). The Consistency Rate (mean share of agreeing format pairs) is 26–34%.

**Learned format selector.** The oracle headroom suggests a cheap remedy: pick the format the model reads best for the *kind* of question at hand. We learn a lookup table of the best rendering per (modality, task) on a training split and apply it to held-out questions. The split is by *source object* (299 tables/series/graphs, 12.7 questions each on average, at most 25), so no held-out question shares its data with training (Table 3). Against the best *fixed* format per modality, task-aware selection adds +2.1 to +8.1pp (+10.6 to +19.5pp over a random format) with no weight update, with object-cluster 95% CIs excluding zero for all seven models; a hand rule that only picks an encoding format (bars for mean comparisons, table image otherwise, text listing for series and graphs) captures +0.4 to +2.7pp of this, and the learned lookup adds +0.7 to +5.4pp on top (CIs exclude zero for six of seven models), recovering about a third of the fixed-to-oracle gap (GPT-4o 50.6→54.6, oracle 63.7). The gain is not uniform: in 1–5 of the 45 (modality, task) cells per model the selector loses to the fixed format—worst −46pp on a 13-question time-series cell, −10pp on an 80-question cell (Gemini, tabular counterfactual); a fixed-format fallback for small cells (not applied) is the obvious safeguard. A Naive-Bayes task classifier is 100% accurate on these templated questions; on free-form inputs it becomes the bottleneck, so we read the gains as a *ceiling*.

**Multi-format fine-tuning, with and without a consistency term.** Building on perturbation-based augmentation with output-level consistency regularization for modality robustness (Cai et al., 2025), we LoRA-tune Qwen2.5-VL-7B (47.6M trainable params, vision tower frozen) jointly on all three modalities and evaluate on a held-out manifest of 554 questions drawn from the same 44 source objects as training (new questions on seen objects; Appendix L). Accuracy uses the tolerance-based numeric metric of Section 4 with no substring credit, and every paired 95% CI is a cluster bootstrap over source objects ($B$=5000). To identify what the consistency term contributes, we also train a $\lambda$=0 control: the same multi-format pairs, schedule, and seed with the KL term removed, i.e. plain multi-format supervised fine-tuning. Plain multi-format SFT does most of the work (Table 4): it raises the Consistency Rate in every modality, mean accuracy by +5 to +22pp, and narrows the best–worst

Table 3: Learned format selector, split by *source object* (148 train / 151 held-out objects; 1,902 held-out questions). Exact match (%) on held-out questions. "random" is the expected accuracy of an arbitrary format; "fixed (mod)" is the best single format per modality, learned on train; "selector" additionally conditions on task. $\Delta$ is the gain over the fixed-per-modality baseline with its 95% object-cluster bootstrap CI (151 held-out objects, $B$=5000).

| Model | random | fixed (mod) | **selector (mod,task)** | $\Delta$ | 95% CI | oracle |
|---|---|---|---|---|---|---|
| GPT-4o | 35.1 | 50.6 | **54.6** | +4.0 | [+2.6, +5.5] | 63.7 |
| Gemini Flash | 37.1 | 52.7 | **56.3** | +3.6 | [+1.9, +5.3] | 68.2 |
| Gemini-2.5 | 36.3 | 51.8 | **53.8** | +2.1 | [+0.9, +3.2] | 66.6 |
| Qwen2.5-VL-7B | 33.4 | 45.0 | **49.0** | +3.9 | [+2.5, +5.3] | 58.3 |
| Claude Sonnet | 26.6 | 36.3 | **44.4** | +8.1 | [+6.5, +9.6] | 52.5 |
| Qwen2.5-VL-32B | 34.3 | 48.5 | **51.5** | +2.9 | [+1.4, +4.4] | 61.2 |
| InternVL2.5-8B | 20.5 | 25.3 | **31.1** | +5.8 | [+4.5, +7.0] | 40.9 |

Table 4: LoRA fine-tuning of Qwen2.5-VL-7B on multi-format pairs: base $\rightarrow$ multi-format SFT ($\lambda$=0) $\rightarrow$ SFT with the consistency term ($\lambda$=1); identical data, schedule, and seed. Held-out questions from seen objects; tolerance-based numeric accuracy; cluster bootstrap over source objects, $B$=5000. "Same wrong" is the share of questions answered incorrectly *and identically* in every format. Bold marks a base$\rightarrow \lambda_0$ change whose CI excludes zero; † marks a $\lambda_0 \rightarrow \lambda_1$ change whose CI excludes zero.

| Modality | $n$ | mean acc. (%) | gap (pp) | CR (%) | same wrong (%) |
|---|---|---|---|---|---|
| Tabular | 236 | $22.4 \rightarrow \mathbf{44.6} \rightarrow 42.3^{\dagger}$ | $16.1 \rightarrow \mathbf{6.4} \rightarrow 6.8$ | $30.8 \rightarrow \mathbf{51.8} \rightarrow 55.5^{\dagger}$ | $7.6 \rightarrow 5.1 \rightarrow 8.5^{\dagger}$ |
| Time series | 149 | $36.0 \rightarrow 40.9 \rightarrow 41.2$ | $19.5 \rightarrow 14.8 \rightarrow 15.4$ | $44.6 \rightarrow \mathbf{59.2} \rightarrow 59.9$ | $6.0 \rightarrow \mathbf{17.4} \rightarrow 14.8$ |
| Graph | 169 | $37.9 \rightarrow \mathbf{48.4} \rightarrow 48.8$ | $18.9 \rightarrow \mathbf{10.7} \rightarrow 10.7$ | $39.5 \rightarrow \mathbf{52.9} \rightarrow 64.4^{\dagger}$ | $1.8 \rightarrow 1.8 \rightarrow 12.4^{\dagger}$ |

gap for tabular (16.1 $\rightarrow$ 6.4pp, CI $[-15.6, -1.9]$) and graph (18.9 $\rightarrow$ 10.7pp, $[-16.7, -0.6]$; not under strict scoring) but not for series ($[-12.7, +3.4]$). The consistency term adds agreement (tabular +3.6, graph +11.5pp) but not accuracy: it narrows no gap further, lowers tabular mean accuracy by 2.3pp ($[-3.5, -1.2]$), and its extra agreement is largely agreement on *errors*—the share of questions answered wrong *and identically* in every format rises from 5.1 to 8.5% (tabular) and from 1.8 to 12.4% (graph): agreement is not robustness. Scope: one open model at reduced resolution, one run per setting, seen objects; not comparable to Table 1.

## 7 Discussion and Limitations

Programmatic ground truth guarantees correctness but limits question surface diversity; a two-rater human audit largely agrees with the code audit ($\kappa$=0.40; its pre-registered legibility bar is not met; Appendix F), and ten near-constant (modality, task) cells (556 questions) are flagged and excluded in a parallel analysis rather than removed. Pre-training frequency of a rendering is characterized rather than controlled; common formats such as bar and scatter charts are among the worst in Table 1, but Section 5 shows that is mostly encoding, not rarity. The answerability audit is the main limitation of the released suite: about half of all (question, rendering) pairs do not encode the answer, so the headline gaps measure rendering design as much as model reading; the corrected suite is the one to evaluate against from now on, and the proprietary models' corrected-suite numbers are the first thing to add. Proprietary results date from March 2026 (Appendix G); Gemini-2.5, the newest, is at least as format-sensitive as GPT-4o (tabular gap 38.5 vs. 36.7pp, n.s.) while more accurate, and across seven models the gap tracks neither accuracy nor recency. Exact match is deliberately strict: free-form answers are never credited, so a few task rows in Appendix E carry little format signal, and 16 of the 25 time-series tasks have at most 20 questions, so their per-task rows are noisy even though the pooled gaps are not. The mitigations are scoped in Section 6: the selector needs the task label, and the fine-tuning result is one open model at reduced resolution on seen objects. Two follow-ups are queued: the proprietary models on the corrected suite, and a re-rating of the corrected renderings by the same two raters (package released). Evaluations of MLLMs on structured data should report *format-conditioned* accuracy, not one aggregate, and should say what each rendering encodes before attributing a gap to the model.

## REPRODUCIBILITY STATEMENT

We release the benchmark, the deterministic generation pipeline, and all evaluation and mitigation code as an anonymized repository: `https://anonymous.4open.science/r/StructViz-Bench-15E1/`. The construction and rendering procedure is described in Section 3, with the complete task inventory, answer functions, and the exact evaluation prompt in Appendix A; the pipeline is deterministic (global seed 42, no model in the data or answer path). Exact model identifiers, including pinned HuggingFace commit revisions for the open-weight models, are in Appendix G. Our statistical procedures (paired permutation tests with Bonferroni correction, and cluster and paired bootstraps) are stated in Sections 4 and 6, and the mitigation experiments report sample sizes and 95% confidence intervals (Table 4). The human-validation protocol is in Appendix F, and a claim-to-artifact map is in Appendix N.

## ETHICS STATEMENT

StructViz-Bench is built from synthetic data produced by our released code and from publicly available datasets (SciTabAlign, ETT, and standard NetworkX graphs) used under their licenses; it contains no personal or sensitive information. The only human involvement is a small answer-legibility audit (Appendix F) by volunteer annotators who give informed consent. We assessed the audit against our institution's human-subjects requirements: it collects no personal or demographic data, records a three-way judgement about a rendered image with an optional own answer and a short note, poses no foreseeable risk, and participation is voluntary and uncompensated. Our findings concern the reliability of MLLMs on structured-data reasoning and are intended to improve evaluation and robustness; we identify no direct dual-use risk. We report negative and non-significant results (the attention probe in Section 5, the time-series gap and the consistency regulariser in Section 6) transparently.

## AI USE STATEMENT

Large language models were used, under author direction, for the following. *Implementing methods*: as coding assistants for the benchmark, evaluation, and mitigation scripts. *Methodology feedback and hypothesis refinement*: to propose and critique analyses (for example the per-task breakdown in Section 5 and the cross-model comparison in Section 7); the research questions and experimental design are the authors'. *Data analysis and interpretation of results*: to write and run the aggregation and statistical scripts behind the reported numbers, and to draft first-pass readings of them; the authors checked every number against the released predictions, and a released script (`scripts/verify_paper_numbers.py`) recomputes each headline figure from the released data. That check caught and corrected an erroneous claim during preparation. *Sourcing information and identifying literature*: to find related work and to verify every bibliography entry against primary sources. *Drafting and editing the paper*: to draft and revise manuscript text, which the authors then edited.

Not used: generating synthetic datasets (the benchmark pipeline is fully programmatic, Section 3; no model output is used as data or as a label), developing theoretical frameworks, formulating or proving mathematical claims, and cleaning or reformatting the released data. The authors take full responsibility for all content and claims.

## REFERENCES

Momina Ahsan, Sarfraz Ahmad, Ming Shan Hee, Roy Ka-Wei Lee, and Preslav Nakov. TABVERSE: Benchmarking cross-format table understanding in LLMs and VLMs. *arXiv preprint arXiv:2606.09578*, 2026.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language

model for few-shot learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, volume 35, pp. 23716–23736. Curran Associates, Inc., 2022. doi: 10.52202/068431-1723. arXiv:2204.14198v2.

Anthropic. Introducing Claude 4. https://www.anthropic.com/news/claude-4, May 2025. Accessed September 2026.

Shir Ashury-Tahan, Yifan Mai, Rajmohan C, Ariel Gera, Yotam Perlitz, Asaf Yehudai, Elron Bandel, Leshem Choshen, Eyal Shnarch, Percy Liang, and Michal Shmueli-Scheuer. The mighty ToRR: A benchmark for table reasoning and robustness in LLMs. In *Proceedings of the First Workshop on Structured Understanding, Retrieval, and Generation in the LLM Era (SURGeLLM 2026)*, pp. 16–51, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.surgellm-1.2. URL https://aclanthology.org/2026.surgellm-1.2/. arXiv:2502.19412v3.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025. v1, February 2025; Qwen Team, Alibaba Group.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models. *arXiv preprint arXiv:2505.19616*, 2025. v4, January 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a. v5, September 2025; technical report for InternVL 2.5.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2024b. Oral; arXiv:2312.14238v3.

Debarati Das, Ishaan Gupta, Jaideep Srivastava, and Dongyeop Kang. Which modality should I use – text, motif, or image? : Understanding graphs with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 503–519. Association for Computational Linguistics, 2024. URL https://aclanthology.org/2024.findings-naacl.34/.

Mayank Daswani, Mathias M. J. Bellaiche, Marc Wilson, Desislav Ivanov, Mikhail Papkov, Eva Schnider, Jing Tang, Kay Lamerigts, Gabriela Botea, Michael A. Sanchez, Yojan Patel, Shruthi Prabhakara, Shravya Shetty, and Umesh Telang. Plots unlock time-series understanding in multimodal models. *arXiv preprint arXiv:2410.02637*, 2024. v2, November 2024.

Naihao Deng, Zhenjie Sun, Ruiqi He, Aman Sikka, Yulong Chen, Lin Ma, Yue Zhang, and Rada Mihalcea. Tables as texts or images: Evaluating the table reasoning ability of LLMs and MLLMs. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024. arXiv:2402.12424.

Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=IuXR1CCrSi. arXiv:2310.04560.

Gemini Team, Google. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. v5, May 2025.

Gemini Team, Google. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. v6, December 2025; Gemini Team, Google.

Google DeepMind. Gemini 2.0 is now available to everyone. https://blog.google/technology/google-deepmind/gemini-model-updates-february-2025/, February 2025. Accessed September 2026.

Aric A. Hagberg, Daniel A. Schult, and Pieter J. Swart. Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference (SciPy 2008)*, pp. 11–15, 2008.

Wei Han, Hui Chen, and Soujanya Poria. Towards robust instruction tuning on multimodal large language models. *arXiv preprint arXiv:2402.14492*, 2024. v2, June 2024.

Yucheng Han, Chi Zhang, Xin Chen, Xu Yang, Zhibin Wang, Gang Yu, Bin Fu, and Hanwang Zhang. ChartLlama: A multimodal LLM for chart understanding and generation. *arXiv preprint arXiv:2311.16483*, 2023. v1, November 2023.

Xanh Ho, Yun-Ang Wu, Sunisth Kumar, Florian Boudin, Atsuhiro Takasu, and Akiko Aizawa. Format matters: The robustness of multimodal LLMs in reviewing evidence from tables and charts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2511.10075.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. DVQA: Understanding data visualizations via question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5648–5656, 2018. doi: 10.1109/CVPR.2018.00592. arXiv:1801.08163v2.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. FigureQA: An annotated figure dataset for visual reasoning. In *ICLR 2018 Workshop Track*, 2018. URL https://openreview.net/forum?id=H1mz0OyDz. arXiv:1710.07300v2.

Donald E. Knuth. *The Stanford GraphBase: A Platform for Combinatorial Computing*. ACM Press / Addison-Wesley, 1993.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023. arXiv:2301.12597v3.

Zhimin Li, Haichao Miao, Xinyuan Yan, Valerio Pascucci, Matthew Berger, and Shusen Liu. See or recall: A sanity check for the role of vision in solving visualization question answer tasks with multimodal LLMs. *arXiv preprint arXiv:2504.09809*, 2025.

Fuxiao Liu, Xiaoyang Wang, Wenlin Yao, Jianshu Chen, Kaiqiang Song, Sangwoo Cho, Yaser Yacoob, and Dong Yu. MMC: Advancing multimodal chart understanding with large-scale instruction tuning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1287–1310, Mexico City, Mexico, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.70. arXiv:2311.10774v2.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, volume 36, pp. 34892–34916. Curran Associates, Inc., 2023. doi: 10.52202/075280-1516. Oral; arXiv:2304.08485v2.

Haoxin Liu, Chenghao Liu, and B. Aditya Prakash. A picture is worth a thousand numbers: Enabling LLMs reason about time series via visualization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7486–7518. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.naacl-long.383/.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*, 2024. Oral; arXiv:2310.02255v3.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. arXiv:2203.10244.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, Megh Thakkar, Md Rizwan Parvez, Enamul Hoque, and Shafiq Joty. ChartQAPro: A more diverse and challenging benchmark for chart question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-acl.978. arXiv:2504.05506v2.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. PlotQA: Reasoning over scientific plots. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1516–1525, 2020. doi: 10.1109/WACV45572.2020.9093523. arXiv:1909.00997v3.

Srija Mukhopadhyay, Adnan Qidwai, Aparna Garimella, Pritika Ramu, Vivek Gupta, and Dan Roth. Unraveling the truth: Do VLMs really understand charts? a deep dive into consistency and robustness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16696–16717. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.973. arXiv:2407.11229.

Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. Towards LLMs robustness to changes in prompt format styles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 529–537, Albuquerque, USA, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-srw.51. URL https://aclanthology.org/2025.naacl-srw.51/. arXiv:2504.06969v1.

OpenAI. GPT-4o system card. Technical report, OpenAI, August 2024. URL https://openai.com/index/gpt-4o-system-card/. Also available as arXiv:2410.21276 (v1, October 2024).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. arXiv:2103.00020.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=RIu5lyNXjT. arXiv:2310.11324v2.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track*, volume 37, pp. 113569–113697. Curran Associates, Inc., 2024. doi: 10.52202/079017-3609. arXiv:2406.18521.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022. arXiv:2201.11903v6.

Yanbin Wei, Shuai Fu, Weisen Jiang, Zejian Zhang, Zhixiong Zeng, Qi Wu, James T. Kwok, and Yu Zhang. GITA: Graph to visual and textual integration for vision-language graph reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2402.02130.

Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. ChartX and ChartVLM: A versatile benchmark and foundation model for complicated chart reasoning. *IEEE Transactions on Image Processing*, 34:7436–7447, 2025. doi: 10.1109/TIP.2025.3607618. arXiv:2402.12185.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. ChartBench: A benchmark for complex visual reasoning in charts. *arXiv preprint arXiv:2312.15915*, 2023. v3, June 2024.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567, 2024. Oral; arXiv:2311.16502v4.

Wayne W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4):452–473, 1977.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11106–11115, 2021.

Luca Zhou, Pratham Yashwante, Marshall Fisher, Alessio Sampieri, Zihao Zhou, Fabio Galasso, and Rose Yu. CaTS-bench: Can language models describe time series? In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34479–34519. Association for Computational Linguistics, 2026. URL https://aclanthology.org/2026.findings-acl.1722/.

## A TASK INVENTORY AND ANSWER FUNCTIONS

**Evaluation prompt.** All models receive the rendered image, the question, and a system prompt with answer-format rules (final value only; yes/no; plain number; name), released with the code; InternVL used the shorter "Answer concisely. Give only the final answer." (Appendix G). Predictions are normalized (numeric equivalence `45.0≡45`; case/whitespace) before scoring exact match, token-level F1, and numeric accuracy (absolute tol. 0.05, relative 0.01).

Table 5: Tabular task templates and answer functions (representative).

| task | question skeleton | answer |
|---|---|---|
| value_extraction | value of 'col' in row 0 | `str(df.iloc[0][col])` |
| comparison | which column has higher average | `a if mean_a>mean_b else b` |
| aggregation | sum/avg/median of 'col' | `round(series.op(),2)` |
| filtering | #rows with col $>$ threshold | `int((series>q).sum())` |
| ranking | which label has highest/lowest value | stable-sort argmax/argmin |
| outlier_detection | entry deviating most from mean | `argmax\|series-mean\|` |
| correlation | positive/negative correlation | sign of Pearson $r$ |

**Time series.** 1-hop: `peak_identification` (argmax/argmin index), `range_query` (max−min), `change_point`, `seasonality_detection`, value/endpoint/trend lookups; 2-hop: `volatility`, `mean_half_comparison`, `threshold_count`, `oscillation_count`; 3-hop: `forecasting` (naive), `anomaly_detection`. *Near-constant cells.* Ten (modality, task) cells have one answer covering ≥90% of their questions and are flagged rather than removed: tabular `counterfactual` (146; 92% "yes"); time-series `seasonality_detection` and `change_point` (107 each; 93%), `normalized_amplitude`, `pattern_classification`, `counterfactual_scale_peak` (20 each; 100%), `counterfactual_half_shift` (20; 90%), `counterfactual_sign_flip` (15; 100%), `local_peak_median_relation` (5); graph `connectivity` (96; all "yes"). `pattern_label_lookup` (107) asks for the

generator's internal pattern label, which no rendering encodes, and is likewise flagged. Section 4 reports the format gaps with the ten cells (556 questions) excluded; `pattern_label_lookup` is excluded from the answerability subsets of Section 5 as unanswerable by design. The tabular `counterfactual` template is also mis-worded: its key encodes "is the ordering of the two column means unchanged after scaling", whereas the wording ("would $A$ remain above $B$") presupposes $A > B$, which fails for 61 of the 146 questions, and a literal reading ("yes" only if $A > B$ before and after scaling) gives the opposite answer for 49; re-keying literally moves the tabular gaps by at most 0.8pp. The wording is corrected in the released generator.

**Graph.** `degree_query`, `centrality` (betweenness), `clustering`, `diameter`, `bipartite_check`, `cycle_detection`, `edge_count`/density, `connectivity` (`nx.has_path`), `shortest_path`, `community` (largest-component fraction). Answers computed via NetworkX.

Table 6 gives one example question per task.

## B PER-FORMAT F1 (METRIC ROBUSTNESS)

Token-level F1 by visualization format tracks the exact-match pattern. F1 normalises text but not numbers (`45.0` vs. `45` scores 0), so this is a consistency check on the format ordering, not an independent metric; the main text relies on exact match and the tolerance-based numeric metric instead.

## C EVERY FORMAT FOR ONE OBJECT PER MODALITY

Figure 3 shows one source object per modality rendered in every format of the original suite (v1, the images all seven models were evaluated on) and of the corrected suite (v2, Section 5(iv)). The three objects are the ones whose per-format predictions are listed in Appendix D. The v1 tabular `bar_chart` draws one bar per numeric column (the column mean); v2 draws one panel per column with one bar per row. The v1 `scatter_plot` plots the first two numeric columns without row labels; v2 labels every point. The v1 time-series `text_only` lists 180 points of which ≈43 lines are visible; v2 prints every point in columns. The v1 graph `text_only` prints $\mathrm{degree}(u), \mathrm{degree}(v)$ beside each edge (visibly cut after ≈43 edges); v2 lists every edge without degrees, and a separate `text_only_deg` variant keeps the degree annotation so that its effect can be isolated. Graph drawings in v1 label nodes only below 30 nodes; v2 labels every node.

## D PER-FORMAT PREDICTIONS FOR THREE EXAMPLE ITEMS

Table 8 lists what every evaluated model answered, format by format, for the three objects of Figure 3 (the tabular one is also Figure 1). The three mechanisms of Section 5 are all visible in one table. *Absent answers*: on the v1 `bar_chart` every model reports the drawn column mean (75.77) instead of the sum; on the unlabelled graph drawings no model recovers node 0's degree, and the corrected `text_only` listing without degree annotations loses it too (Qwen answers 12). *Present answers, misread*: the time-series `heatmap`, `gaf`, and `recurrence_plot` all encode the minimum's position, yet no model finds it, whereas every model but one reads it off the `line_plot`. *Truncation*: the v1 time-series `text_only` shows about 43 of 166 points, so the minimum at index 25 is visible and two models find it; the same rendering fails for the row-limited tables.

Table 6: One example question per (modality, task) with its key; $n$ = questions in the released evaluation file. Questions are template-generated and the key is computed by the task's answer function on the source object.

| Modality | Task | Difficulty | Example question | Answer | $n$ |
|---|---|---|---|---|---|
| tabular | value_extraction | 1-hop | What is the value of 'ticker' in row 0? | BET | 468 |
| tabular | trend_analysis | 2-hop | Is 'close_price' generally increasing or decreasing across rows? | increasing | 247 |
| tabular | outlier_detection | 3-hop | Which 'ticker' entry deviates most from the mean of 'close_price'? | BET | 230 |
| tabular | comparison | 3-hop | Which column has the higher average value: 'close_price' or 'trade_volume_k'? | trade_volume_k | 224 |
| tabular | aggregation | 2-hop | What is the sum of 'close_price' across all rows? Round to 2 decimals. | 2826.23 | 207 |
| tabular | counterfactual | counterfactual | If all values in 'close_price' were multiplied by 1.2, would 'close_price' remain above 't... | yes | 146 |
| tabular | filtering | 2-hop | How many rows have 'close_price' strictly greater than 119.37? | 13 | 104 |
| tabular | ranking | 1-hop | Which 'ticker' has the highest 'close_price' value? | BET | 90 |
| tabular | correlation | 3-hop | Do 'close_price' and 'trade_volume_k' appear positively or negatively correlated? | negative | 49 |
| timeseries | peak_identification | 1-hop | At which timestep does the maximum value occur? | 4 | 214 |
| timeseries | value_lookup | 1-hop | What is the first value in the series? | 34.778 | 214 |
| timeseries | change_point | 1-hop | At approximately which timestep does a significant change occur? | none | 107 |
| timeseries | pattern_label_lookup | 1-hop | What is the pattern type label for this dataset? | linear_trend | 107 |
| timeseries | range_query | 2-hop | What is the range (max - min) of values? | 17.038 | 107 |
| timeseries | seasonality_detection | 1-hop | Does this series show periodic/seasonal behavior? | no | 107 |
| timeseries | volatility | 2-hop | Is this series more volatile in the first or second half? | second half | 107 |
| timeseries | mean_half_comparison | 2-hop | Is the mean of the second half higher than the first half? | no | 62 |
| timeseries | endpoint_comparison | 1-hop | Is the final value higher or lower than the first value? | lower | 55 |
| timeseries | anomaly_detection | 3-hop | Which timestep appears most anomalous relative to the sequence average? | 4 | 20 |
| timeseries | change_volatility | 2-hop | Is the average absolute change larger in first or second half? | second half | 20 |
| timeseries | counterfactual_half_shift | counterfactual | If we add 5 only to second half, would second-half mean exceed first-half mean? | no | 20 |
| timeseries | counterfactual_scale_peak | counterfactual | If every value increased by 10%, would max index change? | no | 20 |
| timeseries | forecasting | 3-hop | Based on the most recent change, what is the next expected value? | 17.382 | 20 |
| timeseries | half_change_intensity | 3-hop | Which half has larger mean absolute first difference? | second half | 20 |
| timeseries | mean_shift_magnitude | 2-hop | By how much does second-half mean exceed first-half mean? | -6.330 | 20 |
| timeseries | median_mean_relation | 2-hop | Is the median value above or below the overall mean? | above | 20 |
| timeseries | normalized_amplitude | 3-hop | Is normalized amplitude (range/|mean|) greater than 1.0? | no | 20 |
| timeseries | pattern_classification | counterfactual | If each point increased by 5, would the trend direction change? | no | 20 |
| timeseries | peak_trough_order | 3-hop | Does the peak occur after the trough? | no | 20 |
| timeseries | split_trend_consistency | 3-hop | Do both halves move in the same trend direction? | yes | 20 |
| timeseries | threshold_count | 2-hop | How many timesteps are above mean plus one standard deviation? | 10 | 20 |
| timeseries | counterfactual_sign_flip | counterfactual | If the series were multiplied by -1, would max and min indices swap? | yes | 15 |
| timeseries | oscillation_count | 2-hop | How many sign changes are in first-order differences? | 40 | 10 |
| timeseries | local_peak_median_relation | 3-hop | Is the median of local maxima greater than overall median? | yes | 5 |
| graph | degree_query | 1-hop | What is the degree of node 0? | 5 | 208 |
| graph | connectivity | 1-hop | Are node 0 and node 23 connected? | yes | 96 |
| graph | community | 3-hop | What fraction of nodes belong to the largest connected component? | 0.833 | 92 |
| graph | counterfactual | counterfactual | If an edge were added between node 0 and node 23, would they be directly connected? | yes | 76 |
| graph | centrality | 3-hop | Which node has the highest betweenness centrality? | 4 | 36 |
| graph | clustering | 2-hop | What is the approximate clustering coefficient of the graph? | 0.096 | 36 |
| graph | shortest_path | 2-hop | What is the shortest path length between node 0 and node 23? | 2 | 36 |
| graph | bipartite_check | 2-hop | Is this graph bipartite? | no | 20 |
| graph | cycle_detection | 2-hop | Does this graph contain a cycle? | yes | 20 |
| graph | diameter | 3-hop | What is the diameter of the graph (or largest component)? | 5 | 20 |
| graph | edge_count | 1-hop | How many edges does this graph have? | 35 | 20 |

Table 7: F1 (%) by format. Tabular / time-series columns share the same 5-format layout.

| | | | | | |
|---|---|---|---|---|---|
| *Tabular*: bar_chart / heatmap / scatter_plot / table_image / text_only | | | | | |
| GPT-4o | 22.2 | 35.7 | 23.2 | 49.1 | 47.5 |
| Gemini Flash | 24.4 | 36.1 | 19.4 | 49.5 | 51.1 |
| Qwen2.5-VL-7B | 25.4 | 34.3 | 17.3 | 44.3 | 40.7 |
| Claude Sonnet | 18.0 | 26.4 | 14.8 | 33.7 | 26.2 |
| *Time series*: gaf / heatmap / line_plot / recurrence_plot / text_only | | | | | |
| GPT-4o | 19.6 | 20.0 | 29.4 | 12.9 | 31.9 |
| Gemini Flash | 21.8 | 26.2 | 34.5 | 22.2 | 34.4 |
| Qwen2.5-VL-7B | 22.8 | 22.7 | 28.8 | 20.9 | 28.4 |
| Claude Sonnet | 15.5 | 17.8 | 23.7 | 12.8 | 16.8 |
| *Graph*: adjacency_matrix / circular_layout / node_link / text_only | | | | | |
| GPT-4o | 32.4 | 35.3 | 37.6 | 53.2 | |
| Gemini Flash | 30.8 | 33.2 | 32.0 | 51.7 | |
| Qwen2.5-VL-7B | 27.9 | 33.8 | 35.6 | 42.4 | |
| Claude Sonnet | 35.8 | 35.0 | 31.7 | 48.0 | |




Figure 3: One object per modality in every format of the original (v1) and corrected (v2) suites. The time-series charts are identical across suites by construction (only `text_only` changed); v2 images are 1024 × 1024, v1 images 1024 × 768.

Table 8: Per-format predictions (bold = exact match with the key) for the three example items; the last row of each block is Qwen2.5-VL-7B re-queried on the corrected suite.

*tabular_001326* — What is the sum of 'Acc' across all rows? Round to 2 decimals. key: `303.1`

| model | table_image | text_only | heatmap | bar_chart | scatter_plot |
|---|---|---|---|---|---|
| GPT-4o | **303.1** | **303.10** | **303.1** | 75.77 | 292.5 |
| Gemini-2.0-Flash | **303.10** | **303.10** | **303.10** | 75.77 | 327.10 |
| Gemini-2.5-Flash | **303.10** | **303.10** | **303.10** | 75.77 | 303.20 |
| Claude Sonnet 4 | **303.10** | 283.40 | **303.10** | 75.77 | 303.00 |
| Qwen2.5-VL-7B | **303.10** | 305.95 | 303.12 | 75.77 | 390.50 |
| Qwen2.5-VL-32B | 283.10 | 292.90 | **303.10** | 75.77 | 380.00 |
| InternVL2.5-8B | 303.7 | 150000.00 | 316.50 | 75.77 | 80.00 |
| Qwen2.5-VL-7B, corrected suite | 303.90 | 303.90 | 303.15 | **303.1** | 253.00 |

*timeseries_000027* — At which timestep does the minimum value occur? key: `25`

| model | line_plot | heatmap | gaf | recurrence_plot | text_only |
|---|---|---|---|---|---|
| GPT-4o | **25** | 91 | 0 | none | 6 |
| Gemini-2.0-Flash | **25** | 0 | 0 | none | **25** |
| Gemini-2.5-Flash | **25** | 18 | none | none | 6 |
| Claude Sonnet 4 | **25** | 100 | 80 | 1 | **25** |
| Qwen2.5-VL-7B | **25** | 0 | 0 | 0 | 39 |
| Qwen2.5-VL-32B | **25** | 165 | 160 | 0 | 39 |
| InternVL2.5-8B | 100 | 165 | 160 | 0 | 24 |
| Qwen2.5-VL-7B, corrected suite | **25** | 0 | 0 | 0 | 65 |

*graph_000076* — What is the degree of node 0? key: `11`

| model | node_link | circular_layout | adjacency_matrix | text_only |
|---|---|---|---|---|
| GPT-4o | 6 | 6 | 5 | **11** |
| Gemini-2.0-Flash | 21 | 13 | 12 | **11** |
| Gemini-2.5-Flash | none | 10 | 10 | **11** |
| Claude Sonnet 4 | 0 | 33 | 15 | **11** |
| Qwen2.5-VL-7B | 5 | 5 | 2 | **11** |
| Qwen2.5-VL-32B | 6 | 10 | 10 | **11** |
| InternVL2.5-8B | 3 | 4 | 3 | **11** |
| Qwen2.5-VL-7B, corrected suite | 6 | 6 | 5 | 12 |

## E STATISTICAL PROCEDURES AND FULL PER-TASK RESULTS

**Unit of analysis.** Every question carries a `data_id` naming its source object (299 objects, 12.7 questions per object on average, at most 25); questions from one object share data and are not independent. All confidence intervals and tests in the paper therefore treat the *object* as the resampling unit unless stated otherwise. Point estimates are ratio estimators $\sum_c d_c / \sum_c n_c$ over object clusters $c$, so they coincide with plain question-level means.

**Best–worst gaps.** For a model and modality the best and worst formats are chosen once on the full sample and held fixed. The gap's 95% interval is a percentile bootstrap over objects ($B$=5000). Its $p$-value is a two-sided sign-flip permutation test over objects ($B$=10,000): the per-object paired difference between the two formats is multiplied by a random sign, and $p = (\text{hits}+1)/(B+1)$ where hits counts permuted statistics at least as extreme as the observed one, so the smallest reportable $p$ is $1/10001$. The 12 main model×modality contrasts are corrected with Holm's step-down procedure (all 12 remain below 0.01). Because the formats are chosen post hoc, we also report an object-level split-half check: the best and worst formats are selected on one random half of the objects and tested on the other; all 24 such tests (two halves × 12 contrasts) are significant after correction.

**Other quantities.** The flip rate is the share of base questions whose exact-match outcome differs between at least two formats; the consistency rate is the share answered identically (all formats agree on the same normalised answer); the oracle is correct if any format is, and the random-format baseline is the mean over formats. The question-only prior (Section 4) is the majority answer per (modality, task) cell, learned on one half of the objects and scored on the other, averaged over twenty random splits. Mitigation intervals (Section 6) use the same object-cluster bootstrap over held-out objects (selector) or over the 44 fine-tuning objects (LoRA).

**Full per-task results.** Tables 9–15 give the numbers behind Figure 2 and Table 1: every model's best and worst format per modality with the object-cluster interval and permutation $p$-value (Table 9), flip, consistency, oracle, and random-format rates (Table 10), the gap by data source (Table 11), and exact match per task and format for the four core models (Tables 12–15; the time-series table is split in two).

Table 9: Best–worst format gap per model and modality (exact match, pp), all seven models. Best and worst formats are chosen once on the full sample; the 95% interval is a percentile bootstrap over source objects ($B$=5000) and $p$ a two-sided sign-flip permutation over objects ($B$=10,000). $n_q$ / $n_{obj}$ = questions / objects.

| Model | Best | Worst | Best EM | Worst EM | Gap | 95% CI | $n_q$ / $n_{obj}$ | $p$ |
|---|---|---|---|---|---|---|---|---|
| *tabular* | | | | | | | | |
| GPT-4o | table_image | bar_chart | 59.5 | 22.8 | 36.7 | [33.6, 40.1] | 1765 / 156 | $< 10^{-4}$ |
| Gemini-2.0-Flash | text_only | scatter_plot | 61.1 | 20.6 | 40.5 | [37.0, 44.3] | 1765 / 156 | $< 10^{-4}$ |
| Qwen2.5-VL-7B | table_image | scatter_plot | 54.6 | 17.5 | 37.1 | [34.1, 40.3] | 1765 / 156 | $< 10^{-4}$ |
| Claude Sonnet 4 | table_image | scatter_plot | 44.5 | 16.1 | 28.4 | [24.8, 32.4] | 1765 / 156 | $< 10^{-4}$ |
| Gemini-2.5-Flash | text_only | bar_chart | 59.7 | 21.2 | 38.5 | [34.8, 42.2] | 1765 / 156 | $< 10^{-4}$ |
| Qwen2.5-VL-32B | table_image | scatter_plot | 53.3 | 18.6 | 34.7 | [31.0, 38.6] | 1765 / 156 | $< 10^{-4}$ |
| InternVL2.5-8B | table_image | scatter_plot | 30.9 | 11.7 | 19.2 | [16.3, 22.2] | 1765 / 156 | $< 10^{-4}$ |
| *timeseries* | | | | | | | | |
| GPT-4o | text_only | recurrence_plot | 35.8 | 10.1 | 25.7 | [23.4, 28.1] | 1370 / 107 | $< 10^{-4}$ |
| Gemini-2.0-Flash | text_only | gaf | 38.7 | 19.5 | 19.2 | [16.5, 22.3] | 1370 / 107 | $< 10^{-4}$ |
| Qwen2.5-VL-7B | text_only | recurrence_plot | 32.8 | 19.2 | 13.6 | [11.0, 16.6] | 1370 / 107 | $< 10^{-4}$ |
| Claude Sonnet 4 | line_plot | recurrence_plot | 23.3 | 11.3 | 12.0 | [10.3, 13.6] | 1370 / 107 | $< 10^{-4}$ |
| Gemini-2.5-Flash | text_only | recurrence_plot | 39.5 | 13.2 | 26.3 | [23.9, 28.7] | 1370 / 107 | $< 10^{-4}$ |
| Qwen2.5-VL-32B | text_only | recurrence_plot | 35.2 | 17.2 | 18.0 | [15.9, 20.2] | 1370 / 107 | $< 10^{-4}$ |
| InternVL2.5-8B | line_plot | gaf | 19.7 | 11.1 | 8.6 | [6.7, 10.6] | 1370 / 107 | $< 10^{-4}$ |
| *graph* | | | | | | | | |
| GPT-4o | text_only | adjacency_matrix | 56.4 | 33.3 | 23.0 | [18.8, 27.6] | 660 / 36 | $< 10^{-4}$ |
| Gemini-2.0-Flash | text_only | adjacency_matrix | 53.5 | 31.8 | 21.7 | [17.9, 25.6] | 660 / 36 | $< 10^{-4}$ |
| Qwen2.5-VL-7B | text_only | adjacency_matrix | 43.0 | 28.2 | 14.8 | [12.1, 18.3] | 660 / 36 | $< 10^{-4}$ |
| Claude Sonnet 4 | text_only | node_link | 49.5 | 34.8 | 14.7 | [11.3, 18.3] | 660 / 36 | $< 10^{-4}$ |
| Gemini-2.5-Flash | text_only | node_link | 59.8 | 29.8 | 30.0 | [24.2, 34.7] | 660 / 36 | $< 10^{-4}$ |
| Qwen2.5-VL-32B | text_only | adjacency_matrix | 57.1 | 34.7 | 22.4 | [18.2, 27.2] | 660 / 36 | $< 10^{-4}$ |
| InternVL2.5-8B | text_only | node_link | 30.9 | 23.3 | 7.6 | [4.0, 11.5] | 660 / 36 | 0.0003 |

Table 10: Per-question consistency across formats, all seven models (%). Random = expected exact match when one format is drawn uniformly per question (equal to the mean over formats); Oracle = correct in at least one format; All-correct = correct in every format; Flip = share of questions whose correctness differs across formats.

| Model | Modality | $n$ | Mean EM | Random | Oracle | All-correct | Flip |
|---|---|---|---|---|---|---|---|
| GPT-4o | tabular | 1765 | 41.4 | 41.4 | 74.4 | 13.0 | 61.4 |
| GPT-4o | timeseries | 1370 | 21.8 | 21.8 | 45.3 | 4.6 | 40.7 |
| GPT-4o | graph | 660 | 43.6 | 43.6 | 68.8 | 24.2 | 44.5 |
| GPT-4o | all | 3795 | 34.4 | 34.7 | 62.9 | 11.9 | 51.0 |
| Gemini-2.0-Flash | tabular | 1765 | 42.4 | 42.4 | 77.2 | 9.8 | 67.4 |
| Gemini-2.0-Flash | timeseries | 1370 | 27.2 | 27.2 | 53.8 | 8.5 | 45.3 |
| Gemini-2.0-Flash | graph | 660 | 39.3 | 39.3 | 66.5 | 19.2 | 47.3 |
| Gemini-2.0-Flash | all | 3795 | 36.2 | 36.3 | 66.9 | 11.0 | 55.9 |
| Qwen2.5-VL-7B | tabular | 1765 | 37.7 | 37.7 | 69.1 | 9.0 | 60.1 |
| Qwen2.5-VL-7B | timeseries | 1370 | 24.3 | 24.3 | 43.1 | 10.8 | 32.3 |
| Qwen2.5-VL-7B | graph | 660 | 35.3 | 35.3 | 54.7 | 22.1 | 32.6 |
| Qwen2.5-VL-7B | all | 3795 | 32.4 | 32.5 | 57.2 | 11.9 | 45.2 |
| Claude Sonnet 4 | tabular | 1765 | 30.2 | 30.2 | 61.6 | 3.2 | 58.4 |
| Claude Sonnet 4 | timeseries | 1370 | 17.5 | 17.5 | 36.7 | 5.6 | 31.1 |
| Claude Sonnet 4 | graph | 660 | 39.7 | 39.7 | 63.9 | 20.6 | 43.3 |
| Claude Sonnet 4 | all | 3795 | 26.8 | 27.3 | 53.0 | 7.1 | 45.9 |
| Gemini-2.5-Flash | tabular | 1765 | 41.0 | 41.0 | 75.1 | 12.6 | 62.5 |
| Gemini-2.5-Flash | timeseries | 1370 | 26.6 | 26.6 | 53.4 | 5.1 | 48.3 |
| Gemini-2.5-Flash | graph | 660 | 40.5 | 40.5 | 67.4 | 16.7 | 50.8 |
| Gemini-2.5-Flash | all | 3795 | 35.5 | 35.7 | 66.0 | 10.6 | 55.3 |
| Qwen2.5-VL-32B | tabular | 1765 | 37.0 | 37.0 | 68.7 | 8.3 | 60.3 |
| Qwen2.5-VL-32B | timeseries | 1370 | 25.2 | 25.2 | 46.4 | 10.1 | 36.2 |
| Qwen2.5-VL-32B | graph | 660 | 41.8 | 41.8 | 64.4 | 25.5 | 38.9 |
| Qwen2.5-VL-32B | all | 3795 | 33.3 | 33.6 | 59.9 | 12.0 | 47.9 |
| InternVL2.5-8B | tabular | 1765 | 21.7 | 21.7 | 47.8 | 3.3 | 44.4 |
| InternVL2.5-8B | timeseries | 1370 | 16.1 | 16.1 | 29.5 | 3.4 | 26.1 |
| InternVL2.5-8B | graph | 660 | 26.2 | 26.2 | 39.7 | 15.5 | 24.2 |
| InternVL2.5-8B | all | 3795 | 20.3 | 20.5 | 39.8 | 5.5 | 34.3 |

Table 11: Best–worst gap (pp) by data source, four core models.

| Model | Modality | Source | $n$ | Best | Worst | Best EM | Worst EM | Gap |
|---|---|---|---|---|---|---|---|---|
| GPT-4o | tabular | synthetic | 500 | table_image | scatter_plot | 39.8 | 16.6 | 23.2 |
| GPT-4o | tabular | real-world | 1265 | table_image | bar_chart | 67.3 | 24.7 | 42.5 |
| GPT-4o | timeseries | synthetic | 500 | text_only | recurrence_plot | 42.6 | 16.2 | 26.4 |
| GPT-4o | timeseries | real-world | 870 | text_only | recurrence_plot | 32.0 | 6.7 | 25.3 |
| GPT-4o | graph | synthetic | 500 | text_only | adjacency_matrix | 54.8 | 35.8 | 19.0 |
| GPT-4o | graph | real-world | 160 | text_only | adjacency_matrix | 61.2 | 25.6 | 35.6 |
| Gemini-2.0-Flash | tabular | synthetic | 500 | text_only | scatter_plot | 42.0 | 16.6 | 25.4 |
| Gemini-2.0-Flash | tabular | real-world | 1265 | table_image | scatter_plot | 69.3 | 22.2 | 47.1 |
| Gemini-2.0-Flash | timeseries | synthetic | 500 | line_plot | gaf | 42.4 | 33.6 | 8.8 |
| Gemini-2.0-Flash | timeseries | real-world | 870 | text_only | recurrence_plot | 36.8 | 11.3 | 25.5 |
| Gemini-2.0-Flash | graph | synthetic | 500 | text_only | adjacency_matrix | 54.2 | 34.4 | 19.8 |
| Gemini-2.0-Flash | graph | real-world | 160 | text_only | adjacency_matrix | 51.2 | 23.8 | 27.5 |
| Qwen2.5-VL-7B | tabular | synthetic | 500 | text_only | scatter_plot | 40.0 | 11.0 | 29.0 |
| Qwen2.5-VL-7B | tabular | real-world | 1265 | table_image | scatter_plot | 61.3 | 20.1 | 41.3 |
| Qwen2.5-VL-7B | timeseries | synthetic | 500 | line_plot | recurrence_plot | 38.4 | 31.2 | 7.2 |
| Qwen2.5-VL-7B | timeseries | real-world | 870 | text_only | recurrence_plot | 32.3 | 12.3 | 20.0 |
| Qwen2.5-VL-7B | graph | synthetic | 500 | text_only | adjacency_matrix | 43.6 | 30.8 | 12.8 |
| Qwen2.5-VL-7B | graph | real-world | 160 | text_only | adjacency_matrix | 41.2 | 20.0 | 21.2 |
| Claude Sonnet 4 | tabular | synthetic | 500 | table_image | scatter_plot | 21.4 | 11.2 | 10.2 |
| Claude Sonnet 4 | tabular | real-world | 1265 | table_image | scatter_plot | 53.7 | 18.1 | 35.6 |
| Claude Sonnet 4 | timeseries | synthetic | 500 | line_plot | recurrence_plot | 29.6 | 18.2 | 11.4 |
| Claude Sonnet 4 | timeseries | real-world | 870 | text_only | recurrence_plot | 24.1 | 7.4 | 16.8 |
| Claude Sonnet 4 | graph | synthetic | 500 | text_only | node_link | 48.8 | 36.0 | 12.8 |
| Claude Sonnet 4 | graph | real-world | 160 | text_only | adjacency_matrix | 51.9 | 28.1 | 23.8 |

Table 12: Exact match (%) per task and format, tabular questions, four core models; $n$ = questions per task; bold = best format in the row.

| Task | $n$ | table_image | text_only | heatmap | bar_chart | scatter_plot |
|---|---|---|---|---|---|---|
| *GPT-4o* | | | | | | |
| value_extraction | 468 | **72.0** | 69.9 | 29.7 | 1.7 | 2.8 |
| trend_analysis | 247 | 62.3 | **72.5** | 67.2 | 19.4 | 48.2 |
| outlier_detection | 230 | **37.8** | 33.5 | 0.0 | 0.0 | 0.0 |
| comparison | 224 | 87.5 | 82.1 | 87.9 | **95.1** | 80.8 |
| aggregation | 207 | 44.4 | 35.3 | **54.6** | 16.9 | 4.3 |
| counterfactual | 146 | 67.1 | **73.3** | 71.2 | 61.0 | 58.2 |
| filtering | 104 | 29.8 | 27.9 | **49.0** | 8.7 | 8.7 |
| ranking | 90 | **61.1** | 44.4 | 0.0 | 0.0 | 0.0 |
| correlation | 49 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| *Gemini-2.0-Flash* | | | | | | |
| value_extraction | 468 | **68.8** | 62.4 | 29.9 | 1.7 | 6.0 |
| trend_analysis | 247 | 69.6 | **75.3** | **75.3** | 35.2 | 47.0 |
| outlier_detection | 230 | 37.4 | **47.4** | 0.0 | 0.0 | 0.0 |
| comparison | 224 | 89.3 | 88.4 | 93.3 | **94.2** | 53.6 |
| aggregation | 207 | 48.8 | 47.3 | **53.1** | 16.9 | 9.2 |
| counterfactual | 146 | 67.8 | **71.9** | 65.1 | 60.3 | 43.2 |
| filtering | 104 | 35.6 | 31.7 | **46.2** | 7.7 | 16.3 |
| ranking | 90 | 53.3 | **64.4** | 0.0 | 0.0 | 0.0 |
| correlation | 49 | 0.0 | 0.0 | 2.0 | **10.2** | 2.0 |
| *Qwen2.5-VL-7B* | | | | | | |
| value_extraction | 468 | **68.6** | 56.6 | 29.9 | 1.7 | 2.8 |
| trend_analysis | 247 | **74.9** | 71.3 | 69.6 | 38.1 | 46.2 |
| outlier_detection | 230 | 22.2 | **24.3** | 1.3 | 0.0 | 0.0 |
| comparison | 224 | 82.1 | 79.0 | 91.5 | **96.4** | 62.5 |
| aggregation | 207 | 26.6 | 10.1 | **34.8** | 16.9 | 1.4 |
| counterfactual | 146 | 61.6 | 61.6 | **65.8** | 63.0 | 23.3 |
| filtering | 104 | 24.0 | 19.2 | **40.4** | 9.6 | 2.9 |
| ranking | 90 | 48.9 | **54.4** | 1.1 | 1.1 | 1.1 |
| correlation | 49 | 16.3 | **22.4** | 10.2 | 4.1 | 2.0 |
| *Claude Sonnet 4* | | | | | | |
| value_extraction | 468 | **68.2** | 60.5 | 29.9 | 1.7 | 4.9 |
| trend_analysis | 247 | 71.3 | 64.8 | **76.1** | 22.3 | 50.2 |
| outlier_detection | 230 | **3.9** | 1.3 | 0.0 | 0.0 | 0.0 |
| comparison | 224 | 24.1 | 8.9 | 33.5 | **58.5** | 17.4 |
| aggregation | 207 | 41.1 | 23.2 | **50.7** | 20.3 | 11.1 |
| counterfactual | 146 | 35.6 | 30.1 | 34.2 | **58.2** | 43.2 |
| filtering | 104 | 33.7 | 35.6 | **54.8** | 11.5 | 12.5 |
| ranking | 90 | **62.2** | 54.4 | 0.0 | 0.0 | 0.0 |
| correlation | 49 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Table 13: Exact match (%) per task and format, time-series questions, GPT-4o and Gemini-2.0-Flash (continued in Table 14).

| Task | $n$ | line_plot | heatmap | gaf | recurrence_plot | text_only |
|---|---|---|---|---|---|---|
| *GPT-4o* | | | | | | |
| peak_identification | 214 | 7.5 | 0.9 | 2.3 | 0.5 | **23.8** |
| value_lookup | 214 | 4.7 | 3.3 | 1.4 | 0.0 | **41.6** |
| change_point | 107 | 0.0 | 0.9 | 0.9 | 0.9 | **1.9** |
| pattern_label_lookup | 107 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| range_query | 107 | 1.9 | 2.8 | 0.0 | 0.0 | **4.7** |
| seasonality_detection | 107 | 51.4 | 8.4 | 8.4 | 11.2 | **65.4** |
| volatility | 107 | 48.6 | **52.3** | 43.9 | 41.1 | 45.8 |
| mean_half_comparison | 62 | **75.8** | 35.5 | 35.5 | 37.1 | 61.3 |
| endpoint_comparison | 55 | **87.3** | 34.5 | 50.9 | 0.0 | 70.9 |
| anomaly_detection | 20 | 10.0 | 0.0 | 0.0 | 0.0 | **25.0** |
| change_volatility | 20 | **55.0** | 45.0 | 50.0 | 30.0 | 45.0 |
| counterfactual_half_shift | 20 | **90.0** | 85.0 | 35.0 | 0.0 | **90.0** |
| counterfactual_scale_peak | 20 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| forecasting | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| half_change_intensity | 20 | 40.0 | 50.0 | **60.0** | **60.0** | 55.0 |
| mean_shift_magnitude | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| median_mean_relation | 20 | **50.0** | **50.0** | 45.0 | 10.0 | 45.0 |
| normalized_amplitude | 20 | 35.0 | 40.0 | 20.0 | 0.0 | **55.0** |
| pattern_classification | 20 | **100.0** | 95.0 | 95.0 | 25.0 | **100.0** |
| peak_trough_order | 20 | **85.0** | 55.0 | 60.0 | 0.0 | 65.0 |
| split_trend_consistency | 20 | **70.0** | 65.0 | 35.0 | 50.0 | 60.0 |
| threshold_count | 20 | **5.0** | **5.0** | 0.0 | 0.0 | 0.0 |
| counterfactual_sign_flip | 15 | **100.0** | **100.0** | **100.0** | 20.0 | **100.0** |
| oscillation_count | 10 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| local_peak_median_relation | 5 | **100.0** | 20.0 | **100.0** | 0.0 | **100.0** |
| *Gemini-2.0-Flash* | | | | | | |
| peak_identification | 214 | 25.7 | 5.1 | 3.7 | 0.0 | **29.9** |
| value_lookup | 214 | 4.7 | 3.7 | 1.9 | 1.4 | **41.6** |
| change_point | 107 | 0.0 | **13.1** | 1.9 | 6.5 | 3.7 |
| pattern_label_lookup | 107 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| range_query | 107 | 0.9 | 0.9 | 0.0 | 0.0 | **5.6** |
| seasonality_detection | 107 | 64.5 | 21.5 | 7.5 | 13.1 | **80.4** |
| volatility | 107 | **64.5** | **64.5** | 47.7 | 47.7 | 53.3 |
| mean_half_comparison | 62 | **80.6** | 53.2 | 50.0 | 62.9 | 53.2 |
| endpoint_comparison | 55 | **96.4** | 63.6 | 54.5 | 49.1 | 80.0 |
| anomaly_detection | 20 | 10.0 | 15.0 | 5.0 | 0.0 | **20.0** |
| change_volatility | 20 | 45.0 | 50.0 | **55.0** | 45.0 | 45.0 |
| counterfactual_half_shift | 20 | **90.0** | **90.0** | **90.0** | 85.0 | **90.0** |
| counterfactual_scale_peak | 20 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| forecasting | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| half_change_intensity | 20 | 35.0 | 45.0 | **65.0** | 60.0 | 45.0 |
| mean_shift_magnitude | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| median_mean_relation | 20 | **60.0** | 40.0 | 45.0 | 25.0 | 45.0 |
| normalized_amplitude | 20 | 85.0 | 65.0 | 0.0 | 60.0 | **90.0** |
| pattern_classification | 20 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| peak_trough_order | 20 | **65.0** | **65.0** | **65.0** | 55.0 | **65.0** |
| split_trend_consistency | 20 | 35.0 | **65.0** | 40.0 | 35.0 | 35.0 |
| threshold_count | 20 | 0.0 | **5.0** | 0.0 | 0.0 | 0.0 |
| counterfactual_sign_flip | 15 | **100.0** | **100.0** | **100.0** | 73.3 | **100.0** |
| oscillation_count | 10 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| local_peak_median_relation | 5 | **100.0** | 80.0 | **100.0** | **100.0** | **100.0** |

Table 14: Exact match (%) per task and format, time-series questions, Qwen2.5-VL-7B and Claude Sonnet 4.

| Task | $n$ | line_plot | heatmap | gaf | recurrence_plot | text_only |
|---|---|---|---|---|---|---|
| *Qwen2.5-VL-7B* | | | | | | |
| peak_identification | 214 | 12.1 | 6.5 | 1.9 | 1.9 | **16.4** |
| value_lookup | 214 | 4.7 | 3.3 | 1.9 | 1.4 | **41.6** |
| change_point | 107 | 0.0 | 0.9 | 0.0 | 0.0 | **1.9** |
| pattern_label_lookup | 107 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| range_query | 107 | 1.9 | 0.9 | 0.0 | 0.0 | **3.7** |
| seasonality_detection | 107 | 50.5 | 15.9 | 47.7 | 14.0 | **73.8** |
| volatility | 107 | **54.2** | 52.3 | 53.3 | 53.3 | 49.5 |
| mean_half_comparison | 62 | **67.7** | 43.5 | 35.5 | 58.1 | 54.8 |
| endpoint_comparison | 55 | **87.3** | 60.0 | 40.0 | 45.5 | 65.5 |
| anomaly_detection | 20 | 5.0 | 5.0 | 0.0 | 0.0 | **10.0** |
| change_volatility | 20 | **40.0** | 35.0 | 0.0 | 0.0 | 25.0 |
| counterfactual_half_shift | 20 | 90.0 | 90.0 | 90.0 | 90.0 | 90.0 |
| counterfactual_scale_peak | 20 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| forecasting | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| half_change_intensity | 20 | **5.0** | 0.0 | 0.0 | 0.0 | 0.0 |
| mean_shift_magnitude | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| median_mean_relation | 20 | 55.0 | 55.0 | 55.0 | 55.0 | 55.0 |
| normalized_amplitude | 20 | 45.0 | **100.0** | **100.0** | **100.0** | **100.0** |
| pattern_classification | 20 | **100.0** | **100.0** | **100.0** | **100.0** | 10.0 |
| peak_trough_order | 20 | **95.0** | 40.0 | 65.0 | 65.0 | 70.0 |
| split_trend_consistency | 20 | 35.0 | 35.0 | **40.0** | 30.0 | 35.0 |
| threshold_count | 20 | 0.0 | 0.0 | **5.0** | 0.0 | 0.0 |
| counterfactual_sign_flip | 15 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| oscillation_count | 10 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| local_peak_median_relation | 5 | **100.0** | **100.0** | **100.0** | 0.0 | 80.0 |
| *Claude Sonnet 4* | | | | | | |
| peak_identification | 214 | 19.2 | 3.7 | 1.9 | 1.9 | **23.8** |
| value_lookup | 214 | 5.1 | 3.3 | 1.4 | 0.9 | **41.6** |
| change_point | 107 | 0.0 | 0.9 | 0.0 | 0.9 | **2.8** |
| pattern_label_lookup | 107 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| range_query | 107 | 2.8 | 1.9 | 0.0 | 0.0 | **5.6** |
| seasonality_detection | 107 | 10.3 | 9.3 | 9.3 | 7.5 | **13.1** |
| volatility | 107 | **63.6** | 60.7 | 54.2 | 43.0 | 42.1 |
| mean_half_comparison | 62 | **58.1** | 30.6 | 1.6 | 0.0 | 1.6 |
| endpoint_comparison | 55 | **96.4** | 60.0 | 54.5 | 45.5 | 83.6 |
| anomaly_detection | 20 | 10.0 | 0.0 | 0.0 | 0.0 | **35.0** |
| change_volatility | 20 | 0.0 | 0.0 | **5.0** | 0.0 | 0.0 |
| counterfactual_half_shift | 20 | 90.0 | **95.0** | 80.0 | 60.0 | 0.0 |
| counterfactual_scale_peak | 20 | 95.0 | 95.0 | **100.0** | **100.0** | 0.0 |
| forecasting | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| half_change_intensity | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| mean_shift_magnitude | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| median_mean_relation | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| normalized_amplitude | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| pattern_classification | 20 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| peak_trough_order | 20 | **75.0** | 40.0 | 30.0 | 25.0 | 0.0 |
| split_trend_consistency | 20 | 35.0 | 35.0 | **45.0** | 30.0 | 25.0 |
| threshold_count | 20 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| counterfactual_sign_flip | 15 | **100.0** | **100.0** | **100.0** | 40.0 | **100.0** |
| oscillation_count | 10 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| local_peak_median_relation | 5 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |

Table 15: Exact match (%) per task and format, graph questions, four core models.

| Task | $n$ | node_link | circular_layout | adjacency_matrix | text_only |
|---|---|---|---|---|---|
| *GPT-4o* | | | | | |
| degree_query | 208 | 13.9 | 11.5 | 13.5 | **54.8** |
| connectivity | 96 | **78.1** | 77.1 | 47.9 | 77.1 |
| community | 92 | **62.0** | 55.4 | 38.0 | 47.8 |
| counterfactual | 76 | **98.7** | **98.7** | 93.4 | 96.1 |
| centrality | 36 | 5.6 | 5.6 | 0.0 | **33.3** |
| clustering | 36 | **11.1** | 8.3 | **11.1** | **11.1** |
| shortest_path | 36 | 16.7 | 22.2 | 0.0 | **38.9** |
| bipartite_check | 20 | **85.0** | 70.0 | 80.0 | 70.0 |
| cycle_detection | 20 | **90.0** | 85.0 | 80.0 | **90.0** |
| diameter | 20 | **20.0** | **20.0** | 15.0 | 10.0 |
| edge_count | 20 | 5.0 | 0.0 | 5.0 | **15.0** |
| *Gemini-2.0-Flash* | | | | | |
| degree_query | 208 | 6.7 | 9.1 | 11.5 | **58.2** |
| connectivity | 96 | 64.6 | **91.7** | 57.3 | 76.0 |
| community | 92 | **55.4** | 47.8 | 48.9 | 47.8 |
| counterfactual | 76 | **81.6** | 59.2 | 65.8 | 69.7 |
| centrality | 36 | 8.3 | 5.6 | 5.6 | **25.0** |
| clustering | 36 | **8.3** | 5.6 | 5.6 | 5.6 |
| shortest_path | 36 | 11.1 | 8.3 | 2.8 | **47.2** |
| bipartite_check | 20 | **80.0** | 70.0 | 65.0 | 65.0 |
| cycle_detection | 20 | **95.0** | 90.0 | 80.0 | 90.0 |
| diameter | 20 | 10.0 | **20.0** | 10.0 | 10.0 |
| edge_count | 20 | 0.0 | 0.0 | 0.0 | **5.0** |
| *Qwen2.5-VL-7B* | | | | | |
| degree_query | 208 | 9.1 | 4.3 | 6.7 | **41.3** |
| connectivity | 96 | 93.8 | **94.8** | 72.9 | 71.9 |
| community | 92 | **40.2** | 35.9 | 35.9 | 39.1 |
| counterfactual | 76 | **76.3** | 75.0 | 43.4 | 60.5 |
| centrality | 36 | 11.1 | 5.6 | 8.3 | **33.3** |
| clustering | 36 | 0.0 | 0.0 | **5.6** | 2.8 |
| shortest_path | 36 | 0.0 | 2.8 | 0.0 | **13.9** |
| bipartite_check | 20 | 65.0 | 55.0 | **70.0** | 65.0 |
| cycle_detection | 20 | **90.0** | **90.0** | 80.0 | 80.0 |
| diameter | 20 | 0.0 | **10.0** | 5.0 | 0.0 |
| edge_count | 20 | 0.0 | 0.0 | 0.0 | 0.0 |
| *Claude Sonnet 4* | | | | | |
| degree_query | 208 | 7.7 | 6.7 | 14.4 | **45.7** |
| connectivity | 96 | 56.2 | **82.3** | 66.7 | 68.8 |
| community | 92 | 40.2 | 33.7 | 30.4 | **42.4** |
| counterfactual | 76 | 97.4 | **100.0** | 93.4 | 96.1 |
| centrality | 36 | 8.3 | 5.6 | 11.1 | **22.2** |
| clustering | 36 | 5.6 | 2.8 | **8.3** | 5.6 |
| shortest_path | 36 | 19.4 | 44.4 | 16.7 | **50.0** |
| bipartite_check | 20 | 60.0 | 60.0 | **80.0** | 20.0 |
| cycle_detection | 20 | **90.0** | 80.0 | 85.0 | 85.0 |
| diameter | 20 | **35.0** | 15.0 | 10.0 | 25.0 |
| edge_count | 20 | 0.0 | 0.0 | 0.0 | 0.0 |

## F HUMAN IMAGE-ANSWERABILITY AUDIT

Ground truth is correct with respect to the underlying data by construction; this audit asks the separate question of whether each answer is legible from the rendered image *alone*. A stratified sample of 100 (question, format) pairs (34/33/33 tabular / series / graph, covering all 14 formats and every reasoning type) was rated independently, with the source data and other formats hidden, by two annotators with quantitative backgrounds (a third did not complete in time) as *correct* / *ambiguous* / *incorrect*, recording their own answer and a free-text note. The released sheets carry the ratings and own answers (free-text answers in English translation); the notes are withheld until de-anonymisation.

**Results.** Rater 1: 44 correct, 7 ambiguous, 49 incorrect; Rater 2: 34/25/41. Three-way agreement is 62% (Cohen's $\kappa = 0.40$, bootstrap 95% CI $[0.27, 0.53]$; 0.54/0.29/0.33 for tables/series/graphs); collapsed to legible vs. not, 76% ($\kappa = 0.50$); 26 of the 38 disagreements involve the ambiguous label, which the raters used at different rates. Both raters marked 27 items correct and 32 incorrect. Rater 2's own answers match the ground truth on 42 of the 89 items they answered, including 10 of the 25 they marked ambiguous. The pre-registered bar of $\geq$95% majority-correct is therefore *not met*, and we report this as a finding rather than a validation, because the code audit of Section 5 predicts it: of the 61 sampled pairs the audit marks *absent*, both raters marked 26 incorrect and at least one rater 45; of the 26 marked visible or derivable, both marked 10 correct and only 2 incorrect (overlapping nodes in a dense drawing; a correlation read from a heatmap). The raters' notes name the same causes as the audit: unlabelled graph nodes (11 items), undecodable GAF and recurrence plots ($\approx$10), truncated tables or listings (10), a bar chart that shows only column means (4), a scatter plot that lacks the queried column (4).

**Disputed ground truths.** The raters flagged ten items as having a wrong answer key. Recomputing each answer function on the full data confirms the key in all ten; in eight the rater's answer equals the answer computed on the *truncated* view (e.g. a column sum over the 12 visible rows of a 17-row table), one is a rounding difference, and one is a time-series counterfactual answered from the 43 visible points. One further item has a tied maximum degree, so two answers are correct. Conversely, the audit's adjacency-matrix rule is too strict: six of nine adjacency items were rated correct by both raters although the rule marks them absent, because the raters counted unlabelled ticks. Ratings, the anonymised notes, the recomputation script, and its output are released; the corrected suite of Section 5(iv) was built in direct response to this audit.

**Examples.** Figure 4 shows three of the ten items both raters flagged, each explained by the rendering rather than by the key: a `bar_chart` of column means asked for a column sum, a graph `text_only` listing cut after $\approx$43 edges when the question concerns later nodes, and a `circular_layout` of an unlabelled graph asked about two named nodes.

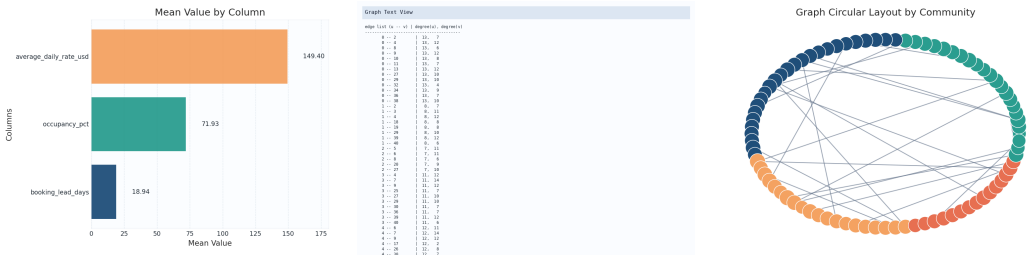


Figure 4: Three items both raters marked as unreadable. Left (H020): "What is the sum of 'occupancy_pct' across all rows?"—the chart shows column means. Middle (H007): "Which node has the highest betweenness centrality?"—the edge list is cut after the visible lines. Right (H011): "Is there a path between node 0 and node 3?"—nodes are unlabelled.

## G MODEL IDENTIFIERS AND EVALUATION SETUP

Table 16 pins the exact model versions. Open-weight models are pinned to specific HuggingFace commit revisions and are fully reproducible. The closed API models were queried through

their providers in March 2026 with the identifiers hard-coded in the released evaluation scripts; `claude-sonnet-4-20250514` is a dated snapshot, whereas `gpt-4o`, `gemini-2.0-flash`, and `gemini-2.5-flash` are pointers that resolve to a provider-side snapshot at query time, and the provider request logs were not retained, so those three runs are reproducible only up to the provider's snapshot policy. (An earlier draft listed `gemini-1.5-flash` and `claude-3-5-sonnet-20241022` from a configuration file the scripts do not read; both had been retired by the providers before the runs.) We flag this as the main reproducibility limitation for the closed models and recommend dated snapshots for any re-run; the open-model findings, which carry all mitigation and corrected-suite results, are fully pinned. The proprietary models and Qwen use one released system prompt (a short answer-format instruction: final value only, yes/no, plain number, or name); InternVL used the shorter "Answer concisely. Give only the final answer." All decoding is greedy.

**Compute.** Open-model runs used single NVIDIA A100 80GB GPUs. The corrected-suite re-evaluation of Qwen2.5-VL-7B (18,975 image queries at 999–1,369 visual tokens each) took 17 GPU-hours of model time (mean 3.2 s per query); the no-image, raw-text, and calculation-aid runs are smaller (3,795, 3,795, and 1,020 queries). Each LoRA fine-tuning run (Appendix L) takes about 80 minutes on one GPU. The proprietary models were queried through their public APIs; we did not retain per-request cost logs.

Table 16: Exact model versions. Open-weight models are pinned to HuggingFace commit revisions; closed API pointers resolve to a provider snapshot at query time (see text).

| Model (paper name) | Identifier | Version / revision |
|---|---|---|
| GPT-4o | `gpt-4o` (OpenAI API) | pointer, queried 2026-03 |
| Gemini Flash | `gemini-2.0-flash` (Google API) | pointer, queried 2026-03 |
| Gemini-2.5 | `gemini-2.5-flash` (Google API) | pointer, queried 2026-03 |
| Claude Sonnet | `claude-sonnet-4-20250514` | 2025-05-14 (dated) |
| Qwen2.5-VL-7B | `Qwen/Qwen2.5-VL-7B-Instruct` | `cc594898` |
| Qwen2.5-VL-32B | `Qwen/Qwen2.5-VL-32B-Instruct` | `7cfb30d7` |
| InternVL2.5-8B | `OpenGVLab/InternVL2_5-8B` | `e9e4c0dc` |

## H Related-benchmark comparison

Table 17: StructViz-Bench versus representative related benchmarks. Only our setting fixes data *and* question while varying the visual rendering across multiple structured modalities.

| Benchmark | Varied axis | Modalities | Format-controlled? |
|---|---|---|---|
| ChartQA (Masry et al., 2022) | question | charts | no |
| CharXiv (Wang et al., 2024) | question type; real charts | charts | no |
| ChartX (Xia et al., 2025) | chart type/task/topic | charts | no[†] |
| ChartQAPro (Masry et al., 2025) | question type | charts | no |
| ToRR (Ashury-Tahan et al., 2026) | *text* serialization | tables | yes (text) |
| FormatSpread (Sclar et al., 2024) | non-semantic formatting | text LLM | yes (text) |
| Single-modality studies[‡] | visual rendering | one modality each | yes (visual) |
| **StructViz-Bench (ours)** | **visual rendering (14)** | **table/series/graph** | **yes (visual)** |

[†]ChartX varies chart type across items, but each item is drawn once, so chart type is confounded with the item's data and question rather than manipulated within it. [‡]Mukhopadhyay et al. (2024), Deng et al. (2024), Wei et al. (2024), Daswani et al. (2024), Das et al. (2024), Liu et al. (2025), Ahsan et al. (2026); each controls the rendering within one modality.

## I Prompt-style ablation

Table 18 gives the prompt-style ablation summarised in Section 4.

Table 18: Prompt-style ablation on a stratified 500-question subsample (2,413 rendered instances per cell). Each entry is overall exact match (%) with the tabular best–worst format gap (pp) in parentheses; **bold** marks each model's best style. Prompt style matters, but the format gap persists in all 16 cells, and CoT never beats the best non-CoT prompt for any model.

| Model | minimal | concise | detailed | CoT |
|---|---|---|---|---|
| GPT-4o | 19.3 (20.2) | 23.6 (25.3) | **35.4** (36.5) | 22.0 (21.9) |
| Gemini Flash | 22.1 (19.7) | **36.4** (33.5) | 36.2 (36.9) | 25.8 (15.5) |
| Qwen2.5-VL-7B | 19.2 (17.6) | **34.1** (32.6) | 32.4 (33.5) | 19.5 (17.2) |
| Claude Sonnet | 19.4 (18.5) | 22.9 (21.0) | **23.6** (22.7) | 19.2 (20.6) |

## J RENDERING AND ANSWERABILITY AUDIT

Fixing the data does not by itself fix what is *visible*, so we audited each renderer against the stored images and the source objects. Only the time-series charts (`line_plot`, `gaf`, `recurrence_plot`, `heatmap`) and the graph drawings draw the whole object; the rest summarise, project, or truncate it, and several renderings encode the object without making the queried entity identifiable. Stated exactly because Sections 4 and 5 depend on them: the tabular `bar_chart` draws *one bar per numeric column showing that column's mean*, no rows; `scatter_plot` draws the first two numeric columns only, with no row labels; the tabular `heatmap` drops categorical and date columns; graph `node_link` and `circular_layout` label nodes only when a graph has fewer than 30 nodes (185 of the 208 `degree_query` questions concern larger graphs), and the `adjacency_matrix` labels every node only up to 24 nodes; `table_image` draws the first 12 rows, so for 708 of the 1,765 tabular questions the image withholds rows and 1,057 questions are shown in full; the tabular `text_only` rendering prints the first 14 rows and the tabular `heatmap` the first 20 (351 questions exceed it); time-series and graph `text_only` list the first 180 points or edges, but on the $1024 \times 768$ canvas only ≈43 lines are visible, so the listing is visibly cut for 1,350 of 1,370 series and 600 of 660 graphs. The graph text rendering additionally prints $\mathrm{degree}(u), \mathrm{degree}(v)$ beside every edge, which hands the 208 `degree_query` questions their answer directly. Two consequences follow and are applied throughout the paper: the tabular gap is analysed within one source on the fully shown subset (Section efsec:mechanism), and `text_only` is treated as a text-faithful but truncated rendering whose graph advantage is a degree-annotation effect, not evidence about reading edges. The released harness recomputes these counts from the released benchmark file, and the renderers expose the row, point, and edge limits as named constants so that an untruncated re-rendering is a one-line change; the corrected suite of Section 5(iv) is that re-rendering, re-queried for one open model.

**Answerability rules.** Each (question, format) pair is classified V/D/A/C from these facts and the question text. Tabular: `value_extraction` (all rows <12) is V in `table_image`/`text_only` and in `heatmap` when the column is numeric; `bar_chart` is V for mean comparisons and mean aggregations, D for counterfactuals about means, A otherwise; `scatter_plot` is V only for correlation questions about its two plotted columns (26 of 49); the whole-table tasks are D in `table_image`/`text_only`/`heatmap` when the rows fit (12/14/20) and the needed columns are drawn, A otherwise. Time series: `line_plot` is V or D for every task; `heatmap` D for ordinal tasks (peak, change point, seasonality, halves), A for numeric read-outs; `gaf`/`recurrence_plot` D only for seasonality and change point; `text_only` as `line_plot` when the series fits (18 questions), else V only for "first value"; `pattern_label_lookup` and the constant counterfactual templates are C. Graph: node-referencing tasks are D in the drawings only when nodes are labelled (<30 nodes; ≤24 for the adjacency matrix), global tasks D; `text_only` is V for `degree_query` when the node appears among the visible edges, D for everything when the graph has ≤43 edges, A otherwise; `connectivity` and add-edge counterfactuals (constant "yes") are C. Categorical-answer scoring: normalising "positively"→"positive" and "increasing/rising" variants raises correlation accuracy from ≈0 to 25–69% on scatter plots but changes no tabular best–worst gap by more than 1.2pp and no best or worst format.

**Corrected suite (v2).** The released renderers gained a `suite="v2"` option that leaves every v1 image byte-identical (a regression script re-renders 56 stored images and compares bytes) and changes only the following. Tabular: `bar_chart` draws one panel per numeric column (at most six)

with one bar per row, labelled "index: first categorical value" and annotated with the value when a table has $\leq$35 rows; `scatter_plot` labels every point the same way and adds up to a $2\times2$ grid of further column pairs; `table_image`, `text_only`, and `heatmap` draw all rows and all columns (wide tables are split into column groups). Time series: `text_only` lists every point in up to four columns; the four charts are unchanged (GAF and recurrence plots are kept as deliberately lossy transforms). Graph: `node_link` and `circular_layout` label every node with a size-scaled font; `adjacency_matrix` labels every row and column; `text_only` lists every edge *without* degrees, and a second listing `text_only_deg` adds the degree columns. All images are at most $1024\times1024$ pixels (999–1,369 visual tokens for Qwen2.5-VL-7B, which never downscales them; the mean image cost rises from 0.79 to 0.80–1.05 Mpx). The calculation-aid $2\times2$ renders the 255 tabular questions that ask for a mean comparison or a column mean in four conditions: the full table with and without an appended "mean" row, and the row-level bar chart with and without a panel listing the column means. Evaluation uses the released local runner on Qwen2.5-VL-7B (revision `cc594898`, greedy, same system prompt, raw responses stored). Because the four time-series charts are byte-identical in v1 and v2, their 5,480 pairs are a built-in control for the software stack between the March and September runs: exact-match values agree on 99.3–99.8% and prediction strings on 96.7–98.1% of pairs, and the paired change is $-0.1$ to $+0.1$pp. Two example strings in the shared system prompt ("east_grid", "first half") coincide with the answer of 4 and 74 questions respectively (and "second half" with 73); we did not control this and note it as a possible, direction-unknown bias. The no-image run sends the question alone; the raw-text run sends the data serialised as text (CSV with a row index; "t=i: value" lines; "u – v" edge lines, 720/3,842/1,109 prompt tokens on average, none truncated) followed by the question; both use the system prompt without the phrase that tells the model to look at an image (all prompts are released with their hashes).

**Second open model on the corrected suite.** InternVL2.5-8B (revision `e9e4c0dc`, the released 4-bit path with one $448\times448$ tile, 256 visual tokens, and its short prompt) was evaluated on the 1,096 questions answerable in every v2 format (5,480 pairs), the $2\times2$, and a no-image run. Every v2 image is downscaled to $448\times448$, so on the same 354 tabular questions `table_image` falls from 33.9% (v1) to 7.3% and `text_only` from 7.3 to 3.7%, while the coarser `heatmap` (+10.2pp) and row-level bars (+9.3pp) gain; the time-series charts are byte-identical between suites and their predictions agree on 214/214 questions. The graph listing loses 9.3pp $[-12.5, -6.2]$ without degrees (+6.7pp $[1.4, 12.4]$ on `degree_query` when added back), the residual on the 186 strict tabular questions is 6.5pp $[2.5, 10.9]$ (`heatmap` vs. `text_only`, $p$=0.007) and on the 528 graphs +1.7pp $[-1.6, 4.9]$ (n.s.); no graph rendering reaches the prior. In the $2\times2$ the printed mean row lifts the 31 mean-aggregation questions from 3 to 35% on the table but only to 13% on the bars (interaction +22.6pp $[6.5, 38.7]$), and the comparisons are flat (30–32%). We report this model as a resolution-bound replication, not as a second estimate of the residual: an evaluation setting that feeds 256 visual tokens cannot read a 35-row table whatever its format. Its v1 file contains 219 API-style error rows (tabular), scored as wrong here as in Section 4. Object-level statistics as in Section 4.

## K Format ensemble table

Table 20 gives the majority-vote ensemble summarised in Section 6.

## L Mitigation training details

**Consistency-regularized LoRA.** For a base question $q$ with answer $y$ rendered in two formats $a, b$, the loss is $\mathcal{L} = \mathrm{CE}(y \mid R_a, q) + \mathrm{CE}(y \mid R_b, q) + \lambda\,\mathrm{KL}_{\mathrm{sym}}\big(p(\cdot \mid R_a, q) \,\|\, p(\cdot \mid R_b, q)\big)$, where CE is the autoregressive cross-entropy on the answer tokens and $\mathrm{KL}_{\mathrm{sym}}$ is the symmetric KL (mean of the two directions) between the two answer-token distributions; $\lambda = 1$. The two CE terms are summed, not averaged, exactly as in the released script, so at $\lambda = 1$ the regulariser carries half the relative weight a mean-CE formulation would give it. Base model Qwen2.5-VL-7B-Instruct (vision tower frozen); LoRA on the language model with $r = 16$, $\alpha = 32$, dropout $0.05$, on the q/k/v/o and gate/up/down projections (47.6M trainable parameters). One epoch over 565 training pairs drawn from all three modalities, learning rate $10^{-4}$, gradient accumulation over 8 pairs, seed 42, images at most $200{,}704$ pixels. Evaluation uses a disjoint held-out manifest of 574 rows, 554 unique base questions after removing 20 duplicated time-series rows (paired statistics in Table 4 are over

Table 19: Answerability audit: exact match (%) on (question, format) pairs whose answer is *present* in the image (visible or derivable) / *absent* by design. Pair counts are the same for every model; e.g. tabular `bar_chart` present $n$=401, absent $n$=1,364; `scatter_plot` present $n$=26; graph `text_only` present $n$=146. "–" marks an empty cell. Constant-answer pairs (12% of series, 20% of graph pairs) are excluded. Full counts per task and format are in the released `answerability_matrix.md`.

| *Tabular*: bar_chart / heatmap / scatter_plot / table_image / text_only | | | | | |
|---|---|---|---|---|---|
| GPT-4o | 83 / 5 | 74 / 8 | 0 / 24 | 70 / 37 | 68 / 33 |
| Gemini Flash | 82 / 8 | 76 / 8 | 4 / 21 | 72 / 36 | 70 / 40 |
| Qwen2.5-VL-7B | 85 / 9 | 70 / 8 | 0 / 18 | 64 / 35 | 54 / 36 |
| Claude Sonnet | 62 / 6 | 61 / 3 | 0 / 16 | 58 / 17 | 45 / 17 |
| *Time series*: gaf / heatmap / line_plot / recurrence_plot / text_only | | | | | |
| GPT-4o | 5 / 17 | 20 / 11 | 27 / – | 6 / 10 | 80 / 31 |
| Gemini Flash | 5 / 20 | 31 / 12 | 33 / – | 10 / 20 | 81 / 35 |
| Qwen2.5-VL-7B | 24 / 19 | 22 / 14 | 26 / – | 7 / 19 | 80 / 29 |
| Claude Sonnet | 5 / 13 | 20 / 6 | 22 / – | 4 / 10 | 80 / 16 |
| *Graph*: adjacency_matrix / circular_layout / node_link / text_only | | | | | |
| GPT-4o | 34 / 20 | 43 / 20 | 47 / 22 | 73 / 41 | |
| Gemini Flash | 37 / 19 | 40 / 17 | 43 / 15 | 78 / 39 | |
| Qwen2.5-VL-7B | 30 / 12 | 29 / 10 | 34 / 11 | 58 / 31 | |
| Claude Sonnet | 32 / 23 | 30 / 22 | 37 / 17 | 67 / 33 | |

Table 20: Majority-vote format ensemble vs. two single-format references, exact match (%) over all base questions with a per-question denominator: "random" is an arbitrary format, "fixed (mod)" the best single format per modality (chosen in-sample), "oracle" correct-if-any-format-correct. The vote beats random but not the fixed format.

| Model | random | fixed (mod) | ensemble | Δ vs. fixed | oracle |
|---|---|---|---|---|---|
| GPT-4o | 34.7 | 50.4 | 37.0 | −13.4 | 62.9 |
| Gemini Flash | 36.3 | 51.7 | 41.4 | −10.3 | 66.9 |
| Gemini-2.5 | 35.7 | 52.4 | 38.6 | −13.8 | 66.0 |
| Qwen2.5-VL-7B | 32.5 | 44.7 | 36.1 | −8.6 | 57.2 |
| Claude Sonnet | 27.3 | 37.7 | 34.3 | −3.4 | 53.0 |
| Qwen2.5-VL-32B | 33.6 | 47.4 | 35.4 | −12.0 | 59.9 |
| InternVL2.5-8B | 20.5 | 26.9 | 20.6 | −6.2 | 39.8 |

unique questions), decoded greedily with 16 new tokens, at the same resolution for base and adapted models. The adapter configuration and training script are released; the $\lambda = 0$ control in Table 4 uses the identical script with `-lambda_consistency 0`. Two scope notes. (i) The training and held-out manifests are disjoint in questions but are drawn from the same 44 source objects, so Table 4 measures generalisation to new questions on *seen* objects, not to new objects (the selector split of Section 6 is by object; the LoRA split is not). All CIs in Table 4 are therefore cluster bootstraps over the 44 objects; question-level bootstraps are narrower and reach the same conclusions. Scoring: Table 4 uses the tolerance-based numeric metric (abs. 0.05 / rel. 0.01) and exact normalised string match otherwise. The evaluation script originally also credited a gold string contained anywhere in a longer answer; we removed that rule (it had credited 57 of 1,180 base tabular cells and none in the other modalities, so the time-series and graph rows are unchanged and the tabular gaps shrink from $20.3 \to 11.9 \to 10.2$ to $16.1 \to 6.4 \to 6.8$pp). The released records store normalised answers, so strict exact match cannot be recomputed from them without re-inference; `rescore_lora.py` reproduces every number in the table. (ii) The seed fixes the order of training pairs (Python RNG); adapter initialisation and dropout use the framework's default unseeded state, so each setting is a single training run and is not bitwise reproducible from the seed alone.

**Cross-modal rendering pairs.** Each Level-2 problem is rendered in five paired formats. Table+series: bar+line, heatmap+GAF, scatter+recurrence, table-image+heatmap, text+text. Table+graph: bar+node-link, heatmap+adjacency, scatter+circular, table-image+node-

link, text+text. Series+graph: GAF+adjacency, heatmap+node-link, line+node-link, recurrence+circular, text+text.

## M DATA SOURCES AND LICENSES

Table 21 lists every data source. The synthetic generators are ours and are released with the benchmark; the real-world objects are drawn from public datasets and are redistributed only as rendered images and question–answer pairs produced by our pipeline. The released benchmark (images, questions, keys, predictions) is licensed CC BY 4.0 and the code Apache 2.0. No source contains personal or sensitive information (Ethics Statement).

Table 21: Data sources. Counts are base questions in the released evaluation file.

| Modality | Source | Questions | Notes |
|---|---|---|---|
| Tabular | synthetic generators (20 domains) | 500 | seeded factories, 16–24 rows |
| Tabular | SciTabAlign tables (Ho et al., 2026) | 1,265 | scientific-paper tables; 1,057 questions on tables with $\leq$12 rows |
| Time series | synthetic generators (20 patterns) | 500 | trend/seasonal/anomaly families |
| Time series | ETT (Zhou et al., 2021) | 870 | electricity-transformer load, windowed |
| Graph | synthetic generators | 500 | Watts–Strogatz, Barabási–Albert, caveman, random geometric |
| Graph | NetworkX (Hagberg et al., 2008) | 160 | Karate (Zachary, 1977), Les Misérables (Knuth, 1993), Florentine families, Davis Southern women, plus generator families; 15–100 nodes |

## N REPRODUCIBILITY CHECKLIST

Table 22 maps each main claim and component to where it is described and to the released artifact that reproduces it.

Table 22: Claim/component $\rightarrow$ location and reproducing artifact.

| Claim / component | Where | Reproducing artifact |
|---|---|---|
| Deterministic data + QA generation | Sec. 3 | released generator (seed 42) |
| Task templates + answer functions | App. A | released code |
| Exact model versions | App. G | pinned HF revisions |
| Format-sensitivity gaps + tests | Sec. 4, Tab. 1 | released predictions + stats |
| Metric robustness (F1) | App. B | released predictions |
| Mechanism (contrast / probe) | Sec. 5 | predictions + verify script |
| Ensemble / selector / LoRA | Sec. 6, Tab. 20–4 | scripts (object-split selector) + LoRA config |
| Bootstrap CIs / sample sizes | Sec. 6, Tab. 4 | released stats script |
| Human-validation protocol | App. F | released sheets + aggregator |