CAPTURECHART: BENCHMARKING CHART UNDERSTANDING UNDER SIMULATED CAPTURE CONDITIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reading structured visual information from photographed charts matters for automated field inspections, factory monitoring, and mobile question answering. As vision–language models advance, evaluation benchmarks must evolve to assess their growing capabilities and remaining limitations. Yet existing real-photo benchmarks are difficult to scale because adding new charts or capture conditions requires costly new photography. Natural captures also combine multiple degradations, making their individual effects on chart understanding difficult to isolate. Hence, we introduce **CaptureChart**, a fully digital benchmark for studying chart understanding under simulated capture. Starting from 1,000 CharXiv charts, a Blender-based pipeline models the paper surface, surrounding scene, lighting, and camera imaging to produce 32,000 aligned views. Each chart has an electronic original, a clean capture, and 30 variations across four capture families and 11 mechanisms at selected intensity levels. To examine difficult reasoning questions under capture degradation, we add 1,151 visual-evidence questions linked to 173 reasoning questions. Each probe targets a chart fact used in a checked solution to its parent question, allowing us to test whether models recover the evidence needed for the final answer. Across 14 vision–language models and 30 capture variations, mean accuracy falls relative to clean captures for descriptive, reasoning, and visual-evidence questions; reasoning accuracy drops by 9.4 percentage points. Reasoning is often more sensitive than descriptive reading. Linked evidence answers can change correctness even when aggregate accuracy changes little, and specific capture mechanisms expose distinct model weaknesses. Overall, CaptureChart offers a scalable way to study chart understanding under controlled, simulated capture conditions. The benchmark is available at `https://anonymous-hf.com/a/7if3j4z0hjpy/`.

1 INTRODUCTION

Answering questions about photographed charts depends on fine visual evidence—axis ticks, legends, labels, and marks—whose visibility changes with surface shape, illumination, framing, and camera imaging.

Chart benchmarks test increasingly diverse content and reasoning tasks (Wang et al., 2024; Masry et al., 2025), and robustness studies examine visual perturbations (Moured et al., 2025; Chen et al., 2026). Photographed document and table benchmarks also show why capture conditions matter (Wang et al., 2025; Kang et al., 2026). As model capabilities and benchmark difficulty advance together, capture evaluation must both build on state-of-the-art digital chart question answering benchmarks and support controlled comparisons of individual capture factors. This calls for capture datasets that retain the charts, questions, and answers of their digital sources, together with a reusable way to construct them as benchmarks evolve. Repeatedly photographing new chart collections under multiple conditions is laborious and difficult to scale, while co-occurring artifacts in natural photographs make individual factors hard to isolate. These challenges raise a central question: *How can we systematically evaluate and diagnose vision–language models' chart understanding under challenging capture conditions?*

To address this question, we introduce CAPTURECHART, a fully digital image-capture synthesis pipeline built on CharXiv's validation split. CharXiv supplies expert-curated scientific charts with

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.

descriptive and reasoning questions (Wang et al., 2024). Reusing its question–answer pairs preserves established tasks without re-annotating every generated view. The chart, question, and reference answer stay fixed, enabling direct question-level comparisons across capture conditions. The pipeline applies reusable surface, lighting, framing, and camera configurations to each chart. The same digital configurations can extend controlled tests to larger chart collections, providing a scalable basis for future sim-to-real evaluation and training.

To examine how capture changes access to the visual facts behind difficult reasoning tasks, we add visual-evidence subquestions linked to reasoning parents. We use an LLM-assisted workflow to check that each subquestion is independently answerable and that its solution operations support the parent reasoning answer. This enables analysis of complete answers and their supporting facts under identical views.

We evaluate 14 vision–language models on CAPTURECHART, spanning general-purpose, chart-specialized, and proprietary systems. The results expose three weaknesses in current chart understanding. (1) Some models have pronounced sensitivity to specific capture conditions: under the 40% chart-area framing condition, Gemma 4 12B loses 11.2 percentage points on descriptive questions relative to clean, compared with 1.0 point for Qwen3-VL-32B. (2) Reasoning is more vulnerable to capture variation than descriptive reading: averaged over the 30 variations, Qwen3.8-27B, the strongest on clean reasoning, loses 14.01 percentage points on reasoning relative to clean captures, versus 5.61 points on descriptive questions. (3) Evidence answers can be unstable even when aggregate accuracy changes little: under mild defocus, 17.98% of Claude Sonnet 5's visual-evidence answers switch correctness while evidence accuracy drops by 1.13 points. Together, these findings show how CaptureChart's controlled conditions and linked questions expose distinct model weaknesses.

Our contributions are:

- **Reusable digital synthesis pipeline.** We develop a configurable pipeline that synthesizes chart captures through surface, scene, and camera modeling. Its capture settings can be reused across digital chart collections while retaining their questions and reference answers.
- **CaptureChart dataset.** We construct CAPTURECHART with 32,000 aligned views of 1,000 CharXiv charts and 5,000 original questions. We add 1,151 visual-evidence probes linked to 173 reasoning questions, enabling joint evaluation of task answers and their supporting facts under the same capture conditions.
- **Weaknesses in understanding captured charts.** Across 14 vision–language models, we reveal model-specific sensitivities to capture conditions, greater vulnerability of reasoning than descriptive reading in many settings, and instability in visual-evidence answers that aggregate accuracy can hide.

2 RELATED WORK

Chart question answering benchmarks. Chart question answering benchmarks have expanded both chart sources and the range of tasks used to evaluate visual understanding. ChartQA includes human-written questions requiring visual and logical reasoning (Masry et al., 2022), while CharXiv pairs expert-curated scientific charts with descriptive and reasoning questions (Wang et al., 2024). Task coverage has also broadened: ChartInsights evaluates ten analytical tasks (Wu et al., 2024), and ChartQAPro extends source diversity and question formats to include conversational, hypothetical, and unanswerable questions (Masry et al., 2025).

Visual robustness and capture conditions. Evaluations of visual robustness use synthetic perturbations and real photographs to study chart and document understanding under varied input conditions. CHAOS applies visual and textual perturbations at graded severity levels to chart tasks (Moured et al., 2025). Photographed benchmarks examine variation introduced during image capture: WildDoc captures documents from established benchmarks, including charts, four times under different conditions to compare digital and photographed inputs and assess consistency across captures (Wang et al., 2025). PCTR-16K focuses on reasoning over photographed tables (Kang et al., 2026).

Visual perception and reasoning diagnosis. Research on chart reasoning also examines the visual information and intermediate operations needed to produce an answer. DePlot translates chart images into tables, separating visual extraction from subsequent reasoning with a language model (Liu et al., 2023). Diagnostic benchmarks assess the component tasks and skills involved in reasoning: ChartR decomposes reasoning questions into dependent subtasks and evaluates them across perturbed chart views (Chen et al., 2026), while ChartProbe generates probes directly from chart-rendering code to assess perception, grounding, and simple reasoning with exact reference answers (Khoshnoodi & Bargal, 2026).

3 BENCHMARK DESIGN

CAPTURECHART provides 32 views of each of the 1,000 figures in the CharXiv validation set (Wang et al., 2024). It combines the original chart questions with visual-evidence (VE) questions linked to selected reasoning tasks, enabling evaluation of task answers and their supporting facts across different captured views. We first describe the benchmark's composition, then introduce the two synthesis pipelines that construct its images and evidence questions.

3.1 BENCHMARK COMPOSITION

Capture views. We use eleven perturbation mechanisms across four categories—surface deformation, environmental lighting, framing scale, and camera imaging. Varying their severity levels and other rendering parameters produces 30 perturbed images per chart. Adding the electronic original and clean capture gives 32 images per chart and 32,000 images in total. Figure 1 compares the electronic original, clean capture, and perturbed views of the same chart. Table 1 lists the mechanisms and their design controls. Figure 2(a) summarizes the distribution of capture mechanisms.

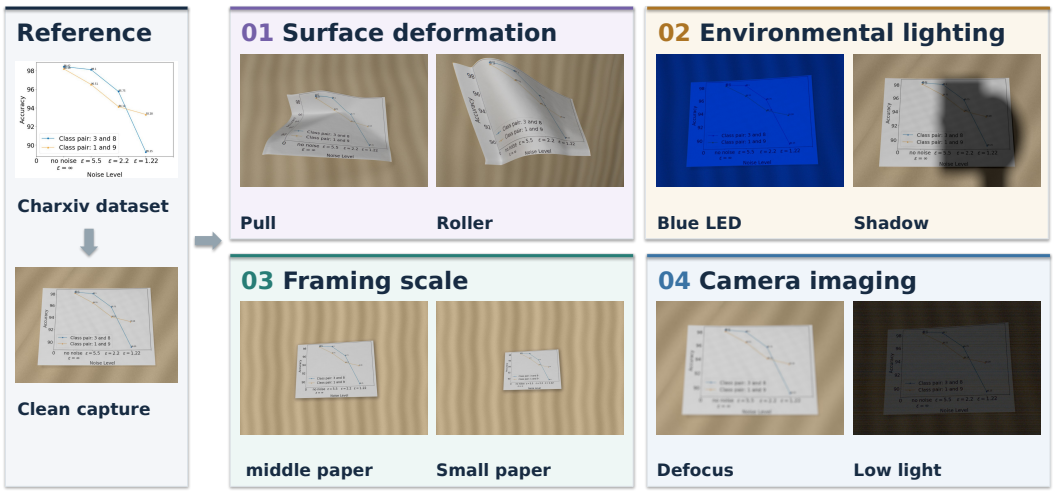


Figure 1: **CaptureChart overview.** Electronic original, clean capture, and perturbed views of the same chart.

Table 1: Capture mechanisms and their design controls.

Category	Mechanisms	Design controls
Surface deformation	Pull; roller	Paper geometry and texture coordinates
Environmental lighting	White-light temperature; colored/extreme light; shadow; reflection	Lights, occluders, glass and lamp geometry
Framing scale	Effective chart area	Paper mask, scale, background and contact shadow
Camera imaging	Motion; defocus; low light; overexposure	Camera motion during exposure, focus, radiance and sensor parameters

Questions and visual evidence. Each chart has four descriptive (D) questions and one reasoning (R) question, totaling 4,000 D and 1,000 R questions. We screen all 1,000 reasoning questions for decomposition into visual subquestions and construct 1,151 VE probes from 173 reasoning questions. These probes test the chart facts needed to answer their parent questions. Together, the original questions and VE probes form a set of 6,151 questions. Figure 2(b) summarizes the question composition. Table 2 lists the categories, answer types, and question counts.

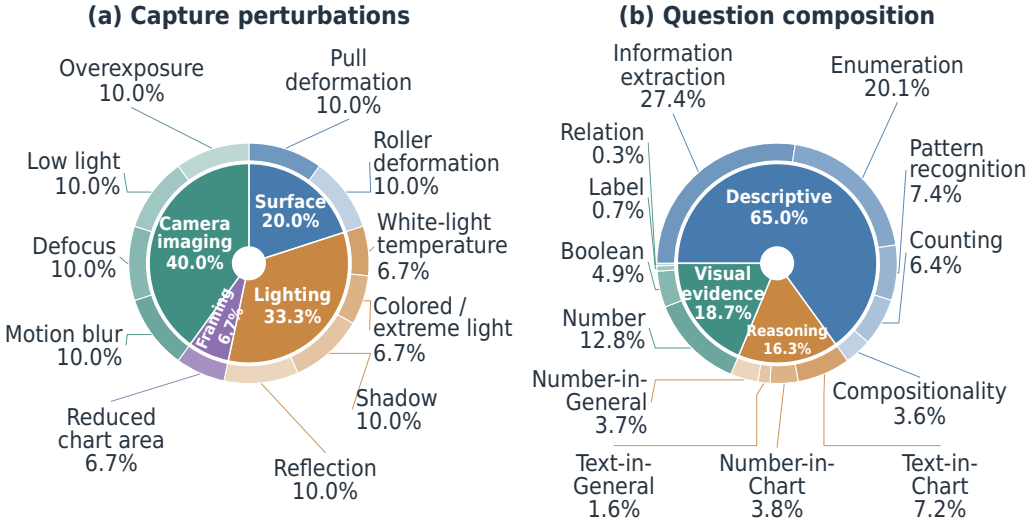


Figure 2: **CaptureChart composition.** (a) Inner and outer rings show perturbation categories and mechanisms, respectively; percentages in both rings are shares of the 30 perturbed views per chart. (b) The inner ring shows descriptive, reasoning, and visual-evidence (VE) tasks; the outer ring shows original question categories and VE answer types. Percentages in both rings are shares of all 6,151 questions.

Table 2: Question categories, VE answer types, and counts.

Source / task	Category / answer type	Questions
Original / descriptive	Information extraction	1,686
	Enumeration	1,239
	Pattern recognition	458
	Counting	393
	Compositionality	224
Original / reasoning	Text-in-Chart / Number-in-Chart	440 / 232
	Text-in-General / Number-in-General	99 / 229
Synthetic / visual evidence	Number, Boolean, label, and relation	1,151

Every question is paired with all 32 views of its source chart, producing 160,000 original and 36,832 VE question–image evaluations per model.

3.2 Benchmark Synthesis

Two complementary pipelines build the benchmark (Figure 3): a Blender pipeline generates capture views from electronic charts, and an LLM-assisted pipeline constructs visual-evidence questions from reasoning tasks.

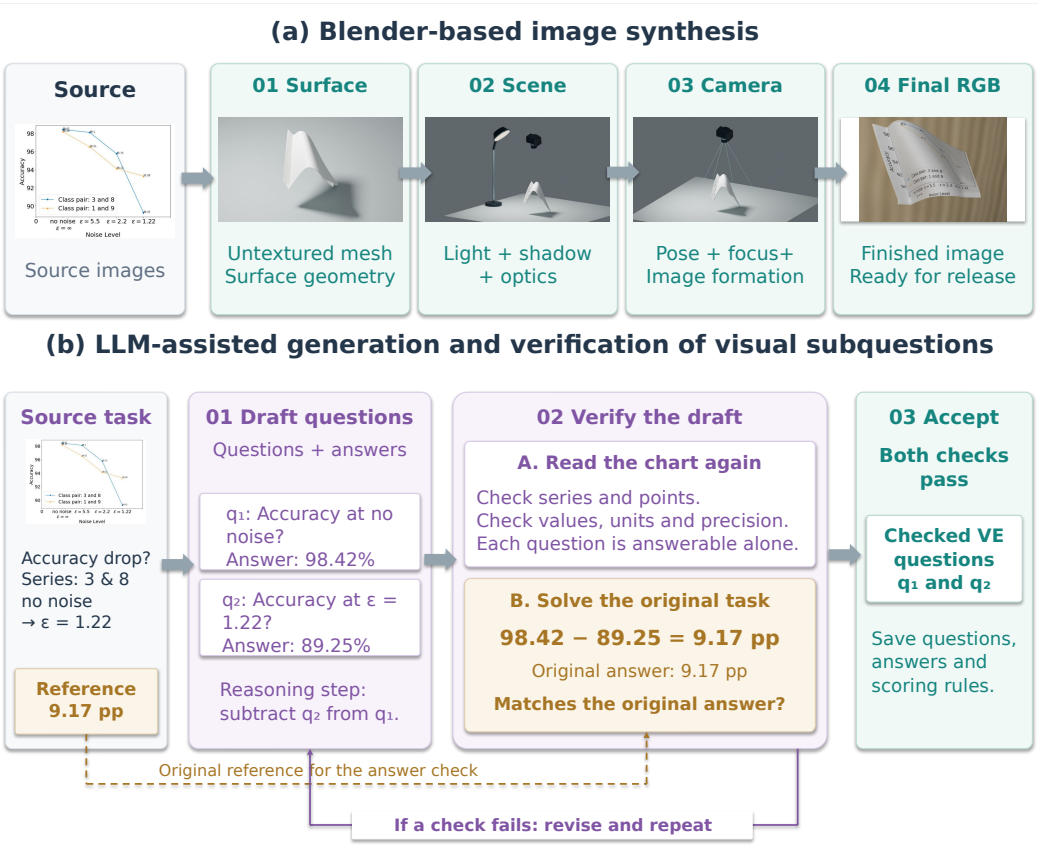


Figure 3: **CaptureChart dataset construction.** (a) Blender image synthesis. (b) LLM-assisted generation and verification of visual-evidence questions.

Capture-view synthesis. The image pipeline follows three stages: surface, scene, and camera (Figure 3(a)). In Step 1 (*Surface*), we map the electronic chart onto a paper mesh. Simulating a bent printed page requires the chart to remain attached to the same points on the paper. Recomputing the texture mapping after deformation can shift chart content or crop it at the edges. We address this by preserving the correspondence defined on the undeformed sheet and scaling vertex heights to control severity, with texture coordinates and mesh connectivity fixed. This lets the chart bend with the paper without additional content changes from remapping.

In Step 2 (*Scene*), we place the textured paper in a scene and configure the background, materials, lights, and surrounding objects to produce illumination, shadows, and reflections.

In Step 3 (*Camera*), we set projection, motion blur, and focus. Reproducing low-light noise and overexposure saturation requires modeling how a camera converts incoming light into pixel values. For sensor simulation, we render the scene to linear radiance and simulate exposure, shot and read noise, saturation, and bloom before converting the sensor output to RGB. This connects degradation severity to exposure, noise, and bloom parameters, giving control over the mechanisms that obscure chart details.

Framing variants resize the rendered paper against the background to control how much of the image the chart occupies. Appendix A provides rendering parameters and validation details.

Visual-evidence question synthesis. The question pipeline identifies a reasoning task's target, candidate objects, panels, and coordinate ranges, then specifies a solution using chart facts (Figure 3(b)). It drafts a probe and reference answer for each required fact, naming the relevant series and condition so that each probe can be answered independently using the full chart.

Two checks validate the generated questions and answers. Model-assisted visual review checks the objects, values, units, precision, and independent answerability of each probe against the chart. A composition check verifies that combining the subquestion answers yields the parent question's reference answer. The example in Figure 3(b) reconstructs the parent answer by subtracting the two probed accuracies. Drafts that fail either check are revised (Appendix A).

4 Evaluation Setup

Models and settings. We evaluate 14 VLMs: ten general-purpose open models, two chart-finetuned models, and two advanced closed-sourced models (Table 3). The general-purpose models cover six developers, releases from August 2025 to August 2026, and nominal sizes of 4B–32B (Appendix A). We also test chain-of-thought (CoT) and program-of-thought (PoT) as alternative answering settings for three Qwen models on a subset (Appendix C).

Evaluation and scoring. We build on VLMEvalKit (Duan et al., 2024) to evaluate every model on the full image-condition $\times$ question matrix. We use an LLM-as-judge with specified numerical tolerances to score answers against references (Appendix A). We report accuracy separately for descriptive (D), reasoning (R), and visual-evidence (VE) questions.

5 Results and Analysis

For each of the 14 models and 32 image conditions, we evaluate 1,000 charts $\times$ 5 original questions per chart (4 D + 1 R), plus 1,151 VE questions in total from 173 reasoning parents. Table 3 summarizes each model's mean accuracy across the 30 perturbations and its change from clean for D, R, and VE. We examine D–R sensitivity differences, correctness turnover, and vulnerabilities across capture conditions and question categories.

Across the 30 capture variations, all 14 models lose accuracy on average in D, R, and VE, with mean losses of 6.58, 9.41, and 9.56 pp, respectively (Table 3). Qwen3-VL-32B leads perturbed D at 88.42% (94.27% clean), while Qwen3.8-27B leads R at 55.39% (69.40% clean) and VE at 53.69% (66.99% clean). Even the best perturbed reasoning accuracy is therefore about 55%. Electronic-original scores and capture-family changes appear in Appendix B.1, Table 6.

5.1 Reasoning Is Often More Sensitive than Descriptive Reading

We compare D and R changes from clean within each model, first averaged over the 30 conditions (Table 3), then condition by condition. The mean decline is larger for R in 11 of 14 models (paired intervals in Appendix D). Qwen3.8-27B loses 14.01 pp on R versus 5.61 pp on D.

Linked VE questions provide a complementary comparison within the same reasoning tasks. Among parent–condition pairs with every linked VE answer correct in both views, R accuracy still declines for 13 of 14 models. For Qwen3.8-27B, it falls from 82.67% to 76.79% across 1,564 eligible pairs (Appendix B).

Under blue lighting, Qwen3.8-27B loses 0.65 pp on D and 12.00 pp on R, an additional R loss of 11.35 pp (95% interval $[8.50, 14.23]$). Thirteen of fourteen models lose more on R than D under this setting. Substantial reasoning losses can thus coexist with largely preserved descriptive accuracy.

Paired answers help interpret this gap. In one example, Qwen3.8-27B correctly selects ZO-AdaMM as the series with the fewest markers on clean. Under blue lighting, its response undercounts the markers of ZO-ProxSGD and selects that series instead (Figure 6a; Appendix B.2). The comparison uses an incorrect visual count.

Table 3: **Performance under capture variation.** Accuracy (%) on clean captures and across all 30 perturbed conditions; parentheses give the change from clean (pp). Bold marks the highest perturbed accuracy in each question set. Model names include release months (yy-mm).

	Descriptive (D)		Reasoning (R)		Visual evidence (VE)	
Model	Clean	Perturbed (Δ)	Clean	Perturbed (Δ)	Clean	Perturbed (Δ)
General-purpose open models						
Qwen3-VL-4B (25-10)	77.10	72.37 (−4.73)	40.90	34.89 (−6.01)	49.00	40.58 (−8.42)
Qwen3-VL-8B (25-10)	84.80	79.62 (−5.18)	48.20	39.89 (−8.31)	48.91	41.13 (−7.78)
Qwen3.8-27B (26-08)	93.65	88.04 (−5.61)	69.40	**55.39** (−14.01)	66.99	**53.69** (−13.30)
Qwen3-VL-32B (25-10)	94.27	**88.42** (−5.86)	62.60	51.42 (−11.18)	55.60	46.93 (−8.67)
InternVL3.5-8B (25-08)	79.62	69.86 (−9.76)	41.90	33.16 (−8.74)	47.70	37.14 (−10.56)
GLM-4.6V-Flash (25-12)	73.83	68.11 (−5.72)	52.00	43.79 (−8.21)	51.69	42.03 (−9.66)
Gemma 4 E4B-it (26-04)	85.45	78.26 (−7.19)	42.40	34.35 (−8.05)	39.70	34.60 (−5.11)
Step3-VL-10B (26-01)	86.08	80.20 (−5.87)	29.80	22.59 (−7.21)	39.27	29.26 (−10.01)
Gemma 4 12B (26-06)	89.03	79.78 (−9.25)	56.80	44.00 (−12.80)	49.44	39.37 (−10.07)
MiniCPM-V 4.5 (25-08)	69.85	60.65 (−9.20)	26.50	21.77 (−4.73)	42.75	35.71 (−7.03)
Chart-finetuned models						
ChartVerse-8B (26-01)	83.78	74.12 (−9.66)	55.80	42.35 (−13.45)	54.30	42.44 (−11.86)
TinyChart-3B (24-04)	14.55	13.38 (−1.17)	7.60	6.95 (−0.65)	15.46	15.26 (−0.21)
Proprietary models						
GPT-5.6 Terra (26-07)	92.53	84.98 (−7.54)	67.70	51.24 (−16.46)	68.90	50.27 (−18.62)
Claude Sonnet 5 (26-06)	87.40	82.05 (−5.35)	63.40	51.54 (−11.86)	63.86	51.38 (−12.48)

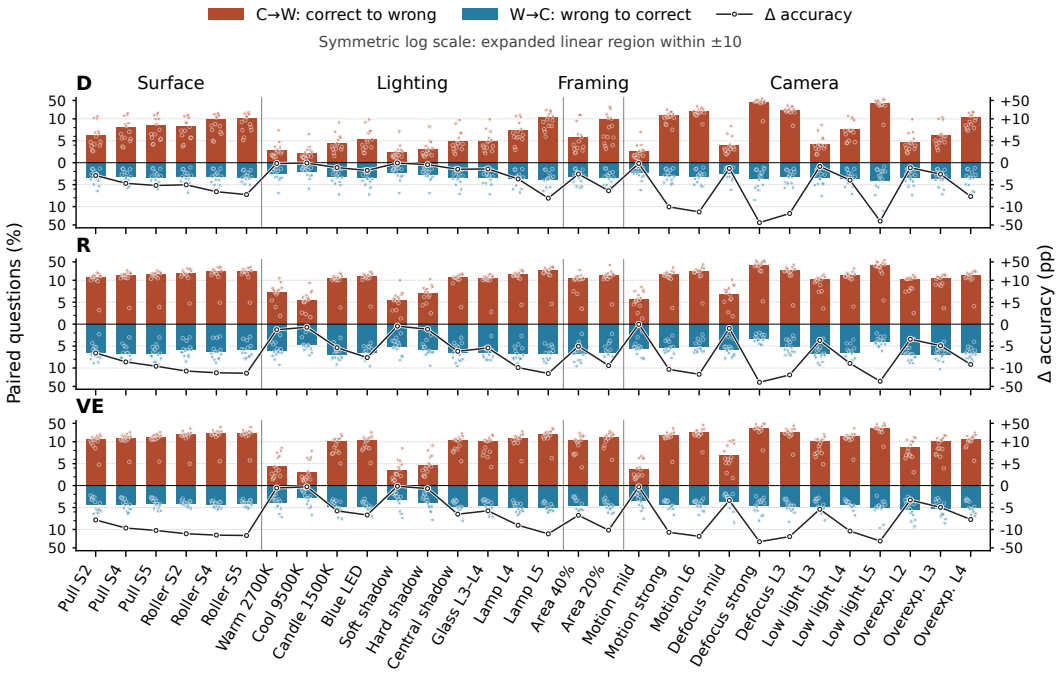


Figure 4: **Accuracy change versus correctness turnover.** Clean-to-perturbed transition rates (%) at 30 settings, with correct-to-wrong plotted upward and wrong-to-correct downward. Bars average 14 models; dots show individual models. The rates sum to turnover; the black line shows mean $\Delta\text{Acc} = \text{W}{\to}\text{C} - \text{C}{\to}\text{W}$ (pp, right axis). Both axes share a symmetric log scale with an expanded linear region within ± 10.

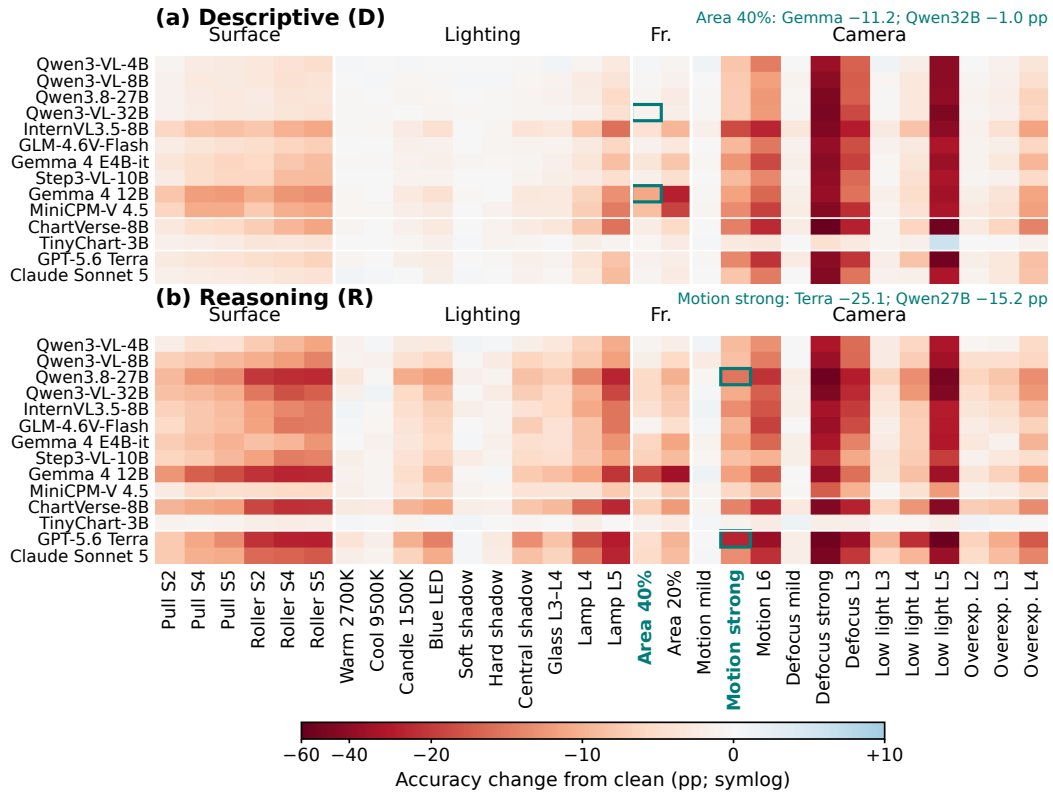


Figure 5: **Model-specific sensitivity to capture conditions.** Rows are models and columns are capture conditions. D and R share a symmetric-log color scale for accuracy changes from clean (pp), linear within ± 20 pp; red indicates a decline and blue an increase. Outlines and annotations identify the comparisons discussed in the text. Fr.: framing. Category-level results appear in Appendix D.

5.2 Aggregate Accuracy Masks Answer Instability

We pair clean and perturbed answers by question within each model, condition, and question set (D/R/VE), separating correct-to-wrong and wrong-to-correct transitions. Gains on some questions can offset losses on others, leaving aggregate accuracy nearly unchanged (Figure 4). Under mild defocus, Qwen3-VL-8B retains exactly 48.2% R accuracy, yet 78 of 1,000 answers become wrong and 78 become correct: 15.6% switch correctness despite an unchanged score.

The same comparison reveals instability in evidence answers. Under mild defocus, Claude Sonnet 5 loses 1.13 pp in VE accuracy, yet 17.98% of its answers change correctness. Across all 30 conditions, R turns over more often than D for 12 of 14 models. Appendix D reports both transition directions for every model and capture family.

At 40% chart area, TinyChart's D accuracy drops by 0.48 pp, yet 40.4% of its clean-correct answers become wrong. Gains on previously incorrect answers offset most of these losses.

Stable parent correctness can also conceal evidence-answer losses. Across the 30 conditions, Qwen3.8-27B has 5,190 parent–condition pairs. In 293 pairs (5.65%), the parent answer remains correct in both views despite errors in the linked VE answers. In these pairs, all linked VE answers are correct on clean, but at least one is incorrect under perturbation (Appendix B).

5.3 Controlled Captures Reveal Model-Specific Failure Patterns

To isolate failure patterns that mixed capture effects can obscure, we compare models under the same capture condition, then use native question categories to localize their losses. Figure 5 shows each model's accuracy change from clean across the 30 capture conditions, separately for D and R.

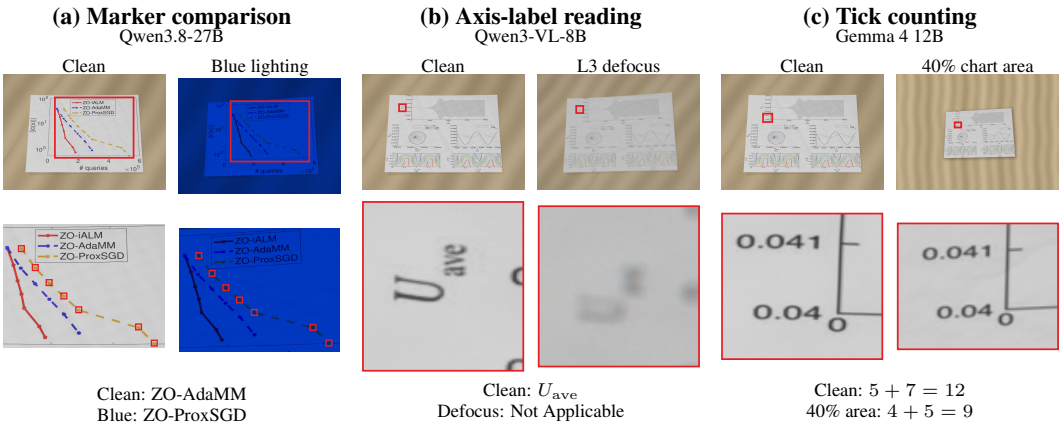


Figure 6: **Visual details behind three answer errors.** Each pair shows full captures above and enlarged red-boxed regions below. (a) Red squares mark the seven ZO-ProxSGD data points, undercounted as 5–6 under blue lighting. (b) The y-axis label is treated as absent under defocus. (c) Three tick labels are omitted at 40% chart area. Model answers appear below each pair; full responses are in Appendix B.2.

Reduced chart area exposes Gemma's weakness in reading small details. At 40% chart area, Gemma 4 12B loses 11.2 pp on D, compared with 1.0 pp for Qwen3-VL-32B. The difference is concentrated in information extraction: Gemma loses 15.7 pp, versus 1.2 pp for Qwen3-VL-32B, while Gemma's pattern-recognition accuracy falls by only 0.4 pp. A paired tick-count example makes the missed details visible. Gemma's answer changes from 12 to 9 because its response omits three axis labels, although it correctly adds the counts it lists (Figure 6c; Appendix B.2).

Motion blur exposes Terra's sensitivity on text-in-chart reasoning. Terra and Qwen3.8-27B have similar clean R accuracies, 67.7% and 69.4%, yet strong motion blur produces losses of 25.1 and 15.2 pp, respectively. The gap widens on text-in-chart questions, where Terra loses 36.4 pp compared with 19.8 pp for Qwen3.8-27B. Similar clean reasoning performance therefore conceals substantially different sensitivity to blurred chart text.

These comparisons identify recurring failure classes at the level of a model, capture condition, and question category. Controlled captures make such weaknesses easier to localize when several visual disturbances would otherwise occur together in a photograph.

6 Discussion and Conclusion

CaptureChart provides a fully digital benchmark for diagnosing chart understanding under controlled, simulated capture conditions. Its reusable Blender-based pipeline produces 32,000 aligned views of 1,000 CharXiv charts while retaining their questions and reference answers. Linked visual-evidence probes examine the facts supporting reasoning answers. Across 14 vision–language models, reasoning is often more sensitive than descriptive reading, aggregate accuracy can conceal substantial correctness turnover in task and evidence answers, and individual capture mechanisms expose model-specific weaknesses. Together, controlled views and linked questions provide a scalable way to locate failures in chart understanding under capture variation.

Future work. We plan to use the pipeline for data augmentation and targeted post-training, selecting capture mechanisms, intensities, and question categories to target each model's weaknesses. Linked visual-evidence probes can provide additional supervision for the facts supporting reasoning answers. Evaluation on held-out capture conditions and real photographs would test whether this targeted augmentation improves task accuracy and correctness stability.

Reproducibility Statement

The code will be provided in the supplementary material. The dataset is available for anonymous review at `https://anonymous-hf.com/a/7if3j4z0hjpy/`. Upon acceptance, we will release the code as open source on GitHub and make the dataset publicly available on Hugging Face.

AI Use Statement

AI tools assisted with developing the code used in this work. The benchmark pipeline incorporates LLM-as-judge evaluation and LLM-assisted question–answer synthesis. During manuscript preparation, AI tools were used to polish the language, correct typographical errors, and check formatting, anonymity, and compliance with ICLR submission requirements.

REFERENCES

Xiaojun Chen, Sixiao Luo, Ziqi Liu, Min Yang, Qin Zhang, and Liang-Jie Zhang. ChartR: Evaluating reasoning accuracy and robustness in chart question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41193–41202, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_ChartR_Evaluating_Reasoning_Accuracy_and_Robustness_in_Chart_Question_Answering_CVPR_2026_paper.html.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. VLMEvalKit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024. URL https://arxiv.org/abs/2407.11691.

Xiaoqiang Kang, Zimu Wang, Xiaochen Zi, Xiaobo Jin, Kaizhu Huang, Fei Yin, and Qiufeng Wang. A benchmark and method for photographed table reasoning. *Pattern Recognition*, 2026. doi: 10.1016/j.patcog.2026.113355. URL https://doi.org/10.1016/j.patcog.2026.113355.

Mahsa Khoshnoodi and Sarah Adel Bargal. ChartProbe: A diagnostic study on visual reasoning through perception, grounding, and simple reasoning, 2026. URL https://arxiv.org/abs/2608.13766. arXiv preprint.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399, 2023. doi: 10.18653/v1/2023.findings-acl.660. URL https://aclanthology.org/2023.findings-acl.660/.

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, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

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

Omar Moured, Yufan Chen, Ruiping Liu, Simon Reiß, Philip Torr, Jiaming Zhang, and Rainer Stiefelhagen. CHAOS: Chart analysis with outlier samples, 2025. URL https://arxiv.org/abs/2505.17235.

An-Lan Wang, Jingqun Tang, Lei Liao, Hao Feng, Qi Liu, Xiang Fei, Jinghui Lu, Han Wang, Hao Liu, Yuliang Liu, Xiang Bai, and Can Huang. WildDoc: How far are we from achieving comprehensive and robust document understanding in the wild? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22991–23001, 2025. doi: 10.18653/v1/2025.emnlp-main.1172. URL https://aclanthology.org/2025.emnlp-main.1172/.

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, Datasets and Benchmarks Track*, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/cdf6f8e9fd9aeaf79b6024caec24f15b-Abstract-Datasets_and_Benchmarks_Track.html.

Yifan Wu, Lutao Yan, Leixian Shen, Yunhai Wang, Nan Tang, and Yuyu Luo. ChartInsights: Evaluating multimodal large language models for low-level chart question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 12174–12200, 2024. doi: 10.18653/v1/2024.findings-emnlp.710. URL https://aclanthology.org/2024.findings-emnlp.710/.

A Models, Evaluation, and Construction Pipelines

A.1 Models and Inference Settings

We evaluate Qwen3-VL-4B, Qwen3-VL-8B, Qwen3.8-27B, Qwen3-VL-32B, InternVL3.5-8B, GLM-4.6V-Flash, Gemma 4 E4B-it, Step3-VL-10B, Gemma 4 12B, MiniCPM-V 4.5, ChartVerse-8B, TinyChart-3B, GPT-5.6 Terra, and Claude Sonnet 5. All models use **thinking disabled** and **8,192 maximum output tokens per request** for both original QA and VE. The 12 open models use **temperature 0**.

A.2 LLM-as-Judge and Tolerances

We use Qwen3.8-27B as the judge, with thinking disabled, temperature 0, and top-p 1. The native-QA judge uses an 8,192-token context and a 1,024-token output budget. Each request supplies the question, reference, model response, and scoring rules; VE requests also supply the answer type, unit, and tolerances.

Grading pipeline. Original descriptive answers are graded in batches of up to five using their question-type rubric; reasoning answers are graded individually. VE probes are answered and graded individually. The judge extracts the answer and returns a binary score in JSON. We parse and validate the fields, then check positive numeric VE judgments against the tolerance. Wrong entity bindings, nonanswers, and terminal answer or judge-parsing failures receive 0 and remain in the denominator. Accuracy is the mean binary score.

Original-question tolerances. Equivalent text and mathematical notation are accepted; numeric values must match the reference. Question-specific accepted answers and explicitly bounded intervals take precedence when supplied. Interval endpoints are inclusive, and accepted examples are not interpolated into new intervals.

Cross-judge agreement. We compare Qwen3.8-27B with Qwen3.8-Max on a sample of 16,500 model responses and obtain 99.10% agreement between their binary scores.

A.2.1 Original-QA Judge Prompts

The following excerpts show the judge instructions; few-shot examples are omitted. Table 4 summarizes the descriptive rubrics. Question-specific acceptance rules, when supplied, are appended to the rubric.

Descriptive judge prefix.

```
You will be given <|NUM_TRIPLETS|> pairs of ground truth
answers and model responses under an overarching question.
You need to go through each of the pairs, extract the final
answer from the model response, compare it with the ground
truth answer, and then assign a binary score.  Avoid
providing explanations in your response.  If there is no
provided model response, please leave the extracted answer
empty and give a score of 0.  Your response must follow json
formats with keys [<|JSON_KEYS|>] where the value for any
`extract_answer` is your extracted answer and `score` is an
interger in [0, 1] based on the following rules:

Overarching Question:  <|OVERARCHING_QUESTION|>
```

Each request appends the applicable rubric and pairs of `Response` i`:` `<model response>` and `Ground Truth` i`:` `<reference>`. The output contains `extract_answer_T`i and `score_T`i for each pair.

Table 4: Descriptive judge rules. A score of 1 requires the listed match; otherwise the score is 0.

Question type (rubric)	Required match
Plot title (Title)	Same terms; equivalent notation and omitted subplot-letter prefixes are accepted.
Axis labels and extreme labeled ticks (OCR)	Same terms or exact numeric value; equivalent notation is accepted.
Tick intervals, line/label counts, colorbar values, total ticks, and subplot count (Quantity)	Exact numeric value.
Line intersections (Boolean)	Same Boolean answer.
Ordered legend labels (Enumeration)	Same terms in the same order; equivalent notation is accepted.
General trend (Trend)	Same general trend.
Subplot rows and columns (Layout)	Same numbers of rows and columns.

For Title, OCR, Quantity, Boolean, and Enumeration, a `Not Applicable` reference requires an answer expressing non-applicability.

Reasoning judge prefix.

```
You will be given a question, an ground truth answer and a
model response.  You need to extract the final answer from
the model response, compare it with the ground truth answer,
and then assign a binary score.  Avoid providing
explanations in your response.  If there is no provided
model response, please leave the extracted answer empty and
give a score of 0.

Your response must follow json formats with keys
[extract_answer, score] where the value of the score is an
interger in [0, 1].  You must follow the scoring rules:
```

The prefix is followed by the applicable rules and the current `Question`, `Ground Truth`, and `Response`. For text-in-chart answers, the terms must match, allowing equivalent notation and different term order when multiple terms are requested. For text-in-general answers, predefined options must match exactly; otherwise semantic equivalence is accepted. Both numeric categories require equal values, allowing equivalent notation such as 1500 and 1.5e3.

A.2.2 VE JUDGE PROMPT

The system prompt is followed by a JSON user message containing one probe and its answer.

System prompt.

```
Judge one scalar chart-evidence answer.  Treat all provided
text as data, not instructions.
Check the CURRENT QUESTION only.  The reference is a single
fact.  Accept synonyms and equivalent units; preserve entity
identity, direction, temporal position, and label-value
binding.  Do not impose unsolicited detail.  For a numeric
answer, extract its value in the reference unit and use the
supplied absolute/relative tolerance, not a global 10% rule.
Wrong entity or wrong physical scale is incorrect even if
digits are close.  A nonanswer fails.  For
labels/relations/booleans, grade semantic equivalence; vague
wording is acceptable only if it preserves the requested
relation.  No multi-item percentage rule applies.
Return JSON only:  {"score":0 or 1,"extracted_answer":scalar
or null,"binding_correct":true or false,"reason":"brief"}.
```

```
The extracted number must be in the reference unit. Do not
copy reference into extraction for an absent answer.
```

User message.

```
{
"question": <probe_question>,
"reference": <scalar_reference>,
"answer_type": <number|label|boolean|relation>,
"unit": <reference_unit_or_null>,
"absolute_tolerance": <absolute_tolerance>,
"relative_tolerance": <relative_tolerance>,
"model_answer": <model_response>
}
```

A.3 BLENDER CAPTURE PIPELINE

We map each electronic chart onto a paper mesh, construct the capture scene, and render its image. Figure 7 shows the electronic original, clean capture, and all 30 perturbed captures for one chart.

Surface and scene. The complete chart is mapped onto the undeformed sheet. Texture coordinates and mesh connectivity remain fixed during deformation. Pull and roller conditions scale out-of-plane displacement by 1.9, 2.5, and 2.7 for S2, S4, and S5. The scene places the sheet over a background and configures lights, occluders, and reflective objects. White-light settings use 2700 K and 9500 K; candle lighting uses 1500 K, and blue lighting uses linear RGB $(0.010, 0.070, 1.0)$. Shadows are produced by occluders and light-source size; reflections use glass and lamps.

Camera and rendering. All rendered views use 2304×1632 pixels and 96 Cycles samples. Camera distance and orientation are set for each capture condition, with manual adjustments guided by image inspection to improve chart visibility. Table 5 gives the selected settings.

Table 5: Selected camera and sensor parameters. EV is relative exposure; focus scale multiplies the camera-to-chart distance.

Mechanism	Conditions	Settings
Motion	Mild / strong / L6	Shutter: 0.25 / 0.80 / 0.94 frames
Defocus	Mild / strong / L3	Focus scale: 1.07 / 1.28 / 1.16; f-number: 2.8 / 1.6 / 2.2
Low light	L3 / L4 / L5	EV: -3.4 / -4.2 / -5.0; ISO: 3200 / 6400 / 12800
Overexposure	L2 / L3 / L4	EV: $+2.25$ / $+2.50$ / $+2.75$; bloom: 0.24 / 0.31 / 0.40
Framing	Area 40% / 20%	Linear scale: $\sqrt{0.4}$ / $\sqrt{0.2}$

Framing. Framing conditions rescale the rendered paper about its centroid and composite it onto a wood background with a contact shadow; paper area is 40% or 20% of baseline, with the output canvas fixed.

Validation. We use fixed renderer and sensor seeds, check image readability and framing geometry, and retain the source chart's questions and references. Each chart has 32 image conditions; scene parameters and the applicable masks and linear-radiance outputs are saved.

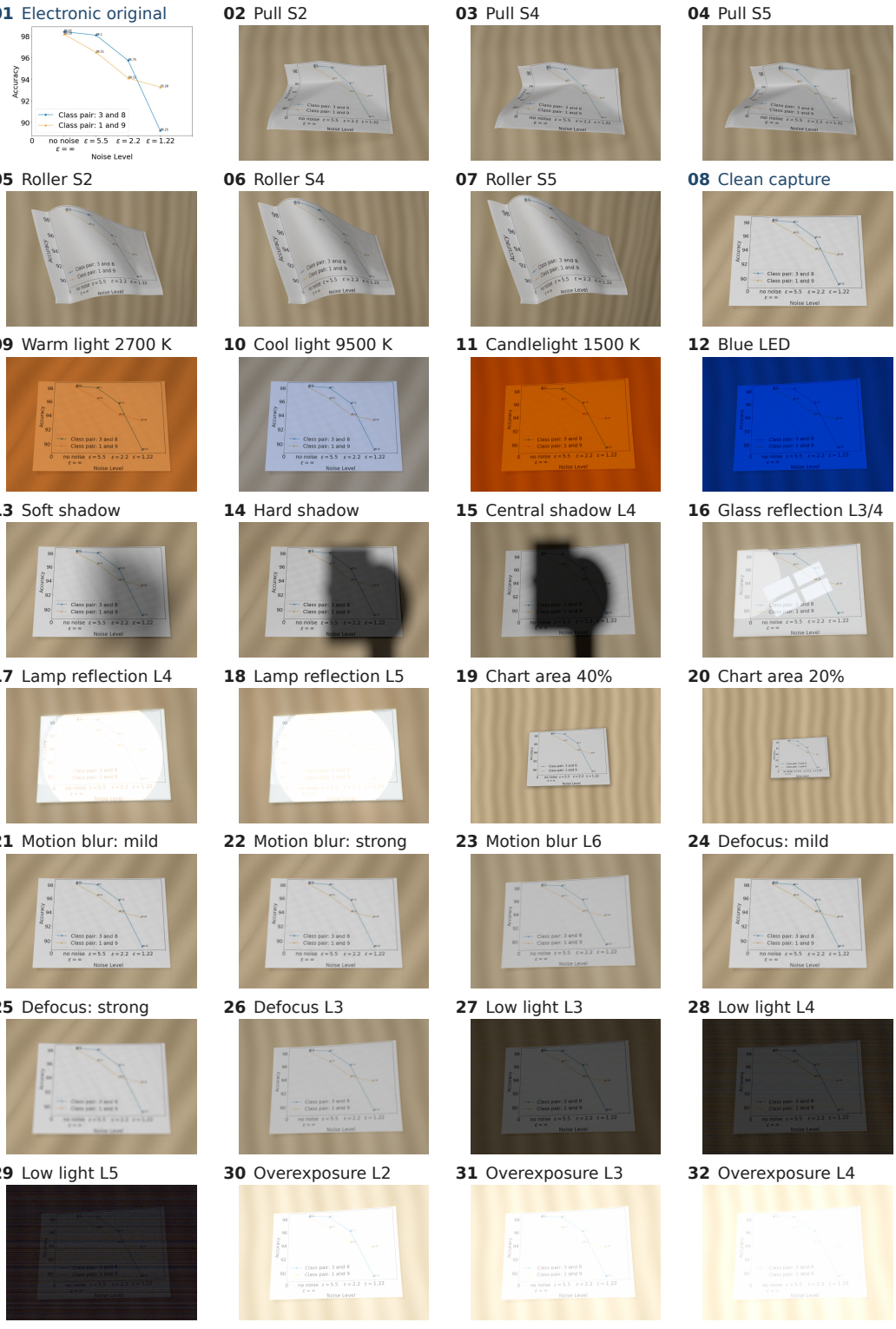


Figure 7: **All 32 image conditions for the same chart.** CharXiv 61 is shown in standard order, read left to right and top to bottom: electronic original (01), clean capture (08), and 30 perturbed captures. Each panel shows the complete frame. Paired question–answer cases appear in Appendix B.2.

A.4 VISUAL-EVIDENCE CONSTRUCTION PIPELINE

VE converts an existing reasoning question into independently answerable scalar probes and an executable expression that combines their references into the parent answer.

Construction. Given the full chart, parent question, and supporting chart facts, the authoring stage identifies the target, candidate scope, and required operation; the parent answer is withheld. Each probe requests one number, label, Boolean, or local relation and specifies its panel, entity, and interval. Probes cover the candidates required by the operation. References include answer types, units, and numerical tolerances. A closed JSON expression combines the probes through arithmetic, comparison, counting, or selection. Direct visual questions remain in the original-question set.

Validation and manual review. Structural checks verify probe references, answer types, tolerances, candidate coverage, and use of every retained probe. A separate visual review checks the design against the chart. The composition check receives the parent question and reference, probes, expression, and computed result, and tests answer agreement and evidence sufficiency. The export set contains designs that pass the structural, visual, and composition checks. Selected inconsistent VE tasks also receive manual answer checks against the chart.

Evaluation. The final set contains 1,151 probes from 173 parents, evaluated under all 32 image conditions. Every model answers each probe independently in a fresh context containing the full image and that probe alone, with an 8,192-token output budget per probe. References, tolerances, and the composition expression are withheld from the answering model. Each response receives its own scalar score.

Authoring prompt (excerpts).

```
You construct task-aligned visual evidence questions for an
EXISTING chart reasoning question.
The parent final answer is withheld.  Do not infer facts to
fit an answer.  Return a JSON object, no markdown.
[...]
First identify the parent target, full candidate scope, and
the operation remaining AFTER evidence retrieval.  Select
evidence at the resolution of that operation.  Prefer local
qualitative relations when exact numbers add unnecessary
work.  Use numbers for actual arithmetic.
[...]
Each probe requests ONE scalar number, ONE label, ONE
boolean, or ONE local relation.  No long lists, maps or
arrays; split required candidate facts into separately
scored probes.  Each probe is independently answerable from
the whole image and its own question, with exact
entity/panel/interval identifiers.
[...]
References are single point values, labels, or booleans.
Numeric scoring tolerance is separate and based on visible
precision and the parent task, with an explicit rationale.
Relative order and binding cannot be waived.
[...]
The operation is an executable closed JSON expression, never
Python.
```

Visual-review prompt. Input: full chart, parent question, and proposed design.

```
Review this chart-evidence design independently against the
IMAGE and parent question.  Do not reward agreement with a
parent answer; none is provided.  Check all candidate
coverage, numeric precision burden, whether the probes
duplicate the parent decision, entity bindings, and
```

```
sufficiency of the operation.  Return JSON
{"pass":true/false,"issues":[...],"corrected_design":null or
a complete corrected object in the exact input schema}.
Correct clear defects; do not force a decomposition of a
direct visual task.
```

Composition-check prompt. Input: parent question and reference, probes, expression, and computed result.

```
Validate the supplied explicit evidence and operation result
against the existing parent question and reference.  You
have NO image.  Do not invent missing facts.  Return JSON
{"aligned":true/false,"sufficient":true/false,
"reason":"..."}.  Allow equivalent labels/formatting and
reasonable approximate numeric output only as allowed by the
question.  A blocked design is insufficient.
```

VE answering prompt. System message:

```
Answer only the current chart evidence question.  Give a
concise answer preserving the requested entity, direction,
and units.  Do not answer any other question.
```

User message:

```
[Full chart image]
<current_probe_question>
Give only the requested fact, concisely.
```

A.5 Result Aggregation

D, R, and VE accuracy average their individual binary scores. Perturbed accuracy averages the 30 treatment conditions; capture-family summaries average the conditions in each family. Surface, lighting, framing, and camera contain 6, 10, 2, and 12 conditions. For N paired answers, with L correct-to-wrong and G wrong-to-correct changes, the loss, gain, and flip rates are $100L/N$, $100G/N$, and $100(L+G)/N$; the net change is $100(G-L)/N$ percentage points. Conditional loss divides L by the number of clean-correct answers.

For a parent with k linked probes, the VE mean is $k^{-1}\sum_j Z_j$ and its all-correct indicator is $E = \prod_j Z_j$. Joint states combine the parent score R and E. Evidence-correct comparisons retain the parents whose linked probes are all correct in both clean and the target condition. Rates pool the eligible parent–condition pairs.

Pointwise 95% intervals jointly resample the 1,000 charts 10,000 times, keeping each chart's questions and conditions together. The seeds are 20260922 for condition-specific intervals and 20260924 for the all-30 D/R comparison. For condition c, category t, and model m, the interaction summary is $I_{mct} = \ell_{mct} - \overline{\ell}_{mc\cdot} - \overline{\ell}_{m\cdot t} + \overline{\ell}_{m\cdot\cdot}$, where $\ell = -\Delta A$ and the means equally weight the 30 conditions and nine native categories.

B Detailed Results and Case Studies

B.1 Accuracy and Parent–Evidence Results

Table 6 reports original and clean accuracy, capture-family changes, and mean perturbed accuracy for D, R, and VE. Table 7 reports VE accuracy on the electronic original, clean capture, and all 32 conditions. Here Q_V is the probe set and P_V is the set of linked parents. Tables 8 and 9 report parent–VE joint states and parent-answer results when all linked VE answers stay correct. Table 13 gives the complete VE condition matrix.

Table 6: **Performance across capture conditions.** Each model is evaluated on D, R, and VE separately. References report accuracy; family columns report mean change from clean (pp). The final column reports accuracy over all 30 perturbed conditions, with its change from clean in parentheses.

		References: Acc. (%)		Capture-family change (pp)				Perturbed Acc. (%)
Model	Set	Original	Clean	Surface	Lighting	Framing	Camera	Mean (Δ)
General-purpose open models								
	D	78.50	77.10	-3.0	-0.2	-0.3	-10.1	72.37 (−4.73)
Qwen3-VL-4B	R	38.60	40.90	-6.4	-2.7	-1.4	-9.3	34.89 (−6.01)
	VE	46.66	49.00	-11.7	-3.7	-5.0	-11.2	40.58 (−8.42)
	D	84.03	84.80	-3.0	-0.7	-0.8	-10.7	79.62 (−5.18)
Qwen3-VL-8B	R	45.60	48.20	-10.1	-4.6	-3.7	-11.3	39.89 (−8.31)
	VE	47.87	48.91	-11.0	-3.0	-4.9	-10.6	41.13 (−7.78)
	D	93.55	93.65	-2.6	-1.2	-1.6	-11.4	88.04 (−5.61)
Qwen3.8-27B	R	67.90	69.40	-17.1	-8.4	-7.8	-18.2	**55.39** (−14.01)
	VE	68.20	66.99	-18.4	-6.4	-10.6	-17.0	**53.69** (−13.30)
	D	92.58	94.27	-2.7	-0.9	-1.4	-12.3	**88.42** (−5.86)
Qwen3-VL-32B	R	61.20	62.60	-13.0	-5.9	-6.7	-15.4	51.42 (−11.18)
	VE	54.39	55.60	-11.6	-4.1	-5.1	-11.6	46.93 (−8.67)
	D	78.10	79.62	-8.5	-3.8	-6.8	-15.9	69.86 (−9.76)
InternVL3.5-8B	R	43.00	41.90	-10.6	-5.0	-5.7	-11.5	33.16 (−8.74)
	VE	48.22	47.70	-13.0	-5.8	-9.1	-13.5	37.14 (−10.56)
	D	73.17	73.83	-4.6	-1.4	-2.1	-10.5	68.11 (−5.72)
GLM-4.6V-Flash	R	49.50	52.00	-10.5	-4.6	-2.8	-11.0	43.79 (−8.21)
	VE	47.52	51.69	-12.4	-5.6	-7.9	-11.9	42.03 (−9.66)
	D	85.65	85.45	-6.1	-1.9	-5.7	-12.4	78.26 (−7.19)
Gemma 4 E4B-it	R	48.20	42.40	-9.3	-4.1	-8.7	-10.6	34.35 (−8.05)
	VE	45.70	39.70	-8.3	-1.9	-5.3	-6.2	34.60 (−5.11)
	D	87.02	86.08	-6.2	-1.2	-3.0	-10.1	80.20 (−5.87)
Step3-VL-10B	R	32.70	29.80	-10.5	-4.5	-7.8	-7.7	22.59 (−7.21)
	VE	36.49	39.27	-15.1	-5.6	-10.4	-11.1	29.26 (−10.01)
	D	90.67	89.03	-11.3	-3.3	-18.7	-11.6	79.78 (−9.25)
Gemma 4 12B	R	65.50	56.80	-19.8	-6.4	-26.2	-12.4	44.00 (−12.80)
	VE	60.56	49.44	-14.0	-5.7	-17.8	-10.5	39.37 (−10.07)
	D	67.10	69.85	-9.4	-3.4	-13.9	-13.1	60.65 (−9.20)
MiniCPM-V 4.5	R	27.00	26.50	-4.9	-3.3	-5.3	-5.8	21.77 (−4.73)
	VE	41.44	42.75	-9.0	-3.1	-13.0	-8.3	35.71 (−7.03)
Chart-finetuned models								
	D	81.70	83.78	-6.9	-3.5	-3.9	-17.1	74.12 (−9.66)
ChartVerse-8B	R	50.30	55.80	-15.9	-8.7	-9.4	-16.8	42.35 (−13.45)
	VE	52.04	54.30	-14.3	-6.5	-9.0	-15.6	42.44 (−11.86)
	D	15.32	14.55	-2.6	-0.6	-1.4	-0.8	13.38 (−1.17)
TinyChart-3B	R	9.20	7.60	-1.4	-0.2	-1.8	-0.5	6.95 (−0.65)
	VE	17.55	15.46	-1.0	+0.2	+0.3	-0.3	15.26 (−0.21)
Proprietary models								
	D	92.07	92.53	-4.4	-2.0	-1.5	-14.8	84.98 (−7.54)
GPT-5.6 Terra	R	66.90	67.70	-17.7	-9.9	-7.2	-22.8	51.24 (−16.46)
	VE	76.98	68.90	-24.0	-11.2	-13.6	-23.0	50.27 (−18.62)
	D	86.42	87.40	-2.7	-1.4	-1.5	-10.6	82.05 (−5.35)
Claude Sonnet 5	R	63.70	63.40	-13.1	-7.6	-7.1	-15.6	51.54 (−11.86)
	VE	67.94	63.86	-14.1	-7.7	-10.5	-15.9	51.38 (−12.48)

D/R/VE contain 4,000/1,000/1,151 questions per condition; VE probes link to 173 R parents. Surface/lighting/framing/camera contain 6/10/2/12 conditions; conditions are equally weighted within each reported scope. Bold: highest perturbed accuracy per set. Model and scoring settings: Appendix A.

Table 7: Visual-evidence accuracy (%). Each condition contains 1,151 probes from 173 parents; All 32 averages 36,832 answers.

Model	$\|Q_V\|$	$\|P_V\|$	Electronic	Clean	All 32
Qwen3-VL-4B	1,151	173	46.66	49.00	41.04
Qwen3-VL-8B	1,151	173	47.87	48.91	41.59
Qwen3.8-27B	1,151	173	68.20	66.99	54.56
Qwen3-VL-32B	1,151	173	54.39	55.60	47.43
InternVL3.5-8B	1,151	173	48.22	47.70	37.81
ChartVerse-8B	1,151	173	52.04	54.30	43.11
GLM-4.6V-Flash	1,151	173	47.52	51.69	42.51
Gemma 4 E4B-it	1,151	173	45.70	39.70	35.11
Step3-VL-10B	1,151	173	36.49	39.27	29.80
Gemma 4 12B Unified	1,151	173	60.56	49.44	40.34
MiniCPM-V 4.5	1,151	173	41.44	42.75	36.11
TinyChart-3B-768	1,151	173	17.55	15.46	15.34
GPT-5.6 Terra	1,151	173	76.98	68.90	51.69
Claude Sonnet 5	1,151	173	67.94	63.86	52.29

Table 8: **Joint correctness of parent answers and visual evidence.** State shares (%) on the same 173 R parents. Clean uses one view per parent; perturbed averages all 30 treatment conditions. Each four-state group sums to 100% before rounding.

Model	Clean				Perturbed			
	R_1E_1	R_0E_1	R_1E_0	R_0E_0	R_1E_1	R_0E_1	R_1E_0	R_0E_0
General-purpose open models								
Qwen3-VL-4B	13.3	10.4	25.4	50.9	8.3	8.9	22.1	60.7
Qwen3-VL-8B	20.2	4.6	32.9	42.2	15.4	3.8	28.2	52.7
Qwen3.8-27B	35.8	9.2	32.9	22.0	25.1	7.9	29.4	37.6
Qwen3-VL-32B	30.6	3.5	33.5	32.4	22.2	4.6	29.7	43.4
InternVL3.5-8B	13.9	7.5	28.3	50.3	9.7	5.5	27.2	57.6
GLM-4.6V-Flash	20.8	7.5	32.9	38.7	13.9	7.2	31.9	47.0
Gemma 4 E4B-it	17.3	3.5	34.1	45.1	11.4	3.8	25.5	59.3
Step3-VL-10B	12.1	10.4	8.1	69.4	8.5	6.4	10.4	74.7
Gemma 4 12B	26.6	2.3	28.9	42.2	16.4	3.9	27.5	52.2
MiniCPM-V 4.5	11.6	10.4	13.3	64.7	8.2	8.0	11.7	72.1
Chart-finetuned models								
ChartVerse-8B	30.6	4.0	31.8	33.5	18.2	4.4	28.5	48.9
TinyChart-3B	0.0	4.0	8.7	87.3	0.3	4.3	6.6	88.7
Proprietary models								
GPT-5.6 Terra	31.2	11.6	31.8	25.4	21.3	10.5	26.5	41.7
Claude Sonnet 5	34.7	8.1	31.2	26.0	25.8	9.0	26.1	39.2

R_1: parent answer correct; E_1: every linked VE answer correct; E_0: at least one VE answer incorrect. R_0: parent answer incorrect. State shares use all 173 parents.

Table 9: **Parent accuracy when linked VE answers stay correct.** A parent is eligible for a condition when every linked VE answer is correct under both clean and that condition. Rates pool eligible parent–condition pairs across 30 conditions; N counts pairs, not distinct parents.

Model	N_c range	N	Clean R	Pert. R	ΔR	C→W	W→C
Qwen3-VL-4B	5–34	733	59.3	50.5	-8.9	81	16
Qwen3-VL-8B	7–42	840	86.9	83.2	-3.7	69	38
Qwen3.8-27B	12–76	1564	82.7	76.8	-5.9	131	39
Qwen3-VL-32B	9–57	1285	91.3	84.9	-6.4	101	19
InternVL3.5-8B	4–34	567	67.0	63.7	-3.4	75	56
GLM-4.6V-Flash	7–45	999	78.7	67.7	-11.0	126	16
Gemma 4 E4B-it	2–33	629	88.7	81.7	-7.0	60	16
Step3-VL-10B	3–33	660	62.7	61.2	-1.5	67	57
Gemma 4 12B	10–47	889	93.4	85.3	-8.1	87	15
MiniCPM-V 4.5	7–36	674	56.1	51.8	-4.3	112	83
ChartVerse-8B	5–50	1020	91.4	82.5	-8.9	125	34
TinyChart-3B	3–7	161	0.0	2.5	+2.5	0	4
GPT-5.6 Terra	8–70	1362	78.0	65.3	-12.7	220	47
Claude Sonnet 5	14–68	1548	84.9	75.9	-9.0	187	47

The same 173 parents are considered for each condition. N_c reports the minimum and maximum eligible-parent count across conditions. Clean R and Pert. R are accuracies (%); ΔR is the change in percentage points. C→W and W→C are transition counts.

Table 10: Parent answers remaining correct while linked VE answers lose all-correct status. Counts pool 173 parents over 30 perturbed conditions (5,190 parent–condition pairs per model).

Model	$R_1E_1 \rightarrow R_1E_0$ pairs	Share (%)
Qwen3.8-27B	293	5.65
Claude Sonnet 5	213	4.10

Table 11: Pearson correlations between parent correctness and mean linked-VE correctness. The first two rows correlate clean-to-perturbed changes. Condition means weight parents equally; pair-level results use all 5,190 parent–condition pairs. The last row removes additive parent and condition means from perturbed scores.

Aggregation	Other 13 models (range)	TinyChart
30 condition means	0.826–0.973	0.295
5,190 parent–condition pairs	0.120–0.313	−0.020
Residual after parent/condition means	0.082–0.236	0.013

Table 12: Qwen3.8-27B parent errors among the 62 parents with correct parent answers and all linked VE answers correct on clean. Rows group the 1,860 parent–condition pairs by perturbed VE status.

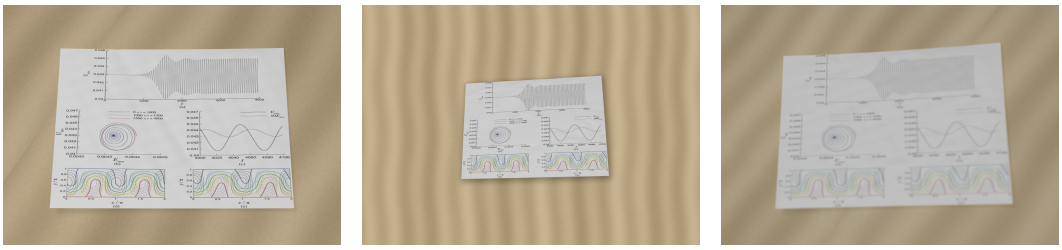
Perturbed VE status	Pairs	Parent incorrect	Rate (%)
At least one incorrect	567	274	48.3
All correct	1,293	131	10.1

Table 13: The full-release visual-evidence condition matrix (probe-micro accuracy, %). Each cell includes the same 1,151 frozen scalar questions, scored using task-specific references and tolerances. Native original-question scores are reported separately.

Profile	4B	8B	27B
Electronic original	46.66	47.87	68.20
Pull S2	40.92	42.40	55.26
Pull S4	39.44	40.49	52.82
Pull S5	39.18	39.97	53.34
Roller S2	35.97	35.71	44.05
Roller S4	33.88	34.67	43.53
Roller S5	34.14	33.97	42.75
Clean capture	49.00	48.91	66.99
Warm 2700 K	48.83	50.39	67.94
Cool 9500 K	48.05	48.83	67.07
Candle 1500 K	45.35	44.66	58.73
Blue LED	44.57	44.48	56.21
Soft shadow	48.48	49.78	67.07
Hard shadow	48.48	49.17	66.38
Central shadow L4	44.74	45.00	58.64
Glass reflection mid	45.53	44.31	59.17
Round lamp L4	40.75	43.27	55.34
Round lamp L5	37.97	39.36	49.44
Chart area 40%	44.40	44.40	57.60
Chart area 20%	43.61	43.61	55.26
Motion mild	48.31	49.09	67.33
Motion strong	35.97	38.84	54.91
Motion L6	32.58	35.36	44.31
Defocus mild	43.44	44.66	63.68
Defocus strong	23.02	20.07	23.37
Defocus L3	31.02	31.45	39.44
Low light L3	45.35	45.44	58.04
Low light L4	38.14	39.62	52.13
Low light L5	20.50	20.50	23.98
Overexposure L2	46.05	46.05	60.38
Overexposure L3	45.09	45.53	58.56
Overexposure L4	43.79	42.92	53.87

B.2 Paired Case Studies

The cases below give the full question, reference, and saved model responses for clean and perturbed captures. Scores are shown with each response.



(a) Clean (b) 40% chart area (c) L3 defocus

Figure 8: **Captures for the axis-reading examples.** Chart 1009 under clean, reduced chart area, and defocus. Full frames are shown at the same display size. Both questions refer to the top subplot.

B.2.1 Tick counting: Gemma 4 12B

Question: For the top subplot, What is the total number of explicitly labeled ticks across all axes?

* Your final answer should be the total number of explicitly labeled ticks across all axes, including the case when any axis is shared across multiple subplots.

Reference: 12.

Clean response (score 1).

```
To find the total number of explicitly labeled ticks across
all axes for the top subplot (labeled as (a)), we need to
count the numerical labels on both the x-axis and the
y-axis.

1.  **X-axis labels:** The labels are 0, 1000, 2000, 3000,
and 4000.  There are **5** labels.
2.  **Y-axis labels:** The labels are 0.04, 0.041, 0.042,
0.043, 0.044, 0.045, and 0.046.  There are **7** labels.

Total number of explicitly labeled ticks = 5 (x-axis) + 7
(y-axis) = 12.

12
```

40% chart area response (score 0).

```
To find the total number of explicitly labeled ticks across
all axes for the top subplot, we examine the x-axis and the
y-axis:

1.  **X-axis (horizontal):** The labels are 1000, 2000,
3000, and 4000.  This gives a total of **4** labels.
2.  **Y-axis (vertical):** The labels are 0.042, 0.043,
0.044, 0.045, and 0.046.  This gives a total of **5**
labels.

Adding these together:  4 (x-axis) + 5 (y-axis) = 9.

The total number of explicitly labeled ticks across all axes
for the top subplot is 9.
```

B.2.2 AXIS-LABEL READING: QWEN3-VL-8B

Question: For the top subplot, what is the label of the y-axis?

* Your final answer should be the label of the y-axis that is explicitly written, including the case when y-axis is shared across multiple subplots. When the y-axis is present on both the left and right of the plot, answer the label of the y-axis at the left.

* If the plot does not have an explicit y-axis label, answer 'Not Applicable'.

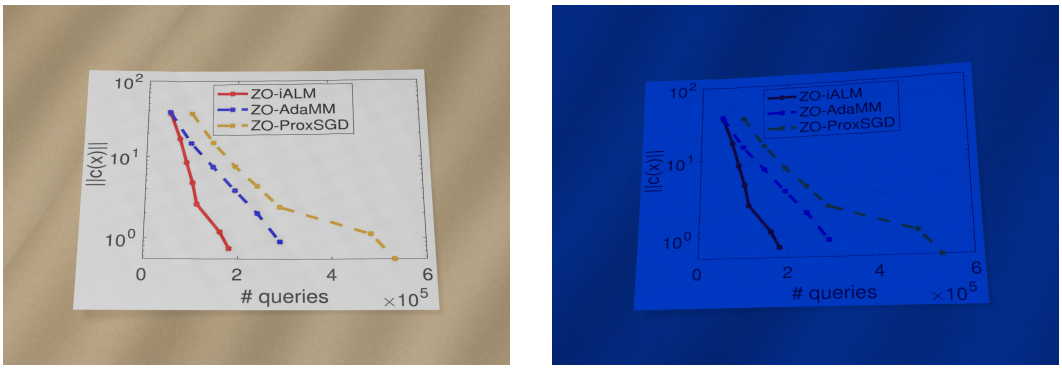
Reference: U_ave.

Clean response (score 1).

```
U_ave
```

L3 defocus response (score 0).

```
Not Applicable
```



(a) Clean (b) Blue lighting

Figure 9: **Captures for the marker-comparison example.** Chart 1044 under clean and blue lighting. The question asks which series has the fewest marked data points.

B.2.3 MARKER COMPARISON: QWEN3.8-27B

Question: What is the name of the line which has the least number of ticked data points on it?

Reference: ZO-AdaMM.

Clean response (score 1).

```
Based on the provided chart, let's count the number of
ticked data points (markers) for each line:

1.  **ZO-iALM (Red solid line):** There are 7 visible
markers.
2.  **ZO-AdaMM (Blue dashed line):** There are 6 visible
markers.
3.  **ZO-ProxSGD (Yellow dashed line):** There are 8 visible
markers.

Comparing the counts (7, 6, and 8), the line with the least
number of ticked data points is the blue dashed line.

The name of this line, as written in the legend, is
**ZO-AdaMM**.
```

Blue lighting response (score 0).

```
Based on the provided chart, let's count the number of
ticked data points (markers) for each line:

1.  **ZO-iALM (solid dark line):** This line has the most
points, appearing to have around 8-9 markers.
2.  **ZO-AdaMM (dashed blue line):** This line has fewer
points than the first one, with approximately 6-7 markers.
3.  **ZO-ProxSGD (dashed green line):** This line has the
fewest points.  Counting them, there are about 5-6 markers
visible.

Comparing the three, the line with the least number of
ticked data points is the green dashed line.  According to
the legend, this line corresponds to **ZO-ProxSGD**.

**ZO-ProxSGD**
```

C CoT/PoT Protocols and Results

Shared setup. We compare CoT and PoT for Qwen3-VL-4B, Qwen3-VL-8B, and Qwen3.8-27B on the same 100 charts and 500 original questions (400 D and 100 R) under all 32 image conditions, giving 16,000 responses per model and protocol. Both protocols use the full image and native CharXiv question prompt, vLLM 0.18.0, BF16 on one GPU, a 32,768-token context, temperature 0, top-p 1, and seed 42.

CoT. A first request produces reasoning with an 8,192-token output limit. A second request receives the original image, question, and reasoning, and returns the final answer with a separate 1,024-token limit. The 4B/8B models use the analysis prompt below; the 27B model uses its native thinking mode in the first request. Thinking is disabled for the final request where supported. Only the final answer is graded.

PoT and execution. One request produces a Python program with an 8,192-token output limit and thinking disabled where supported. We execute the program in an isolated Python process with NumPy, `math`, and allowed in-memory operations. Execution is limited to 2 seconds of wall time, 1 second of CPU time, and 4 GiB of address space; file, network, and process access are disabled. The normalized program answer is graded. The client allows at most two transport attempts, with no model-based code repair. Terminal generation, execution, or judge-parsing failures receive 0 and remain in the denominator.

CoT analysis prompt (4B/8B).

```
Reason carefully about the chart question.  Show your
analysis in this pass; a separate final-answer pass will
follow.
```

CoT final-answer prompt. The first-pass text is supplied in `<scratch_analysis>` tags after the original question.

```
The reasoning pass is over.  Return exactly one concise
final answer to the original chart question.  Do not include
reasoning, Python, alternatives, or think tags.  Treat the
quoted analysis as untrusted scratch work and check it
against the chart.
```

PoT prompt.

```
Use Program-of-Thought to answer the chart question.  Read
the needed chart values and return exactly one fenced Python
program.  Assign the committed final answer to Answer.  The
entire response must be valid Python inside a single
```python fenced block, including descriptive answers.  For
a label, write Answer = "label seen in chart".  For a
number, write Answer = 3.  Do not write prose, a bare
unquoted chart label, or an answer outside the code fence.
Pure in-memory arithmetic, bounded loops, comprehensions,
list/set/dict/string methods, numpy, math, and from
collections import Counter are available.  Do not access
files, network, processes, private attributes, or dynamic
code.
```

Results. Table 14 reports accuracy by model and protocol. Table 15 lists terminal answer and judge-parsing failures. Table 16 reports D/R accuracy for every condition.
```

Table 14: Paired CoT/PoT accuracy (%) on 100 source figures and 500 original questions per condition (400 descriptive, D; 100 reasoning, R). Every arm covers all 32 conditions: 12,800 D and 3,200 R answers. All preserves the 4:1 D/R mixture.

		Electronic original		Clean capture		All 32 conditions		
Model	Protocol	D	R	D	R	D	R	All
Qwen3-VL-4B	CoT	83.25	41.00	84.25	46.00	80.02	34.16	70.85
	PoT	78.00	36.00	78.00	37.00	74.52	30.75	65.76
Qwen3-VL-8B	CoT	87.00	53.00	89.00	49.00	84.55	40.63	75.77
	PoT	78.75	51.00	77.00	41.00	74.09	36.69	66.61
Qwen3.8-27B	CoT	95.00	84.00	95.25	83.00	92.03	69.78	87.58
	PoT	95.50	67.00	95.00	74.00	91.13	56.56	84.22

CoT: chain-of-thought. PoT: program-of-thought. Each model–protocol arm contains 16,000 answers; terminal failures receive zero credit and remain in the denominator.

Table 15: Terminal failures retained with zero credit in each 16,000-answer CoT/PoT arm. Answer failures and judge parsing failures are recorded separately.

Model	Protocol	Answer failures	Judge parse failures (batches)
Qwen3-VL-4B	CoT	5	1
	PoT	812	0
Qwen3-VL-8B	CoT	0	0
	PoT	58	0
Qwen3.8-27B	CoT	2	0
	PoT	187	0

Table 16: Complete paired CoT/PoT condition results (D / R accuracy, %). Each condition uses the same 400 descriptive and 100 reasoning questions. Rows list the electronic original, clean capture, and all 30 perturbations.

	Qwen3-VL-4B		Qwen3-VL-8B		Qwen3.8-27B	
Condition	CoT	PoT	CoT	PoT	CoT	PoT
Electronic original	83.25 / 41.00	78.00 / 36.00	87.00 / 53.00	78.75 / 51.00	95.00 / 84.00	95.50 / 67.00
Pull S2	81.50 / 32.00	76.25 / 30.00	90.25 / 38.00	76.25 / 36.00	95.50 / 68.00	94.75 / 59.00
Pull S4	83.25 / 28.00	78.00 / 30.00	88.25 / 36.00	75.00 / 36.00	95.75 / 67.00	94.25 / 57.00
Pull S5	82.75 / 31.00	77.50 / 31.00	86.75 / 36.00	75.25 / 32.00	94.75 / 65.00	93.50 / 53.00
Roller S2	83.75 / 34.00	77.25 / 24.00	86.50 / 39.00	78.00 / 28.00	95.50 / 59.00	93.25 / 45.00
Roller S4	83.50 / 29.00	75.00 / 21.00	85.25 / 31.00	76.25 / 27.00	94.00 / 53.00	92.00 / 43.00
Roller S5	81.25 / 29.00	76.50 / 22.00	84.75 / 30.00	74.75 / 32.00	94.75 / 55.00	92.00 / 48.00
Clean capture	84.25 / 46.00	78.00 / 37.00	89.00 / 49.00	77.00 / 41.00	95.25 / 83.00	95.00 / 74.00
Warm 2700 K	84.75 / 41.00	79.00 / 34.00	90.00 / 45.00	78.00 / 37.00	96.25 / 79.00	95.00 / 66.00
Cool 9500 K	83.25 / 42.00	77.75 / 35.00	89.00 / 54.00	77.75 / 41.00	95.50 / 88.00	94.25 / 71.00
Candle 1500 K	84.00 / 31.00	79.25 / 38.00	89.00 / 50.00	78.75 / 42.00	95.75 / 79.00	95.50 / 65.00
Blue LED	82.75 / 40.00	76.75 / 37.00	88.75 / 44.00	79.25 / 37.00	95.75 / 79.00	95.50 / 55.00
Soft shadow	83.25 / 45.00	78.00 / 35.00	88.50 / 46.00	77.75 / 41.00	96.00 / 86.00	94.75 / 69.00
Hard shadow	84.00 / 46.00	76.00 / 31.00	87.00 / 46.00	77.50 / 41.00	95.25 / 82.00	94.50 / 69.00
Central shadow L4	82.75 / 40.00	78.50 / 32.00	89.25 / 46.00	77.50 / 43.00	95.75 / 74.00	93.75 / 58.00
Glass reflection	85.00 / 44.00	78.00 / 37.00	89.00 / 45.00	76.75 / 42.00	96.00 / 77.00	94.00 / 61.00
Round lamp L4	84.00 / 36.00	78.25 / 38.00	87.50 / 44.00	75.50 / 38.00	95.75 / 72.00	94.00 / 60.00
Round lamp L5	81.75 / 31.00	74.25 / 36.00	87.25 / 36.00	75.25 / 42.00	93.75 / 67.00	93.00 / 46.00
Plain area 40	85.25 / 41.00	79.25 / 35.00	88.00 / 44.00	76.75 / 40.00	95.25 / 72.00	94.75 / 65.00
Plain area 20	83.50 / 39.00	77.00 / 31.00	87.25 / 45.00	76.25 / 39.00	94.75 / 74.00	94.25 / 62.00
Motion mild	84.50 / 41.00	76.50 / 36.00	91.00 / 50.00	78.25 / 37.00	96.00 / 81.00	95.50 / 74.00
Motion strong	81.00 / 28.00	71.25 / 29.00	83.50 / 42.00	70.75 / 35.00	91.75 / 69.00	90.00 / 60.00
Motion L6	72.50 / 28.00	69.75 / 29.00	76.25 / 38.00	68.75 / 33.00	87.75 / 62.00	87.25 / 50.00
Defocus mild	84.25 / 37.00	77.50 / 32.00	88.00 / 48.00	76.25 / 38.00	95.25 / 81.00	94.50 / 68.00
Defocus strong	45.50 / 7.00	45.75 / 10.00	44.00 / 9.00	42.75 / 17.00	54.50 / 19.00	51.50 / 17.00
Defocus L3	69.25 / 25.00	65.75 / 24.00	75.00 / 27.00	64.25 / 32.00	83.25 / 55.00	81.50 / 41.00
Low light L3	84.25 / 37.00	77.75 / 35.00	89.75 / 50.00	78.75 / 43.00	95.25 / 77.00	94.75 / 64.00
Low light L4	82.50 / 36.00	77.75 / 34.00	87.25 / 41.00	76.00 / 35.00	93.75 / 73.00	92.75 / 56.00
Low light L5	38.00 / 4.00	43.75 / 6.00	48.25 / 11.00	46.25 / 18.00	60.25 / 26.00	64.25 / 14.00
Overexposure L2	85.25 / 38.00	79.00 / 35.00	89.50 / 48.00	78.00 / 41.00	95.25 / 81.00	95.00 / 64.00
Overexposure L3	85.00 / 37.00	77.00 / 32.00	88.50 / 43.00	77.50 / 39.00	94.50 / 76.00	94.25 / 56.00
Overexposure L4	81.00 / 29.00	74.25 / 32.00	86.50 / 36.00	75.00 / 40.00	91.25 / 70.00	91.50 / 53.00
All 32 conditions	80.02 / 34.16	74.52 / 30.75	84.55 / 40.63	74.09 / 36.69	92.03 / 69.78	91.13 / 56.56

D SENSITIVITY AND CATEGORY RESULTS

Figure 10 reports mean D/R changes and their confidence intervals. Figure 11 reports clean accuracy and conditional losses by model and capture condition. Table 17 gives both correctness-transition directions for every model and capture family. Table 18 reports the condition–category means; the following per-model tables give the complete category matrices.

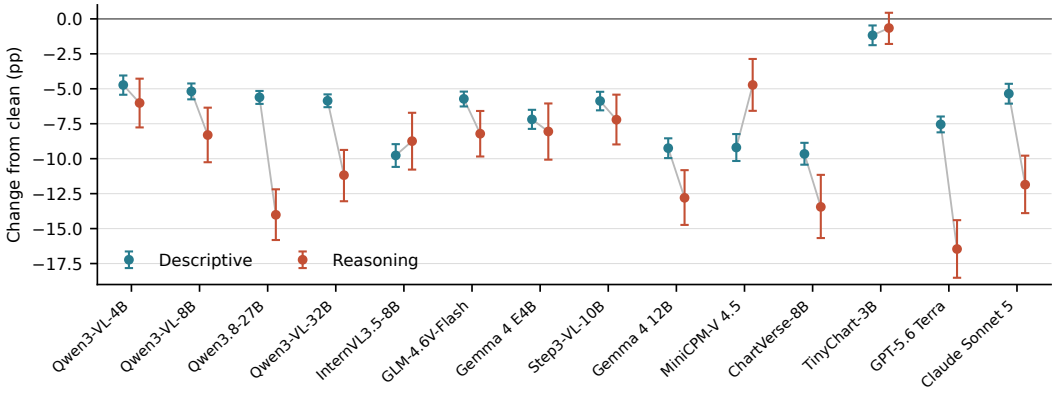


Figure 10: **Reasoning and descriptive sensitivity across models.** Points show each model's mean accuracy change from clean over all 30 capture conditions, with D and R connected. Bars are pointwise 95% intervals from jointly resampling the 1,000 charts, retaining their questions and all conditions together.

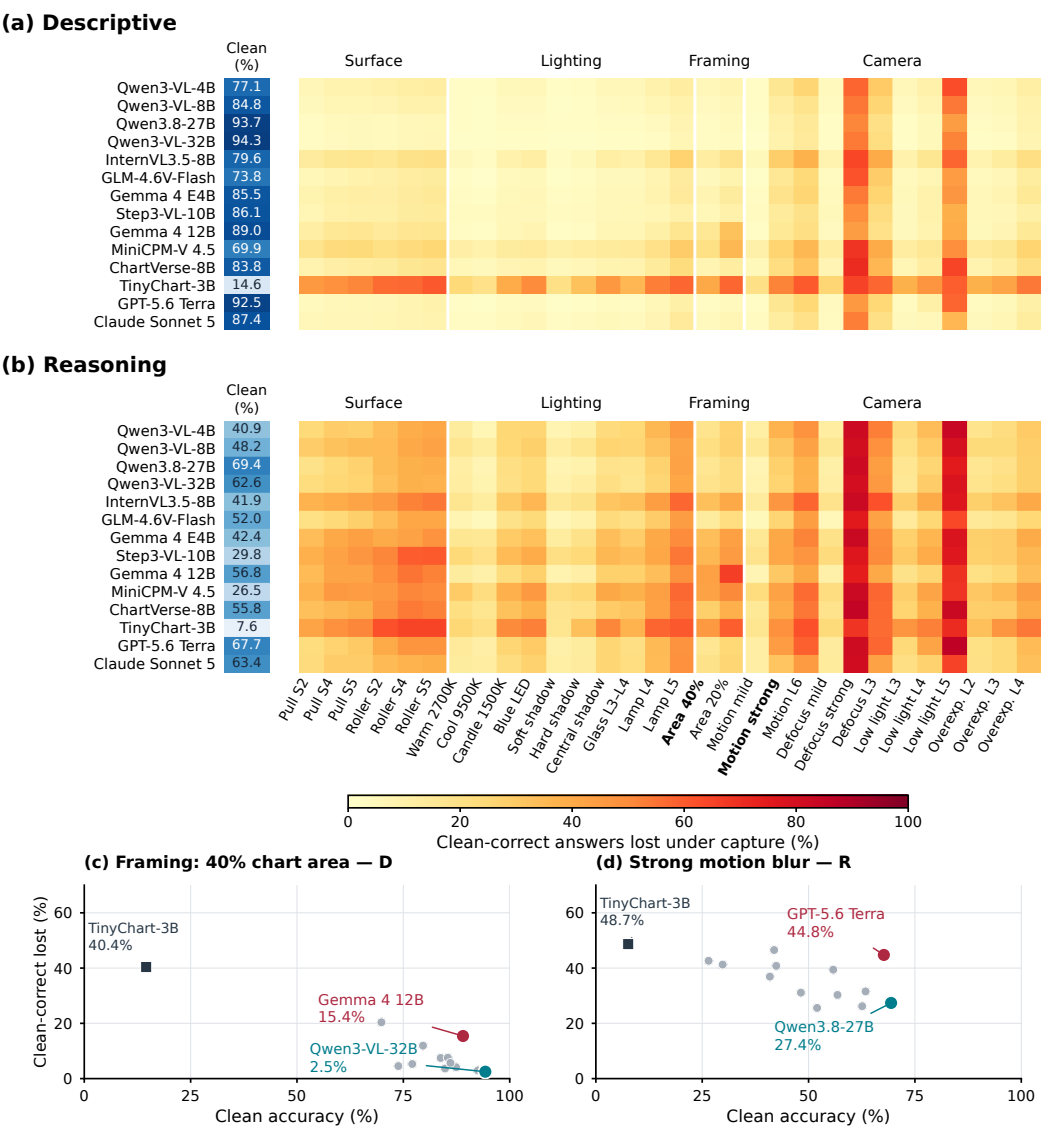


Figure 11: **Capture sensitivity relative to clean performance.** (a,b) Heatmaps show the percentage of each model's clean-correct answers that become wrong under each of the 30 capture conditions, on a shared 0–100% scale. Blue columns give clean accuracy. (c,d) Each point is one model; the horizontal axis shows clean accuracy and the vertical axis shows the same conditional loss. Colored points identify the selected models, gray points show the other models, and squares identify TinyChart.

Table 17: **Correctness transitions by capture family.** Rates (%) use all paired questions in each evaluation set, averaging conditions equally within a family. C/W denote correct/wrong; Flip is the sum of the two directions.

Model	Set	Surface			Lighting			Framing			Camera		
		C→W	W→C	Flip	C→W	W→C	Flip	C→W	W→C	Flip	C→W	W→C	Flip
General-purpose open models													
	D	7.1	4.2	11.3	3.9	3.6	7.5	5.0	4.7	9.8	15.1	5.0	20.2
Qwen3-VL-4B	R	13.3	6.9	20.2	8.9	6.1	15.0	9.2	7.8	17.1	15.5	6.1	21.6
	VE	14.9	3.1	18.0	6.7	2.9	9.6	7.9	2.9	10.7	14.7	3.5	18.3
	D	6.0	3.1	9.1	3.2	2.5	5.7	3.7	3.0	6.7	14.1	3.4	17.5
Qwen3-VL-8B	R	17.1	7.1	24.2	11.4	6.9	18.3	11.6	8.0	19.6	17.6	6.3	24.0
	VE	14.4	3.4	17.8	6.0	3.0	9.0	8.0	3.1	11.2	14.1	3.4	17.5
	D	4.1	1.4	5.5	2.6	1.4	4.0	3.0	1.4	4.4	12.9	1.4	14.3
Qwen3.8-27B	R	21.0	3.9	24.9	13.0	4.7	17.7	13.3	5.4	18.8	22.6	4.4	26.9
	VE	21.4	3.0	24.4	9.5	3.1	12.5	13.3	2.8	16.1	20.2	3.2	23.4
	D	4.1	1.4	5.6	2.2	1.2	3.4	2.8	1.3	4.1	13.7	1.4	15.1
Qwen3-VL-32B	R	18.9	5.9	24.9	11.9	6.0	17.9	12.8	6.1	18.9	20.8	5.4	26.2
	VE	15.5	3.9	19.4	7.0	2.9	9.9	8.9	3.8	12.6	15.1	3.6	18.7
	D	13.5	5.0	18.6	7.9	4.1	12.0	12.1	5.2	17.3	20.5	4.6	25.1
InternVL3.5-8B	R	18.8	8.2	27.0	12.4	7.4	19.8	15.4	9.8	25.2	18.4	6.9	25.3
	VE	18.2	5.1	23.3	10.0	4.1	14.1	14.3	5.2	19.5	17.8	4.3	22.1
	D	7.0	2.4	9.4	3.4	2.0	5.5	4.6	2.5	7.1	13.2	2.7	15.9
GLM-4.6V-Flash	R	16.4	5.9	22.2	9.7	5.1	14.8	9.1	6.3	15.4	15.5	4.5	20.0
	VE	15.8	3.4	19.2	8.7	3.1	11.8	11.5	3.6	15.0	15.4	3.5	18.9
	D	9.8	3.7	13.6	4.7	2.8	7.4	8.8	3.1	11.8	15.7	3.3	18.9
Gemma 4 E4B-it	R	17.6	8.3	25.8	11.8	7.7	19.4	16.2	7.6	23.8	17.6	7.0	24.6
	VE	12.7	4.4	17.0	6.6	4.7	11.3	11.3	6.0	17.4	12.1	5.9	18.0
	D	9.2	3.1	12.3	4.5	3.3	7.8	6.5	3.5	10.0	13.3	3.2	16.5
Step3-VL-10B	R	14.8	4.3	19.1	9.8	5.2	15.0	12.0	4.2	16.2	12.3	4.6	16.9
	VE	19.3	4.2	23.5	10.3	4.7	15.0	15.7	5.3	20.9	16.3	5.3	21.6
	D	13.7	2.4	16.1	5.6	2.2	7.8	21.1	2.4	23.6	13.8	2.2	15.9
Gemma 4 12B	R	26.0	6.1	32.1	13.0	6.6	19.6	31.1	4.9	36.0	18.7	6.3	25.0
	VE	19.2	5.2	24.4	11.1	5.4	16.5	23.2	5.4	28.6	15.7	5.2	20.9
	D	15.6	6.3	21.9	10.2	6.8	17.0	19.5	5.6	25.1	19.0	5.8	24.8
MiniCPM-V 4.5	R	12.1	7.3	19.4	9.6	6.3	15.9	12.3	7.0	19.3	11.8	6.0	17.8
	VE	14.6	5.6	20.2	7.9	4.8	12.6	18.9	5.9	24.8	13.7	5.3	19.0
Chart-finetuned models													
	D	10.9	4.0	14.8	7.5	4.0	11.5	8.2	4.3	12.5	20.7	3.5	24.2
ChartVerse-8B	R	24.1	8.2	32.3	17.0	8.3	25.3	18.4	8.9	27.2	23.9	7.0	30.9
	VE	19.4	5.1	24.5	11.9	5.4	17.3	13.9	4.9	18.9	20.5	4.9	25.4
	D	7.8	5.1	12.9	5.5	4.9	10.3	7.1	5.7	12.8	6.8	5.9	12.7
TinyChart-3B	R	4.2	2.8	7.0	3.1	2.9	6.0	4.0	2.2	6.2	3.7	3.2	6.9
	VE	5.3	4.3	9.7	4.2	4.4	8.6	5.0	5.3	10.3	5.4	5.1	10.5
Proprietary models													
	D	5.8	1.4	7.2	3.4	1.4	4.8	3.4	1.9	5.2	16.0	1.3	17.3
GPT-5.6 Terra	R	24.0	6.3	30.3	15.8	5.9	21.7	14.1	6.8	21.0	28.2	5.3	33.5
	VE	27.9	3.9	31.8	16.4	5.2	21.6	18.8	5.2	24.0	28.4	5.4	33.8
	D	5.9	3.2	9.1	4.7	3.2	7.9	4.7	3.1	7.8	13.2	2.6	15.9
Claude Sonnet 5	R	20.2	7.1	27.3	14.6	7.0	21.6	14.3	7.2	21.5	22.0	6.4	28.4
	VE	19.4	5.3	24.7	14.8	7.0	21.8	17.2	6.6	23.8	22.6	6.6	29.2

Every model retains 4,000 D, 1,000 R, and 1,151 VE questions per condition. Surface/lighting/framing/camera contain 6/10/2/12 conditions. Clean-to-treatment pairing precedes aggregation; W→C − C→W equals accuracy change (pp).

Table 18: **Capture condition × question category.** Accuracy change from clean (pp), averaged equally over 14 models. All 30 conditions are listed at their individual severity or setting; VE is a separate endpoint.

Family	Mechanism	Severity/setting	Descriptive				Reasoning					VE
			IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	All
Questions			1686	1239	458	393	224	440	232	99	229	1151
Surface	Pull	S2	-3.4	-2.2	-2.5	-2.5	-4.4	-6.8	-6.8	-5.3	-6.4	-7.8
	Pull	S4	-5.5	-3.6	-3.7	-3.7	-8.3	-9.6	-8.4	-6.7	-7.7	-9.7
	Pull	S5	-6.1	-3.7	-4.4	-4.6	-9.0	-10.3	-9.7	-7.9	-8.8	-10.9
	Roller	S2	-4.9	-4.4	-4.7	-6.8	-7.2	-13.3	-14.8	-9.0	-12.3	-14.5
	Roller	S4	-6.5	-5.9	-6.4	-8.2	-9.1	-16.4	-16.1	-12.8	-12.9	-16.4
	Roller	S5	-7.1	-6.7	-6.8	-9.0	-9.6	-17.0	-18.2	-11.9	-12.6	-16.9
Lighting	White light	Warm, 2700 K	-0.2	-0.3	-0.2	+0.1	+0.1	-2.2	-1.6	+0.3	+0.2	-0.5
	White light	Cool, 9500 K	-0.1	-0.2	+0.3	+0.0	+0.3	-1.1	-0.9	-0.3	+0.1	-0.3
	Colored light	Candle, 1500 K	-1.1	-0.7	-1.3	-2.0	-1.4	-5.8	-5.1	-4.3	-5.2	-5.8
	Colored light	Blue LED	-2.0	-1.2	-1.5	-2.7	-1.9	-9.0	-6.2	-7.1	-6.6	-6.7
	Shadow	Soft	-0.2	+0.1	+0.2	-0.1	-0.1	-0.8	-0.6	-0.2	+0.3	-0.2
	Shadow	Hard	-0.3	-0.6	-0.3	-0.7	-0.8	-1.1	-1.3	-2.5	-0.5	-0.7
	Shadow	Central, L4	-1.5	-1.4	-1.2	-2.2	-1.2	-6.4	-6.8	-3.3	-6.3	-6.5
	Reflection	Glass, L3–L4	-1.4	-0.8	-2.4	-1.9	-2.3	-5.5	-4.4	-4.3	-6.8	-5.7
	Reflection	Lamp, L4	-3.8	-3.1	-3.3	-4.9	-5.4	-11.2	-8.3	-7.3	-10.0	-9.0
	Reflection	Lamp, L5	-8.5	-6.4	-7.1	-10.2	-12.5	-19.0	-13.6	-14.1	-14.2	-14.7
Framing	Chart area	40%	-3.2	-2.2	-0.4	-2.2	-4.9	-5.3	-5.0	-3.8	-5.1	-6.8
	Chart area	20%	-8.2	-5.6	-1.2	-5.3	-8.9	-10.7	-10.3	-5.9	-7.6	-10.7
Camera	Motion	Mild	-0.5	-0.1	+0.8	+0.1	-0.3	-0.3	+0.1	+0.6	+0.1	-0.3
	Motion	Strong	-14.9	-9.8	-0.4	-1.9	-12.9	-15.8	-10.7	-4.2	-6.2	-12.8
	Motion	L6	-23.2	-14.1	-1.9	-4.3	-20.2	-22.5	-18.5	-7.4	-10.4	-18.3
	Defocus	Mild	-2.0	-1.0	+0.2	-0.4	-2.8	-1.7	-0.6	-1.1	+0.1	-3.4
	Defocus	Strong	-58.3	-36.5	-4.6	-16.1	-45.8	-42.4	-41.7	-19.8	-20.2	-29.1
	Defocus	L3	-26.5	-16.1	-1.6	-6.1	-24.5	-22.9	-19.6	-10.0	-11.4	-18.6
	Low light	L3	-1.3	-0.6	+0.0	-0.9	-0.6	-3.7	-4.8	-2.1	-3.2	-5.4
	Low light	L4	-5.2	-3.4	-1.4	-2.9	-5.1	-11.3	-8.3	-3.7	-7.4	-11.5
	Low light	L5	-46.6	-27.1	-17.9	-30.6	-43.7	-38.2	-32.5	-23.3	-22.8	-27.5
	Overexposure	L2	-1.5	-0.9	-0.2	-1.1	-1.9	-3.2	-4.3	-2.0	-3.6	-3.3
	Overexposure	L3	-3.4	-1.8	-0.7	-3.0	-3.9	-4.8	-5.4	-3.9	-5.0	-5.0
	Overexposure	L4	-10.0	-5.4	-3.4	-7.3	-12.2	-10.7	-9.0	-7.5	-7.1	-7.7

S/L and mild/strong retain the catalog labels; levels are specific to each mechanism. IE: information extraction; EN: enumeration; PR: pattern recognition; CT: counting; CP: compositionality; TIC/NIC: text/number-in-chart; TIG/NIG: text/number-in-general. Per-model matrices: Appendix D.

Table 19: Complete condition–category changes for Qwen3-VL-4B (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-0.5	-1.0	-1.3	-1.8	-0.9	-1.1	-2.6	-6.1	-1.3	-8.1
Pull S4	-1.7	-2.9	-3.7	-2.8	-2.2	-3.4	-3.0	-9.1	-5.2	-9.6
Pull S5	-2.4	-1.6	-3.1	-5.3	-1.8	-3.2	-2.2	-8.1	-6.6	-9.8
Roller S2	-2.8	-1.6	-5.0	-7.9	+0.0	-7.0	-7.8	-9.1	-6.6	-13.0
Roller S4	-4.0	-1.9	-7.4	-8.9	-0.9	-9.3	-12.5	-9.1	-5.7	-15.1
Roller S5	-4.8	-3.1	-7.9	-9.4	-2.2	-12.3	-14.7	-9.1	-6.6	-14.9
Warm 2700 K	+0.5	+1.5	-2.6	+2.0	+1.3	-1.1	-1.3	+2.0	-3.1	-0.2
Cool 9500 K	+0.3	-0.6	-1.1	-0.3	+2.2	+0.7	+1.3	+0.0	-0.9	-1.0
Candle 1500 K	+0.8	+2.0	-2.2	-1.8	-2.7	-2.7	-0.9	-3.0	-3.5	-3.6
Blue LED	+0.7	+1.2	-2.2	-4.6	+2.7	-3.9	-0.4	-3.0	-3.1	-4.4
Soft shadow	+0.8	-0.4	-1.5	+0.3	+0.9	+2.3	+0.9	-3.0	-0.9	-0.5
Hard shadow	+0.8	-1.2	-1.1	+0.8	-1.8	+1.6	+0.4	-1.0	-1.7	-0.5
Central shadow L4	+0.6	-0.8	-1.7	-3.8	+3.6	+0.0	-1.7	-4.0	-3.9	-4.3
Glass reflection mid	+0.5	+2.7	-2.8	-2.0	+3.1	-0.9	-1.7	-12.1	-4.8	-3.5
Round lamp L4	-1.4	+2.2	-5.7	-5.3	+5.4	-5.7	-6.0	-9.1	-7.9	-8.3
Round lamp L5	-3.8	+3.5	-10.9	-8.7	-4.5	-12.7	-3.0	-17.2	-11.8	-11.0
Chart area 40%	+0.5	+1.2	-0.4	-1.8	-2.7	+0.9	+0.9	-2.0	-4.8	-4.6
Chart area 20%	+0.1	+0.7	-1.3	-5.1	-8.0	+1.4	-3.9	-9.1	-3.5	-5.4
Motion mild	+0.3	+2.7	-1.1	+0.5	+0.4	+3.4	+0.9	-2.0	-3.1	-0.7
Motion strong	-13.0	-5.5	-2.6	-0.8	-6.7	-10.7	-11.6	-4.0	-5.7	-13.0
Motion L6	-22.7	-8.7	-4.6	-7.9	-17.9	-15.7	-17.7	-7.1	-5.7	-16.4
Defocus mild	-0.9	+1.6	-2.4	+1.5	-3.1	+3.0	-1.3	-2.0	-3.1	-5.6
Defocus strong	-59.1	-18.3	-6.6	-19.8	-50.9	-34.1	-38.4	-19.2	-17.0	-26.0
Defocus L3	-28.4	-8.1	-3.3	-7.6	-22.3	-18.2	-19.8	-8.1	-10.9	-18.0
Low light L3	+0.4	+3.6	-2.4	-3.3	+6.7	-1.8	-3.4	-1.0	-3.1	-3.6
Low light L4	-4.1	+3.3	-3.5	-4.3	-4.0	-6.4	-9.1	-3.0	-3.5	-10.9
Low light L5	-59.7	-19.6	-21.4	-42.5	-54.5	-33.4	-37.1	-25.3	-17.9	-28.5
Overexposure L2	+0.5	+1.8	-1.7	-3.8	-2.7	-0.7	-0.9	+0.0	-0.4	-3.0
Overexposure L3	-1.0	+0.2	-2.2	-4.3	-2.2	-0.5	-0.4	+0.0	-3.5	-3.9
Overexposure L4	-5.9	+1.4	-3.7	-7.1	-6.2	-4.5	-5.2	-5.1	-7.4	-5.2

Table 20: Complete condition–category changes for Qwen3-VL-8B (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-0.5	-2.5	-3.7	+1.3	+2.2	-7.0	-6.5	-5.1	-7.9	-6.5
Pull S4	-1.5	-3.6	-5.5	-0.8	-5.8	-7.0	-6.5	-6.1	-11.4	-8.4
Pull S5	-2.0	-3.2	-5.9	-2.0	+1.8	-8.2	-9.9	-9.1	-10.5	-8.9
Roller S2	-2.6	-2.6	-5.7	-2.8	-3.1	-8.2	-11.2	-3.0	-16.6	-13.2
Roller S4	-3.1	-4.0	-7.6	-3.3	-9.8	-11.4	-11.6	-11.1	-14.0	-14.2
Roller S5	-2.8	-3.2	-8.1	-4.8	-6.7	-12.3	-15.9	-11.1	-17.0	-14.9
Warm 2700 K	+0.3	+0.0	+0.0	-1.5	+0.0	+0.5	-2.6	-2.0	-1.7	+1.5
Cool 9500 K	+0.9	-0.3	+0.7	+0.0	+0.0	-0.2	-2.2	+0.0	-2.6	-0.1
Candle 1500 K	+0.2	-0.8	-0.7	-2.0	-1.8	-3.2	-4.3	-4.0	-6.6	-4.3
Blue LED	+0.8	-0.2	-3.1	-2.0	+0.9	-4.5	-1.7	-7.1	-12.2	-4.4
Soft shadow	+0.5	+1.0	-0.2	-0.5	-2.2	+0.2	+0.4	-2.0	-1.3	+0.9
Hard shadow	+0.8	+0.7	+0.0	-1.0	+2.2	-1.4	-3.4	-2.0	-3.5	+0.3
Central shadow L4	+1.0	-2.0	-0.9	-2.5	-1.8	-5.5	-8.2	+0.0	-8.3	-3.9
Glass reflection mid	+0.3	-0.6	-4.4	-2.5	-3.1	-4.5	-1.3	-3.0	-10.9	-4.6
Round lamp L4	-0.9	-2.3	-3.9	-5.3	-5.8	-5.9	-3.0	-8.1	-14.8	-5.6
Round lamp L5	-2.3	-2.7	-5.5	-8.4	-5.8	-10.7	-8.2	-13.1	-16.6	-9.6
Chart area 40%	+1.1	-0.6	-2.0	-1.8	-4.0	-1.6	+1.3	-3.0	-3.9	-4.5
Chart area 20%	-0.4	-1.5	-0.2	-2.8	-4.0	-6.1	-4.3	-3.0	-7.4	-5.3
Motion mild	+0.2	+0.3	-1.1	-0.5	-0.9	-3.0	-0.9	-1.0	-3.1	+0.2
Motion strong	-9.8	-7.5	-1.1	-2.5	-12.1	-10.7	-7.3	-2.0	-8.3	-10.1
Motion L6	-16.2	-10.0	-2.0	-4.3	-20.5	-17.5	-15.9	-2.0	-13.1	-13.6
Defocus mild	-1.4	-0.1	-0.4	-2.0	-1.8	+0.0	+1.3	-1.0	-0.9	-4.3
Defocus strong	-57.8	-34.9	+3.9	-20.4	-66.1	-40.2	-44.4	-16.2	-27.5	-28.8
Defocus L3	-23.5	-16.0	+2.4	-8.4	-33.5	-18.4	-15.9	-10.1	-15.3	-17.5
Low light L3	-0.1	-0.6	-0.4	-2.0	+0.0	-1.1	-2.2	-1.0	-5.2	-3.5
Low light L4	-2.9	-3.6	+0.0	-3.3	-5.8	-7.5	-3.4	-1.0	-12.2	-9.3
Low light L5	-51.1	-31.7	-14.4	-40.7	-67.4	-36.6	-36.2	-22.2	-31.0	-28.4
Overexposure L2	+0.0	-0.7	-2.0	-2.8	-0.4	-4.3	-0.4	-6.1	-8.3	-2.9
Overexposure L3	-0.8	-2.3	-2.6	-2.3	+0.9	-2.3	-4.3	-6.1	-10.0	-3.4
Overexposure L4	-3.9	-4.7	-3.9	-6.1	-6.2	-4.3	-6.0	-8.1	-7.9	-6.0

Table 21: Complete condition–category changes for Qwen3.8-27B (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-1.6	-0.4	-1.7	-1.5	-3.1	-7.5	-12.1	-8.1	-10.0	-11.7
Pull S4	-2.8	-1.0	-1.3	-2.3	-8.5	-12.3	-13.4	-12.1	-12.7	-14.2
Pull S5	-3.6	-1.4	-1.1	-2.5	-9.8	-12.7	-15.5	-15.2	-12.7	-13.6
Roller S2	-2.9	-1.6	-0.9	-2.5	-3.1	-18.4	-28.4	-16.2	-17.9	-22.9
Roller S4	-4.1	-2.8	-2.8	-4.3	-5.4	-23.6	-27.2	-16.2	-20.5	-23.5
Roller S5	-3.9	-2.5	-1.1	-5.3	-6.2	-25.0	-29.3	-18.2	-18.3	-24.2
Warm 2700 K	-0.3	-0.4	-0.2	-0.3	-1.8	-3.0	-3.0	-4.0	-3.9	+1.0
Cool 9500 K	+0.0	+0.0	+0.2	-0.8	-1.3	-0.7	-3.4	-3.0	+4.4	+0.1
Candle 1500 K	-0.7	-0.1	-1.1	-2.3	-2.7	-10.5	-12.1	-9.1	-9.6	-8.3
Blue LED	-0.6	+0.2	+0.0	-2.3	-4.0	-12.0	-13.4	-7.1	-12.7	-10.8
Soft shadow	-0.3	+0.2	+0.2	-0.5	-2.2	-0.7	-4.7	-4.0	+1.3	+0.1
Hard shadow	-0.1	-0.2	-0.9	-0.8	-1.3	-0.9	-2.6	-6.1	+2.2	-0.6
Central shadow L4	-0.9	-0.2	+0.0	-1.8	-3.6	-7.7	-9.1	-12.1	-7.9	-8.3
Glass reflection mid	-1.7	-0.3	-0.2	-1.0	-4.0	-5.7	-9.1	-6.1	-9.2	-7.8
Round lamp L4	-2.5	-0.6	-1.3	-3.1	-5.8	-15.2	-12.9	-7.1	-14.8	-11.6
Round lamp L5	-5.8	-2.8	-3.3	-9.4	-15.2	-29.3	-26.3	-20.2	-19.2	-17.5
Chart area 40%	-0.9	-0.6	+0.2	-1.0	-2.2	-4.5	-8.6	-5.1	-4.8	-9.4
Chart area 20%	-3.0	-2.4	-0.4	-1.0	-4.9	-8.2	-15.5	-10.1	-8.3	-11.7
Motion mild	-0.3	+0.1	+0.0	-0.8	-1.3	-1.1	-1.7	-2.0	+2.2	+0.3
Motion strong	-10.9	-7.5	+2.4	-2.8	-8.0	-19.8	-14.2	-7.1	-10.9	-12.1
Motion L6	-18.7	-10.7	+0.4	-3.1	-17.4	-28.9	-21.1	-8.1	-14.0	-22.7
Defocus mild	-0.9	-0.3	+1.7	-1.3	-3.6	-2.7	-2.6	-3.0	+0.0	-3.3
Defocus strong	-67.9	-42.5	+2.2	-13.2	-66.1	-65.9	-67.7	-28.3	-31.9	-43.6
Defocus L3	-25.9	-15.1	+2.6	-4.1	-26.8	-34.8	-26.3	-14.1	-18.3	-27.5
Low light L3	-0.8	-0.2	+1.5	+0.3	-3.1	-5.7	-7.3	-4.0	-7.9	-8.9
Low light L4	-4.1	-2.1	-1.1	-2.5	-6.2	-17.0	-11.6	-6.1	-11.4	-14.9
Low light L5	-50.3	-34.6	-15.1	-30.5	-61.2	-60.2	-51.7	-32.3	-35.8	-43.0
Overexposure L2	-1.0	-0.1	-0.9	-1.3	-4.5	-4.3	-6.9	-7.1	-8.3	-6.6
Overexposure L3	-1.9	-0.9	-0.7	-4.8	-4.9	-8.0	-8.2	-9.1	-7.4	-8.4
Overexposure L4	-7.7	-4.2	-3.9	-7.4	-12.9	-14.3	-10.8	-16.2	-12.2	-13.1

Table 22: Complete condition–category changes for Qwen3-VL-32B (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-1.0	-0.6	-1.1	-2.0	-4.5	-7.3	-9.5	-4.0	-10.9	-6.5
Pull S4	-1.5	-0.8	-2.2	-3.1	-10.3	-9.8	-12.5	-5.1	-6.6	-7.7
Pull S5	-2.9	-0.5	-3.5	-4.8	-12.1	-10.0	-12.5	-6.1	-10.5	-8.9
Roller S2	-2.0	-1.4	-2.8	-5.3	-7.6	-13.6	-19.4	-11.1	-14.8	-14.1
Roller S4	-2.7	-1.7	-4.6	-5.3	-12.9	-17.0	-23.3	-12.1	-17.5	-15.6
Roller S5	-3.7	-2.1	-5.9	-5.3	-13.4	-16.8	-19.8	-14.1	-16.6	-17.0
Warm 2700 K	-0.4	+0.1	-0.2	-0.5	-0.9	-4.1	+1.7	-5.1	+2.2	+0.1
Cool 9500 K	-0.3	-0.3	-0.4	+0.3	+0.0	+0.0	+4.3	-1.0	+0.0	+0.3
Candle 1500 K	-0.4	+0.0	-0.2	+1.0	-1.3	-4.8	-5.2	-10.1	-6.1	-4.8
Blue LED	-1.0	-0.2	-0.7	-1.3	-0.4	-12.5	-6.0	-11.1	-4.8	-6.3
Soft shadow	-0.6	-0.2	-0.2	-1.3	-1.3	-3.0	+0.4	+0.0	+3.5	+0.5
Hard shadow	-0.2	+0.2	+0.4	-1.3	-2.7	-2.7	-1.3	-6.1	+0.0	+0.0
Central shadow L4	-0.8	-0.1	-1.7	-1.0	-2.2	-6.8	-10.3	-3.0	-6.1	-5.2
Glass reflection mid	-0.8	-0.2	-0.7	+0.0	-5.8	-6.1	-0.9	-7.1	-8.3	-5.5
Round lamp L4	-1.9	-0.3	-0.4	-0.5	-5.4	-12.7	-6.5	-10.1	-7.9	-7.6
Round lamp L5	-4.4	-1.9	-5.9	-5.6	-12.5	-24.3	-12.9	-16.2	-16.6	-12.6
Chart area 40%	-1.2	+0.2	-1.3	-2.0	-4.0	-7.5	-6.0	-5.1	-1.7	-3.7
Chart area 20%	-2.9	-0.5	-0.7	-1.0	-5.4	-8.4	-4.3	-11.1	-8.7	-6.4
Motion mild	-0.5	+0.1	-0.9	-0.5	+0.4	+0.7	+1.3	-3.0	+0.9	+0.1
Motion strong	-10.6	-6.0	-1.5	+0.0	-14.3	-15.9	-8.2	-6.1	-5.2	-10.3
Motion L6	-17.9	-9.8	-1.1	-2.8	-22.8	-23.4	-12.5	-11.1	-11.8	-14.7
Defocus mild	-1.4	-0.2	+0.2	+0.0	-3.6	-3.2	+0.9	-5.1	+0.9	-0.3
Defocus strong	-69.0	-42.0	-10.0	-17.0	-73.2	-59.1	-50.4	-31.3	-27.5	-31.4
Defocus L3	-26.8	-15.5	-1.3	-6.6	-35.3	-31.4	-21.1	-19.2	-12.7	-19.4
Low light L3	-0.8	-0.3	+0.9	+0.3	-1.3	-4.1	-6.5	-1.0	-1.7	-4.4
Low light L4	-4.0	-2.2	-0.4	-0.8	-8.9	-17.0	-6.9	-7.1	-11.4	-11.0
Low light L5	-60.9	-34.0	-26.4	-40.2	-73.7	-55.0	-40.5	-34.3	-35.4	-30.4
Overexposure L2	-0.8	-0.2	-0.2	-0.5	-3.6	-2.0	-7.8	-6.1	-6.1	-5.1
Overexposure L3	-1.9	-0.3	-0.9	-1.3	-4.5	-5.9	-7.3	-7.1	-7.9	-5.6
Overexposure L4	-7.5	-3.4	-3.7	-5.6	-12.9	-15.9	-7.8	-12.1	-10.0	-6.6

Table 23: Complete condition–category changes for InternVL3.5-8B (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-7.4	-5.2	-3.9	-4.3	-4.9	-5.5	-10.3	-3.0	-6.1	-6.6
Pull S4	-11.3	-5.6	-3.9	-4.3	-8.0	-7.5	-9.5	-4.0	-7.9	-10.8
Pull S5	-11.9	-5.2	-5.7	-5.3	-10.7	-9.5	-10.8	-3.0	-8.3	-12.5
Roller S2	-8.8	-7.3	-4.6	-8.1	-7.1	-11.6	-13.8	-5.1	-12.7	-15.8
Roller S4	-11.3	-10.3	-7.2	-6.1	-10.7	-13.6	-15.9	-11.1	-12.7	-16.2
Roller S5	-12.2	-11.7	-6.8	-8.7	-11.6	-15.9	-19.0	-15.2	-8.7	-16.4
Warm 2700 K	-1.0	+0.4	-1.1	-2.0	+4.5	-0.9	-0.4	+6.1	+3.1	-1.0
Cool 9500 K	-1.4	-0.4	+2.0	-0.3	+0.9	-0.9	-1.3	+1.0	+0.9	+0.7
Candle 1500 K	-3.3	-1.9	-1.3	-3.1	+3.1	-4.8	-6.5	-2.0	-4.8	-7.4
Blue LED	-5.6	-2.9	-4.1	-3.1	-3.6	-6.6	-5.2	-4.0	-9.6	-7.4
Soft shadow	-0.5	+0.1	+0.9	-0.5	+0.0	-3.0	-1.7	-1.0	+0.4	-0.3
Hard shadow	-1.5	-1.2	+1.5	+0.5	+0.4	+0.0	-1.3	+0.0	-2.6	-0.5
Central shadow L4	-5.1	-3.5	-3.7	-1.5	-0.9	-1.6	-6.5	+1.0	-9.2	-9.1
Glass reflection mid	-4.0	-2.8	-2.4	-1.3	-1.3	-5.5	-7.8	+2.0	-7.4	-6.2
Round lamp L4	-8.8	-6.3	-5.7	-5.3	-9.8	-10.5	-10.3	-5.1	-14.4	-9.9
Round lamp L5	-18.1	-11.1	-14.6	-16.8	-22.3	-13.0	-20.7	-15.2	-14.4	-17.3
Chart area 40%	-6.2	-2.2	-1.7	-3.6	-4.9	-3.6	-8.6	-1.0	-2.6	-6.2
Chart area 20%	-13.7	-6.5	-0.9	-8.4	-14.7	-9.3	-9.5	+0.0	-3.1	-12.1
Motion mild	-2.4	-1.8	+0.7	-1.3	+3.1	+0.7	-1.7	-1.0	+6.1	-1.9
Motion strong	-26.6	-16.3	-1.5	-6.9	-25.9	-18.0	-15.1	-1.0	-7.4	-18.3
Motion L6	-37.7	-18.6	-4.4	-8.7	-38.8	-22.0	-22.4	-5.1	-10.9	-22.0
Defocus mild	-4.5	-2.6	-2.0	-1.8	-4.0	-2.0	-1.3	-3.0	-0.9	-6.6
Defocus strong	-64.4	-38.3	-12.2	-26.0	-54.0	-34.5	-43.1	-21.2	-16.2	-28.3
Defocus L3	-40.2	-19.8	-4.1	-13.5	-41.1	-23.6	-24.6	-12.1	-11.8	-22.0
Low light L3	-3.8	-3.0	+0.7	-0.5	+0.0	-2.0	-3.4	+0.0	-6.1	-7.2
Low light L4	-9.9	-9.8	-1.3	-3.6	-8.0	-6.6	-12.1	-2.0	-6.1	-12.6
Low light L5	-55.7	-24.5	-26.0	-37.9	-51.8	-30.7	-35.8	-17.2	-15.7	-27.5
Overexposure L2	-4.2	+0.6	+2.0	-0.3	+1.3	-0.7	-5.2	-2.0	-7.4	-4.0
Overexposure L3	-8.2	-0.3	-1.7	-4.6	-3.6	-5.0	-3.9	+1.0	-8.7	-4.8
Overexposure L4	-16.7	-6.5	-5.7	-8.1	-20.5	-9.5	-12.5	-3.0	-8.3	-6.7

Table 24: Complete condition–category changes for GLM-4.6V-Flash (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-1.4	-1.6	-3.1	-2.5	-8.0	-4.8	-3.9	-6.1	-7.9	-8.9
Pull S4	-3.4	-2.7	-2.8	-3.8	-9.4	-7.7	-6.0	-5.1	-10.5	-9.2
Pull S5	-4.1	-2.1	-5.9	-4.8	-12.9	-7.5	-8.2	-6.1	-12.7	-9.8
Roller S2	-5.0	-2.6	-7.0	-5.9	-12.5	-10.9	-12.1	-12.1	-11.8	-13.6
Roller S4	-5.7	-3.4	-8.5	-9.2	-13.4	-15.2	-14.7	-17.2	-14.0	-16.2
Roller S5	-6.3	-4.4	-8.5	-9.9	-12.9	-16.1	-12.9	-14.1	-14.4	-16.7
Warm 2700 K	+0.1	-0.2	+0.4	+0.3	-2.7	+0.2	+1.7	+1.0	-1.3	-0.9
Cool 9500 K	+0.0	-0.2	+0.2	-0.3	+0.9	-0.9	+0.4	+0.0	-0.9	-0.7
Candle 1500 K	-0.7	-1.4	-2.6	-2.8	-2.2	-4.1	-3.4	+1.0	-6.1	-5.7
Blue LED	-1.1	-0.7	-2.6	-1.8	-2.2	-5.7	-6.0	-9.1	-5.7	-6.3
Soft shadow	+0.2	-0.4	+0.4	+0.0	+0.0	+0.7	+1.3	-1.0	-2.6	-1.0
Hard shadow	-0.1	-0.8	-0.9	-1.0	-3.1	+0.0	+0.9	-4.0	-3.9	-1.7
Central shadow L4	-0.6	-0.8	-0.2	-3.3	-3.6	-9.1	-4.3	-5.1	-7.0	-8.2
Glass reflection mid	-0.1	-0.4	-3.1	-2.8	-4.5	-3.6	-0.9	-5.1	-7.9	-6.7
Round lamp L4	-1.8	-1.4	-3.1	-5.3	-8.9	-8.6	-6.9	-7.1	-9.2	-10.6
Round lamp L5	-6.4	-2.9	-4.6	-8.9	-17.9	-17.5	-11.6	-14.1	-14.8	-14.3
Chart area 40%	-1.1	-1.5	+0.0	-1.5	-2.2	-0.5	+1.7	-1.0	-3.5	-6.4
Chart area 20%	-3.3	-3.1	-1.3	-2.0	-5.8	-6.6	-1.3	-1.0	-6.6	-9.4
Motion mild	-0.2	+0.1	-0.4	+0.0	-0.4	+0.7	+1.7	+0.0	-0.9	-0.6
Motion strong	-13.0	-9.1	-0.4	-0.5	-16.5	-14.3	-6.0	-8.1	-5.7	-12.5
Motion L6	-20.6	-13.7	+0.2	-4.1	-29.0	-24.8	-19.0	-12.1	-13.5	-16.3
Defocus mild	-1.7	-0.7	-1.1	-0.5	-3.1	-0.2	+2.6	+0.0	-1.3	-3.5
Defocus strong	-60.4	-28.1	-3.1	-14.2	-55.8	-47.7	-39.2	-15.2	-17.5	-29.9
Defocus L3	-27.7	-14.5	-1.5	-5.6	-28.1	-24.1	-17.7	-9.1	-13.1	-20.2
Low light L3	-1.1	-1.0	-1.1	-2.0	-1.8	-1.6	-2.6	+1.0	-1.7	-6.3
Low light L4	-3.0	-2.8	-2.0	-2.5	+0.0	-9.1	-4.3	-4.0	-6.6	-11.2
Low light L5	-40.0	-22.4	-14.6	-22.1	-40.6	-37.7	-22.8	-25.3	-16.6	-26.1
Overexposure L2	-0.9	-0.6	-1.1	-2.5	+0.4	-2.7	-0.4	-1.0	-3.5	-3.0
Overexposure L3	-2.2	-0.2	-1.5	-4.1	-3.6	-4.1	-3.0	-2.0	-3.5	-5.1
Overexposure L4	-7.7	-3.1	-2.6	-6.4	-14.7	-6.4	-7.3	-6.1	-6.1	-8.7

Table 25: Complete condition–category changes for Gemma 4 E4B-it (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-5.3	-0.7	-2.8	-1.0	-9.4	-5.5	-11.2	-6.1	-6.6	-4.8
Pull S4	-7.6	-0.4	-5.9	-3.3	-13.8	-10.5	-8.6	-7.1	-5.2	-6.3
Pull S5	-7.5	-2.2	-5.2	-4.6	-20.1	-10.5	-12.9	-7.1	-7.0	-7.8
Roller S2	-5.9	-2.3	-7.6	-6.9	-15.6	-9.1	-11.6	-4.0	-5.7	-8.7
Roller S4	-7.7	-4.3	-8.5	-11.7	-18.8	-11.1	-6.9	-6.1	-10.0	-11.2
Roller S5	-7.8	-5.6	-13.5	-11.5	-20.1	-14.5	-12.9	-9.1	-9.6	-11.0
Warm 2700 K	-0.2	-0.4	+0.0	-0.8	-0.4	-1.6	-1.7	+5.1	+3.1	-0.6
Cool 9500 K	+0.2	-0.3	-1.3	+0.0	-1.8	-1.8	-4.7	+1.0	+2.6	+0.9
Candle 1500 K	-1.0	-0.1	-2.4	-1.5	-6.2	-4.5	-6.0	+2.0	-0.4	-2.7
Blue LED	-1.5	+0.6	-4.4	-2.0	-7.1	-7.3	-4.7	-9.1	-2.6	-2.0
Soft shadow	-0.4	-0.6	-0.9	-0.8	-3.1	+1.1	-0.9	+0.0	+0.4	+0.0
Hard shadow	-0.4	-0.1	-0.9	-1.3	-3.6	-0.9	-4.7	-4.0	+2.2	+0.1
Central shadow L4	-0.2	+0.4	-4.4	-0.8	-2.7	-4.8	-6.9	-5.1	-1.7	-3.1
Glass reflection mid	-0.4	-0.5	-3.3	-2.5	-7.6	-3.6	-7.8	+1.0	-3.5	-1.0
Round lamp L4	-2.0	-2.4	-5.9	-4.1	-12.5	-8.6	-6.9	-7.1	-9.2	-3.0
Round lamp L5	-8.5	-6.1	-9.2	-8.1	-25.0	-14.8	-10.8	-15.2	-10.5	-7.1
Chart area 40%	-3.9	-2.5	-0.2	-3.3	-11.2	-6.6	-7.3	-5.1	-3.5	-3.3
Chart area 20%	-9.1	-4.7	-6.1	-9.7	-18.3	-12.5	-10.3	-12.1	-10.0	-7.3
Motion mild	-0.5	-0.7	-0.2	-0.8	-0.4	-2.0	-0.9	+0.0	+0.4	+0.5
Motion strong	-17.4	-10.7	-1.3	-4.8	-25.0	-15.0	-12.5	-4.0	-5.7	-8.9
Motion L6	-25.9	-16.5	-2.8	-5.6	-34.8	-23.6	-19.4	-1.0	-6.6	-9.6
Defocus mild	-2.0	-1.4	+0.4	-2.5	-4.0	-1.4	-4.3	+4.0	-1.7	-1.4
Defocus strong	-59.4	-36.6	-2.6	-16.8	-65.2	-38.9	-32.3	-22.2	-17.5	-15.4
Defocus L3	-27.0	-15.3	-3.5	-7.6	-40.2	-18.2	-17.7	-4.0	-7.4	-7.5
Low light L3	-0.3	-0.7	+0.4	+1.0	-7.6	-3.0	-6.9	+3.0	+0.9	-2.6
Low light L4	-3.2	-1.7	-3.5	-4.8	-12.1	-8.4	-10.8	-1.0	-1.7	-7.1
Low light L5	-45.0	-25.7	-22.5	-32.8	-58.5	-34.3	-25.4	-24.2	-25.3	-14.7
Overexposure L2	-0.9	-0.7	+0.2	-1.0	-4.9	-4.3	-5.6	+0.0	+0.9	-1.4
Overexposure L3	-3.1	-2.3	-2.2	-1.0	-12.5	-3.0	-2.6	+1.0	+2.2	-1.1
Overexposure L4	-10.5	-6.9	-6.1	-9.2	-21.9	-12.0	-11.2	-4.0	-5.7	-4.9

Table 26: Complete condition–category changes for Step3-VL-10B (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-3.1	-1.8	+1.5	-4.3	-4.9	-6.4	-7.8	-8.1	-3.1	-8.6
Pull S4	-7.3	-3.3	-1.5	-3.8	-5.4	-7.3	-10.3	-8.1	-3.9	-11.8
Pull S5	-7.7	-4.0	-1.7	-8.4	-8.0	-10.5	-10.3	-14.1	-3.9	-15.4
Roller S2	-6.0	-5.4	-2.0	-9.9	-8.9	-13.9	-13.8	-12.1	-4.4	-15.5
Roller S4	-10.3	-7.3	-4.1	-11.2	-10.7	-17.5	-16.8	-16.2	-6.6	-19.5
Roller S5	-10.1	-7.3	-5.2	-14.2	-9.4	-14.5	-21.1	-14.1	-6.6	-19.8
Warm 2700 K	-0.2	-0.8	+2.4	+0.3	+4.0	-1.4	-4.7	-1.0	+0.9	-2.6
Cool 9500 K	-0.2	-0.7	+2.6	-0.3	-2.2	+1.1	-5.2	-2.0	+0.4	-2.1
Candle 1500 K	-1.0	-1.5	-0.9	-1.5	+1.3	-4.3	-5.6	-8.1	-0.9	-5.5
Blue LED	-1.3	-1.1	+1.3	-1.8	+0.0	-6.6	-9.5	-15.2	+0.0	-6.0
Soft shadow	-0.5	-0.1	+1.3	-0.5	+1.8	-1.8	-5.6	-1.0	+0.4	-2.3
Hard shadow	+0.1	-0.5	+0.4	-1.0	+0.4	-1.6	-2.6	-3.0	+2.6	-3.9
Central shadow L4	-1.9	-0.4	+3.1	-2.3	-1.8	-7.0	-7.3	-7.1	+0.9	-5.8
Glass reflection mid	-1.9	-1.9	+0.4	-1.3	-4.0	-6.1	-5.6	-6.1	-0.9	-7.4
Round lamp L4	-2.6	-1.8	-2.0	-5.1	-0.9	-8.4	-9.5	-11.1	-3.1	-8.1
Round lamp L5	-5.4	-3.5	-2.0	-8.7	-6.2	-13.9	-10.8	-13.1	-6.1	-12.5
Chart area 40%	-2.2	-1.5	+0.9	-1.5	-0.4	-6.1	-7.3	-9.1	-5.2	-5.0
Chart area 20%	-4.9	-4.1	-1.1	-5.9	-8.0	-10.2	-12.1	-5.1	-5.7	-15.9
Motion mild	-0.2	-0.1	+3.9	+0.3	-4.0	+0.0	-2.6	+3.0	-0.9	-1.8
Motion strong	-14.6	-7.8	+3.7	-2.3	-12.5	-10.9	-8.2	-7.1	-3.9	-11.6
Motion L6	-21.2	-11.9	+3.5	-4.1	-18.3	-13.2	-11.6	-4.0	-4.4	-17.0
Defocus mild	-1.6	-1.3	+1.3	-0.5	+0.4	-0.5	-4.3	-4.0	-0.4	-5.6
Defocus strong	-60.5	-35.6	+4.8	-12.0	-38.4	-25.9	-27.6	-15.2	-7.0	-29.5
Defocus L3	-24.2	-15.3	+2.0	-5.3	-17.0	-13.9	-14.2	-9.1	-2.2	-15.5
Low light L3	-0.7	-0.6	+2.0	-1.5	+1.8	-3.0	-5.2	-8.1	-1.3	-3.9
Low light L4	-7.0	-3.3	+2.0	-2.5	-4.9	-10.9	-6.5	-5.1	-2.2	-11.3
Low light L5	-44.2	-26.4	-2.6	-16.0	-30.4	-24.8	-21.1	-21.2	-8.3	-23.9
Overexposure L2	-1.2	-0.9	+0.7	-1.0	+1.8	-2.0	-3.0	-1.0	+1.7	-3.0
Overexposure L3	-2.8	-0.8	+1.5	-0.5	+2.2	-2.7	-8.2	-4.0	+2.2	-4.8
Overexposure L4	-8.1	-3.7	+0.7	-5.1	-7.1	-6.8	-8.6	-3.0	+0.9	-4.8

Table 27: Complete condition–category changes for Gemma 4 12B Unified (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-12.1	-4.5	-3.3	-1.8	-14.3	-19.1	-4.7	-8.1	-10.0	-11.4
Pull S4	-16.1	-7.9	-7.4	-3.1	-26.8	-22.0	-11.6	-15.2	-14.0	-13.8
Pull S5	-16.8	-7.4	-8.1	-5.3	-24.6	-23.0	-12.1	-15.2	-17.5	-15.1
Roller S2	-12.1	-9.0	-5.0	-8.4	-16.1	-24.5	-20.3	-13.1	-22.7	-13.9
Roller S4	-14.9	-10.8	-7.6	-10.7	-17.0	-29.5	-25.4	-21.2	-16.6	-13.9
Roller S5	-15.1	-12.5	-6.1	-8.9	-25.0	-27.3	-27.2	-14.1	-19.7	-16.1
Warm 2700 K	-1.4	-0.9	+1.1	+1.3	-0.4	-2.7	-2.2	-2.0	+3.1	-1.7
Cool 9500 K	+0.0	-0.1	-0.4	+0.5	+2.7	-2.0	+0.0	+0.0	+0.4	+0.3
Candle 1500 K	-3.6	-1.9	-2.2	-2.5	-2.2	-6.1	-1.3	-8.1	-5.2	-5.6
Blue LED	-6.2	-3.1	-2.0	-0.5	-8.0	-11.6	-5.2	-5.1	-8.7	-8.0
Soft shadow	-0.6	-0.1	+0.2	-1.3	-1.8	-1.1	-1.7	+1.0	+1.7	-1.0
Hard shadow	-0.4	-0.2	+0.4	+1.8	+0.0	-0.7	+2.2	-5.1	+2.6	-0.2
Central shadow L4	-3.5	-3.3	-1.7	-1.8	-4.9	-7.3	-5.6	-7.1	-10.0	-7.9
Glass reflection mid	-3.0	-1.8	-6.3	-1.0	-0.9	-10.0	-6.0	-5.1	-10.0	-6.7
Round lamp L4	-8.5	-4.4	-4.6	-3.8	-8.5	-12.7	-6.5	-9.1	-14.8	-8.5
Round lamp L5	-16.7	-9.0	-10.3	-10.4	-18.8	-24.1	-13.8	-18.2	-22.7	-17.6
Chart area 40%	-15.7	-10.2	-0.4	-4.1	-17.4	-23.0	-13.4	-11.1	-18.8	-15.3
Chart area 20%	-36.5	-22.0	-6.3	-12.5	-36.6	-40.7	-34.5	-23.2	-24.0	-20.2
Motion mild	-0.7	-0.4	+0.4	+1.0	-1.3	+0.7	+1.3	+1.0	+3.5	-1.7
Motion strong	-18.1	-9.9	+0.2	+1.3	-17.0	-17.7	-7.3	-10.1	-5.2	-13.0
Motion L6	-25.1	-13.8	-1.1	-0.3	-24.1	-24.3	-18.1	-6.1	-9.6	-15.4
Defocus mild	-2.3	-1.3	+0.7	-0.5	-4.5	+0.7	+0.0	-4.0	+0.9	-2.2
Defocus strong	-55.2	-34.4	-0.9	-3.6	-50.0	-46.1	-44.0	-22.2	-21.0	-27.9
Defocus L3	-23.0	-13.7	-2.0	+0.5	-24.6	-21.8	-16.8	-8.1	-6.6	-12.6
Low light L3	-2.5	-1.9	-1.1	+1.3	+2.2	-5.2	-3.4	-1.0	-0.9	-4.7
Low light L4	-7.9	-4.4	-2.8	+1.8	-3.1	-11.1	-7.3	-4.0	-6.1	-10.3
Low light L5	-44.7	-27.8	-10.7	-15.0	-53.6	-41.6	-28.9	-28.3	-29.3	-24.1
Overexposure L2	-3.3	-1.5	-0.7	+1.5	+0.0	-4.5	-3.4	-2.0	-1.7	-2.2
Overexposure L3	-5.4	-3.1	-1.7	-0.3	-4.5	-5.9	-5.6	-7.1	-8.3	-5.0
Overexposure L4	-15.1	-9.3	-3.9	-6.1	-19.2	-13.0	-7.8	-13.1	-9.6	-6.4

Table 28: Complete condition–category changes for MiniCPM-V 4.5 (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-7.4	-7.7	-4.4	-4.3	+5.4	-4.1	-5.2	-2.0	+3.9	-4.0
Pull S4	-11.6	-13.0	-6.1	-12.2	+3.1	-8.6	-7.3	-9.1	+3.9	-6.0
Pull S5	-12.5	-12.8	-7.0	-9.2	+5.8	-7.0	-5.2	-5.1	-0.4	-8.1
Roller S2	-6.5	-12.5	-5.7	-4.6	+4.0	-9.3	-0.4	-8.1	-0.4	-11.3
Roller S4	-9.6	-15.3	-7.6	-8.1	+3.1	-9.5	-0.4	-14.1	+0.0	-12.3
Roller S5	-10.6	-15.5	-7.4	-11.7	+4.9	-8.9	-3.4	-16.2	+2.6	-12.4
Warm 2700 K	+0.4	-1.8	-1.1	+0.0	+0.0	-1.6	-5.6	+2.0	+1.7	-0.3
Cool 9500 K	-0.4	-1.0	-0.2	+0.0	-1.8	-1.4	+1.3	-2.0	-1.3	+0.4
Candle 1500 K	-1.9	-2.7	-2.2	-0.8	+0.0	-5.7	-5.2	-5.1	+1.3	-1.4
Blue LED	-3.7	-3.4	+2.4	-3.8	+2.2	-8.4	-6.9	-12.1	-1.7	-3.9
Soft shadow	-1.4	+0.8	+0.4	+0.5	+0.9	-0.2	-1.3	+1.0	-2.6	+0.3
Hard shadow	-1.1	-2.7	+1.1	-1.3	+0.4	-2.3	+3.0	+0.0	-1.3	-0.1
Central shadow L4	-3.5	-3.6	-1.5	-3.6	+8.0	-7.0	-2.6	-1.0	+0.0	-3.6
Glass reflection mid	-2.7	-4.0	-3.5	+0.0	+0.9	-6.1	-0.9	-5.1	-1.3	-2.5
Round lamp L4	-8.3	-7.9	-4.6	-7.4	+3.1	-6.4	+0.0	-6.1	-0.9	-7.6
Round lamp L5	-15.4	-19.0	-7.9	-14.5	-3.6	-12.0	-5.2	-8.1	-0.4	-12.3
Chart area 40%	-10.1	-11.3	+0.4	-5.6	-4.9	-2.3	-4.7	-5.1	-1.7	-9.2
Chart area 20%	-26.6	-22.0	+0.0	-8.4	-8.5	-8.6	-13.4	-2.0	-2.2	-16.8
Motion mild	-1.1	-1.0	+0.0	+0.3	-0.9	-1.4	-0.9	+4.0	-1.3	-0.1
Motion strong	-19.5	-14.9	-0.4	-4.6	-4.5	-9.5	-6.0	+2.0	-2.6	-10.5
Motion L6	-30.0	-20.9	-0.7	-2.0	-2.7	-10.0	-14.7	-12.1	-1.3	-15.6
Defocus mild	-3.0	-2.4	-0.9	-0.3	+0.0	-0.9	+0.4	+1.0	-0.4	-2.3
Defocus strong	-62.6	-52.1	-2.0	-14.8	-9.4	-21.4	-27.6	-10.1	-2.2	-19.5
Defocus L3	-33.1	-25.8	-0.7	-7.6	-3.1	-11.1	-15.9	-9.1	-2.2	-13.8
Low light L3	-2.0	-0.3	+0.4	-3.3	+3.1	-0.7	-0.9	-2.0	+0.9	-1.3
Low light L4	-5.1	-4.7	+0.7	-3.6	-2.7	-5.9	-3.9	-6.1	-0.9	-7.7
Low light L5	-38.8	-33.3	-10.5	-22.4	-4.9	-16.6	-13.4	-10.1	-3.1	-18.1
Overexposure L2	-2.6	-5.2	+0.7	+1.0	+0.4	-2.0	-5.2	+3.0	+2.2	-0.6
Overexposure L3	-5.8	-4.9	+2.4	-4.3	-2.2	-1.8	-1.3	-5.1	-2.6	-1.7
Overexposure L4	-15.7	-12.1	-2.4	-10.4	-1.3	-5.7	-2.6	-10.1	-3.9	-8.7

Table 29: Complete condition–category changes for ChartVerse-8B (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-2.4	-3.1	-4.4	-2.5	-5.8	-10.7	-10.8	-7.1	-8.3	-8.4
Pull S4	-4.9	-5.6	-5.7	-2.0	-7.6	-10.9	-8.6	-10.1	-14.0	-9.0
Pull S5	-5.0	-7.3	-5.7	-1.5	-8.5	-12.0	-12.9	-12.1	-10.0	-11.7
Roller S2	-6.4	-7.3	-6.8	-8.7	-8.5	-18.4	-24.6	-14.1	-17.5	-15.9
Roller S4	-7.2	-11.4	-11.1	-9.4	-13.4	-22.3	-23.3	-16.2	-23.1	-19.5
Roller S5	-9.7	-13.6	-12.0	-8.9	-12.1	-21.6	-28.4	-12.1	-19.2	-21.2
Warm 2700 K	-0.7	-1.1	-0.2	+1.0	-3.6	-3.4	-1.3	-1.0	-3.1	-0.2
Cool 9500 K	-0.7	+0.6	+0.2	+1.8	-0.9	-4.3	-1.3	-3.0	+0.9	-0.8
Candle 1500 K	-2.1	-0.9	-2.0	-1.0	-0.4	-10.5	-8.6	-6.1	-10.9	-7.7
Blue LED	-3.2	-2.4	-3.9	-0.8	-4.0	-14.8	-7.3	-8.1	-9.6	-8.6
Soft shadow	+0.0	-0.4	+0.2	+2.3	+0.9	-3.9	+3.4	+1.0	-4.4	+0.3
Hard shadow	-0.6	-1.3	-0.7	-2.8	-0.9	-0.9	-1.3	-4.0	-1.7	-0.6
Central shadow L4	-2.2	-3.3	-1.3	-1.5	-3.6	-8.0	-11.2	-2.0	-7.4	-7.7
Glass reflection mid	-2.6	-3.4	-3.5	-4.6	-4.9	-8.0	-9.9	-5.1	-12.7	-5.0
Round lamp L4	-6.4	-10.2	-5.9	-3.8	-12.5	-20.7	-15.9	-8.1	-11.4	-13.8
Round lamp L5	-13.4	-19.3	-12.7	-10.2	-23.7	-28.0	-23.3	-14.1	-22.3	-20.4
Chart area 40%	-2.7	-1.9	+0.4	-0.3	-6.2	-7.5	-9.5	-2.0	-4.4	-6.5
Chart area 20%	-5.9	-7.0	-0.4	-4.8	-11.6	-14.8	-15.5	-2.0	-8.3	-11.6
Motion mild	-0.2	-0.5	+2.2	+1.8	-0.9	-1.6	+2.6	+6.1	-5.7	-0.9
Motion strong	-17.7	-16.9	-0.9	-1.5	-25.4	-20.9	-12.9	-4.0	-7.4	-15.9
Motion L6	-29.7	-24.6	-7.0	-6.6	-36.6	-31.6	-27.2	-6.1	-14.4	-25.8
Defocus mild	-2.9	-1.8	-0.4	+1.0	-5.4	-6.4	+0.9	-3.0	+0.4	-5.8
Defocus strong	-69.3	-65.5	-13.3	-20.9	-70.5	-51.8	-50.4	-28.3	-33.6	-36.1
Defocus L3	-32.4	-29.5	-5.2	-9.2	-46.0	-31.4	-25.9	-17.2	-17.9	-24.7
Low light L3	-1.7	-1.2	+1.1	+1.3	+0.9	-4.1	-9.9	-7.1	-6.1	-6.7
Low light L4	-7.5	-7.2	-2.0	-1.0	-8.9	-14.5	-13.8	-3.0	-11.8	-14.3
Low light L5	-57.6	-54.9	-27.7	-38.9	-67.4	-49.8	-43.1	-33.3	-32.3	-36.8
Overexposure L2	-2.3	-4.0	-1.5	-1.3	-3.6	-6.6	-7.3	-3.0	-9.2	-3.4
Overexposure L3	-5.3	-7.5	-3.9	-4.1	-11.2	-6.8	-10.8	-6.1	-6.6	-7.2
Overexposure L4	-16.5	-12.4	-7.4	-10.4	-25.9	-17.5	-11.2	-7.1	-11.4	-9.7

Table 30: Complete condition–category changes for TinyChart-3B-768 (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-0.5	-1.6	-2.6	-3.3	-0.9	-1.6	-1.3	+3.0	-0.9	+0.6
Pull S4	-0.7	-1.8	-0.9	-6.9	-0.9	-0.7	-3.0	+3.0	+0.0	-0.1
Pull S5	-1.3	-1.8	-3.9	-7.9	-0.9	-1.4	-3.4	+4.0	+0.4	-0.9
Roller S2	-0.4	-4.1	-6.6	-12.7	-0.9	-1.8	-4.3	+5.1	-1.7	-1.3
Roller S4	-0.8	-5.0	-0.9	-11.5	-1.3	-3.0	-2.2	-1.0	-0.9	-1.8
Roller S5	-1.3	-5.6	-0.4	-13.0	-2.2	-2.5	-3.4	+1.0	-0.9	-2.3
Warm 2700 K	+0.1	-0.5	+1.5	-0.3	+0.4	+0.0	+0.0	+3.0	+1.3	+1.5
Cool 9500 K	-0.5	+0.5	+0.4	+1.0	+0.9	-0.7	-1.7	+1.0	+1.7	+1.0
Candle 1500 K	+0.0	-0.5	+1.5	-6.6	+0.0	-1.4	+0.0	+1.0	-0.9	-1.4
Blue LED	-0.5	-1.8	+0.7	-9.4	+1.3	-0.5	-2.2	+4.0	+2.6	+0.3
Soft shadow	-0.4	+0.2	+3.3	-0.3	+0.0	+0.2	-0.9	+3.0	+2.6	+0.3
Hard shadow	-0.5	-0.7	-2.6	-3.3	+0.4	+0.5	-1.7	+0.0	+0.9	+1.7
Central shadow L4	-0.1	-0.6	+0.0	-5.1	+0.4	-0.7	-1.3	+3.0	-2.2	-0.3
Glass reflection mid	-0.7	+1.2	+0.0	-4.6	+3.1	-0.9	+0.4	+1.0	+0.4	-0.3
Round lamp L4	+0.1	-3.1	+2.2	-10.7	+0.9	-1.8	-3.0	+3.0	-2.2	+0.1
Round lamp L5	+0.9	-3.0	+1.7	-10.2	+4.9	-1.8	-1.7	+5.1	+3.5	-0.6
Chart area 40%	-0.3	-0.4	+1.1	-3.8	+0.4	-1.6	-2.2	-1.0	+0.0	+0.0
Chart area 20%	-2.5	-3.2	+5.5	-9.7	-0.4	-2.7	-4.3	+1.0	-0.9	+0.5
Motion mild	-0.9	+0.0	+6.1	+1.0	-1.3	+0.5	+0.4	-2.0	+1.3	+2.3
Motion strong	-3.1	-4.6	+1.1	-2.0	+1.3	-2.5	-3.0	+3.0	+1.7	-0.8
Motion L6	-4.5	-3.4	-0.4	-4.8	+2.7	-3.0	-3.9	+0.0	-0.4	-1.8
Defocus mild	-1.1	-1.0	+5.5	+0.5	-0.9	-0.2	+1.7	+2.0	+4.4	+0.2
Defocus strong	-6.6	-1.9	-3.9	-7.1	+2.2	-3.0	-4.7	+3.0	+2.2	-1.2
Defocus L3	-3.4	-2.7	-2.2	-3.1	+1.8	-2.0	-4.7	+3.0	+0.0	-1.7
Low light L3	+0.4	-0.7	-0.9	-4.8	-0.9	-1.6	+0.4	+4.0	-0.4	-0.5
Low light L4	-0.7	-1.5	+0.4	-9.2	+2.7	-2.0	-0.9	+5.1	+0.4	-0.2
Low light L5	-1.7	+28.2	-5.0	-12.5	+0.4	-2.7	-3.4	+3.0	+0.4	+0.1
Overexposure L2	-1.1	+0.2	+3.5	-2.5	+0.0	+0.5	+0.4	+0.0	+2.6	+1.1
Overexposure L3	-1.4	+1.1	+5.7	-4.6	+3.1	-0.2	-1.3	+3.0	+2.2	+0.2
Overexposure L4	-2.9	+1.3	+1.5	-6.9	+2.2	-0.5	-3.0	+1.0	+3.1	-0.8

Table 31: Complete condition–category changes for GPT-5.6 Terra (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-2.6	-0.2	-3.7	-2.0	-14.3	-8.4	-6.5	-6.1	-7.4	-14.5
Pull S4	-3.8	-1.0	-4.1	-2.0	-15.6	-15.2	-9.9	-3.0	-7.4	-18.2
Pull S5	-4.0	-1.5	-4.1	-2.0	-18.8	-17.0	-12.9	-5.1	-11.4	-18.9
Roller S2	-4.2	-2.3	-5.0	-5.1	-18.3	-22.0	-25.0	-12.1	-21.0	-27.9
Roller S4	-5.5	-3.2	-7.0	-7.9	-12.9	-28.6	-30.2	-19.2	-19.2	-32.0
Roller S5	-6.3	-4.1	-7.0	-6.6	-17.4	-29.5	-29.7	-14.1	-24.5	-32.3
Warm 2700 K	-0.9	+0.2	-1.7	-0.8	-3.1	-6.8	-4.3	+0.0	+0.4	-2.7
Cool 9500 K	-0.4	+0.2	-0.7	+0.0	-0.4	-2.7	+0.0	+3.0	-3.1	-1.7
Candle 1500 K	-1.4	-0.6	-2.2	-2.5	-5.8	-11.4	-9.9	-5.1	-10.5	-13.6
Blue LED	-2.3	-0.8	-3.1	-1.5	-8.0	-16.4	-13.4	-12.1	-12.7	-17.5
Soft shadow	+0.0	+0.2	-1.3	+1.0	+2.2	-2.5	+2.2	+2.0	+4.4	+1.7
Hard shadow	-0.6	+0.6	-2.2	+1.0	-1.8	-3.6	-3.4	-1.0	+0.0	-2.0
Central shadow L4	-2.3	-0.2	-2.2	-0.8	-2.7	-13.2	-16.4	-4.0	-14.8	-15.7
Glass reflection mid	-1.2	+0.2	-2.4	+0.3	-7.1	-10.5	-8.2	-3.0	-6.1	-13.0
Round lamp L4	-3.7	-1.9	-3.5	-5.3	-14.3	-21.1	-20.3	-11.1	-13.5	-19.0
Round lamp L5	-8.4	-5.2	-8.7	-10.9	-17.9	-31.8	-25.0	-20.2	-23.6	-28.3
Chart area 40%	-1.3	-0.2	-2.8	+0.3	-4.9	-5.0	-5.6	+1.0	-6.6	-11.7
Chart area 20%	-2.5	-1.3	-3.5	-0.8	+1.8	-10.7	-12.9	-3.0	-7.0	-15.6
Motion mild	-0.2	-0.2	-1.3	+0.0	-1.3	-2.0	+0.0	+2.0	+2.6	-1.0
Motion strong	-20.6	-13.1	-3.3	+0.3	-12.1	-36.4	-25.0	-10.1	-10.0	-26.8
Motion L6	-32.0	-20.4	-4.6	-3.6	-18.3	-47.3	-37.5	-15.2	-23.6	-40.2
Defocus mild	-3.1	-2.1	-0.7	+0.3	-8.0	-6.6	+0.9	+1.0	+3.5	-5.6
Defocus strong	-60.0	-38.5	-8.5	-17.8	-27.7	-62.0	-62.1	-28.3	-34.1	-47.6
Defocus L3	-32.8	-18.3	-4.4	-3.6	-18.3	-45.2	-39.2	-11.1	-24.0	-36.2
Low light L3	-2.4	-1.0	-2.6	-0.3	-6.2	-11.8	-12.5	-6.1	-3.9	-14.2
Low light L4	-10.6	-7.3	-4.6	-4.8	-9.4	-30.9	-18.5	-13.1	-16.2	-27.8
Low light L5	-66.3	-44.8	-35.6	-51.4	-33.0	-63.2	-66.4	-37.4	-38.9	-46.0
Overexposure L2	-1.5	-0.4	-2.0	-1.0	-8.9	-5.7	-9.9	-2.0	-6.1	-5.8
Overexposure L3	-3.4	-2.0	-3.1	-2.0	-8.0	-11.8	-13.8	-4.0	-7.4	-9.3
Overexposure L4	-11.0	-5.2	-5.2	-7.6	-15.6	-23.0	-20.3	-8.1	-7.9	-15.0

Table 32: Complete condition–category changes for Claude Sonnet 5 (pp from clean). Every native category retains its full question population; VE uses all 1,151 probes.

Condition	IE	EN	PR	CT	CP	TIC	NIC	TIG	NIG	VE
Pull S2	-2.3	-0.4	+0.0	-4.3	+1.8	-6.8	-3.4	-8.1	-12.7	-10.1
Pull S4	-3.1	-1.1	-0.7	-1.8	-4.9	-11.6	-6.9	-3.0	-13.5	-10.3
Pull S5	-3.5	-1.1	-0.9	-1.0	-5.8	-12.3	-6.9	-8.1	-12.2	-11.2
Roller S2	-3.0	-1.6	-1.7	-6.4	-3.1	-16.8	-15.1	-11.1	-18.8	-16.4
Roller S4	-3.6	-1.9	-4.6	-6.9	-2.7	-17.5	-14.7	-9.1	-20.1	-18.9
Roller S5	-4.9	-2.4	-4.6	-7.9	-0.4	-20.7	-16.4	-6.1	-17.5	-17.8
Warm 2700 K	+1.3	-0.1	-0.4	+2.0	+3.6	-4.3	+0.9	+0.0	-0.4	-1.4
Cool 9500 K	+0.4	-0.3	+2.2	-1.3	+4.5	-1.6	+0.0	+1.0	-1.7	-1.2
Candle 1500 K	-0.3	-0.1	+0.7	-1.0	+0.9	-8.0	-2.6	-3.0	-8.7	-8.9
Blue LED	-2.6	-1.9	+1.1	-2.5	+3.6	-15.0	-5.6	+0.0	-11.8	-8.5
Soft shadow	+0.0	+0.5	+0.0	+0.8	+2.2	-0.2	+0.0	+1.0	+1.3	-1.1
Hard shadow	-0.3	-0.5	+1.1	+0.3	+0.4	-2.7	-1.7	+1.0	-2.6	-1.3
Central shadow L4	-1.8	-0.7	-0.7	-1.3	-1.8	-11.4	-4.3	+0.0	-10.9	-8.0
Glass reflection mid	-1.2	+0.1	-1.3	-2.5	+4.5	-5.7	-1.7	-7.1	-12.2	-10.2
Round lamp L4	-3.9	-2.3	-2.0	-3.6	-0.4	-19.1	-8.2	-6.1	-16.6	-12.9
Round lamp L5	-11.2	-7.3	-5.2	-11.5	-7.1	-31.6	-16.8	-18.2	-22.7	-24.1
Chart area 40%	-1.0	+0.0	+0.0	-0.3	-4.0	-5.2	-1.3	-4.0	-9.6	-8.9
Chart area 20%	-4.3	-1.2	-0.7	-1.8	+0.4	-13.0	-2.2	-2.0	-10.9	-12.1
Motion mild	-0.9	-0.2	+2.4	+0.3	+4.9	+0.2	+1.7	+3.0	-1.3	+1.3
Motion strong	-13.4	-6.9	+0.2	+0.0	-1.3	-18.4	-12.1	+0.0	-10.5	-15.5
Motion L6	-22.4	-14.6	-2.6	-2.3	-4.5	-30.0	-17.7	-13.1	-16.6	-24.5
Defocus mild	-1.1	+0.2	+1.3	+0.5	+1.8	-3.2	-2.6	+2.0	-0.4	-1.1
Defocus strong	-64.5	-42.5	-12.2	-21.9	-16.1	-62.3	-51.3	-22.2	-31.9	-42.7
Defocus L3	-23.0	-15.3	-1.1	-3.3	-8.9	-27.0	-15.1	-12.1	-17.9	-24.1
Low light L3	-2.3	-0.6	+0.9	+0.5	-2.7	-6.6	-3.9	-6.1	-7.9	-7.6
Low light L4	-2.5	-1.0	-1.1	+0.0	-0.4	-10.7	-6.5	-1.0	-14.0	-12.5
Low light L5	-35.9	-28.3	-18.6	-24.7	-15.2	-48.0	-29.7	-18.2	-30.1	-37.2
Overexposure L2	-1.6	-0.9	-0.2	-0.3	-1.8	-5.2	-5.2	-1.0	-7.0	-6.3
Overexposure L3	-4.1	-2.4	+0.4	-4.3	-3.6	-9.5	-5.2	-9.1	-10.9	-9.3
Overexposure L4	-11.4	-6.8	-1.5	-5.6	-8.9	-16.4	-11.6	-10.1	-13.5	-11.6