Scale Is Not Symmetric: Directional Scale Reliability in Chart Question Answering

Anonymous ACL submission

Abstract

Chart question answering models must preserve the numerical scale expressed by visual marks, axes, labels, and units. We study a sparse but high-impact reliability failure in which a model predicts a numerically related answer at the wrong scale, such as *0.61* instead of *61*. The tempting repair is to normalize numbers or train a one-direction scale adapter; our central finding is that this is the wrong abstraction. Scale correction is directional and regime-dependent. We construct ScaleQA-Natural, an author-adjudicated and independently inclusion-audited diagnostic set of 520 natural scale-conflict candidates, and use adapters, prompts, structured verifiers, calculators, and policy selectors as probes. In an answer-tuned Qwen2.5-VL-7B setting, a large-positive scale-contrastive adapter improves held-out ChartQA test accuracy from 88.60 to 89.32 and recovers 49.52% of strict test-set gold-scale errors. Yet on ScaleQA-Natural, self-check prompting and a lightweight no-leak selector outperform a global adapter, while candidate oracles reach 90–92%. The bottleneck is therefore not simply generating scaled alternatives, but selecting the scale regime required by the chart and question. Cross-backbone and external numeric checks show the same boundary: scale-like failures transfer more robustly than a particular correction recipe. Our results argue for regime-aware scale reliability evaluation and caution against unconditional numeric correction.

1 Introduction

Charts make quantitative relations visible, but they also make scale a moving target. The same visual evidence can be verbalized as a percentage, a decimal, a raw value, or a value in thousands or millions. A chart question answering model must therefore do more than find the relevant bar, line, or label: it must decide which scale the answer should use.

Modern multimodal large language models (MLLMs) achieve strong aggregate accuracy on chart QA benchmarks such as ChartQA (Masry et al., 2022). Aggregate accuracy can nevertheless hide rare failures that are costly when they occur. We study *scale inconsistency*: a model gives a numerically related answer under the wrong scale, such as predicting *0.61* when the chart question requires *61*. This is not just a generic numeric miss. The model has often found a related quantity, but has failed at the last reliability step: choosing the scale in which to express it.

The obvious fix is tempting and wrong. One might normalize all percentages, train an adapter that prefers larger numbers, or ask the model to self-check units. Each helps some cases and hurts others. A legitimate answer such as *0.0593* should not be expanded into *6.0333*; an axis in millions should not be treated like a percent display; and a prediction that is already too large needs the opposite intervention from a prediction that is too small. The central question is therefore not simply whether a model can produce a scaled alternative. It is whether a system can select the correct *scale regime*: the direction, unit context, and output convention required by the chart and question.

Natural strict scale errors are sparse under any single model and split, so the evidence must be read at two levels. Across the available ChartQA validation/test outputs we audit, strict factor-of-100 errors account for 0.12–1.09% of examples (Table 12); our original ChartQA validation analysis contains 23 such examples. We use these subsets as natural recovery probes, not as independent benchmarks. To understand the mechanism behind those

sparse failures, we construct ScaleQA-Natural, a mined, machine-audited, author-adjudicated, and independently inclusion-audited set of 520 natural scale-conflict candidates. ScaleQA-Natural is deliberately framed as a diagnostic set, not as a replacement for official ChartQA evaluation.

Our main empirical finding is that scale correction is directional and regime-dependent. In an answer-tuned Qwen2.5-VL-7B setting, a large-positive scale-contrastive adapter improves held-out ChartQA test accuracy from 88.60 to 89.32 and recovers 49.52% of strict test-set gold-scale errors. This +0.72 point gain is real, but it is not the paper's main claim: it is measured after answer tuning and should not be confused with the much larger raw-to-answer-tuned improvement. The more important result appears when the same idea is stress-tested. On ScaleQA-Natural, a self-check prompt and a lightweight no-leak policy selector outperform the global adapter, while high candidate oracles show that the correct scaled answer is often present in the candidate pool. The bottleneck is selection.

This changes the role of adaptation in the paper. We use LoRA adapters as diagnostic probes for directional preferences, not as a universal repair mechanism. Their successes and failures reveal when a scale preference is helpful, when it is harmful, and why unconditional correction is unsafe. Cross-backbone and external numeric checks reinforce the same lesson: the phenomenon transfers more robustly than the intervention.

Our contributions are:

- a strict natural scale-error recovery probe and ScaleQA-Natural, an audited diagnostic set for natural scale conflicts;
- evidence that scale correction is directional: a large-positive adapter repairs nearly half of strict held-out test scale errors, but is unsafe as a global policy;
- a comparison of prompts, verifiers, calculators, likelihood scoring, and no-leak selectors that separates candidate generation from candidate selection;
- a persistent bottleneck: the best deployable selector modestly beats self-check, but large policy and candidate-oracle gaps remain.

2 Related Work

Chart question answering. Chart question answering requires models to combine visual perception, OCR, structural grounding, and numerical reasoning. Earlier benchmarks such as DVQA, FigureQA-style synthetic chart reasoning, and PlotQA emphasize controlled visual reasoning over chart data (Kafle et al., 2018; Methani et al., 2020). ChartQA (Masry et al., 2022) evaluates both human-written and machine-generated chart questions that involve value lookup, comparison, and arithmetic. Recent chart-specialized systems add plot-to-table reasoning, chart derendering, chart-specific pretraining, and visual instruction tuning (Liu et al., 2023a,b; Masry et al., 2023; Han et al., 2023; Masry et al., 2024). Recent benchmarks also stress more complex and realistic charts (Xu et al., 2023; Wang et al., 2024). Our work studies a narrower reliability issue within this setting: the model may identify a related numerical value but report it under the wrong scale.

Numerical reliability in MLLMs. Numerical errors in multimodal models can arise from visual grounding failures, OCR mistakes, wrong axis interpretation, arithmetic errors, or answer normalization artifacts. Scale inconsistency differs from a generic numeric miss because the prediction is close under a simple multiplicative transformation. This makes it easy to miss in aggregate error analysis and risky in deployment, since an answer like 0.61 can look superficially related to 61 while being semantically wrong. The distinction also changes the evaluation target: the question is not only whether a model can compute or retrieve a number, but whether it can choose the output convention supported by the chart's axis, unit, and question wording.

Diagnostic evaluation and correction policies. Most chart QA work reports aggregate task accuracy, while reliability failures are often analyzed post hoc. We instead treat scale conflicts as a diagnostic slice with inclusion criteria, an audit trail, and fix–break accounting. A useful correction must recover scale errors, preserve non-scale examples, and avoid applying the wrong direction. The high oracle gaps in our experiments further distinguish

candidate generation from candidate selection, a failure mode not captured by single-answer accuracy alone. We use TextVQA (Singh et al., 2019) only as an adjacent text-rich safety check.

Parameter-efficient and preference-style adaptation. LoRA updates a model through low-rank trainable matrices while keeping the base weights fixed (Hu et al., 2021). This makes it attractive for targeted reliability interventions in large MLLMs. Contrastive and preference-style objectives, including direct preference optimization (Rafailov et al., 2023), optimize relative preferences between candidate outputs. Our objective is more local: for a chart and question, we contrast a preferred numeric answer with a scale-related alternative, and study how the direction of this contrast changes behavior.

3 Diagnosing Scale Inconsistency

We use three views of the same reliability problem, because no single metric tells the whole story: official ChartQA accuracy measures task utility, strict natural scale-error recovery measures whether real scale mistakes can be repaired, and ScaleQA-Natural exposes why a repair that helps one regime can damage another.

Strict natural scale errors. Let y be the gold answer and $\hat{y}$ a model prediction. We extract a single numeric value $n(\cdot)$ when possible and call an example a *strict gold-scale error* if the baseline prediction is incorrect, both answers contain one nonzero number, and their ratio is approximately 100 or 1/100:

$$\begin{aligned} &\frac{|n(\hat{y})|}{|n(y)|} \in [80, 125], \\ \text{or} \quad &\frac{|n(\hat{y})|}{|n(y)|} \in [0.008, 0.0125]. \end{aligned} \tag{1}$$

The baseline accuracy on this subset is zero by construction, so we treat it as a recovery probe, not as an independent benchmark.

ScaleQA-Natural. Because strict errors from a single model and split are sparse, we mine scale-conflict candidates from ChartQA predictions across model outputs and split shards, then review them with a machine-assisted annotation workflow. ScaleQA-Natural makes this rare failure mode measurable without turning it into a hidden training set or leaderboard. It contains 520 author-adjudicated positives, of which 473 are directly image-evaluable with available ChartQA images. Released items include question, gold answer, source row, model prediction, subtype, direction, review label, and audit metadata; images are referenced through original ChartQA rows. We also retain the 403-item seed set, *ScaleQA-Core*, and a 192-example InternVL-mined subset for source-distribution analysis.

Group	Item	Count
Reviewed diagnostic set	Annotated ScaleQA-Natural items	520
Reviewed diagnostic set	Directly image-evaluable items	473
Reviewed diagnostic set	Core reviewed seed set	403
Reviewed diagnostic set	Image-assisted additions	86
Reviewed diagnostic set	Other self-annotated additions	31
Diagnostic sets	ScaleQA-Natural-520	520
Diagnostic sets	ScaleQA-Natural-520, image-evaluable	473
Diagnostic sets	ScaleQA-Natural-Core	403
Diagnostic sets	InternVL-mined subset	192
Subtypes	Value scaled down	120
Subtypes	Decimal to percent	114
Subtypes	Value scaled up	104
Subtypes	Percent to decimal	96
Subtypes	Minor factor or unit candidate	42
Subtypes	Minor factor percentage candidate	28
Subtypes	Untyped or mixed factor candidate	16

Table 1: ScaleQA-Natural construction statistics. We use it as a diagnostic set; official ChartQA validation/test remain the task evaluations.

Split hygiene and audit. The main Qwen scale adapters are trained from ChartQA training rows and training-set prediction/log-probability caches. ChartQA validation is used for development; held-out ChartQA test is reserved for the main task result. ScaleQA-Natural is used for diagnosis and policy selection analysis, not for training the reported Qwen scale adapters. Because ScaleQA-Natural is mined rather than independently authored, we run two checks after the first-pass review: a deterministic metadata audit that flags suspicious inclusions, and an independent human audit of the final positive set. Appendix B gives the data-flow table, audit statistics, and release details.

Table 2 makes the reliability boundary explicit. The metadata audit has moderate agreement with the first-pass labels ($\kappa = 0.443$) and flags 108 final positives for manual adjudication. After adjudication, an independent second reviewer audited the 520 final positives and made no label changes. We therefore describe ScaleQA-Natural as an author-adjudicated and independently inclusion-audited diagnostic set,

Audit item	Count/Score	Notes
Candidate pool	739	First-pass reviewed candidates
Observed agreement	0.633	Three-way labels
Cohen's κ	0.443	Include/review/reject
Conflicts for review	271	Exported separately
Final positives	520	Author-adjudicated
Second-pass include	412	Deterministic audit
Second-pass review/reject	108	Manually adjudicated
Independent inclusion audit	520	Second human reviewer
Second-review agreement	1.000	Final positives; zero label changes

Table 2: ScaleQA-Natural audit. The metadata pass flags conflicts for author adjudication; a second human reviewer independently audited the final positive inclusions.

not a fully independent human-authored benchmark.

We also audit the negative side of the mining process. Automatically mined candidates that remain uncertain or are rejected often involve plausible numeric ratios but insufficient evidence that the prediction and gold answer refer to the same chart quantity. These cases are excluded from ScaleQA-Natural rather than treated as positives. Table 12 reports strict-error prevalence across the available model/split outputs, and Table 13 gives representative uncertain/rejected examples.

Annotation guideline. Annotators mark an item as a scale diagnostic only when the prediction and gold answer are both grounded in the chart question and differ primarily by a simple multiplicative scale relation, such as percentage versus decimal form or a visible unit-scale mismatch. Items are rejected when the prediction reflects the wrong visual entity, wrong operation, OCR failure, ambiguous wording, duplicate evidence, or an arbitrary numeric miss. Ambiguous cases are assigned to a review/uncertain queue and are not included without adjudication. This guideline is designed to make ScaleQA-Natural a reusable diagnostic artifact while keeping the benchmark claim narrow.

Why this decomposition matters. Overall accuracy can remain stable while scale reliability changes sharply, and a correction can recover a few strict errors while breaking many non-scale examples. This decomposition is the bridge between the dataset and the experiments: ChartQA test asks whether an intervention helps the task, strict scale subsets ask whether it repairs natural scale errors, and ScaleQA-Natural asks why the same intervention helps one regime but harms another. We therefore report full accuracy, strict scale-error recovery, non-scale accuracy, and safety checks whenever possible.

Chart + question → Base answer (e.g., *0.61*)

Candidate pool: keep base, self-check answer, directional adapter answers, and explicit scale variants.
Regime evidence: question form, answer magnitude, percent/unit tokens, axis labels, tick values, and candidate likelihoods.
Policy decision: base, self-check, large-positive, small-positive, decimal guard, or unit guard.

Prediction too small: 0.61 → 61
Prediction too large: 61 → 0.61
True small decimal: keep 0.0593
Unit mismatch: millions vs. raw values

Figure 1: Regime-aware diagnostic view. Our experiments test whether scale-related candidates can be generated and, more importantly, whether a deployable policy can select the scale regime required by the chart and question.

4 Diagnostic Probes and Policy Selection

We freeze the MLLM backbone and train LoRA parameters Δ (Hu et al., 2021). For a chart image x, question q, and answer y, the adapted model defines $p_{\theta,\Delta}(y \mid x, q)$, with frozen θ. The methods below are organized as a probe ladder rather than a single proposed fix. Directional adapters test whether scale preference can be moved; clean-direct controls test whether that movement is safe without answer tuning; policy selectors test the harder question of whether candidate evidence can be turned into a deployable regime decision.

Directional adapters as probes. For numeric training examples, we construct scale-related pairs (y^+, y^-). *Large-positive* pairs prefer the larger-scale value; *small-positive* pairs prefer the smaller-scale value. We also evaluate balanced mixtures, all mined pairs, and real-conflict pairs mined from model disagreements. The names describe the training preference, not the observed error direction: an answer that is too small requires a large-positive correction, while an answer that is too large requires the opposite.

Objective. We compute mean token negative log-likelihood $\mathrm{NLL}(y \mid x, q)$ and minimize

$$\mathcal{L}_{\text{scale}} = \mathrm{softplus}\big(-\beta[\mathrm{NLL}(y^-) - \mathrm{NLL}(y^+) - m]\big), \tag{2}$$

where m is a margin and β controls sharpness. The loss encourages the preferred answer to have lower NLL than its scale-related alternative.

Adaptation regimes. The main positive result uses an *answer-tuned* starting point: scale adaptation continues from an answer-only LoRA adapter in the Qwen2.5-VL-7B pipeline. We also test *clean direct* scale-only tuning from the raw backbone to avoid conflating answer tuning with scale tuning. Direct scale-only tuning is unsafe, so we report it as a negative control. At inference time, main results use the same deterministic short-answer decoding template as the baseline, without candidate reranking or external calculators.

Regime-aware policy selection. ScaleQA-Natural suggests that no single direction is safe across all regimes. We therefore evaluate policy selectors that choose among base answering, self-check prompting, unit/axis guards, and directional adapters. The deployable selectors use question text, base prediction text, numeric features, and chart-context strings extracted from the image, including title, axis labels, tick values, legends, and units. To avoid reporting a selector that memorizes the diagnostic labels, the main no-leak selector trains and selects hyperparameters on held-out splits without using gold subtype or direction fields. Oracle subtype, policy-pool oracle, and candidate-pool oracle results are reported only as upper bounds; their role is to show how much headroom remains if regime detection were solved.

5 Experiments

Setup. The main experiments use Qwen2.5-VL-7B-Instruct (Qwen Team, 2025). We use ChartQA validation for development and held-out ChartQA test for the main task result (Masry et al., 2022). We also evaluate InternVL2.5-1B/8B (OpenGVLab, 2024) and numeric subsets of PlotQA (Methani et al., 2020), ChartBench (Xu et al., 2023), MathVista (Lu et al., 2023), and InfographicVQA (Mathew et al., 2022). Numeric answers are scored with the repository evaluator: normalized exact match for text answers and 5% tolerance for numeric answers. The final Qwen scale adapters use $K = 300$ pairs, 300 steps, learning rate 10^{-6}, $\beta = 2.0$, and margin $m = 0.10$.

System	Starting point	Test acc.	Std.	Delta
Raw Qwen2.5-VL-7B	Raw backbone	0.7804	–	–
Answer-tuned LoRA	Answer-only	0.8860	–	+0.1056
Answer-tuned + large-positive scale	Answer-tuned	0.8932	0.0003	+0.0072
Clean direct large-positive scale	Raw backbone	0.7456	–	-0.0348
Clean direct small-positive scale	Raw backbone	0.7136	–	-0.0668

Table 3: Held-out ChartQA test attribution for Qwen2.5-VL-7B. The scale update is measured on top of the answer-tuned adapter; clean direct scale-only tuning is a negative control.

The experiments follow Figure 1. Held-out ChartQA asks whether a directional update can repair real task errors; ScaleQA-Natural asks whether it remains safe across regimes. Prompting, verification, and policy selection then test whether the correct answer is missing from the candidate set or merely not selected.

5.1 Held-Out ChartQA

The first question is deliberately narrow: can a scale-directed intervention fix natural scale errors without damaging the main task?

Table 3 separates answer tuning from scale tuning. The raw Qwen2.5-VL-7B pipeline obtains 0.7804 on ChartQA test. Answer tuning reaches 0.8860. Continuing from that adapter, large-positive scale-contrastive tuning reaches 0.8932 over five seeds, a +0.0072 gain over the answer-tuned baseline, with paired bootstrap 95% CI $[+0.0038, +0.0109]$. In contrast, clean scale-only adaptation from the raw backbone hurts test accuracy. Scale supervision can repair a real direction of error once the model is in a stable chart-answering regime, but it is not a standalone numeric fix.

This attribution also explains why the answer-tuned setting is the only setting where the scale update is useful. The answer-only adapter already improves short-answer formatting, OCR-sensitive answer extraction, and ChartQA-style response alignment. The scale adapter then makes a small preference update inside this stabilized answer space. Starting from the raw backbone, the same scale-only objective is too narrow: it changes numeric preferences without first improving the model's general answer format and grounding behavior.

Split	Base	Ours	Std.	Delta	N
All test examples	0.8860	0.8932	0.0003	+0.0072	2500
Strict gold-scale errors	0.0000	0.4952	0.0543	+0.4952	21
Non-gold-scale examples	0.8935	0.8966	0.0005	+0.0031	2479

Table 4: Held-out ChartQA test subset analysis. The strict subset is baseline-zero by construction and is used only as a recovery probe.

Table 4 shows where the gain comes from. On strict test-set gold-scale errors, the answer-tuned baseline scores 0 by construction, while the scale adapter recovers 49.52% of the 21 examples. Non-gold-scale accuracy also increases slightly, from 0.8935 to 0.8966. TextVQA-500 remains unchanged at 0.9420, suggesting no measurable degradation on this adjacent text-rich VQA safety set.

5.2 ScaleQA-Natural

The second question is whether the same intervention is safe when scale conflicts are collected across regimes rather than filtered to one held-out recovery subset.

ScaleQA-Natural reveals why a single aggregate number is insufficient. On ScaleQA-Natural-520, answer tuning improves the raw backbone from 0.4567 to 0.5095, while a scale-aware prompt drops to 0.4419 and the global large-positive adapter reaches 0.4951. On ScaleQA-Core, balanced and all-pair variants outperform the answer-tuned base; on the InternVL-mined subset, small-positive adaptation is much stronger than every other global policy. Thus, directionality is real, but it is not simply "large is good, small is bad." The safe direction depends on the source and subtype of the scale conflict.

Table 6 adds a stronger non-training baseline. A self-check prompt reaches 0.5455, outperforming the direct answer prompt and the global large-positive adapter on ScaleQA-Natural-520. Majority voting gives little improvement, while oracle prompt choice reaches 0.6744. These results weaken any claim that a global adapter is the best intervention; they strengthen the diagnostic conclusion that scale correction depends on selecting the right regime or prompt behavior.

5.3 Cross-Backbone and External Checks

The final experimental question is scope: does the reliability phenomenon extend beyond the main Qwen setting, and does the intervention transfer with it?

Table 7 summarizes the broader transfer picture. InternVL2.5-1B and InternVL2.5-8B can both be adapted with the same scale-pair recipe, showing that the experiment is not Qwen-only. We also evaluate InternVL2.5-26B as a stronger non-Qwen diagnostic backbone on ScaleQA-Natural, PlotQA, ChartBench, MathVista, and InfographicVQA subsets. The result is not a clean transfer win for the adapter: InternVL2.5-8B shows only tiny differences among large, small, and balanced policies, all Qwen7 stress-set variants tie, and Qwen7 external ChartBench/MathVista/InfographicVQA scores are unchanged across adapters. The 26B base is useful as a stronger diagnostic model, but these checks diagnose prevalence and difficulty rather than adaptation success. We therefore use cross-backbone results to define the method's scope: the phenomenon transfers more robustly than the intervention.

6 Analysis

The experiments leave a consistent pattern: scale interventions can create useful alternatives, but the system often fails at deciding when to trust them. This section turns that pattern into the paper's mechanism account.

6.1 Pair Direction Is Not Symmetric

Table 8 reports the development ablation that motivated the method. Large-positive pairs give the best full-validation score among these single-policy variants, while small-positive pairs reduce full accuracy by 1.51 points. The key point is not that small-positive supervision is useless: ScaleQA-Natural shows regimes where it is best. Rather, pair direction changes both which scale conflicts are repaired and which non-scale examples are broken. On ChartQA validation, the large-positive adapter fixes an average of 56.52% of the 23 strict gold-scale errors across seeds, while breaking 1.27% of base-correct examples. Appendix D gives the full fix–break table, subtype table, and case studies.

Set	N	Raw	Ans.	Prompt	Large	Bal.	All	Real	Small
ScaleQA-Natural-520	473	0.4567	0.5095	0.4419	0.4951	–	–	–	–
ScaleQA-Core	403	0.4591	0.4963	0.4293	0.4844	0.5092	0.5047	0.5002	0.4174
InternVL-mined	192	0.1510	0.3281	0.0990	0.1896	0.3323	0.2771	0.2510	0.7490

Table 5: ScaleQA-Natural diagnostic results. ScaleQA-Natural-520 reports the image-evaluable subset; ScaleQA-Core and InternVL-mined report corrected five-seed pair-mode means.

Prompting strategy	Accuracy
Direct answer	0.5243
Axis-aware prompt	0.5307
Math-check prompt	0.5137
Scale prompt	0.4905
Self-check prompt	0.5455
Unit-guard prompt	0.5180
Majority vote	0.5264
Oracle prompt choice	0.6744

Table 6: Training-free prompting baselines on ScaleQA-Natural-520 using the answer-tuned Qwen2.5-VL-7B pipeline.

Check	Result	Takeaway
InternVL2.5-8B ScaleQA	0.4672 → 0.4723	tiny adapter gain
InternVL2.5-26B external	0.4080–0.7200	stronger diagnostic backbone
Qwen7 external adapters	no change	no broad transfer win
ScaleQA-stress1000	all 0.6940	global policies tie

Table 7: Transfer boundary summary. Full cross-backbone and external rows are in Appendix D.

Method	Accuracy	Delta
Large-positive	0.863500	0.003100
Real conflicts	0.862500	0.002100
Balanced	0.860900	0.000500
All pairs	0.860900	0.000500
Base	0.860400	0.000000
Small-positive	0.845300	-0.015100

Table 8: Pair-mode ablation on ChartQA validation. Small-positive adaptation is locally useful in some regimes but globally unsafe on this split.

Set	Fixed best	Pred. router	Router acc.	Oracle subtype	Instance oracle
ScaleQA-Core	0.4997	0.5071	0.5444	0.8077	0.8081
InternVL-mined	0.7292	0.7292	0.6918	0.7741	0.7792

Table 9: Subtype routing on ScaleQA-Natural diagnostics. Fixed best is the best single global policy. Oracle columns use gold information and are upper bounds.

6.2 Regime Selection Is the Bottleneck

The subtype reversal suggests that scale correction should be selected rather than applied globally. Table 9 first evaluates a simple subtype router. On ScaleQA-Core, the best fixed policy scores 0.4997 and the predicted subtype router reaches 0.5071. This gain is small because subtype prediction is still weak. In contrast, the oracle subtype router reaches 0.8077, nearly matching the instance oracle. The gap identifies the missing component: reliable scale-regime detection, not another unconditional scale adapter. This also explains why training-free self-check prompts are competitive on ScaleQA-Natural; different examples require different correction behavior.

The predicted router uses only cheap question/base-answer TF–IDF and numeric features, so its small gain is a lower-bound diagnostic rather than a strong routing result. The oracle gap suggests a concrete path for future systems: add chart-side evidence such as OCR text, axis labels, units, tick values, or visual embeddings before applying a scale correction.

We next test stronger deployable selectors: multi-candidate verifiers, learned and rule-based routers, consensus guards, structured chart-context verification, likelihood scoring, Bayesian fusion, and a DePlot-style table-extraction-plus-calculation baseline (Table 25). The progression is important: v8 is clean but weak, v9 raises the policy pool but ties self-check, and v10b is the first no-leak selector that clears self-check. Later variants add log-probability, rationale-verifier, candidate-likelihood, and top-8 evidence, but preserve rather than improve v10b. The best no-leak selector remains 0.5667, modestly above split-matched self-check at 0.5514. This shows that deployable regime evidence is useful, but its distance from the 0.9004–0.9260 oracles is much larger than its gain over self-check.

The gain is small, but its shape is useful. The selector wins by being conservative: most examples still use a robust fixed behavior, and only a minority switch. More ambitious statistical fusion, bucketed Bayesian priors, and Bayesian-MLP variants over-switch away from self-check. Extra features are predictive on development

partitions, but not calibrated enough to override a safe fixed prompt.

This gives the main methodological lesson. Regime-aware selection is necessary, but more candidate evidence is not automatically better. The useful selector changes behavior only when the scale evidence is strong enough. The experiments therefore separate two capabilities that are easy to conflate: producing a plausible scaled answer and knowing whether it should be used.

The remaining failures are structured. In the image-evaluable ScaleQA-Natural subset, most candidates involve percent evidence in the question or axis context (92.8%). Question forms are diverse: percent questions, averages, direct value lookup, differences, and extrema all appear. The large-positive adapter fixes 5.71% of examples that the answer-tuned model misses, but also breaks 7.40% of examples that it gets right. The structured verifier exposes the same asymmetry at the example level: it performs well when the original prediction is too small, but fails when the original prediction is too large, despite seeing the same image and candidate list. This fix–break asymmetry explains why global adaptation helps strict ChartQA test errors but does not dominate on ScaleQA-Natural: the correct policy depends on question form, unit evidence, and the direction of the original scale conflict.

6.3 What the Negative Results Establish

The negative results rule out simple explanations for the oracle gap. Table extraction plus calculation tests whether the issue is only missing arithmetic; it is not. Structured verifiers test whether exposing chart context and explicit scale variants is enough; it is not. Candidate-likelihood and pairwise-verifier variants test whether the model internally prefers the right candidate once shown a compact list; again, the answer is no. The repeated pattern is sharper: evidence and alternatives can be present without a reliable decision about output scale.

This is the reliability boundary we want to expose. Global adapters can repair one direction of scale error, self-check prompts can be strong fixed policies, and lightweight selectors can improve slightly over them. But when the candidate pool expands, most additional headroom appears only in the oracle. The missing capability is calibrated regime detection: deciding when *0.61* is a decimal answer, when it is a percent written as a decimal, and when a larger or smaller unit is required by the axis. Cross-backbone and external checks add the same caution: the phenomenon travels farther than the specific correction recipe.

7 Conclusion

We studied directional scale inconsistency in chart question answering. The main empirical finding is not that one adapter universally fixes scale errors, but that scale correction is regime-dependent. In an answer-tuned Qwen2.5-VL-7B setting, large-positive scale-contrastive adaptation gives a small held-out ChartQA test gain and recovers nearly half of strict test-set scale errors. ScaleQA-Natural shows why the same intervention is not a general repair: self-check is stronger than a global adapter, small-positive policies help different regimes, and aggressive selectors can get worse by switching too often.

The strongest evidence is the gap between candidate availability and candidate choice. Structured verifiers, likelihood scoring, pairwise comparison, and top-8 candidate pools frequently contain a correct scaled answer, yet deployable selectors often fail to choose it. Our best no-leak selector improves over self-check, but remains far below policy and candidate oracles. Cross-backbone and external numeric checks reinforce the same boundary: scale-like failures are not unique to one Qwen setting, while the correction recipe does not transfer automatically.

Directional adapters reveal which regimes are helped or broken, and selectors test whether evidence becomes reliable policy. Chart QA systems should evaluate scale reliability by regime, with fix–break accounting and candidate-selection analysis alongside aggregate accuracy.

Limitations

The strict naturally occurring scale-error subsets on official ChartQA are small. We therefore use them as recovery probes and rely on ScaleQA-Natural for broader regime analysis, but ScaleQA-Natural is still a mined diagnostic set rather than a fully independent human-authored benchmark. Its final 520 positive inclusions are author-adjudicated and independently audited by a second human reviewer, but prevalence estimates remain tied to the mining sources. The main positive adaptation result is obtained with Qwen2.5-VL-7B-Instruct in an answer-tuned setting. We add InternVL2.5-1B, 8B, and 26B diagnostics, including five-seed 8B adaptation, but these experiments show bounded transfer rather than universal improvement. The external PlotQA, ChartBench, MathVista, and InfographicVQA subsets are useful robustness checks, but they are not full-dataset benchmark submissions and should not be interpreted as state-of-the-art comparisons. Our stronger baselines include prompt variants, multi-candidate verification, learned and rule-based selectors, statistical policy fusion, Bayesian bucket/MLP selectors, structured chart-context verification, token-likelihood candidate scoring, and a DePlot-style table-extraction-and-calculation pipeline, but these are still lightweight baselines rather than specialized chart parsers. The scale definition focuses primarily on factor-of-100 style errors, especially percentage-decimal confusions, although ScaleQA-Natural includes smaller groups of 10x, 1000x, and unit-like candidates. Finally, the no-leak selector improves over self-check only modestly and remains far below oracle candidate selection. A deployable system would need stronger scale-regime detection using OCR, axis/unit evidence, visual features, or calibrated verifier probabilities before applying scale correction.

References

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

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. DVQA: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Fangyu Liu, Julian Martin Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. 2023a. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Martin Eisenschlos. 2023b. MatCha: Enhancing visual language pretraining with math reasoning and chart derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Ahmed Masry, Parsa Kavehzadeh, Xuan Long Do, Enamul Hoque, and Shafiq Joty. 2023. UniChart: A universal vision-language pretrained model for chart comprehension and reasoning. *arXiv preprint arXiv:2305.14761*.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*.

Ahmed Masry, Megh Thakkar, Aayush Bajaj, Aaryaman Kartha, Enamul Hoque, and Shafiq Joty. 2024. ChartGemma: Visual instruction-tuning for chart reasoning in the wild. *arXiv preprint arXiv:2407.04172*.

Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. 2022. InfographicVQA. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. 2020. PlotQA: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

OpenGVLab. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*.

Qwen Team. 2025. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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. 2024. CharXiv: Charting gaps in realistic chart understanding in multimodal LLMs. *arXiv preprint arXiv:2406.18521*.

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

A Implementation Details

Appendix roadmap. The main paper compresses the evidence into the story that scale correction is directional and that the hard part is choosing the right scale regime. This appendix gives the details behind that claim. Section B documents how ScaleQA-Natural is mined, audited, and separated from training and official evaluation. Section C reports development evidence showing that the scale-adapter gain is real but narrow. Section D then follows the negative results: stronger prompts, verifiers, table calculators, candidate likelihood scoring, Bayesian fusion, and cross-backbone checks often expose the right answer in the candidate pool, but still fail to choose it reliably. This is why we frame adapters and verifiers as diagnostic probes rather than as a complete repair.

Adapters. The answer-tuned setting continues from the ChartQA/TextVQA answer-only LoRA adapter used in the evaluation pipeline. The answer-only adapters use rank $r = 16$, LoRA $\alpha = 32$, dropout 0.05, and target the language-model projection and MLP modules `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`. Answer-only training uses AdamW with learning rate 2×10^{-5}. The final Qwen scale adapters use large-positive pairs with $K = 300$, 300 optimization steps, AdamW learning rate 10^{-6}, weight decay 0.01, gradient clipping at 1.0, $\beta = 2.0$, and margin $m = 0.10$. The clean direct setting starts from the raw Qwen2.5-VL-7B-Instruct backbone and applies only the scale-contrastive update. The InternVL2.5-1B adapter check uses the same pair construction, rank $r = 8$, LoRA $\alpha = 16$, dropout 0.05, 200 optimization steps, and the same language-model target modules.

Why separate answer-tuned and clean-direct settings. The answer-only adapter is not a scale intervention. It mostly makes the model better aligned to short chart answers, numeric formatting, and the evaluation prompt. Keeping this setting separate matters because the clean-direct scale-only adapter is unsafe on the raw backbone (Table 33), while the answer-tuned setting can absorb a small scale-specific update without the same global collapse. This difference suggests that scale correction is not a standalone numerical operation. It depends on a model already being in a chart-answering regime where answer extraction, formatting, and short-answer calibration are reasonably stable.

Decoding and scoring. All single-adapter systems use the same short-answer prompt and deterministic short-answer decoding template: the question is followed by "Answer briefly." and generation uses greedy decoding with `do_sample=False` and `max_new_tokens=24`. We evaluate numeric answers with the repository evaluator, which uses normalized exact matching for text and a 5% numerical tolerance for numeric answers.

ScaleQA-Natural annotation. Candidate scale conflicts are mined from model predictions and automatically prefilled with numeric-ratio features. The review workflow assigns labels including scale diagnostic candidate, requires image inspection, non-scale, uncertain, and duplicate. Positive labels require that the gold answer and prediction refer to the same chart quantity but differ primarily by a simple

scale relation; wrong entity, wrong operation, OCR failure, arbitrary numeric miss, ambiguous question, and duplicate cases are excluded or sent to adjudication. The final ScaleQA-Natural-520 set contains author-adjudicated positive candidates after question/gold deduplication. The image-evaluable subset contains 473 examples with available ChartQA image rows. The audit workflow exports three files: the 739-row candidate audit, a conflict queue, and a 108-row final-positive adjudication queue. After this adjudication, an independent second reviewer audited the 520 final positive inclusions and agreed with all inclusion labels. The release package is intended to include item IDs, source split and row index, question, gold answer, model prediction, subtype, direction, review label, audit metadata, and scripts for reproducing the filters. We reference ChartQA images by source row rather than redistributing them.

PlotQA external subset. The PlotQA external check uses a 500-example subset sampled from the official PlotQA test parquet after filtering to examples with a single numeric answer. We use the same relaxed numeric evaluator as the ChartQA experiments. This subset is intended as an external numeric-answer transfer check rather than a full PlotQA benchmark.

B Data Flow and Annotation Audit

ScaleQA-Natural is useful only if it is not mistaken for an ordinary benchmark split or for hidden training data. Table 10 therefore separates five roles: answer-only tuning, scale-pair training, development analysis, diagnostic set construction, and held-out or external evaluation. The reported Qwen scale adapters are trained from ChartQA training rows and training-time prediction caches; ScaleQA-Natural is used to diagnose regimes and evaluate policy selection, not to train those scale adapters. The held-out ChartQA test result remains the official aggregate task result.

The annotation story has two parts. First, strict factor-of-100 failures are sparse in any single ChartQA output file (Table 12), so a single split underestimates the phenomenon. Second, mining introduces the opposite risk: plausible numeric ratios can be false positives when the model answered the wrong entity or operation. We therefore use ScaleQA-Natural as a diagnostic set whose positives require author adjudication and an independent inclusion audit (Table 11), and we keep uncertain/rejected examples visible (Table 13) rather than silently discarding them. The resulting set is deliberately conservative: it favors precision of the scale-conflict label over treating every ratio-like mismatch as a scale error.

C Additional Results

The development results explain why the paper does not present the large-positive adapter as a general correction method. On ChartQA validation, the adapter recovers strict gold-scale errors much more often than it changes ordinary examples (Tables 14 and 16), but the aggregate confidence interval is small and overlaps zero on the full split. This is exactly the diagnostic pattern we care about: the intervention finds a real failure mode, but it is not a broad accuracy engine.

Table 17 is another reason we avoid a "more LoRA pairs solves scale" claim. Gains appear quickly and then saturate, while the subtype table in the main paper shows that the opposite direction requires a different policy. The ablation in Table 18 should therefore be read as a probe-selection experiment: pair direction changes which mistakes are repaired and which safe cases are put at risk.

D Additional Transfer and Analysis Tables

This section contains the main negative evidence behind the paper's conclusion. If scale correction were a simple numeric normalization problem, stronger backbones, larger candidate pools, chart-to-table calculators, or Bayesian policy fusion should close the gap. Instead, the same pattern recurs. The correct answer is often present under some candidate or policy oracle, but deployable selectors remain directionally biased. This supports the paper's central distinction: the phenomenon transfers more robustly than the intervention.

Tables 19–30 show the transfer boundary. Larger InternVL models improve the base level on ScaleQA-Natural and external numeric subsets, but the same large/small/balanced scale adapters do not become uniformly helpful. On

Component	Source	Used for	Leakage control
Answer-only tuning	ChartQA/TextVQA training data	Starting adapter for the main Qwen setting	Reported separately from scale tuning; not counted as a scale-objective gain
Scale-pair training	ChartQA training rows and training-set prediction/log-prob caches	LoRA scale adapters and pair-mode ablations	Official ChartQA test and PlotQA are not used for scale-adapter gradients
Development analysis	ChartQA validation and ScaleQA-Core diagnostics	Pair-mode selection, seed checks, and failure analysis	Reported as development evidence; held-out ChartQA test is the main official result
ScaleQA-Natural	Mined model outputs, machine audit, and author adjudication	Diagnostic regime analysis and router experiments	Used as a diagnostic set, not as ordinary task training data for the reported Qwen scale adapters
Held-out/external evaluation	ChartQA test, TextVQA-500, PlotQA numeric-500	Final task result, safety, and external robustness check	Not used to train the reported scale adapters

Table 10: Data flow and split hygiene.

Audit item	Count/Score	Notes
Candidate pool	739	First-pass reviewed candidates
Observed agreement	0.633	Three-way labels
Cohen's κ	0.443	Include/review/reject
Conflicts for review	271	Exported separately
Final positives	520	Author-adjudicated
Second-pass include	412	Deterministic audit
Second-pass review/reject	108	Manually adjudicated
Independent inclusion audit	520	Second human reviewer
Second-review agreement	1.000	Final positives; zero label changes

Table 11: Annotation audit for ScaleQA-Natural. The deterministic metadata pass flags conflicts for author adjudication; the final positive inclusions are then independently audited by a second human reviewer.

Model / pipeline	Split	Strict errors	Rate
Qwen2.5-VL-7B + answer/large	ChartQA test	10 / 2500	0.40%
Qwen2.5-VL-3B direct/large	ChartQA test	19 / 2500	0.76%
Qwen2.5-VL-3B direct/large	ChartQA val	21 / 1920	1.09%
InternVL2.5-1B	ChartQA test	3 / 2500	0.12%
InternVL2.5-1B	ChartQA val	17 / 1920	0.89%

Table 12: Strict factor-of-100 scale-error prevalence in available ChartQA validation/test prediction files. These sparse natural subsets motivate the larger ScaleQA-Natural diagnostic set.

ScaleQA-stress1000, for example, InternVL2.5-8B adapters slightly reduce accuracy relative to the base model; on MathVista and InfographicVQA the changes are small; on PlotQA the adapter can hurt. These results do not refute the existence of scale inconsistency. They show that a global scale intervention learned in one model/task setting is not the right abstraction for deployment.

The selector and verifier results give the sharpest version of the bottleneck. Self-check prompting is already a strong fixed policy, and the best no-leak selector improves on it only modestly (Table 25). Yet the candidate and policy oracles are much higher. The gap is not caused by a failure to generate alternatives: it is caused by failures to rank alternatives under chart-specific scale evidence. The structured verifier makes this especially clear. It is strong when the prediction is too small but nearly collapses when the prediction is too large (Table 26), even though the augmented oracle is high in both directions.

The table-calculator baseline is a useful stress test because it is closer to symbolic reasoning than to answer-style prompting. It performs well on percent-to-decimal and value-scaled-down cases, but almost never solves decimal-to-percent or value-scaled-up cases (Table 27). That asymmetry mirrors the adapter story: changing the evidence source does not remove the need to identify the regime first.

The subtype and taxonomy tables explain why the task is hard for deployable selectors. Most image-evaluable examples contain explicit percent evidence, but the correct action is not always "add a percent" or "multiply by 100". In decimal-to-percent and value-scaled-up cases, the small-positive policy is often best; in percent-to-decimal and value-scaled-down cases, the large-positive policy is best (Table 31). The same words—percent, average, difference, value lookup—appear across regimes (Table 32), so a selector needs to reason over the chart, the question, the base answer, and the candidate answer jointly rather than keying off a single unit token.

Finally, Table 33 is the safety counterexample. A scale-only update from the raw

Audit label	Auto subtype	Question snippet
uncertain	percent_to_decimal	How much Good more than Bad?
uncertain	value_scaled_up	How many diamond cutting and polishing jobs were there in the Far East i...
uncertain	decimal_to_percent	How many respondents say no?
uncertain	decimal_to_percent	How many says that internet has good influence on education?
uncertain	decimal_to_percent	How confident are people about Angela Merkel doing the right thing regar...
uncertain	decimal_to_percent	How many more people answered Very acceptable for the most and least pop...

Table 13: Examples from the negative/uncertain audit queue. These candidates contain a plausible scale relation but are not included as positive diagnostics without stronger evidence that the prediction and gold answer refer to the same chart quantity.

Split	N	Base	Ours	Delta	CI low	CI high	$p(\Delta \leq 0)$
All examples	1920	0.860420	0.862500	0.002080	-0.001250	0.005520	0.113500
Numeric-answer examples	1452	0.860880	0.862810	0.001930	-0.002200	0.006060	0.187800
Scale-keyword numeric examples	1179	0.845630	0.847840	0.002210	-0.002880	0.007460	0.202500
Strict gold-scale errors	23	0.000000	0.113040	0.113040	0.000000	0.252170	0.036800
Non-gold-scale examples	1897	0.870850	0.871590	0.000740	-0.002210	0.003800	0.327100

Table 14: Bootstrap analysis on ChartQA validation for one large-positive adapter snapshot. The full-set confidence interval overlaps zero, so we emphasize the directionality and diagnostic analyses rather than overclaiming aggregate accuracy.

Split	Base	Ours	Std	Delta	N
All examples	0.860400	0.864000	0.000600	0.003500	1920
Strict gold-scale errors	0.000000	0.208700	0.036400	0.208700	23
Non-gold-scale examples	0.870800	0.871900	0.000400	0.001100	1897

Table 15: ChartQA validation subset analysis for the answer-tuned large-positive scale adapter. This was used as development evidence; the main official result is the held-out test result in Table 4.

Qwen backbone degrades ChartQA validation, ChartQA test, and ScaleQA-Core, even though it helps a separate InternVL-mined subset in the opposite direction. This is why the paper treats scale adapters as probes of directional asymmetry, not as a universal repair. The case studies in Table 34 give the qualitative version of the same lesson: the model can repair 0.61→61 and still over-correct a true small decimal such as 0.0593.

Overall appendix takeaway. Taken together, the appendix strengthens the main claim in a deliberately conservative way. Scale-like mistakes are not frequent enough in a single benchmark split to support sweeping accuracy claims, but they are systematic enough to reveal a reliability failure once mined and audited. Global adapters expose the directionality, training-free self-checks provide a strong practical baseline, and expanded candidate pools show that many failures are recoverable in principle. The unresolved problem is selecting the right regime without gold subtype labels. This is the reason the paper argues for regime-aware scale reliability rather than presenting unconditional scale correction as solved.

Split	N	Base	Mean fixes	Mean breaks	Any fix	All fix	Any break	All break
All examples	1920	0.860400	0.021400	0.010900	0.005700	0.003100	0.002600	0.001600
Base-wrong examples	268	0.000000	0.153000	0.000000	0.041000	0.022400	0.000000	0.000000
Base-correct examples	1652	1.000000	0.000000	0.012700	0.000000	0.000000	0.003000	0.001800
Strict gold-scale errors	23	0.000000	0.565200	0.000000	0.130400	0.087000	0.000000	0.000000
Non-gold-scale examples	1897	0.870800	0.014800	0.011100	0.004200	0.002100	0.002600	0.001600

Table 16: Seed-level fix and break rates on ChartQA validation.

Method	Pairs per kind	Accuracy	Delta
Base	0	0.850700	0.000000
Large-25	25	0.856000	0.005300
Large-50	50	0.856000	0.005300
Large-100	100	0.855300	0.004700
Large-200	200	0.856000	0.005300
Large-300	300	0.855300	0.004700
Large-400	400	0.855300	0.004700
Large-600	600	0.856000	0.005300
Large-800	800	0.856000	0.005300

Table 17: Large-positive pair-size curve on the ChartQA-1500 setting. Gains appear with few pairs and then saturate.

Method	Overall	Δ	Gold-scale	Δ	Non-scale	Δ
Base	79.00	+0.00	0.00	+0.00	82.81	+0.00
Scale-aware prompt	79.00	+0.00	0.00	+0.00	82.81	+0.00
Direct SFT (800)	78.20	-0.80	4.35	+4.35	81.76	-1.05
Direct SFT (1500)	77.40	-1.60	4.35	+4.35	80.92	-1.89
Mixed contrastive	79.40	+0.40	8.70	+8.70	82.81	+0.00
Balanced contrastive	79.20	+0.20	4.35	+4.35	82.81	+0.00
Large-only contrastive	79.00	+0.00	4.35	+4.35	82.60	-0.21
Real contrastive	79.40	+0.40	8.70	+8.70	82.81	+0.00
Real contrastive (5 seeds)	**79.52**	**+0.52**	11.30	+11.30	82.81	+0.00

Table 18: Ablation results on ChartQA-500. Gold-scale denotes strict 100× scale-error cases where the base model predicts an answer at the wrong numerical scale.

Backbone	Split	N	Accuracy	Strict scale errors
Qwen2.5-VL-7B	ChartQA test	2500	0.7804	10
InternVL2.5-1B	ChartQA validation	1920	0.7745	17
InternVL2.5-1B	ChartQA test	2500	0.7524	3
InternVL2.5-1B	ScaleQA-Natural-520	473	0.3340	–

Table 19: Cross-backbone strict scale-like error diagnostic.

Set	Model	Base	Large	Small	Bal.
ScaleQA-Natural-520	InternVL2.5-1B	0.3340	0.3446	0.3446	–
ScaleQA-Natural-520	InternVL2.5-8B	0.4672	0.4685	0.4723	0.4689
ScaleQA-Natural-520	InternVL2.5-26B	0.4989	–	–	–
ScaleQA-stress1000	InternVL2.5-8B	0.7320	0.7272	0.7276	0.7256
ScaleQA-stress1000	Qwen2.5-VL-7B	0.6940	0.6940	0.6940	0.6940
PlotQA-557	InternVL2.5-8B	0.6912	0.6829	0.6862	0.6840
PlotQA-500	InternVL2.5-26B	0.7200	–	–	–
ChartBench-500	Qwen2.5-VL-7B	0.4820	0.4820	0.4820	0.4820
ChartBench-500	InternVL2.5-8B	0.3520	0.3512	0.3516	0.3520
ChartBench-500	InternVL2.5-26B	0.4080	–	–	–
MathVista-243	Qwen2.5-VL-7B	0.6708	0.6708	0.6708	0.6708
MathVista-243	InternVL2.5-8B	0.6337	0.6379	0.6387	0.6387
MathVista numeric	InternVL2.5-26B	0.6749	–	–	–
InfographicVQA-300	Qwen2.5-VL-7B	0.7767	0.7767	0.7767	0.7767
InfographicVQA-300	InternVL2.5-8B	0.6067	0.6087	0.6073	0.6093
InfographicVQA-300	InternVL2.5-26B	0.6300	–	–	–

Table 20: Full cross-backbone and external numeric transfer table.

Backbone and setting	N	Accuracy	Std.
InternVL2.5-1B base	473	0.3340	–
InternVL2.5-1B large-positive LoRA	473	0.3446	–
InternVL2.5-1B small-positive LoRA	473	0.3446	–
InternVL2.5-8B base	473	0.4672	–
InternVL2.5-8B large-positive LoRA	473	0.4685	0.0017
InternVL2.5-8B small-positive LoRA	473	0.4723	0.0035
InternVL2.5-8B balanced LoRA	473	0.4689	0.0025

Table 21: InternVL scale adaptation on ScaleQA-Natural-520.

InternVL2.5-8B setting	N	Accuracy	Std.
Base	1000	0.7320	–
Large-positive adapter	1000	0.7272	0.0016
Small-positive adapter	1000	0.7276	0.0034
Balanced adapter	1000	0.7256	0.0023

Table 22: InternVL2.5-8B on a synthetic ScaleQA-stress1000 set.

Prompting strategy	Accuracy
Direct answer	0.5243
Axis-aware prompt	0.5307
Math-check prompt	0.5137
Scale prompt	0.4905
Self-check prompt	0.5455
Unit-guard prompt	0.5180
Majority vote	0.5264
Oracle prompt choice	0.6744

Table 23: Training-free prompting baselines on ScaleQA-Natural-520.

Method	Overall	Pred. too large	Pred. too small
Candidate majority	0.5159	0.3029	0.7371
Free-form verifier	0.4588	0.1411	0.7888
Letter verifier	0.4651	0.1411	0.8017
Scale-guard letter verifier	0.4609	0.1452	0.7888
Candidate oracle	0.8288	0.7593	0.9009

Table 24: Multi-candidate verifier on ScaleQA-Natural-520. Candidate oracle selects the best answer from the same candidate pool and is an upper bound.

Baseline or selector	Accuracy	Notes
Self-check prompt	0.5514	Best fixed prompt on held-out splits
v8 no-leak selector	0.5432 ± 0.0501	Question, answer, numeric, and chart-context features
v9 expanded selector	0.5501 ± 0.0460	Adds structured/pairwise policies but does not improve
v10b no-leak selector	0.5667 ± 0.0561	Best clean selector; adds corrected TTA-v2 evidence
v11 log-prob selector	0.5667 ± 0.0561	Log-prob evidence preserves but does not improve v10b
v13 rationale selector	0.5667 ± 0.0561	Rationale-verifier evidence preserves v10b
v14/v15 candidate-likelihood selectors	0.5667 ± 0.0561	Answer/large/base likelihood evidence preserves v10b
v16/v17 top-8 selectors	0.5667 ± 0.0561	Larger candidate pools preserve v10b but raise only oracle
Statistical policy fusion	0.5309 ± 0.0470	Beta/Wilson reliability fusion
Bucketed Bayesian selector	0.5239 ± 0.0480	Bucketed policy priors; over-switches
Bayesian-MLP selector	0.5218 ± 0.0540	Bucket priors plus hidden-layer selector; over-switches
Pairwise top-8 verifier	0.5433	Candidate comparison prompt
Candidate likelihood top-8	0.4503	Log-probability candidate scoring
Structured-context verifier	0.4841	Image, chart context, and top-4 candidates
Structured scale-audit verifier	0.4376	Explicit unit/scale audit prompt
Table extraction + calculator	0.4228	DePlot-style chart-to-table evidence, then symbolic calculation
Original policy-pool oracle	0.7023 ± 0.0384	Upper bound over early policy pool
Expanded policy-pool oracle	0.9004	Upper bound after adding prompts/verifiers
Candidate top-8 oracle	0.9175	Correct answer often appears in candidate set
Augmented structured oracle	0.9260	Upper bound after adding scaled candidate variants

Table 25: Deployable selectors and strong baselines on ScaleQA-Natural. Learned and rule-based selectors use held-out splits, while fixed prompting, verifier, and table-calculator rows are full-set scores. The gap to the oracle rows shows that policy selection remains the bottleneck.

Method	Overall	Pred. too large	Pred. too small
Structured-context verifier	0.4503	0.0954	0.8190
Structured scale-audit verifier	0.4376	0.0996	0.7888
Augmented structured oracle	0.9260	0.9419	0.9095

Table 26: Structured verifier by scale direction. Adding chart context and scaled candidate variants makes the oracle high, but deployable verifiers remain directionally biased.

Subset	Reason	Calculator	Either
All examples	0.3890	0.4080	0.4228
Percent-to-decimal	0.8378	0.8243	0.8378
Value scaled down	0.6949	0.8220	0.8475
Decimal-to-percent	0.0275	0.0183	0.0183
Value scaled up	0.0098	0.0098	0.0098

Table 27: DePlot-style table-extraction plus calculator baseline by subtype on ScaleQA-Natural. The method is strong for some regimes but fails on the opposite direction, again showing that scale correction is regime-dependent.

Policy pool	Best fixed	Learned	+Chart ctx.	Subtype oracle	Inst. oracle
Prompt regimes (473)	0.5408	0.5366	0.5358	0.6192	0.6701
Adapters+prompts (403)	0.5032	0.4467	0.4615	0.8010	0.8410

Table 28: Learned policy routing with question/base-answer numeric features and optional chart-side context extracted from the image. Values are 50-seed means.

Backbone	Method	Accuracy
Qwen2.5-VL-7B	Answer-tuned	0.902
Qwen2.5-VL-7B	Answer-tuned + large-positive scale	0.898
Qwen2.5-VL-7B	Base + scale prompt	0.832
Qwen2.5-VL-7B	TTA majority	0.892
Qwen2.5-VL-7B	TTA oracle prompt choice	0.904
InternVL2.5-1B	Base	0.600

Table 29: External numeric PlotQA-500 check.

Dataset	N	Base	Large	Small	Balanced
PlotQA	557	0.6912	0.6829	0.6862	0.6840
ChartBench	500	0.3520	0.3512	0.3516	0.3520
MathVista	243	0.6337	0.6379	0.6387	0.6387
InfographicVQA	300	0.6067	0.6087	0.6073	0.6093

Table 30: External numeric checks with InternVL2.5-8B.

Subtype	N	Base	Large-pos.	Small-pos.	Best direction
Decimal to percent	109	0.0917	0.1560	0.7890	Small
Value scaled up	102	0.0784	0.0784	0.7059	Small
Percent to decimal	74	0.9054	0.9324	0.0270	Large
Value scaled down	118	0.8475	0.8559	0.0593	Large

Table 31: ScaleQA-Natural subtype analysis on the 403-example reviewed set.

Group	Count	Share
Question pattern		
Percent / percentage	133	28.1
Average / mean	75	15.9
Value lookup	75	15.9
Difference / comparison	65	13.7
Maximum / highest	44	9.3
Minimum / lowest	30	6.3
Unit evidence		
Percent in axis or question	439	92.8
No explicit unit	14	3.0
Large-unit word	13	2.7
Ratio or rate	6	1.3
Currency	1	0.2

Table 32: Fine-grained taxonomy of the image-evaluable ScaleQA-Natural subset. Counts summarize question patterns and unit evidence; categories are not mutually exclusive across the two blocks.

Evaluation set	N	Base	Direct-large	Delta	Direct-small	Delta
ChartQA validation	1920	0.7536	0.7219	-0.0318	0.6948	-0.0589
ChartQA test	2500	0.7804	0.7456	-0.0348	0.7136	-0.0668
ScaleQA-Core	403	0.4591	0.4367	-0.0223	0.4243	-0.0347
InternVL-mined	192	0.1510	0.0938	-0.0573	0.3542	+0.2031

Table 33: Clean direct scale-only adaptation from the raw Qwen2.5-VL-7B backbone.

Type	Question	Gold	Base	Calibrated
Recovered	How much times Madagascar greater than Central Europe data?	32	0.032	32
Recovered	What the difference in value between Asia and Caribbean?	1.18	0.18	1.18
Recovered	What is the highest value of gray bar?	61	0.61	61
Recovered	What is the difference in % between Poland and Australia for people who are worried about losing their job?	17.05	0.1705	17.05
Over-correction	Find out the average of the bottom three countries ??	0.0593	0.06	6.03333

Table 34: Representative recoveries and an over-correction failure.