UniBound-Bench: Diagnosing Visual Decision Boundaries in Multimodal Language Models

Anonymous ACL submission

Abstract

Recent multimodal benchmarks largely report aggregate accuracy, which often hides where models fail around perceptual decision boundaries. We present **UniBound-Bench**, a unified benchmark for *visual boundary sensitivity* in multimodal language models (MLLMs). Each sample is associated with an explicit boundary distance δ, enabling boundary-centric evaluation via $\mathrm{Acc}(\delta)$, $\mathrm{Threshold}@\tau$, and AUC-δ. UniBound-Bench currently contains 19,388 examples across 17 task types covering geometry, quantity, depth, occlusion, temporal ordering, and visual illusions. We further provide a standardized yes/no inference protocol, data validators, and reproducible evaluation scripts. An eight-model evaluation panel shows strong task heterogeneity: count-related boundaries are relatively robust, multiple geometry and illusion categories remain near chance, and stronger frontier APIs still exhibit substantial near/far boundary gaps. A Gemma4-E4B LoRA pilot further shows that overall accuracy and near-boundary robustness need not move together. Surprisingly, the strongest frontier APIs achieve only about 51% near-boundary accuracy despite about 63% aggregate accuracy, and boundary-aware finetuning trades aggregate gains for near-threshold robustness, exposing failure modes that standard accuracy reporting would substantially understate.

1 Introduction

Multimodal large language models (MLLMs) have rapidly improved across OCR, VQA, and general visual instruction following (Wang et al., 2024; Li et al., 2024; Bai et al., 2025). Yet, many practical errors arise not from gross scene misunderstanding but from *near-threshold perception*: when two candidate decisions are visually close, a small perturbation can flip the ground-truth label while model outputs remain unstable.

This motivates a shift from purely aggregate evaluation to *boundary-aware evaluation*. Existing benchmark lines have exposed important weaknesses in spatial relation understanding, perspective, dynamic scene understanding, and occlusion perception (Liu et al., 2025a; Zhang et al., 2024; Wu et al., 2025a; Tang et al., 2026; Lin et al., 2026; Liu et al., 2025b), but their results are still difficult to compare under one consistent notion of "distance to boundary." While prior benchmarks such as SpatialScore and HallusionBench reveal that models struggle with spatial and illusory stimuli, none provides a calibrated scalar distance to the decision boundary, making threshold-aware comparison across tasks effectively impossible. As a result, we still lack a unified answer to a simple question: *how far from a perceptual boundary must an MLLM operate before becoming reliably correct?*

To answer this question, we propose UniBound-Bench, a unified benchmark and evaluation protocol centered on a calibrated boundary distance variable δ. The benchmark integrates multiple visual boundary families under one output space (yes/no), one data schema, and one metric suite. This design allows both fine-grained within-task analysis and cross-family comparisons.

From our eight-model evaluation panel, four empirical patterns emerge. First, even the strongest frontier APIs reach only around 51% near-boundary accuracy despite more than 61–63% aggregate accuracy, showing that aggregate visual competence does not imply boundary reliability. Second, boundary robustness is strongly task-dependent rather than a single scalar "vision quality" property. Third, reasoning-style output behavior does not automatically translate into stronger visual boundary performance. Fourth, $\mathrm{Acc}(\delta)$ curves and per-task rankings reveal qualitative differences that are invisible in one-number accuracy reporting.

Our contributions are as follows. Rather than treating visual robustness as a single ag-

gregate score, UniBound-Bench reframes multimodal evaluation around calibrated distance to perceptual decision boundaries. This boundary-aware evaluation paradigm enables a threshold-aware diagnostic view that existing spatial, illusion, and VQA-style benchmarks do not provide. We release UniBound-Bench (19,388 samples; 17 tasks; train/val/test_id/test_ood) spanning geometry, quantity, depth, temporal, and illusion boundaries under one data schema. We additionally provide a reproducible toolkit covering generation, validation, inference normalization, and boundary-centric metrics, and we report initial baseline findings plus a lightweight finetuning pilot that reveal structured failure regimes and training-time trade-offs for future model design.

2 Related Work

A large body of recent work studies spatial weaknesses in MLLMs, including relation understanding, perspective reasoning, and geometry-intensive settings (Liu et al., 2025a; Zhang et al., 2024; Chen et al., 2025; Fu et al., 2024; Wu et al., 2025a; Tang et al., 2026; Wang et al., 2025). Complementary studies examine shape sensitivity and cross-modal consistency (Rudman et al., 2025; Yan et al., 2025; Wu et al., 2025b), while dynamic and perception-centric benchmarks probe online scene understanding and occlusion robustness (Ou et al., 2025; Lin et al., 2026; Liu et al., 2025b). Collectively, these results show that spatial-perceptual competence remains a bottleneck even as model scale and general capability increase.

Another related line uses illusion-focused diagnostics to expose subtle perception failures (Guan et al., 2024; Zhang et al., 2025). Classical perception work also demonstrates that contextual cues can systematically bias perceived size or relation (Mruczek et al., 2015), suggesting that evaluating only gross object recognition is insufficient for robust visual reasoning. In parallel, method-side progress has explored synthetic data and prompting strategies for improving spatial reliability (Ogezi and Shi, 2025; Tang et al., 2024; Ou et al., 2025), but cross-paper comparison remains difficult because benchmarks often differ in task definition, label space, and reporting granularity.

Very recent studies further broaden the landscape to GUI-oriented visual perception and multi-hop video spatial reasoning (Kanade and Ganu, 2026; Zhu et al., 2026). Our position in this ecosystem is benchmark-centric: UniBound-Bench does not propose a new reasoning architecture; rather, it provides a unified boundary formalism and a threshold-aware evaluation interface that can absorb heterogeneous visual boundary tasks within one consistent protocol.

Boundary-aware evaluation gap. Existing spatial and illusion benchmarks identify whether models fail on particular perceptual categories, but they usually do not ask how model reliability changes as samples move continuously away from the decision boundary. This distinction matters because two models with similar aggregate accuracy can have very different near-boundary behavior. UniBound-Bench is designed to make this hidden dimension measurable under one task-normalized metric interface.

3 UniBound-Bench: A Scalable Data Engine for Visual Boundary Tasks

3.1 Task taxonomy

UniBound-Bench includes 17 tasks organized into five families. The geometry family covers left-right, distance boundary, inside-outside, and overlap; the quantity/scale family includes count boundary and size compare; the depth/occlusion family includes depth order and occlusion visibility; the temporal family includes temporal order; and the illusion family includes Ebbinghaus, Delboeuf, Müller-Lyer, Ponzo, Poggendorff, Zollner, Rod-Frame, and Vertical-Horizontal illusions. This taxonomy is designed to balance broad visual coverage with a consistent binary decision interface.

Figure 2 provides a family-wise illustrative overview of the task space. We place this visualization in the main paper because benchmark papers are easier to audit when readers can immediately inspect what kinds of decision boundaries the benchmark covers, rather than inferring them only from task names.

3.2 Four-step data engine

UniBound-Bench is produced by a task-oriented four-step engine. First, we collect visual assets, including source objects and controlled backgrounds, for each task family. Second, we synthesize boundary instances by varying task-relevant factors to create paired or structured samples

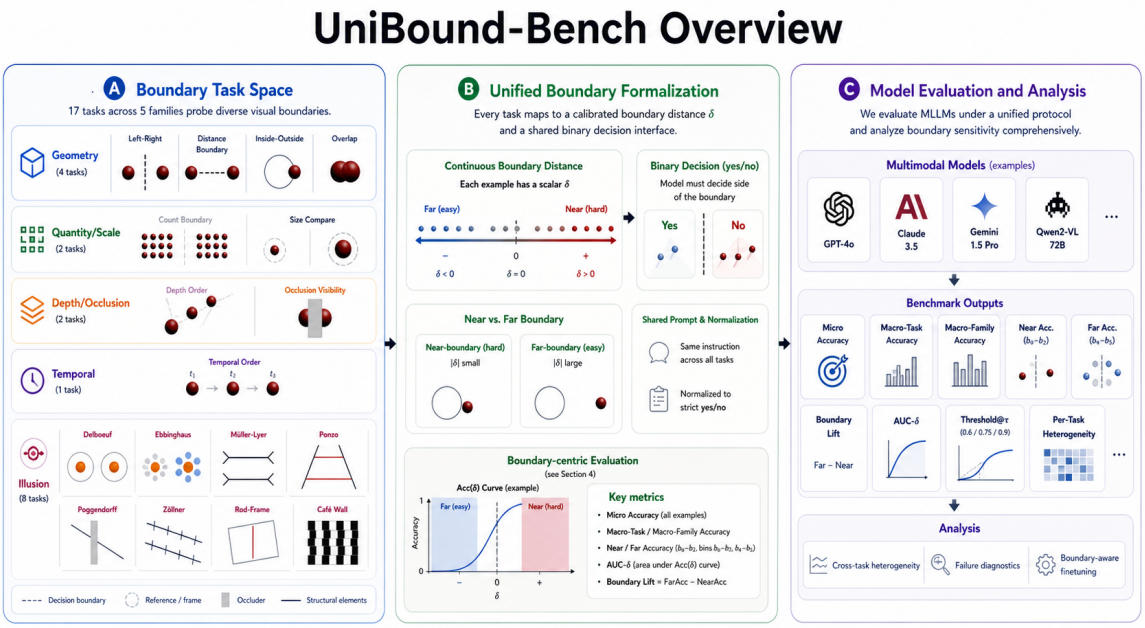

Figure 1: Conceptual overview of UniBound-Bench. Left: the benchmark spans 17 tasks across five visual boundary families. Middle: heterogeneous tasks are unified by a calibrated boundary-distance variable δ and a shared binary decision interface. Right: models are evaluated under one protocol and summarized through aggregate, family-level, and boundary-centric analyses.

around target boundaries. Third, we assign boundary distance by annotating each sample with δ and `delta_bin`, where smaller δ indicates a harder near-boundary instance. Fourth, we validate and split the dataset through automatic schema/path checks and deterministic generation of `train`, `val`, `test_id`, and `test_ood` partitions.

3.3 Unified sample schema

Each JSONL record contains `sample_id`, `split`, `task_id`, `delta`, `delta_bin`, `question`, `answer`, and metadata fields (generation parameters, nuisance controls, provenance). The schema is intentionally model-agnostic: any MLLM or agent that can consume an image and produce text can be evaluated after binary answer normalization.

3.4 Dataset statistics

The current release has 19,388 examples: `train` 12,736, `val` 2,854, `test_id` 1,858, `test_ood` 1,940. We provide automatic schema/path validators and deterministic data merge scripts. In addition, we expose per-task and per-δ summaries so that users can audit coverage and difficulty balance before benchmarking. As a lightweight human sanity check, we also asked five researchers to manually inspect representative samples across all five families, including near-boundary cases, and verify that the visual question and binary label were interpretable under the intended yes/no protocol. This check is meant to confirm task legibility rather than to serve as a formal psychophysical study.

Family	#Tasks	#Samples
Geometry	4	5,438
Quantity/Scale	2	2,028
Depth/Occlusion	2	2,604
Temporal	1	1,342
Illusion	8	7,976
Total	17	19,388

Table 1: Task families in UniBound-Bench.

4 Evaluation Protocol

4.1 Boundary formalization

Let x be an image-question pair and $y \in \{0, 1\}$ be the binary label (yes/no). Each sample has boundary distance $\delta(x) \geq 0$, where smaller values indicate harder, near-boundary instances. Because δ is task-normalized before aggregation, we do not claim that raw physical units are directly comparable across tasks; instead, each task maps its raw monotonic hardness scalar to six task-local bins (`b0`–`b5`) through fixed generation-time thresholds, and evaluation aggregates these task-local bound-

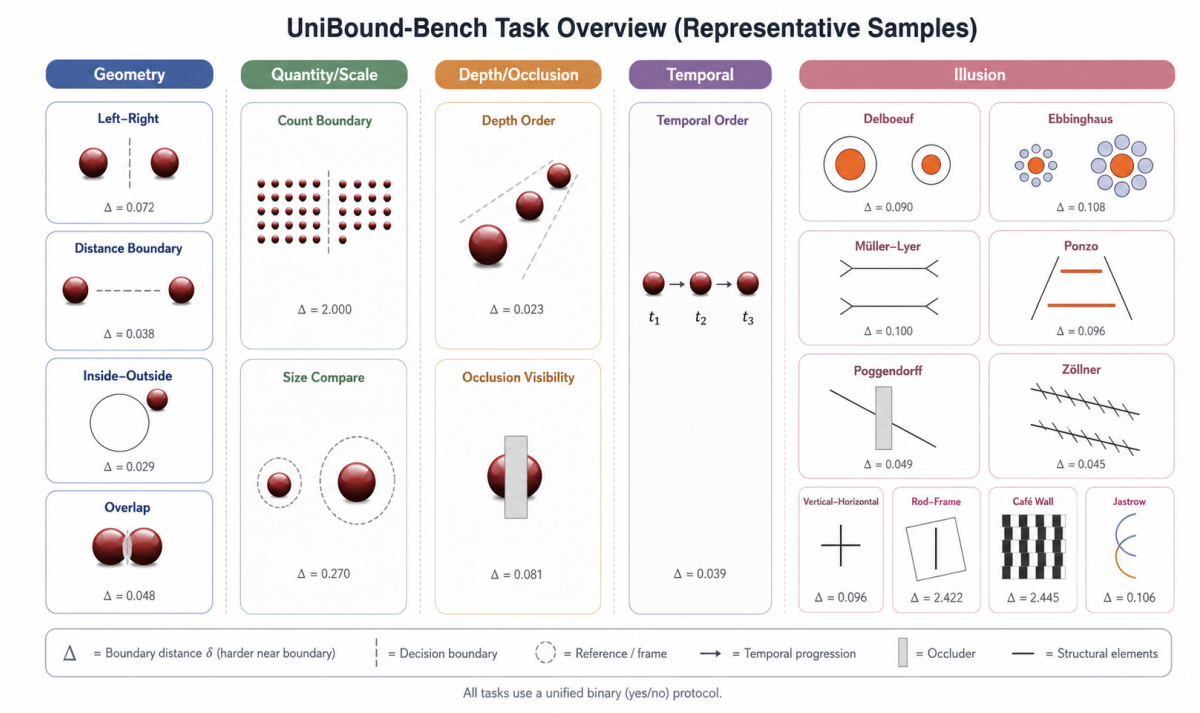

Figure 2: Family-wise illustrative overview of UniBound-Bench. The schematic highlights coverage across geometry, quantity/scale, depth/occlusion, temporal ordering, and visual illusion boundaries under a unified yes/no decision interface.

ary coordinates rather than raw cross-task physical units.

Given model $f_\theta(x) \in \{0, 1\}$, we use:

$$\mathrm{Acc}(\delta = b) = \mathbb{E}[\mathbf{1}(f_\theta(x) = y) \mid \delta(x) \in b], \tag{1}$$

$$\mathrm{Threshold@}\tau = \min\{d : \mathrm{Acc}(\delta \geq d) \geq \tau\}, \tag{2}$$

$$\mathrm{AUC}\text{-}\delta = \int \mathrm{Acc}(\delta)\, d\delta \text{ (normalized)}. \tag{3}$$

4.2 Inference normalization

To reduce prompt-format artifacts, we enforce strict binary outputs. Raw generations are normalized by regex into `yes/no`; non-matching outputs are marked invalid. This rule is shared across tasks and models, preventing format variance from inflating or deflating boundary metrics. For reasoning-style checkpoints that emit long `think`-prefixed traces, we additionally report a forced-binary variant that strips reasoning wrappers and applies a conservative yes/no fallback, so that format-only verbosity does not dominate benchmark conclusions.

4.3 Metrics

UniBound-Bench is intended to be used as a benchmark rather than only as a visualization tool, so we define a default scorecard in addition to task-wise curves. The first component is *Micro-Accuracy*, computed over all evaluated examples in a split, and *Macro-Task Accuracy*, computed as the unweighted mean of per-task accuracies. The latter prevents large tasks from dominating the benchmark summary and makes cross-model comparison less sensitive to task cardinality. We also report *Macro-Family Accuracy* when users want a coarser, family-level view.

The second component is explicitly boundary-centric. We define *Near-Accuracy* on near-boundary bins (`b0–b2`) and *Far-Accuracy* on easier bins (`b4–b5`), and summarize their difference as *Boundary Lift*:

$$\mathrm{BoundaryLift} = \mathrm{FarAcc}_{b4\text{–}b5} - \mathrm{NearAcc}_{b0\text{–}b2}.$$

This score captures how strongly a model benefits from moving away from the decision boundary. A large positive value indicates substantial brittleness near threshold, while a smaller value suggests flatter degradation across difficulty.

The third component measures full-curve behavior and operational usability. We report nor-

malized AUC-δ per task and aggregate it as *Macro-AUC-δ*. We further report Threshold@τ with $\tau \in \{0.6, 0.75, 0.9\}$, and use *Mean Threshold Delta@0.75* as the default threshold summary at $\tau = 0.75$. The released evaluator writes these benchmark-level metrics into a canonical `benchmark_scorecard.csv`, making the benchmark directly usable for future leaderboard-style comparison.

5 Experiments

5.1 Evaluation setup

We report five open-source baselines, Qwen2-VL-2B, Qwen3-VL-8B, Gemma4-E4B, Kimi-VL-A3B, and GLM-4.1V-9B, together with three frontier API models, GPT-5.5, Claude Opus 4.7, and Gemini 3.1 Pro, on the vision-balanced expanded-17 `test_id` split (1,858 examples). We use the same binary prompt protocol and the same normalization rules across tasks; for reasoning-heavy checkpoints (Kimi/GLM), we additionally report a forced-binary normalization variant to handle think-style outputs. For frontier APIs, we use the provider-reported model identifiers at evaluation time and record these identifiers together with prompt variants in the saved prediction artifacts. This evaluation subset is intentionally smaller than the full `test_id` corpus reported in Section 3, because it balances task and family contributions for fairer cross-model comparison; this is why the sample counts shown in the result tables are much smaller than the corpus-wide counts in Table 1. In addition to per-task tables and Acc(δ) plots, our evaluator emits the benchmark scorecard defined in Section 4, so future work can compare models through a fixed set of benchmark-native numbers rather than ad hoc selections of task-level results.

5.2 Benchmarking results on UniBound-Bench

Table 2 and Table 3 now summarize both open-weight baselines and frontier API runs on the same 1,858-example vision-balanced expanded-17 `test_id` subset. Reporting the results in this decomposed form is important: UniBound-Bench is not intended to collapse model behavior into a single headline number, but to expose whether a model is uniformly reliable, boundary-sensitive, or operationally brittle under a fixed binary protocol.

Figure 3 is one of the clearest pieces of evidence for the benchmark's diagnostic value: the per-task heatmap shows that boundary sensitivity is not a monotonic function of model strength, because models that dominate one task family can remain near chance on another, and the ranking among models changes materially across task types.

On the shared 1,858-example vision-balanced expanded-17 `test_id` split, the frontier APIs occupy the top three positions in headline micro-accuracy: Claude Opus 4.7 reaches **63.3%**, GPT-5.5 reaches **62.1%**, and Gemini 3.1 Pro reaches **61.4%**. The best open-weight baseline, Qwen3-VL-8B, reaches **58.4%**, followed by Gemma4-E4B at **56.7%** and Qwen2-VL-2B at **53.4%**. Under forced-binary normalization for reasoning-style outputs, Kimi-VL-A3B and GLM-4.1V-9B remain substantially lower at **50.7%** and **49.6%**, respectively. This ranking already shows that the benchmark is not saturated even under a strict yes/no interface.

Boundary-centric metrics reveal a more important pattern than the headline ranking alone. Even the strongest frontier APIs remain much weaker near the boundary than far from it: Claude Opus 4.7 rises from **51.0%** near-boundary accuracy to **71.9%** far-boundary accuracy, GPT-5.5 rises from **50.6%** to **70.1%**, and Gemini 3.1 Pro rises from only **43.8%** to **74.1%**. Their Boundary Lift values (**+20.8**, **+19.4**, and **+30.3** points) show that stronger aggregate accuracy does not eliminate threshold sensitivity. The same pattern appears among open models: Qwen3-VL-8B and Gemma4-E4B improve sharply from **53.7%/50.8%** near-boundary accuracy to **62.9%/62.1%** far-boundary accuracy, confirming that far-from-boundary competence and near-boundary robustness remain distinct capabilities.

To check whether the same signal appears beyond the in-distribution split, we additionally ran a small `test_ood` audit with Claude Opus 4.7 on 84 completed samples from a 17-task, five-per-task OOD probe. Table 5 shows that aggregate accuracy remains similar to `test_id`, while the near/far boundary gap persists. We treat this as a sanity audit rather than a full OOD leaderboard: it supports the boundary-distance signal but does not replace the larger eight-model `test_id` comparison above.

Task-level performance is highly non-uniform, indicating that "visual boundary capability" is not a single scalar trait; the main text therefore empha-

Model	Type	Eval Subset	Micro Acc.	Macro Task Acc.	Macro Family Acc.	Eval. N
Claude Opus 4.7	Closed	balanced-exp17 `test_id`	0.633	0.634	0.633	1,858
GPT-5.5	Closed	balanced-exp17 `test_id`	0.621	0.625	0.648	1,858
Gemini 3.1 Pro	Closed	balanced-exp17 `test_id`	0.614	0.598	0.635	1,858
Qwen3-VL-8B	Open	balanced-exp17 `test_id`	0.584	0.582	0.589	1,858
Gemma4-E4B	Open	balanced-exp17 `test_id`	0.567	0.566	0.570	1,858
Qwen2-VL-2B	Open	balanced-exp17 `test_id`	0.534	0.536	0.540	1,858
Kimi-VL-A3B	Open	balanced-exp17 `test_id`	0.507	0.509	0.516	1,858
GLM-4.1V-9B	Open	balanced-exp17 `test_id`	0.496	0.495	0.497	1,858

Table 2: Unified headline model comparison on UniBound-Bench under the shared 1,858-example vision-balanced expanded-17 `test_id` subset.

Model	Type	Macro AUC-δ	Mean Threshold@0.75	Eval. N
Claude Opus 4.7	Closed	0.672	0.069	1,858
GPT-5.5	Closed	0.638	0.069	1,858
Gemini 3.1 Pro	Closed	0.600	0.072	1,858
Qwen3-VL-8B	Open	0.542	0.185	1,858
Gemma4-E4B	Open	0.518	0.185	1,858
Qwen2-VL-2B	Open	0.514	0.191	1,858
Kimi-VL-A3B	Open	0.453	0.303	1,858
GLM-4.1V-9B	Open	0.426	0.185	1,858

Table 3: Detailed operational benchmark metrics on UniBound-Bench under the shared vision-balanced expanded-17 `test_id` subset, sorted by headline micro-accuracy from Table 2.

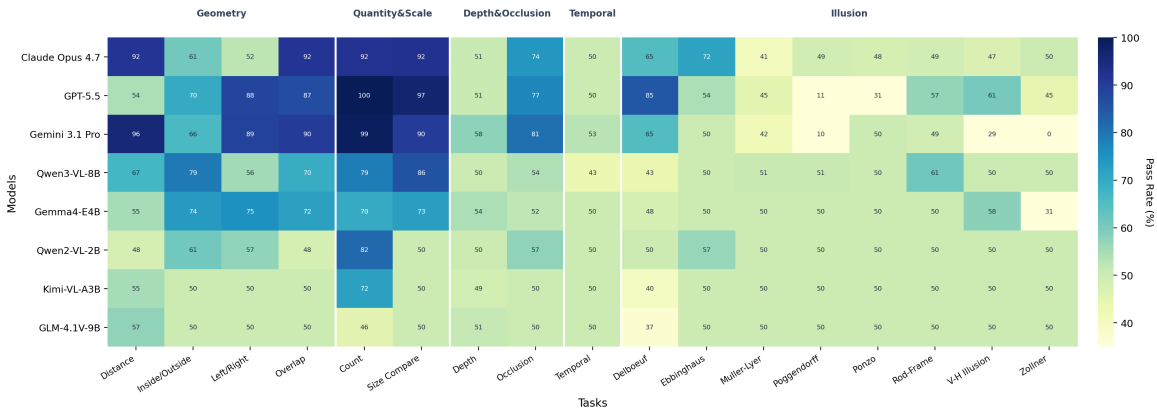

Figure 3: Per-task pass-rate heatmap on the vision-balanced expanded-17 `test_id` evaluation subset across both open-weight and frontier API models. The heatmap makes cross-task heterogeneity explicit by showing how model rankings change across task families.

sizes the unified heatmap and boundary-gap summary rather than a long per-task score table.

Figure 4 shows $\mathrm{Acc}(\delta)$ for `count_boundary`: performance improves as samples move away from the boundary, supporting the validity of the difficulty design.

5.3 Analysis

Four conclusions stand out from the eight-model panel. First, stronger closed-source systems improve aggregate performance but do not solve the boundary problem. Claude Opus 4.7, GPT-5.5, and Gemini 3.1 Pro lead the overall table, yet all three still show large near/far gaps (Table 4), with Gemini exhibiting the largest Boundary Lift in the entire comparison. UniBound-Bench therefore distinguishes *overall visual competence* from *stability near a decision threshold*: the former can improve while the latter remains brittle.

Second, the open-weight models reveal a simi-

Model	Type	Near Acc.	Far Acc.	Lift
Claude Opus 4.7	Closed	0.510	0.719	+0.208
GPT-5.5	Closed	0.506	0.701	+0.194
Gemini 3.1 Pro	Closed	0.438	0.741	+0.303
Qwen3-VL-8B	Open	0.537	0.629	+0.092
Gemma4-E4B	Open	0.508	0.621	+0.113

Table 4: Compact boundary-gap summary on the shared expanded-17 `test_id` subset. Even the strongest frontier APIs retain large near/far gaps, while the best open-weight models show smaller lift but lower far-boundary ceilings.

Model	N	Micro	Near	Far
Claude Opus 4.7	84	0.667	0.640	0.740

Table 5: Small `test_ood` audit over 17 task types. The probe is intentionally compact, but it confirms that near-boundary samples remain harder than far-boundary samples outside the main `test_id` split.

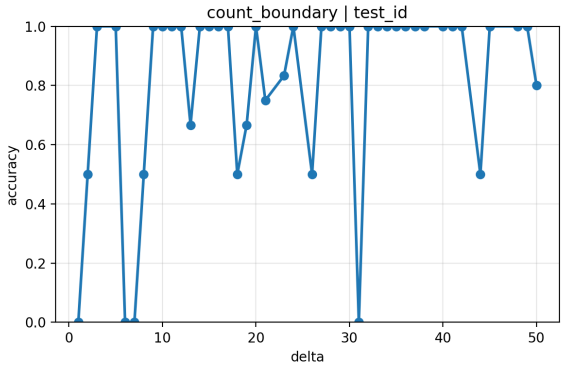

Figure 4: $\mathrm{Acc}(\delta)$ on `count_boundary` (Qwen2-VL-2B).

lar separation. Qwen3-VL-8B and Gemma4-E4B are the strongest open baselines in aggregate score and macro AUC-δ, but their near-boundary accuracies remain only **53.7%** and **50.8%**. Their much higher far-boundary accuracies (**62.9%** and **62.1%**) indicate that they benefit from easier instances without becoming uniformly reliable on fine-grained threshold cases. This is exactly the regime a boundary-centric benchmark should expose, and it would be largely hidden by a single aggregate accuracy number.

Third, reasoning-style output behavior should not be conflated with visual boundary ability. Kimi-VL-A3B and GLM-4.1V-9B require forced-binary normalization to control verbose generations, yet after format normalization they still lag behind the strongest open baselines on both micro accuracy and macro AUC-δ. This suggests that long-form reasoning traces, by themselves, are not evidence of stronger perceptual discrimination near difficult visual boundaries.

Fourth, the per-task picture is not a simple rescaling of the overall leaderboard. Figure 3 shows that model ranking changes materially across task families: frontier APIs are very strong on `count_boundary` and several geometry tasks, but multiple illusion categories remain close to chance for most of the panel; meanwhile, some open models are competitive on specific relations such as `inside_outside` or `left_right`. This task-dependent rank reshuffling is central to the benchmark's value: UniBound-Bench is not merely another leaderboard, but a diagnostic lens for *which* visual boundaries a model handles well and which ones remain unstable.

Fifth, the task-level heterogeneity suggests that boundary robustness is not uniformly distributed across perceptual modalities. As Figure 3 shows, count-boundary cases are among the strongest regions of the panel, likely because discrete count differences align well with strong linguistic and compositional priors from pretraining, making suprathreshold cases easier to verbalize. In contrast, illusion tasks such as `illusion_delboeuf` and `illusion_zollner` remain near chance for most of the panel, consistent with the hypothesis that contextual manipulations bias low-level visual features before they can be stably mapped into the language decoder. On this view, model scale can improve aggregate accuracy without eliminating the perceptual distortion itself, which is exactly the kind of failure mode a boundary-centric benchmark should isolate.

6 Boundary-aware Finetuning Pilot

To test whether UniBound-Bench can serve not only as an evaluator but also as a training signal, we ran two single-seed Gemma4-E4B LoRA pilots on the full 17-task training split and evaluated them on the same expanded-17 `test_id` split (1,858 examples) used for the unified benchmarking tables. The first uses standard instruction finetuning (VANILLA-LORA); the second upweights near-boundary examples during training (BOUNDARY-LORA). The benchmark construction uses globally unique `sample_ids` and deterministic split assignment from the hashed sample identifier, so the training and `test_id` examples are strictly disjoint at the sample level. We report

Setting	Micro Acc.	Near Acc.	Far Acc.	Macro AUC-δ
VANILLA-LORA	0.865	0.718	0.929	0.859
BOUNDARY-LORA	0.854	0.722	0.913	0.844

Table 6: Gemma4-E4B LoRA pilots on the full 17-task `test_id` split. Boundary-aware loss weighting slightly improves near-boundary accuracy but reduces aggregate and full-curve metrics. A paired bootstrap gives a near-accuracy delta of +0.4 points with a 95% CI of [-4.4, +5.2] points, so we interpret the pilot as a trade-off diagnostic rather than a significant gain claim.

these numbers separately from Tables 2 and 3 because they are adaptation pilots rather than zero-shot/open-weight baseline runs.

This pilot makes two points. First, UniBound-Bench is already usable as supervision for lightweight adaptation without changing the evaluation interface. Second, the negative result is itself informative: compared with VANILLA-LORA, BOUNDARY-LORA raises near-boundary accuracy by +0.4 points, but lowers micro accuracy and macro AUC-δ. We do not interpret this single-seed difference as a statistically significant improvement claim; the paired bootstrap interval overlaps zero, so the result is best read as a qualitative trade-off illustration showing that optimizing for near-boundary examples can change the robustness profile without uniformly improving all metrics. This is precisely the kind of trade-off a boundary-centric benchmark should make visible. Together with the zero-shot comparison panel, the finetuning pilot shows that the benchmark can support both diagnosis and intervention.

7 Limitations

UniBound-Bench prioritizes controlled decision-boundary manipulation over photorealistic scene diversity. This improves interpretability, but some tasks are visually simpler than natural-image benchmarks and therefore emphasize perceptual thresholding more than open-world visual complexity. In addition, several balanced evaluation slices remain modest in size at the per-task level, which increases variance for task-specific estimates and makes single-task rank differences less stable than the benchmark-level trends. Finally, our current finetuning evidence is still limited to a single-model LoRA pilot, so stronger claims about boundary-aware training methods will require broader multi-model adaptation studies. The main text includes a compact `test_ood` sanity audit, but the full eight-model OOD panel is left to the public leaderboard release because OOD generalization requires additional controls for stimulus-level distribution shift that fall outside the scope of this paper. More broadly, the paper's main claim is about boundary-distance sensitivity rather than open-world OOD generalization.

8 Ethical Considerations

UniBound-Bench is a diagnostic benchmark and does not use private user data. Still, boundary errors in vision models can transfer to high-impact applications (e.g., assistive systems). We therefore emphasize transparent reporting of failure modes, release reproducible scripts, and discourage overclaiming from aggregate metrics alone.

9 Reproducibility Statement

We release construction scripts, validation tools, inference utilities, and evaluation code. Results in this paper are computed from saved prediction JSONL and metric CSV artifacts. We also record model identifiers and prompt variants to support exact reruns.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Guizhen Chen, Weiwen Xu, Hao Zhang, Hou Pong Chan, Deli Zhao, Anh Tuan Luu, and Yu Rong. 2025. Geopqa: Bridging the visual perception gap in mllms for geometric reasoning. *ArXiv, abs/2509.17437, 2025a. URL https://arxiv.org/abs/2509.17437*.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

Aditya Sanjiv Kanade and Tanuja Ganu. 2026. Do you see me: A multidimensional benchmark for evaluating visual perception in multimodal llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7285–7326.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Jingli Lin, Chenming Zhu, Runsen Xu, Xiaohan Mao, Xihui Liu, Tai Wang, and Jiangmiao Pang. 2026. Ost-bench: Evaluating the capabilities of mllms in online spatio-temporal scene understanding. *Advances in Neural Information Processing Systems*, 38.

Jingping Liu, Ziyan Liu, Zhedong Cen, Yan Zhou, Yinan Zou, Weiyan Zhang, Haiyun Jiang, and Tong Ruan. 2025a. Can multimodal large language models understand spatial relations? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 620–632.

Zhaochen Liu, Kaiwen Gao, Shuyi Liang, Bin Xiao, Limeng Qiao, Lin Ma, and Tingting Jiang. 2025b. Beyond the visible: Benchmarking occlusion perception in multimodal large language models. *arXiv preprint arXiv:2508.04059*.

Ryan EB Mruczek, Christopher D Blair, Lars Strother, and Gideon P Caplovitz. 2015. The dynamic ebbinghaus: motion dynamics greatly enhance the classic contextual size illusion. *Frontiers in human neuroscience*, 9:77.

Michael Ogezi and Freda Shi. 2025. Spare: Enhancing spatial reasoning in vision-language models with synthetic data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7855–7875.

Siqu Ou, Hongcheng Liu, Pingjie Wang, Yusheng Liao, Chuan Xuan, Yanfeng Wang, and Yu Wang. 2025. Bridging the dynamic perception gap: Training-free draft chain-of-thought for dynamic multimodal spatial reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6560–6578.

William Rudman, Michal Golovanevsky, Amir Bar, Vedant Palit, Yann LeCun, Carsten Eickhoff, and Ritambhara Singh. 2025. Forgotten polygons: Multimodal large language models are shape-blind. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11983–11998.

Yihong Tang, Ao Qu, Zhaokai Wang, Dingyi Zhuang, Zhaofeng Wu, Wei Ma, Shenhao Wang, Yunhan Zheng, Zhan Zhao, and Jinhua Zhao. 2024. Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to composite spatial reasoning.

Yunlong Tang, Pinxin Liu, Mingqian Feng, Zhangyun Tan, Rui Mao, Chao Huang, Jing Bi, Yunzhong Xiao, Susan Liang, Hang Hua, and 1 others. 2026. Mm-perspective: Do mllms understand perspective? a comprehensive benchmark for perspective perception, reasoning, and robustness. *Advances in Neural Information Processing Systems*, 38.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Siting Wang, Luoyang Sun, Cheng Deng, Kun Shao, Minnan Pei, Zheng Tian, Haifeng Zhang, and Jun Wang. 2025. Spatialviz-bench: Automatically generated spatial visualization reasoning tasks for mllms. *arXiv e-prints*, pages arXiv–2507.

Haoning Wu, Xiao Huang, Yaohui Chen, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025a. Spatialscore: Towards unified evaluation for multimodal spatial understanding. *arXiv e-prints*, pages arXiv–2505.

Jiarui Wu, Zhuo Liu, and Hangfeng He. 2025b. Mitigating hallucinations in multimodal spatial relations through constraint-aware prompting. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 3450–3468.

Qianqi Yan, Yue Fan, Hongquan Li, Shan Jiang, Yang Zhao, Xinze Guan, Ching-Chen Kuo, and Xin Eric Wang. 2025. Multimodal inconsistency reasoning (mmir): A new benchmark for multimodal reasoning models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18829–18845.

Wenyu Zhang, Wei En Ng, Lixin Ma, Yuwen Wang, Jungqi Zhao, Boyang Li, and Lu Wang. 2024. Sphere: A hierarchical evaluation on spatial perception and reasoning for vision-language models. *arXiv e-prints*, pages arXiv–2412.

Yiming Zhang, Zicheng Zhang, Xinyi Wei, Xiaohong Liu, Guangtao Zhai, and Xiongkuo Min. 2025. Illusionbench: A large-scale and comprehensive benchmark for visual illusion understanding in vision-language models. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Rui Zhu, Xin Shen, Shuchen Wu, Chenxi Miao, Xin Yu, Yang Li, Weikang Li, Deguo Xia, and Jizhou Huang. 2026. Video-msr: Benchmarking multi-hop spatial reasoning capabilities of mllms. *arXiv preprint arXiv:2601.09430*.

A Appendix Overview

This appendix complements the main paper with implementation-level details that would otherwise interrupt the main narrative. Our goal is not simply to provide extra tables, but to make the benchmark legible as a reusable research artifact. To that end, we organize the appendix around the same logic as the benchmark itself: task definition, evaluation protocol, score interpretation, and empirical diagnostics. The presentation is intentionally paragraph-driven so that readers can follow the design rationale rather than treating the appendix as a checklist of disconnected facts.

We focus on four questions. First, how is the notion of a visual decision boundary instantiated across heterogeneous tasks? Second, what exact prompting and normalization choices are used to make cross-model evaluation comparable? Third, what additional evidence supports the claim that boundary-aware evaluation reveals failure patterns that aggregate accuracy alone would miss? Finally, how should readers interpret the benchmark as both an evaluation interface and a training-time diagnostic tool? The following sections answer these questions in the same prose-first style used in the main body.

B Task Cards and Boundary Variables

The central design constraint of UniBound-Bench is that every task must admit an operational notion of *distance to decision boundary*. In some tasks this quantity is directly geometric, such as horizontal offset, radial displacement, or relative scale ratio. In others it is perceptual or contextual, as in the illusion families where the relevant signal is the strength of the contextual manipulation that biases human or model judgments. What matters for the benchmark is not that all tasks share the same physical unit, but that each task provides a monotonic hardness variable that can be mapped to a scalar δ and therefore supports the same threshold-aware analysis.

This decision has two important consequences. First, it lets us compare qualitatively different tasks under one evaluation interface without forcing them into an unnatural common representation. Second, it makes the benchmark extensible: adding a new task does not require redesigning the full metric stack, only specifying a valid boundary variable and a binary decision protocol. Table 7 summarizes how this principle is instantiated in the current release. Readers can therefore interpret the table not as a static dataset inventory, but as the operational map that connects task semantics, question templates, and threshold-aware evaluation.

C Inference Protocol (Exact Form)

All compared models are queried with the same binary-constrained instruction:

```
You are an evaluator. Read the question
and image.  Answer strictly with one
token: yes or no (lowercase). No extra
words.
```

The purpose of this strict prompt is to minimize prompt-format variance and force all models into the same output space. Because UniBound-Bench is diagnostic rather than generative, we do not want leaderboard differences to be driven by verbosity, explanation style, or model-specific chat templates. The evaluation target is a binary visual decision, so the protocol constrains every model to produce one lowercase token. This makes the benchmark less expressive than free-form VQA, but much more reliable for boundary analysis, where small differences in formatting can otherwise swamp the signal we care about.

After generation, answers are normalized with a simple regex-based rule set: `yes|y|true|1` maps to `yes`, `no|n|false|0` maps to `no`, and all other outputs are treated as invalid. We use this post-processing layer not to rescue arbitrary generations, but to absorb small surface variations that do not change the intended decision.

This evaluation design also supports apples-to-apples comparison across model families. Whether the underlying system is Qwen, Gemma, GLM, Kimi, or another open-weight MLLM, the scoring layer remains identical: same prompt, same normalization, same boundary-aware metrics, and same split definitions. As a result, performance differences are easier to interpret as model capability differences rather than quirks of prompting infrastructure.

D Additional Diagnostics Beyond Main Text

The main paper intentionally reports only a compact slice of the current empirical picture, but auxiliary audits from saved prediction artifacts reveal several additional patterns. The first is family-level heterogeneity. As an auxiliary Qwen2-VL-2B audit, Table 8 uses an earlier vision-balanced

Task ID	Boundary variable (δ semantics)	Canonical yes/no question pattern
`left_right`	Horizontal signed offset magnitude around alignment boundary	"Is object A to the left of object B?"
`distance_boundary`	Inter-object distance gap to target threshold	"Is the distance larger than threshold t?"
`inside_outside`	Signed radial offset to container boundary	"Is object A inside region B?"
`overlap`	Intersection ratio around overlap decision boundary	"Do A and B overlap?"
`count_boundary`	Absolute count difference to query threshold	"Is the number of A greater than k?"
`size_compare`	Relative scale ratio around equality boundary	"Is A larger than B?"
`depth_order`	Relative depth difference (front/back ambiguity)	"Is A in front of B?"
`occlusion_visibility`	Visible-area ratio around visibility threshold	"Is A still visible?"
`temporal_order`	Temporal displacement around event-order boundary	"Did event A happen before event B?"
`illusion_ebbinghaus`	Context-induced perceived-size contrast strength	"Is the center circle larger?"
`illusion_delboeuf`	Ring-context size bias strength	"Is the inner disk larger?"
`illusion_muller_lyer`	Arrowhead-induced line-length bias strength	"Is line A longer than line B?"
`illusion_ponzo`	Perspective-context length bias strength	"Is upper segment longer?"
`illusion_poggendorff`	Misalignment magnitude under occluding context	"Are the two segments collinear?"
`illusion_zollner`	Orientation-context induced angle bias strength	"Are the two long lines parallel?"
`illusion_rod_frame`	Frame-induced apparent rod-tilt deviation from vertical	"Is the central rod tilted to the right?"
`illusion_vertical_horizontal`	Apparent vertical-over-horizontal length ratio bias	"Does the vertical line look longer than the horizontal one?"

Table 7: Unified task cards in UniBound-Bench.

`test_id` diagnostic slice and shows the same family-level heterogeneity pattern seen in the main paper: quantity/scale tasks are materially easier than geometry, depth/occlusion, illusion, and temporal tasks. This matters because a model that looks acceptable on a subset dominated by count-like instances may still be brittle on relation-heavy or context-sensitive boundary problems.

The second pattern is the gap between near-boundary and far-boundary performance. When we pool all evaluated tasks, near-boundary bins (`b0`–`b2`) reach 41/81 = 50.6%, whereas the far-boundary bin (`b5`) reaches 213/346 = 61.6%. In other words, the same model that appears moderately competent under aggregate reporting becomes much less reliable in the regime that is most relevant to threshold sensitivity. This result is precisely the motivation for UniBound-Bench: once evaluation is conditioned on boundary distance, failure modes that are invisible in overall accuracy become immediately measurable.

The third pattern is that different tasks produce qualitatively different $\mathrm{Acc}(\delta)$ curves rather than merely shifted versions of the same profile. Figure 5 illustrates this point with three representative tasks: `left_right`, `occlusion_visibility`, and `illusion_zollner`. The geometry curve highlights instability in relation judgments, the occlusion curve reflects graded visibility sensitivity, and the illusion curve reveals cases where contextual bias suppresses performance across multiple bins. These differences reinforce our claim that boundary robustness is not a single scalar property and should be analyzed as a family of task-specific response curves.

One useful way to read these supplementary results is to treat them as additional evidence for the benchmark's diagnostic value rather than as mere overflow from the main tables. Even a lightweight baseline already separates multiple kinds of errors that standard benchmark reporting would conflate: threshold collapse near difficult bins, family-specific brittleness, and curve-shape differences that persist even when two tasks have similar mean accuracy. In that sense, the appendix does not merely "add more numbers" to the main paper; it shows why the boundary-centric view is analytically richer than conventional aggregate evaluation.

E Interpretation of Benchmark Scope

Because UniBound-Bench is benchmark-centric, the appendix clarifies what the main paper emphasizes and what it deliberately leaves to supplementary inspection. The central claim of the paper is methodological and diagnostic: a unified decision-boundary formulation makes it possible to compare qualitatively different visual tasks under one metric interface and to surface structured failure modes that aggregate accuracy hides. The main text therefore prioritizes the unified scorecard, the per-task comparison figure, and the fine-tuning pilot, while the appendix deepens the evaluation semantics and provides additional diagnostic evidence.

This means the appendix should be read as an analysis supplement rather than as a hidden second experimental section. We prioritize clarifying the evaluation semantics, showing representative quantitative patterns, and documenting how the benchmark should be interpreted. Details that are more appropriate for an artifact package, such as complete script inventories or low-level generation bookkeeping, can be released with the codebase and supplementary materials without occupying central appendix space in the paper itself.

Family	#Samples	Accuracy	Near Acc.	Far Acc.	Boundary Lift
Quantity/Scale	182	0.659	0.650	0.682	+0.032
Geometry	506	0.540	0.405	0.554	+0.148
Depth/Occlusion	278	0.536	0.471	0.564	+0.094
Illusion	636	0.508	0.498	0.489	-0.009
Temporal	102	0.500	0.429	0.526	+0.097

Table 8: Supplementary family-level diagnostic scorecard from an earlier Qwen2-VL-2B audit slice on `test_id`. We include it only to illustrate family-level heterogeneity; all headline comparisons in the main text use the newer 1,858-example expanded-17 subset.

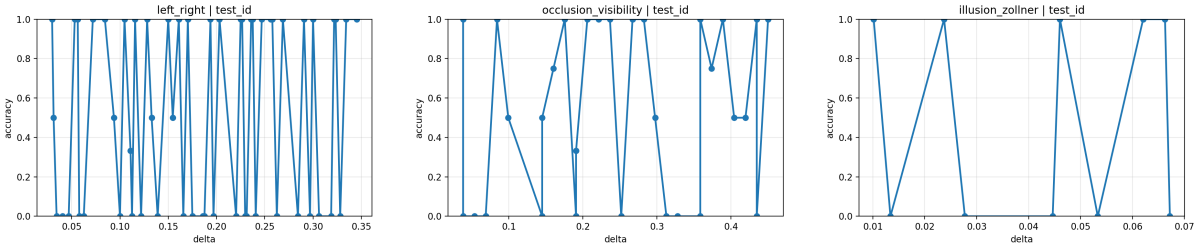

Figure 5: Additional $\mathrm{Acc}(\delta)$ examples on `test_id`: `left_right` (left), `occlusion_visibility` (middle), `illusion_zollner` (right).

This choice keeps the appendix aligned with the style of strong ACL/EMNLP benchmark papers, where supplementary pages deepen argumentation and analysis rather than restating repository documentation.

F Supplementary Material Scope

The appendix is intended to provide enough procedural and empirical detail that a careful reviewer can assess whether the benchmark design is principled, whether the reported results are reproducible, and whether the proposed evaluation lens is genuinely informative. For anonymous review, we keep the manuscript self-contained and avoid repository-specific exposition inside the main body. The core reproducibility ingredients are nevertheless explicit in the paper: split definitions, prompt form, normalization rules, metric definitions, and representative diagnostic outputs.