SHARED-REFERENCE ANCHORING AND REFERENCE SEPARATION FOR CHART REASONING VERIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable multimodal reasoning verification requires judgments to remain grounded in the source image even when intermediate references are imperfect. Existing verifiers increasingly use model-generated references to make visual premises explicit, but reusing the same fallible reference across multiple verification paths can couple their errors and redirect otherwise distinct judgments toward the same false premise. We identify this failure as shared-reference anchoring and examine its causal effect by varying only the reference while holding the chart, question, and candidate trajectories fixed. To mitigate this dependency, we propose RefSplit, a reference-separated verification method that retains reference-conditioned reasoning while withholding the shared reference from a complementary chart-grounding path using the same verifier checkpoint. On 300 Value Position pairs and 300 Legend Binding pairs, RefSplit reduces the preference-accuracy drop under aligned false references from 26.67 to 4.17 points and from 7.50 to 0.33 points, respectively. The same failure also arises under ordinary generation, where 26.46% of 2,509 audited incorrect-reference value-error pools contain an incorrect candidate aligned with the false reference, and RefSplit reduces aligned-error selection by 8.43 points. Across five public chart benchmarks and five candidate policies, RefSplit achieves 53.02% mean Best-of-8 accuracy, the highest among the evaluated prior selectors and within 0.03 points of the matched non-separated control. These results demonstrate that access to the source image alone does not eliminate the influence of a shared fallible reference. By separating reference access across complementary verification paths, RefSplit limits shared error propagation while preserving candidate-selection quality.

1 INTRODUCTION

Multimodal large language models (MLLMs) increasingly rely on explicit reasoning to solve tasks that require more than direct visual recognition. Complex problems in mathematical reasoning, scientific figures, and chart understanding require models to connect multiple visual observations, recover latent relationships, and compose them into a final answer (Lu et al., 2024; Wang et al., 2024a; Yue et al., 2024; Wang et al., 2024c). Reasoning trajectories have therefore become a central interface between visual perception and final prediction. As multimodal reasoning becomes more capable and more widely used, the reliability of these intermediate trajectories becomes increasingly important because a plausible reasoning trace can still be built on an incorrect visual premise.

As reasoning becomes central to multimodal inference, recent work increasingly verifies generated reasoning after it is produced. Process reward models (PRMs) evaluate intermediate reasoning steps and use these judgments for reranking or Best-of-N selection (Lightman et al., 2024; Wang et al., 2024b; Zheng et al., 2025). VisualPRM extends this paradigm to multimodal reasoning, while Perceval places greater emphasis on whether individual reasoning steps remain grounded in the source image (Wang et al., 2026b; Min et al., 2026). Other approaches explicitly recover and verify visual dependencies before accepting a candidate trajectory (Wang et al., 2026a). Across these approaches, different reasoning candidates are ultimately evaluated against the same source image. The central objective is to distinguish reasoning that merely appears coherent from reasoning whose intermediate claims are actually supported by the visual input.

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

Recent approaches further make these visual premises explicit through structured intermediate representations. Chart representations expose numerical values together with their series and category bindings (Liu et al., 2023a;b; Masry et al., 2023; Chen et al., 2024; Xia et al., 2026), while multimodal verifiers use visual constraints or candidate-independent priors to guide judgment (Wang et al., 2026a; Andrade et al., 2026). Unlike the source image, however, these intermediate references are model-generated and therefore fallible. Candidate-independent extraction prevents a trajectory from defining the evidence used to judge itself, but the same extracted record can still be reused across multiple verification paths. When that shared record is incorrect, its false premise becomes a common conditioning signal. Multiple verification outputs may then agree not because they independently recover support from the source image, but because they inherit the same erroneous premise. We refer to this coupling between a fallible shared reference and downstream verification as shared-reference anchoring.

To address this failure, we introduce RefSplit, a reference-separated verification method. RefSplit breaks this dependency by retaining reference-conditioned reasoning while withholding the shared reference from a complementary chart-grounding path. Both paths use the same verifier checkpoint, which isolates information access as the key design choice. Controlled interventions and naturally occurring candidate pools consistently show that reference separation limits the influence of erroneous shared evidence. At the same time, RefSplit preserves competitive Best-of-N selection performance across diverse chart benchmarks and candidate policies. These findings establish reference access as a consequential component of multimodal reasoning verification and motivate separation as a simple way to prevent shared evidence errors from propagating across verification paths.

Our contributions are summarized as follows.

- We identify shared-reference anchoring as a failure mode in multimodal reasoning verification and causally demonstrate that a false shared premise can redirect both reasoning and chart-grounding judgments even when the source chart remains available.
- We introduce RefSplit, a reference-separated verification method that retains reference-conditioned reasoning while preserving a chart-grounded judgment outside the influence of the shared reference. Matched access comparisons isolate the effect of reference access while holding the verifier checkpoint, grounding role, and scoring procedure fixed.
- We establish the mechanism and practical relevance of shared-reference anchoring through reference interventions, reference-correctness ablations, aligned candidate pools, and Best-of-N selection across five public chart benchmarks and five candidate policies.

2 Shared-Reference Anchoring in Chart Reasoning Verification

We characterize shared-reference anchoring and establish the failure mechanism that motivates our method. The analysis separates fallible reference extraction from verification errors caused by repeated access to the same erroneous reference.

Preliminaries. Given a chart image I and a question Q, we extract a candidate-independent structured reference E that summarizes question-relevant visual premises such as series, category, and numerical value. Candidate reasoning trajectories are denoted by r.

For controlled analysis, we construct matched trajectory pairs (r^+, r^-), where r^+ is supported by the chart and r^- changes exactly one visual premise while preserving the remaining reasoning. We evaluate each pair under four reference conditions,

$$E \in \{E^+, E_{\text{random}}, E_{\text{aligned}}, E_{\text{none}}\},$$

where E^+ supports r^+, E_{random} comes from another example, E_{aligned} supports the incorrect premise in r^-, and E_{none} provides no reference content.

To isolate the effect of reference access, we compare three verification paths,

$$R(I, Q, E, r), \qquad V_E(I, Q, E, r), \qquad V(I, Q, r).$$

R performs reference-conditioned reasoning verification. V_E and V use the same chart-grounding interface, but only V_E receives E. Because V_E and V share the verifier checkpoint, grounding instruction, and verdict readout, their comparison isolates the effect of reference access.

2.1 Fallible Shared References

A fallible shared reference can propagate the same extraction error across multiple verification judgments. Figure 1 shows a natural example. A chart-supported trajectory conflicts with an incorrect extracted reference. Both verification paths that receive the reference reject the trajectory, while direct chart verification without the reference accepts it. Candidate-independent extraction prevents candidates from influencing reference construction, but it does not guarantee that the extracted reference is correct. Among 938 unique questions in our development set, 227 extracted references disagree with renderer truth, corresponding to an error rate of 24.2%. These errors include incorrect numerical values, series or category bindings, and combinations of multiple fields.

Reference reuse turns local extraction errors into shared inputs across multiple judgments. Each of the 938 extracted references is reused across two candidate policies, producing 1,876 question-policy cases. An incorrect reference can therefore affect multiple candidate evaluations rather than remaining an isolated extraction error. Figure 1 illustrates this shared exposure, where two verification paths follow the same false premise despite retaining access to the source chart.

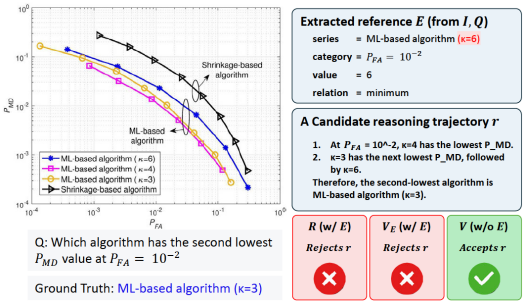

Figure 1: Example of shared-reference anchoring. A chart-supported trajectory is rejected by both verification paths that receive an incorrect shared reference, while direct chart verification without that reference accepts it.

Repeated exposure alone, however, does not establish shared-reference anchoring. An incorrect reference may concern a field that is irrelevant to the available candidates, or it may leave their relative scores unchanged. A reference error and a verification error can therefore co-occur without the former causing the latter. The audit establishes the prerequisite for shared-reference anchoring by showing that erroneous references are repeatedly reused across downstream verification judgments.

2.2 Effect of Reference Replacement on Verification

Changing only the shared reference can redirect verification toward a supplied false premise. For each matched pair, we keep the chart, question, r^+, and r^- fixed and vary only the reference. We evaluate the four conditions defined above using pair preference accuracy, which measures whether a verification path ranks r^+ above r^- with half credit for ties. Figure 2 summarizes the results.

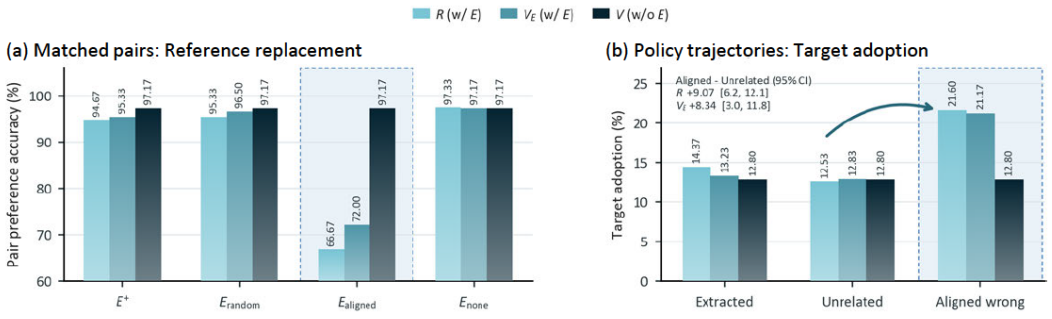

Figure 2: Verification shifts under reference replacement. Panel (a) compares preference accuracy under a correct reference E^+, an unrelated reference E_{random}, a false reference E_{aligned} that supports the incorrect trajectory, and no reference E_{none}. Panel (b) measures whether an aligned false reference increases selection of an incorrect candidate present in frozen policy-generated pools. Reference-free chart verification V remains unchanged because it does not receive the reference.

An aligned false reference substantially reduces preference for the correct trajectory in both paths that receive it. On 300 Value Position pairs, preference accuracy falls from 95.33% to 66.67% for R when moving from an unrelated reference to an aligned false reference. The same replacement reduces accuracy for V_E from 96.50% to 72.00%. The response of V_E is particularly informative because it performs chart grounding rather than reference consistency. Access to the source chart therefore does not prevent the false reference from influencing its judgment.

The reference-free path does not exhibit the same shift. Its preference accuracy remains at 97.17% across reference conditions because its input never includes E. The same contrast appears on 300 Legend Binding pairs. Additional sample sizes and a second renderer reproduce the pattern in Appendix A. The reference-induced shift therefore persists across error types, sample sizes, and rendering configurations.

The same effect also appears in policy-generated candidate pools. We select the first eligible incorrect candidate in the original pool order without using verifier scores and replace the reference value with the number asserted by that candidate. Across 3,000 eligible pools, target adoption increases from 12.53% to 21.60% for R and from 12.83% to 21.17% for V_E. The rate for V remains unchanged. A reference-field ablation further identifies the false numerical value itself as sufficient to induce the shift. Together, the controlled pairs and policy-generated pools show that false reference content can directly redirect verification toward an incorrect candidate.

A complementary verification signal should therefore not depend on the same fallible reference. This motivates separating reference-conditioned reasoning from direct chart verification that does not receive E.

3 Method: Reference-Separated Verification

3.1 Overall Framework

RefSplit separates reference access across complementary reasoning and chart-grounding paths without adding a separate verifier model. Given a chart and question, we first extract a shared reference. For each candidate trajectory, one path performs reference-conditioned reasoning verification while the other checks chart support without the shared reference. We then combine the two judgments into a verification score. Figure 3 shows the information flow. A single trained Qwen3-VL-8B-Instruct checkpoint performs reference extraction and both verification roles (Bai et al., 2025a). Role-specific training procedures and optimization settings are described in Appendix B. Candidate generation is performed by a separate policy and is not part of the verifier.

We define the verifier as the model that produces trajectory scores. When these scores are used for candidate selection, the selector denotes the rule that ranks candidates and returns one trajectory. The verification unit is a completed reasoning trajectory. Each score evaluates the trajectory's reasoning and visual premises rather than only its final answer. The method does not rely on chronological step labels or stepwise rewards. Reference separation therefore concerns the information available to trajectory-level verification rather than the granularity of process supervision.

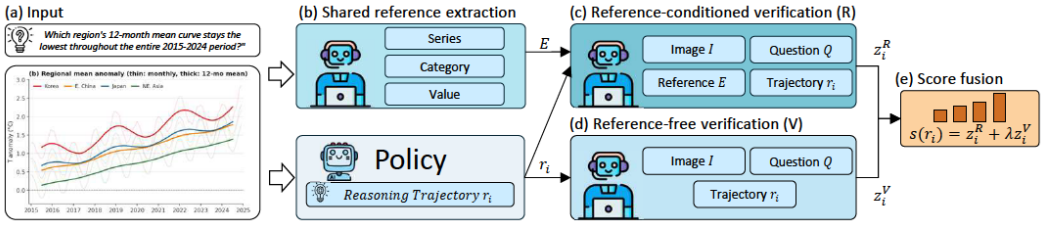

Figure 3: Reference-separated verification. A candidate-independent reference conditions reasoning verification, while direct chart verification evaluates the same trajectory without that reference. The two judgments are combined while the direct grounding path remains unconditioned on the shared reference.

3.2 Shared Reference Extraction

A shared reference provides a candidate-independent summary of the chart evidence relevant to the question. Given a chart I, a question Q, and a frozen candidate pool $\mathcal{R} = \{r_1, \ldots, r_N\}$, we compute

$$E = f_\theta^{\text{ref}}(I, Q). \tag{1}$$

The record contains series, category, and value fields, with an optional relation field for extremum questions. The same E is cached and reused for every candidate in the pool. This construction prevents candidate content from influencing reference extraction. The access pattern determines which verification paths are exposed to errors in E.

3.3 Reference-Conditioned and Direct Chart Verification

Reference-conditioned reasoning verification measures whether a trajectory is consistent with the supplied chart reference. For candidate r_i, we define the probability assigned to a positive verdict as

$$p_i^R = P_\theta(y^R = 1 \mid I, Q, E, r_i). \tag{2}$$

The verification prompt treats the values and bindings in E as given. A high p_i^R therefore indicates consistency with the supplied reference rather than independent validation of that reference. The source chart remains available to this path.

Direct chart verification provides a complementary signal by checking the same trajectory against the source chart without E. Its positive verdict probability is

$$p_i^V = P_\theta(y^V = 1 \mid I, Q, r_i). \tag{3}$$

The complete trajectory is passed to the chart-grounding interface, while the shared reference is omitted. The diagnostic path V_E uses the same grounding interface but additionally receives E. Because V_E and V differ only in reference access, their matched comparison isolates the influence of the shared reference within the same grounding role.

3.4 Score Fusion

Score fusion combines reference consistency with direct chart support so that visual evidence can counteract a candidate favored by a false reference. We clip verdict probabilities between 10^{-6} and $1 - 10^{-6}$, convert them to log odds, and compute

$$z_i^R = \log \frac{p_i^R}{1 - p_i^R}, \qquad z_i^V = \log \frac{p_i^V}{1 - p_i^V}, \qquad s_i = z_i^R + \lambda z_i^V. \tag{4}$$

We set $\lambda = 1$ in all primary results and select $r^* \in \arg\max_{r_i \in \mathcal{R}} s_i$. Evaluation assigns uniform credit across tied highest-scoring candidates. A candidate favored by R can be demoted when its weaker chart support outweighs its reference-consistency advantage. Reference separation keeps this corrective signal unconditioned on the supplied E, while score fusion determines its contribution to the final ranking.

The matched non-separated variant tests the same scoring procedure without reference separation. It replaces z_i^V with $z_i^{V_E}$ while holding the checkpoint, clipping rule, fusion weight, candidate pool, and tie handling fixed. Both selectors are constructed from the same checkpoint and require one shared extraction call and two verification calls per candidate. Their comparison isolates reference access without adding another verifier, a separate training history, or additional verification calls.

4 Experiments

4.1 Experimental Setup

Our experiments test whether reference separation suppresses reference-induced errors while preserving strong candidate selection under ordinary evaluation. All methods operate on identical frozen pools of eight reasoning trajectories generated by InternVL3.5-8B, Qwen3.5-9B, Gemma-4-12B-it, Qwen2.5-VL-7B, and Qwen3-VL-8B (Wang et al., 2025; Gemma Team, 2026; Bai et al.,

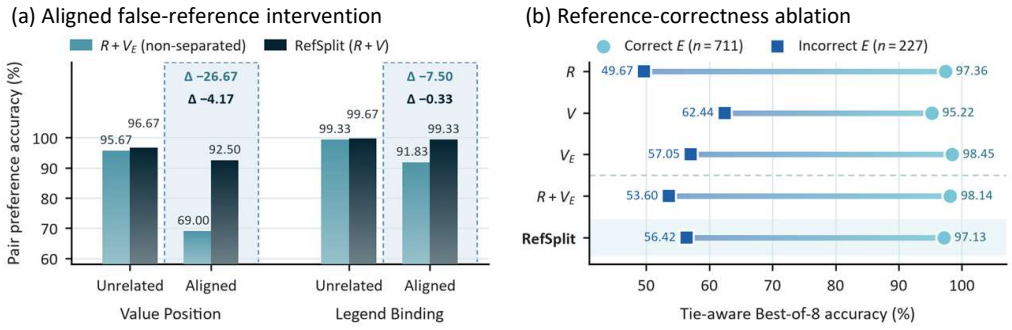

Figure 4: Controlled robustness and component complementarity. Panel (a) measures the effect of aligned false references under matched scoring. Panel (b) compares reference-conditioned and reference-free verification as reference correctness changes.

2025b;a). Using the same candidate pools ensures that performance differences reflect verification and selection rather than candidate generation.

The comparison set covers complementary approaches to multimodal reasoning selection. Self-Consistency (SC) provides a consensus-based baseline without an explicit verifier, VisualPRM represents general multimodal reasoning verification, and Perceval emphasizes verification against visual evidence (Wang et al., 2022; 2026b; Min et al., 2026). All baselines score the same frozen candidate pools. The components R, V_E, and V isolate the contribution of each verification signal, while $R + V_E$ provides the matched non-separated control that differs from RefSplit only in reference access. EVPV and TIM-PRM are discussed in Section 5, since matched end-to-end scores under our frozen-pool protocol are unavailable.

The benchmark suite spans complementary visual and reasoning demands. ChartQAPro covers diverse chart and question formats, CharXiv focuses on realistic scientific charts, and ChartMuseum stresses complex visual reasoning (Masry et al., 2025; Wang et al., 2024c; Tang et al., 2026). ChartX and ChartBench broaden chart structures and numerical reasoning while providing structured source annotations for deterministic reference audits (Xia et al., 2025; Xu et al., 2023). We report Pass@1, Pass@N, and tie-aware Best-of-N accuracy, averaging first across candidate policies and then equally across benchmarks. Appendix C provides the grading and aggregation details.

4.2 Causal Suppression of Aligned False References

We first test whether reference separation prevents a false shared premise from propagating into candidate selection. We compare $R + V_E$ with RefSplit on 300 Value Position pairs and 300 Legend Binding pairs. The unrelated-to-aligned intervention changes only the reference while fixing the chart, trajectories, checkpoint, fusion weight, and scoring rule. The two selectors therefore differ only in whether the grounding path receives the manipulated reference. Figure 4(a) reports the resulting change in preference accuracy.

Reference separation substantially reduces the effect of aligned corruption. On Value Position, preference accuracy drops by 26.67 points for the matched non-separated selector but by only 4.17 points for RefSplit. The observed decrease is therefore 84.4% smaller under reference separation. The paired 95% intervals for the signed changes are -32.50 to -21.17 and -7.17 to -1.33 points, respectively. The same pattern appears on Legend Binding, where the corresponding drops are 7.50 and 0.33 points. Both selectors retain the same reference-conditioned R path, while only RefSplit preserves a grounding signal that does not receive the false reference. The matched intervention therefore attributes the reduction in sensitivity to reference access rather than additional model capacity, training, or verification calls.

4.3 Complementarity Across Reference Correctness

We next examine whether reference-conditioned and reference-free verification remain complementary as reference correctness changes. The controlled stress setting contains 938 unique questions, two candidate policies, and eight candidates per question. Five injected candidates share one incorrect answer, creating repeated competing alternatives while the naturally extracted reference remains unchanged. We divide the questions into 711 correct-reference and 227 incorrect-reference cases. Figure 4(b) compares the verification components across these two conditions.

The results reveal a clear reversal in component strength. On incorrect-reference cases, direct chart verification V reaches 62.44% accuracy, compared with 49.67% for R. On correct-reference cases, V_E reaches 98.45%, compared with 95.22% for V. Explicit references therefore provide useful guidance when their premises are correct, while direct chart grounding provides the stronger signal when those premises are wrong. The crossover shows that reference-conditioned and reference-free verification provide complementary rather than interchangeable evidence.

The fusion results show how this complementarity affects candidate selection. On incorrect-reference cases, RefSplit increases accuracy from 53.60% for $R + V_E$ to 56.42%, a gain of 2.83 points. On correct-reference cases, accuracy changes from 98.14% to 97.13%. The fusion weight remains fixed across both conditions, so the improvement on erroneous references does not rely on condition-specific tuning. Reference separation instead preserves access to explicit premises while retaining a corrective chart signal when those premises are wrong.

4.4 Natural Shared-Reference Anchoring

We next test whether the failure isolated by controlled intervention also arises under ordinary candidate generation. We audit frozen ChartX and ChartBench candidate pools using their structured source annotations. For value errors, an incorrect candidate is defined as aligned when its reasoning or final answer reproduces the false value in the extracted reference. The unit of analysis is a question-policy pool in which eight candidates share the same reference. A manual audit of 200 sampled candidate-reference pairs yields 90.9% precision, 100.0% recall, and 99.0% agreement for the alignment rule after excluding one ambiguous pair.

Aligned incorrect candidates occur frequently without intervention. Among 2,509 audited incorrect-reference value-error pools, 664 contain at least one incorrect candidate that agrees with the false reference, corresponding to 26.46%. The rate is 24.02% on ChartX and 27.94% on ChartBench. Across policies, 51.19% of audited incorrect-reference questions contain at least one aligned pool. The controlled intervention therefore isolates a configuration that also occurs naturally in frozen policy-generated trajectories.

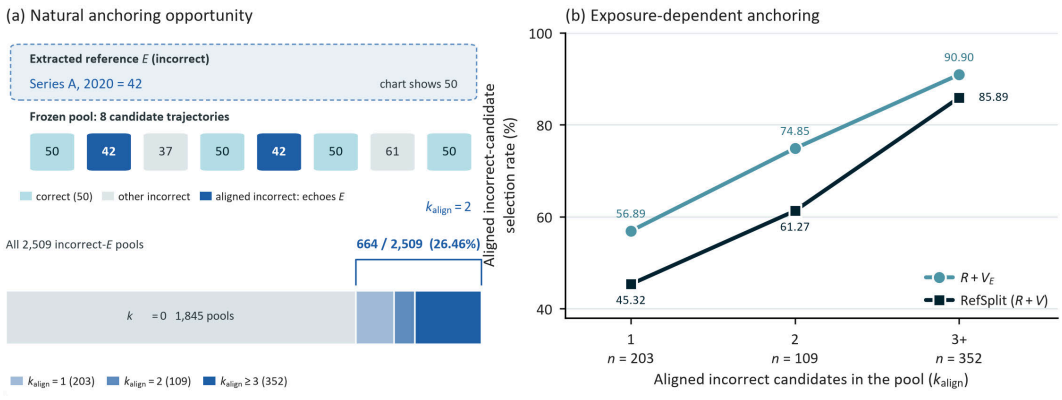

Figure 5: Natural occurrence and suppression of reference-aligned errors. Panel (a) measures how often incorrect-reference pools contain an incorrect candidate aligned with the false reference. Panel (b) compares aligned-error selection under matched non-separated and reference-separated verification across candidate multiplicities.

Reference separation also reduces selection of these naturally occurring aligned errors. Across the 664 aligned pools, the selection rate falls from 77.87% for $R + V_E$ to 69.44% for RefSplit. The reduction of 8.43 points has a 95% question-cluster bootstrap interval of 6.19 to 10.70 points. Figure 5(b) shows lower aligned-error selection under RefSplit in every displayed multiplicity group. Both selectors become more likely to choose an aligned error as more such candidates enter the pool, but reference separation remains less susceptible throughout.

The reduction in aligned-error selection is accompanied by measurable answer recovery. Accuracy on the same aligned pools increases from 4.82% to 6.53%, a gain of 1.71 points with a 95% interval of 0.52 to 3.01 points. The accuracy gain is smaller than the reduction in aligned-error selection because avoiding one aligned error can still lead to another incorrect candidate. We therefore use aligned-error selection as the direct measure of the diagnosed mechanism and accuracy as the resulting downstream outcome. The accuracy interaction between aligned and non-aligned pools is not statistically resolved, so the full accuracy gain cannot be attributed to alignment alone.

Table 1: Best-of-8 chart selection across five candidate policies. Best and second-best selector results are shown in bold and underlined, respectively.

	Baseline	Prior selectors			Ours	Oracle
Benchmark	Pass@1	SC	VisualPRM	Perceval	RefSplit	Pass@8
ChartQAPro (Masry et al., 2025)	38.90	**41.82**	40.38	39.88	<u>41.24</u>	55.53
CharXiv (Wang et al., 2024c)	51.61	<u>55.36</u>	54.08	52.20	**56.50**	78.36
ChartX (Xia et al., 2025)	65.69	<u>67.53</u>	66.70	66.54	**67.98**	79.43
ChartMuseum (Tang et al., 2026)	36.29	**40.18**	37.80	37.46	<u>39.72</u>	64.34
ChartBench (Xu et al., 2023)	57.69	<u>58.64</u>	57.93	57.73	**59.64**	68.16
Average	50.04	<u>52.71</u>	51.38	50.76	**53.02**	69.16

4.5 Public Benchmark Selection

Finally, we test whether reference separation preserves ordinary candidate-selection performance when neither references nor candidate pools are manipulated. Table 1 reports Best-of-8 accuracy across five public chart benchmarks and five candidate policies. Overall, RefSplit achieves the highest mean among the compared selectors. Its 53.02% mean exceeds SC, VisualPRM, and Perceval by 0.31, 1.64, and 2.25 points, respectively. RefSplit achieves the highest selector accuracy on CharXiv, ChartX, and ChartBench and the second-highest on ChartQAPro and ChartMuseum. It also improves over Pass@1 on all five benchmarks, showing that the separated verifier remains effective under ordinary candidate selection.

The matched non-separated control reveals a distinction that aggregate benchmark accuracy alone does not capture. Its mean accuracy is 53.05%, only 0.03 points above the 53.02% obtained by RefSplit. The two selectors therefore perform nearly identically on the unmodified public suite despite the large difference observed under aligned false-reference intervention. Reference separation reduces sensitivity to the diagnosed error pathway while changing the observed five-benchmark mean by only 0.03 points. Together with the natural-pool results, this comparison shows that the targeted robustness gain does not require sacrificing the ordinary selection quality observed across the public benchmarks. Appendix C.2 extends the matched comparison across candidate budgets and separates candidate availability from successful selection.

5 Related Work

5.1 Multimodal Reasoning Verification

Process supervision evaluates intermediate reasoning rather than only final answers (Lightman et al., 2024; Wang et al., 2024b; Zheng et al., 2025). VisualPRM extends this approach to multimodal trajectories, and recent generative verifiers study verification capabilities separately from solution generation (Wang et al., 2026b; Zhang et al., 2025; 2026). Our task is completed-trajectory selection from frozen pools, with explicit checks of reasoning consistency and visual premises. We study the information supplied to those checks rather than proposing a new step-level supervision scheme. The distinction places reference separation alongside, rather than in competition with, the choice of verification granularity.

5.2 Visual Premises and Intermediate References

Visual premise checking reconnects reasoning to image evidence through different access patterns (Xiao et al., 2026; Wu et al., 2025; Zhu et al., 2026). TIM-PRM queries visual tools, while Perceval verifies image-related claims against the source image (Kuang et al., 2025; Min et al., 2026). EVPV instead checks declared visual dependencies against structured constraints extracted once and reused across steps and candidates (Wang et al., 2026a). Our method retains an explicit reference but gives one grounding path direct chart access without that reference. The distinction concerns which intermediate information conditions each judgment, not whether visual premises should be checked.

Reference construction and reference access address different sources of dependence. SGV elicits candidate-independent priors, whereas BIRCH builds an anchor from corrected candidate answers (Andrade et al., 2026; Zou et al., 2026). Prior work also examines noisy priors and corrupted constraints (Andrade et al., 2026; Wang et al., 2026a; Nooralahzadeh et al., 2026; Panickssery et al., 2024). We isolate the directional influence of a reference aligned with a fixed incorrect candidate, then compare grounding paths that differ only in access to that record. The matched design connects reference fallibility to a specific, removable dependency between verification inputs.

5.3 Verifier Dependence and Candidate Selection

Repeated judgments and additional candidates do not guarantee reliable selection when their errors are dependent or their scoring signal is imperfect (Kim et al., 2025; Kohli, 2026; Panickssery et al., 2024; Malik et al., 2026; Li et al., 2026; Yu et al., 2026). Prior work diagnoses reward-model failures and limits of inference-time selection. We identify a concrete common input that can be manipulated while candidates remain fixed. Reference replacement tests its causal influence, and matched access restriction tests a targeted remedy. The resulting contribution is an experimentally isolated dependency in multimodal verification and a matched design that limits its influence on candidate selection.

6 Conclusion

We identify shared-reference anchoring as a causal failure mode in chart reasoning verification. A fallible reference can redirect multiple verification judgments when the same record conditions each path, even when the source chart remains available. To address this dependency, we introduce RefSplit, a reference-separated verification method that retains reference-conditioned reasoning while withholding the shared reference from a complementary chart-grounding path. Controlled interventions show that RefSplit reduces the preference-accuracy drop from 26.67 to 4.17 points on Value Position and from 7.50 to 0.33 points on Legend Binding. The failure also arises under ordinary generation, where 26.46% of audited incorrect-reference value-error pools contain an incorrect candidate aligned with the false reference. Across these pools, RefSplit reduces aligned-error selection by 8.43 points. On five public chart benchmarks and five candidate policies, RefSplit achieves the highest mean among the evaluated prior selectors while remaining within 0.03 points of the matched non-separated control. Reliable multimodal verification therefore depends not only on constructing useful intermediate evidence, but also on controlling which verification paths receive that evidence.

AI USE STATEMENT

We used generative AI tools, including OpenAI ChatGPT and Anthropic Claude, to assist with literature retrieval and discovery, improve writing quality, and generate synthetic training data. Retrieved references were verified against their original sources, and synthetic training examples were filtered and validated using the data construction procedures reported in this paper. The authors reviewed all AI-assisted outputs and take responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

We provide the model interfaces, reference-access definitions, score-fusion rule, intervention procedures, evaluation protocol, training construction, and optimization settings used in our experiments. We will release the training, inference, intervention, and evaluation code used in the reported experiments. The derived training instances will not be publicly released. We instead document their source datasets, construction and filtering procedures, composition, and use to support independent reconstruction of the training pipeline where permitted.

REFERENCES

Moises Andrade, Joonhyuk Cha, Brandon Ho, Vriksha Srihari, Karmesh Yadav, and Zsolt Kira. Let's Think in Two Steps: Mitigating Agreement Bias in MLLMs with Self-Grounded Verification. In *International Conference on Learning Representations*, volume 2026, pp. 36919–36972, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-VL Technical Report. *arXiv preprint arXiv:2511.21631*, 2025a. doi: 10.48550/arXiv.2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025b. doi: 10.48550/arXiv.2502.13923.

Jinyue Chen, Lingyu Kong, Haoran Wei, Chenglong Liu, Zheng Ge, Liang Zhao, Jianjian Sun, Chunrui Han, and Xiangyu Zhang. OneChart: Purify the Chart Structural Extraction via One Auxiliary Token. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 147–155, 2024.

Gemma Team. Gemma 4 Technical Report. *arXiv preprint arXiv:2607.02770*, 2026. doi: 10.48550/arXiv.2607.02770.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated Errors in Large Language Models. *arXiv preprint arXiv:2506.07962*, 2025.

Guneet Kohli. Nine Judges, Two Effective Votes: Correlated Errors Undermine LLM Evaluation Panels. *arXiv preprint arXiv:2605.29800*, 2026.

Peng Kuang, Xiangxiang Wang, Wentao Liu, Jian Dong, and Kaidi Xu. TIM-PRM: Verifying Multimodal Reasoning with Tool-Integrated PRM. *arXiv preprint arXiv:2511.22998*, 2025. doi: 10.48550/arXiv.2511.22998.

Jiachun Li, Pengfei Cao, Zhuoran Jin, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. Fixing the Broken Compass: Diagnosing and Improving Inference-Time Reward Modeling. In *International Conference on Learning Representations*, volume 2026, pp. 102566–102604, 2026.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's Verify Step by Step. In *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot Visual Language Reasoning by Plot-to-Table Translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399, 2023a.

Fangyu Liu, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Yasemin Altun, Nigel Collier, and Julian Eisenschlos. MatCha: Enhancing Visual Language Pretraining with Math Reasoning and Chart Derendering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12756–12770, 2023b.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. In *International Conference on Learning Representations*, volume 2024, pp. 23439–23554, 2024.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah Smith, Hanna Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing Reward Model Evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 144839–144866, 2026.

Ahmed Masry, Jia Qing Tan, Shafiq Joty, Enamul Hoque, et al. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Ahmed Masry, Parsa Kavehzadeh, Enamul Hoque, Shafiq Joty, et al. UniChart: A Universal Vision-Language Pretrained Model for Chart Comprehension and Reasoning. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 14662–14684, 2023.

Ahmed Masry, Mohammed Saidul Islam, Mahir Ahmed, Aayush Bajaj, Firoz Kabir, Aaryaman Kartha, Md Tahmid Rahman Laskar, Mizanur Rahman, Shadikur Rahman, Mehrad Shahmohammadi, et al. ChartQAPro: A More Diverse and Challenging Benchmark for Chart Question Answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 19123–19151, 2025.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. PlotQA: Reasoning over Scientific Plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1516–1525. IEEE, 2020.

Yingqian Min, Kun Zhou, Yifan Li, Yuhuan Wu, Han Peng, Yifan Du, Wayne Xin Zhao, Min Yang, and Ji-Rong Wen. Improving Vision-Language Models with Perception-Centric Process Reward Models. *arXiv preprint arXiv:2604.24583*, 2026.

Farhad Nooralahzadeh, Omid Rohanian, Yi Zhang, Jonathan Fürst, and Kurt Stockinger. Arbitration Failure, Not Perceptual Blindness: How Vision-Language Models Resolve Visual-Linguistic Conflicts. *arXiv preprint arXiv:2604.09364*, 2026.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. LLM Evaluators Recognize and Favor Their Own Generations. *Advances in Neural Information Processing Systems*, 37:68772–68802, 2024.

Liyan Tang, Grace Kim, Xinyu Zhao, Thom Lake, Wenxuan Ding, Fangcong Yin, Prasann Singhal, Manya Wadhwa, Zeyu Liu, Zayne Sprague, et al. ChartMuseum: Testing Visual Reasoning Capabilities of Large Vision-Language Models. *Advances in Neural Information Processing Systems*, 38, 2026.

Junxin Wang, Dai Guan, Weijie Qiu, Zhihang Li, Yongbo Gai, Zhengyi Yang, Mengyu Zhou, Erchao Zhao, Xiaoxi Jiang, and Guanjun Jiang. Grounding the Score: Explicit Visual Premise Verification for Reliable Vision-Language Process Reward Models. *arXiv preprint arXiv:2603.16253*, 2026a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169, 2024a.

Peiyi Wang, Lei Li, Zhihong Shao, Runxin Xu, Damai Dai, Yifei Li, Deli Chen, Yu Wu, and Zhifang Sui. Math-Shepherd: Verify and Reinforce LLMs Step-by-step without Human Annotations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9426–9439, 2024b.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Weiyun Wang, Zhangwei Gao, Lianjie Chen, Zhe Chen, Jinguo Zhu, Xiangyu Zhao, Yangzhou Liu, Yue Cao, Shenglong Ye, Xizhou Zhu, et al. VisualPRM400K: An Effective Dataset for Training Multimodal Process Reward Models. In *The Fourteenth International Conference on Learning Representations*, 2026b.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171*, 2022.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. CharXiv: Charting Gaps in Realistic Chart Understanding in Multimodal LLMs. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024c.

Xueqing Wu, Yuheng Ding, Bingxuan Li, Pan Lu, Da Yin, Kai-Wei Chang, and Nanyun Peng. VISCO: Benchmarking Fine-Grained Critique and Correction Towards Self-Improvement in Visual Reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9527–9537. IEEE, 2025.

Renqiu Xia, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Botian Shi, Junchi Yan, and Bo Zhang. ChartX & ChartVLM: A Versatile Benchmark and Foundation Model for Complicated Chart Reasoning. *IEEE Transactions on Image Processing*, 2025.

Renqiu Xia, Haoyang Peng, Hancheng Ye, Mingsheng Li, Xiangchao Yan, Peng Ye, Botian Shi, Yu Qiao, Junchi Yan, and Bo Zhang. StructChart: On the Schema, Metric, and Augmentation for Visual Chart Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Tong Xiao, Xin Xu, Zhenya Huang, Hongyu Gao, Quan Liu, Qi Liu, and Enhong Chen. Perception-R1: Advancing Multimodal Reasoning Capabilities of MLLMs via Visual Perception Reward. In *International Conference on Learning Representations*, volume 2026, pp. 26868–26898, 2026.

Zhengzhuo Xu, Sinan Du, Yiyan Qi, Chengjin Xu, Chun Yuan, and Jian Guo. ChartBench: A Benchmark for Complex Visual Reasoning in Charts. *arXiv preprint arXiv:2312.15915*, 2023.

Zhuohao Yu, Zhiwei Steven Wu, and Adam Block. From Curiosity to Caution: Mitigating Reward Hacking for Best-of-N with Pessimism. *arXiv preprint arXiv:2604.04648*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9556–9567, 2024.

Di Zhang, Jingdi Lei, Junxian Li, Xunzhi Wang, Yujie Liu, Zonglin Yang, Jiatong Li, Weida Wang, Suorong Yang, Jianbo Wu, et al. Critic-V: VLM Critics Help Catch VLM Errors in Multimodal Reasoning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9050–9061. IEEE, 2025.

Xinchen Zhang, Xiaoying Zhang, Youbin Wu, Yanbin Cao, Renrui Zhang, Ruihang Chu, Ling Yang, Yujiu Yang, and Guang Shi. Generative Universal Verifier as Multimodal Meta-Reasoner. In *International Conference on Learning Representations*, volume 2026, pp. 109211–109243, 2026.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying Process Errors in Mathematical Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025.

Didi Zhu, Changrui Chen, Stefanos Zafeiriou, and Jiankang Deng. VisualFLIP: Do Predictions Depend on Task-Critical Visual Evidence in Multimodal Reasoning? *arXiv preprint arXiv:2606.07872*, 2026.

Table 2: Component preference under reference replacement

Axis	Path	E^+	E_{random}	E_{aligned}	E_{none}
Value Position	R	94.67	95.33	66.67	97.33
	V_E	95.50	96.50	72.00	97.00
	V	97.17	97.17	97.17	97.17
Legend Binding	R	100.00	99.33	89.00	99.67
	V_E	100.00	99.67	94.17	99.67
	V	99.67	99.67	99.67	99.67

Xiaohan Zou, Roshan Sridhar, Mohammadtaher Safarzadeh, and Dan Roth. When Vision-Language Models Judge Without Seeing: Exposing Informativeness Bias. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15417–15448, 2026.

A Reference Intervention Details

A.1 Matched-Pair Construction

The matched-pair intervention tests whether a supplied reference changes judgments over otherwise fixed inputs. We construct 300 Value Position pairs and 300 Legend Binding pairs, each with a correct trajectory r^+ and a trajectory r^- that differs in one visual premise. The correct record supports r^+, the aligned record supports the false premise in r^-, the unrelated record comes from deterministic pairing with another item, and the empty condition supplies no reference content. Pair preference accuracy assigns one unit of credit when the correct trajectory ranks higher, half a unit for a tie, and zero otherwise.

The role comparison separates consistency with a reference from grounding in the chart. The path R receives the reference-conditioned consistency instruction. The paths V_E and V share the chart-grounding instruction, with the reference supplied only to V_E. We reuse cached V scores across reference conditions because its inputs never change. Its constant score therefore verifies the input restriction, while changes in R and V_E measure susceptibility to reference replacement.

Table 2 shows that aligned content, rather than reference presence alone, drives the largest preference loss. Correct, unrelated, and empty records leave both conditioned paths substantially more accurate than the aligned false record on each axis. The result rules out an account based only on the addition of reference text. The matched grounding comparison further locates susceptibility in reference access rather than only in the consistency objective of R.

A.2 Robustness Across Fixtures

Additional fixtures test whether the observed effect persists across sample sizes and rendering configurations. We evaluate Legend Binding on a separate 800-pair fixture and repeat both intervention axes with a second renderer. Unrelated references are paired again within each fixture. The 800-pair evaluation is therefore a separate run rather than a pooled extension of the 300-pair result.

Table 3 reproduces directional steering across the additional fixtures. On the 800-pair Legend Binding fixture, R falls from 99.25% to 88.75% under aligned replacement, while V_E falls from 99.44% to 93.88%. Renderer B preserves the same component ordering and direction on both axes. The stable pattern supports the reference-access mechanism beyond the primary rendering configuration.

A.3 Policy-Trajectory Intervention

Policy-generated candidates test reference influence without constructing both sides of a trajectory pair. Eligibility requires a parsed reference value and an incorrect candidate that states a number. We choose the first eligible incorrect candidate in the original pool order without using verifier scores. The aligned condition replaces the reference value with that candidate's asserted value and leaves

Table 3: Component preference across sample sizes and renderers

Fixture	Path	Pairs	E^+	E_{random}	E_{aligned}	E_{none}
Legend Binding	R	800	100.00	99.25	88.75	99.75
	V_E	800	100.00	99.44	93.88	99.50
	V	800	99.50	99.50	99.50	99.50
Value Position Renderer B	R	300	94.67	95.83	68.00	97.00
	V_E	300	96.00	96.50	74.33	96.83
	V	300	97.00	97.00	97.00	97.00
Legend Binding Renderer B	R	800	100.00	99.44	89.25	99.88
	V_E	800	100.00	99.44	94.56	99.75
	V	800	99.75	99.75	99.75	99.75

Table 4: Target adoption on eligible chart candidate pools

Reference condition	R	V_E	V
Extracted	14.37	13.23	12.80
Unrelated	12.53	12.83	12.80
Aligned false	21.60	21.17	12.80

the chart and all trajectories unchanged. The target therefore precedes the intervention rather than being chosen for a favorable measured response.

Table 4 shows increased target adoption in both conditioned paths across 3,000 eligible pools. Using the unrounded rates, the increases from unrelated to aligned references are 9.07 points for R and 8.33 points for V_E, with reported 95% intervals of 6.2 to 12.1 and 6.0 to 10.8 points. The unchanged V rate is consistent with its fixed input. Target adoption measures the intervention's designated direction rather than all possible answer errors.

Table 5 shows the intervention effect separately for the two reference-conditioned paths. Target adoption increases under aligned false references on all three benchmarks for both R and V_E. The magnitude varies across benchmarks, with the largest increase observed on ChartQAPro. The consistent direction across both conditioned paths supports reference access as the common source of the steering effect.

A.4 Reference-Field Ablation

The field ablation tests whether the complete record is necessary for steering or whether its false numerical value is sufficient. For 200 cases per benchmark from the Qwen2.5-VL-7B pools, we compare the unrelated record, the full aligned record, a value-only input, and the remaining fields after removing the asserted number. The candidates remain fixed throughout the comparison.

Table 6 shows that the false value alone sustains target adoption in all three subsets. On ChartQAPro, adoption is 27.00% with the value alone and 26.50% with the full aligned record, compared with 10.50% after removal of the number. CharXiv and ChartX show the same ordering. The result identifies the numerical premise as sufficient for the observed steering in these subsets rather than attributing the effect to the complete reference schema.

Table 5: Reference-conditioned target adoption by chart benchmark

	R		V_E		
Benchmark	Unrelated	Aligned false	Unrelated	Aligned false	Eligible cases
CharXiv	13.00	16.90	12.10	16.10	1,000
ChartX	13.90	21.10	13.20	21.40	1,000
ChartQAPro	10.70	26.80	13.20	26.00	1,000

Table 6: Target adoption under reference-field interventions for the R path on Qwen2.5-VL-7B pools

Benchmark	Cases	Unrelated	Aligned	Value only	Other fields
CharXiv	200	12.50	14.50	16.50	10.50
ChartX	200	9.00	20.50	23.00	10.00
ChartQAPro	200	4.50	26.50	27.00	10.50

B Implementation and Training

B.1 Role Interfaces, Candidate Generation, and Verdict Readout

The extraction interface returns a compact record of question-relevant chart information. It receives the chart and question and produces series, category, and value fields, with an optional relation field for extremum questions. The parser reads the named fields, converts the numerical value, permits omitted fields, and falls back to a bare number when the structured value does not parse. Candidate trajectories are not inputs to reference extraction.

The verification interfaces control which information each role can inspect. The role R judges consistency with the supplied reference. Direct chart verification V checks the completed trajectory against the chart without that reference, while V_E adds the same reference to the grounding interface. Both grounding variants use the same checkpoint. The deployed procedure passes the complete trajectory and does not invoke the auxiliary claim-recovery role.

Candidate trajectories are generated once and frozen before selector evaluation. We use vLLM with $N = 8$, temperature 0.8, top-p 0.95, a maximum generation length of 1,024 tokens, and seed 0. Gemma 4 is served through an OpenAI-compatible endpoint with thinking disabled. For Qwen3.5, the thinking channel is closed before the final response. The common generation instruction asks the model to reason in at most five short steps and place the final answer on a separate line. For multiple-choice questions, the final response is restricted to the option letter. All selectors receive exactly the same frozen candidate pools.

Verifier inference uses greedy decoding with batch size 8. Reference extraction is limited to 64 generated tokens, and each verification role is limited to 8 tokens. The verdict reader scans backward for the last accepted Yes or No token, takes the maximum logit across accepted variants of each verdict, and normalizes the two values. A missing verdict receives a score of 0.5. The resulting quantity is used for candidate ranking and is not interpreted as an independently calibrated probability of answer correctness.

B.2 Verifier Training

We train a single Qwen3-VL-8B-Instruct verifier by fully fine-tuning the language-model parameters while keeping the vision tower frozen. The training mixture contains 20,000 role-specific task records together with pairwise preference supervision. The task records cover reference extraction, reference-conditioned consistency verification, direct chart support, and auxiliary premise-oriented supervision. All verifier roles used at inference are produced by the same trained checkpoint.

The training data combine ChartQA training examples with locally re-rendered PlotQA charts (Masry et al., 2022; Methani et al., 2020). Chart-support supervision uses positive cells

Table 7: Candidate-generation configuration

Setting	Value
Inference engine	vLLM
Candidates per question	8
Temperature	0.8
Top-p	0.95
Maximum generation length	1,024 tokens
Seed	0
Reasoning format	At most five short steps
Final-answer format	Separate `Answer:` line
Multiple choice	Option letter only

Table 8: Verifier training configuration

Setting	Value
Backbone	Qwen3-VL-8B-Instruct
Training	Full LM fine-tuning with frozen vision tower
Optimizer	8-bit AdamW
Learning rate	1×10^{-5}
Supervised task records	20,000
Preference supervision	Pairwise ranking for R

derived from source tables and deterministic negative construction. The retained negatives include 2,248 wrong-cell examples, 2,058 perturbed-value examples, and 207 series-swap examples. Auxiliary records pair Qwen3-VL-8B-Instruct trajectories with premise records produced by a Qwen3.6-27B teacher. Their acceptance filter requires the numerical value to occur in the trajectory but does not require an oracle match for series or category strings.

Pairwise supervision ranks preferred trajectories above rejected trajectories for the reference-conditioned role R. The auxiliary supervision is used only during training and does not introduce an additional inference stage. The same trained checkpoint is used for reference extraction, R, V_E, and V so that the matched selector comparison changes reference access rather than model identity.

B.3 Optimization and Score Handling

The verifier uses single-stage full language-model fine-tuning with the vision tower frozen. We optimize with 8-bit AdamW and a learning rate of 1×10^{-5}. The training run uses the 20,000 supervised task records together with the pairwise preference samples described above. No additional verifier is trained for the reference-separated path.

Identical score handling is used for the matched fusion comparison. Both selectors clip verdict probabilities between 10^{-6} and $1 - 10^{-6}$ before conversion to log odds and assign uniform evaluation credit across tied highest scores. Raw component baselines instead rank the stored verdict probabilities directly. Clipping can therefore create boundary ties that make a zero-weight fusion differ slightly from the corresponding raw component baseline. We treat this numerical effect separately from the effect of reference access.

C Additional Evaluation Results

C.1 Grading and Paired Uncertainty

All selectors use the same correctness labels for the same frozen candidates. ChartQAPro uses deterministic benchmark grading. CharXiv, ChartX, ChartMuseum, and ChartBench use Qwen2.5-72B-Instruct with greedy decoding and the released grading procedures. ChartBench uses model output for answer extraction before applying the official 5% numerical rule. Table 9 specifies the common configuration.

Table 9: Grading configuration for the public chart suite

Benchmark	Evaluation configuration
ChartQAPro	Deterministic benchmark grading
CharXiv	Qwen2.5-72B-Instruct with the released grading prompt
ChartX	Qwen2.5-72B-Instruct with the released grading prompt
ChartMuseum	Qwen2.5-72B-Instruct with the released grading prompt
ChartBench	Qwen2.5-72B-Instruct answer extraction followed by the official 5% numerical rule

Table 10: Differences on the five-benchmark chart mean

Comparator	RefSplit minus comparator
$R + V_E$	-0.03
V	$+0.09$
VisualPRM	$+1.64$
Perceval	$+2.25$
SC	$+0.31$

The common grading configuration supports paired comparisons within our evaluation. Every selector uses the same frozen candidate labels, so label differences across selectors do not drive the reported contrasts. The absolute scores are not direct reproductions of external leaderboards because the original benchmark judges differ. Cross-judge agreement has not been measured in the reported evaluation.

Each public benchmark row averages accuracy across the five candidate policies. The five-benchmark mean is the macro average of those rows. Question-level bootstrap resampling preserves the paired evaluation unit. Reported intervals resample questions while retaining matched candidate pools and selector outcomes. In the natural analysis, all policy pools associated with one question form one bootstrap cluster. We distinguish questions, question-policy pools, and candidate-reference pairs because the same extracted reference can appear in multiple observations. The natural 26.46% rate uses 2,509 incorrect-reference value-error pools as its denominator rather than all benchmark questions or all generated candidates.

Table 10 reports descriptive differences on the public-suite mean. RefSplit exceeds VisualPRM, Perceval, and SC in the reported mean while remaining 0.03 points below the matched non-separated selector and 0.09 points above V. These point differences are not evidence of statistical equivalence. The controlled intervention and the natural aligned-error analysis provide the direct evidence for the reference-access mechanism.

C.2 Candidate-Budget Scaling

Candidate-budget scaling separates available correct answers from the selector's ability to identify them. We vary N over 1, 2, 4, and 8 by enumerating subsets of the same frozen eight-candidate pools. Figure 6 compares V, $R + V_E$, and RefSplit against the Pass@N oracle on ChartQAPro, CharXiv, and ChartX. The procedure changes the selection budget without regenerating trajectories.

The matched fused selectors remain close across the displayed budgets, while oracle availability increases more than realized selection accuracy. The separation result is therefore not confined to the eight-candidate endpoint. The widening gap to the oracle identifies remaining selection headroom rather than a failure caused by increasing the candidate budget. Candidate availability and reliable candidate identification remain separate requirements for inference-time gains.

C.3 Reference Error Types

The controlled error breakdown tests whether the aggregate separation benefit is shared across reference errors. The 227 incorrect records comprise 163 value errors, 24 series or category binding errors, and 40 multiple-field errors. Table 11 compares the same components and fused selectors

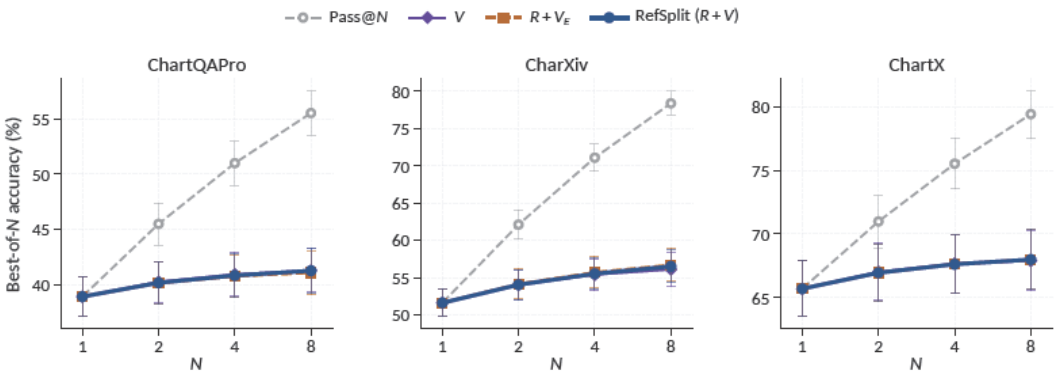

Figure 6: Candidate-budget scaling under matched reference access. Pass@N shows candidate availability, while V, $R + V_E$, and RefSplit select from the same frozen candidate pools. The two fused selectors remain close as oracle headroom increases with N.

Table 11: Controlled selection by audited reference error type

Error type	Questions	R	V_E	V	$R + V_E$	RefSplit
Value	163	49.08	56.90	65.18	53.48	56.80
Series or category binding	24	100.00	100.00	93.75	100.00	100.00
Multiple fields	40	21.88	31.87	32.50	26.25	28.75

within each stratum. The audit keeps this observational breakdown distinct from the designed Value Position and Legend Binding interventions.

Reference separation improves fused selection in the value and multiple-field strata. For value errors, RefSplit reaches 56.80% compared with 53.48% for $R + V_E$, while direct chart verification reaches 65.18%. Multiple-field errors show the same direction of improvement. The 24 binding-error questions show no difference between the fused selectors. The observed aggregate benefit is therefore concentrated mainly in the dominant value-error stratum.