S2R: SELECTIVE-STABILITY REPAIR IN FROZEN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models can be unstable when an answer should stay the same and insensitive when its visual support is weakened. We call the desired behavior *selective stability*: agreement across answer-preserving text and image changes, coupled with reduced confidence when proposed visual evidence is attenuated. **S2R** turns this criterion into a test-time repair signal. It reads answers at several points along the base reasoning trajectory and fits a temporary, hard-sparse edit of visual-token states in an otherwise frozen model. Hyperparameters and gate thresholds are selected on separate calibration data. On Qwen3-VL-8B-Thinking, S2R raises macro-average accuracy from 72.4% to 75.5%, while S2R-Full reaches 75.7%. The improvement transfers to two other frozen 8B backbones, and ablations show that preserving views, the directional evidence constraint, and localized support contribute complementary gains.

1 INTRODUCTION

Accuracy says whether a vision–language model (VLM) answered correctly, but not whether the answer depends on the question and image in the right way. A correct prediction may disappear after an equivalent rewording or a small edit to an uninformative image region. Conversely, a confident prediction may survive even after the region the model presents as evidence is obscured. In both cases, the response changes out of proportion to the information that was changed.

Figure 1 illustrates this mismatch. The original answer is correct, yet a paraphrase and a low-relevance occlusion change it; applying the same occlusion to the proposed-evidence region barely changes the model's confidence. Related failures appear under meaning-preserving prompts, diagram edits, and dynamically generated visual variants (Qiu et al., 2022; Zou et al., 2025; Sadanandan and Behzadan, 2026). Robustness therefore cannot mean invariance to every perturbation. It must distinguish changes that preserve the answer from changes that weaken the support for that answer.

We call this relationship *selective stability*: answer-preserving views should agree, while attenuation of proposed visual evidence should reduce confidence in the anchored answer. The two requirements are complementary. Agreement alone can consolidate an answer that ignores the image; sensitivity alone can make the model react to irrelevant changes. Their combination asks for stability only where the available evidence warrants it, connecting behavioral testing and invariant/covariant analysis to a usable test-time signal (Tramèr et al., 2020; Ribeiro et al., 2020; Goyal et al., 2017; Agarwal et al., 2020).

S2R instantiates this idea on a single request. It creates constrained question variants and matched image pairs: the same operator, strength, and area are applied once to a low-relevance region and once to a region proposed by the unedited model as evidence. The model's answers are read at several points along its fixed base reasoning trace, exposing failures that occur before, during, or after the trace has biased the decision. These controlled views define one answer-level objective rather than a pseudo-label.

Repair changes neither weights nor prompts. S2R selects 128 scalar coordinates from visual-token states, optimizes a shared additive edit on that support, regenerates the reasoning and answer from the repaired visual context, and then discards the edit. S2R-Full evaluates every controlled view;

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

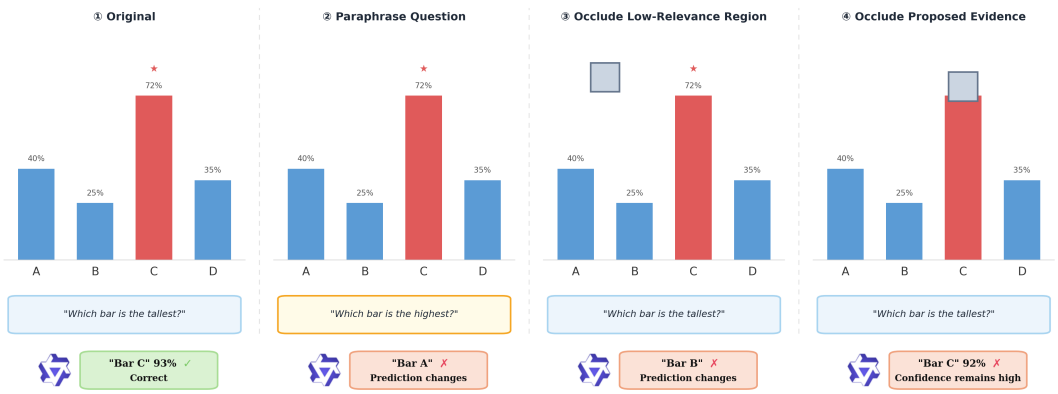


Figure 1: **A selective-stability violation.** An equivalent paraphrase and an occlusion of a low-relevance region change a correct answer, whereas the same occlusion over the proposed-evidence region has little effect. Selective stability preserves the answer in the middle two cases and lowers confidence in the evidence-attenuated view.

default S2R samples views and stops adaptively; S2R-Lite uses a four-update budget. A calibrated cascade runs Lite only when the measured violation warrants its cost.

Across the complete public splits of six visual-reasoning benchmarks, default S2R improves Qwen3-VL-8B-Thinking from 72.4% to 75.5% macro-average accuracy, and S2R-Full reaches 75.7%. The gain transfers to Qwen3-VL-8B-Instruct and InternVL3.5-8B. Forward aggregation, dense editing, and symmetric consistency objectives improve less, while support corruption degrades accuracy in proportion to how much of the selected support is lost.

Contributions.

- We formulate selective stability as the joint requirement of agreement under answer-preserving changes and sensitivity to attenuation of model-proposed visual evidence.
- We derive a trajectory-conditioned repair objective and optimize it through a temporary hard-sparse edit of visual-token states in a frozen VLM.
- We evaluate accuracy, transfer, support quality, optimization depth, cost, and selective execution across six benchmarks and three 8B backbones.

2 Related Work

Behavioral tests of visual grounding. Robust prediction requires invariance to nuisance changes and responsiveness to task-relevant ones (Tramèr et al., 2020). Checklists, counterfactual pairs, and invariant/covariant edits operationalize this idea for VQA (Ribeiro et al., 2020; Goyal et al., 2017; Agarwal et al., 2020); recent benchmarks reveal related failures in diagrams and mathematical figures (Zhang et al., 2024; Zou et al., 2025; Liu et al., 2026). VISE and SIVA-RL use paired invariance and sensitivity signals for training or self-evolution (Venkatraman et al., 2026; Tang et al., 2026). S2R instead uses the same distinction to diagnose and repair one test request without updating model weights.

Test-time adaptation and inference-time search. Test-time methods adapt classifiers, prompts, parameters, or prototypes from unlabeled views (Zhang et al., 2022; Sun et al., 2020; Shu et al., 2022; Zhu et al., 2024; Karmanov et al., 2024). For generative VLMs, self-consistency and best-of-n search over solutions, while TTAug and Test-Time Consistency aggregate or adapt across input variants (Wang et al., 2022; Kaya et al., 2025; Chou et al., 2026; Sheng et al., 2025). Other methods target CLIP-style or streaming adaptation (Zhu et al., 2026; Wang et al., 2026; Cao et al., 2025; Ye et al., 2026). These approaches mainly reward agreement or confidence. S2R also requires a directional change when proposed evidence is attenuated, preventing consistency from becoming the only target.

Intervening on internal representations. Hidden states can be steered by gradient objectives while parameters remain fixed (Dathathri et al., 2019; Yoshida and Gimpel, 2021). Multimodal variants optimize visual prompts, derive steering directions, contrast distorted images, or amplify intermediate visual signals (Wu et al., 2024; Liu et al., 2025; Nguyen et al., 2026; Leng et al., 2024; Li et al., 2025; Yin et al., 2025). S2R differs in how it obtains both target and support: controlled views of the current request define the objective, and a gradient–attention score selects a small set of visual-state coordinates for the edit.

3 Selective Stability as a Repair Principle

Let $c_0 = (x, q)$ denote the original request. The unedited model produces a reasoning trace r and a normalized answer $\hat{y}_0$. We use $\hat{y}_0$ as an *anchor* for measuring how evidence changes, not as a pseudo-label that the repaired model must preserve. The diagnostic asks two questions: do answer-preserving views agree, and does weakening proposed visual evidence reduce support for the anchor? S2R optimizes both conditions jointly.

3.1 Trajectory-conditioned answer distributions

A controlled view must be compared through the model's answer rather than its free-form explanation. For the Thinking backbone, we therefore score answers at three positions along the detached base trace, $\mathcal{T} = \{0, \frac{1}{2}, 1\}$. At position τ, the model receives the corresponding prefix $r_{\leq\tau}$ followed by `Final:`, yielding $\pi_\delta^\tau(\cdot \mid c)$. The direct readout $\tau = 0$ omits the trace; the other two reveal whether an edit survives increasingly committed reasoning context. Non-Thinking backbones use only $\tau = 0$.

For multiple-choice inputs, the distribution is normalized over valid option labels. For free-form inputs, the unedited model constructs a fixed set of at most eight short candidate answers, and we normalize their length-adjusted likelihoods. The candidate set, base trace, and anchor stay fixed during repair. They are discarded for evaluation, where the model generates a new trace and answer autoregressively. Appendix A.2 gives the exact scoring rule.

3.2 A model-derived evidence proposal

No localization annotation is available at test time, so the intervention is grounded in the unedited model's own computation. Each visual token t receives

$$\rho_t = 0.6\, g_t + 0.4\, a_t,$$

where g_t is the local sensitivity of the trajectory-averaged anchor score and a_t is attention-based information flow (Simonyan et al., 2013; Abnar and Zuidema, 2020). The gradient term identifies tokens that can change the answer locally; attention supplies broader spatial context. The smallest top-ranked patch set containing 25% of the relevance mass becomes the *proposed-evidence region*; a matched bottom-ranked set becomes the *low-relevance region*. Both are computed once at $\delta = 0$ and detached.

3.3 Controlled views

Question views use deterministic, rule-based edits: wh-form alternation, clause reordering, active–passive alternation when applicable, and question-type-specific synonyms. A filter rejects any rewrite that changes a number, comparison, negation, entity, unit, relation, option identifier, or answer type. Three variants are retained for every request and a fourth when another rule applies; image-only tasks apply the same procedure to their fixed task cue.

Image views are matched by construction. We sample three operator families from Gaussian blur, patch masking, block occlusion, additive noise, local contrast, and patch shuffle. For each family, the same strength and perturbed area are applied once to the low-relevance region and once to the proposed-evidence region. Text variants and low-relevance views should preserve the answer distribution. Proposed-evidence views should instead reduce confidence in the anchor relative to their matched low-relevance views. The rules and operator strengths are fixed before evaluation (Appendix B).

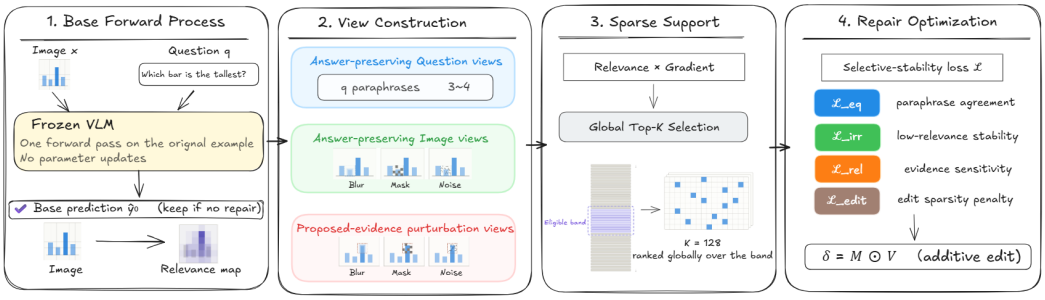


Figure 2: **S2R pipeline.** The frozen VLM produces an anchor answer and evidence proposal. Matched views measure disagreement under preserving changes and the confidence drop under proposed-evidence attenuation. S2R selects a hard-sparse support, optimizes one shared visual-state edit, and regenerates the answer. The online gate can bypass repair.

4 S2R: Sparse Visual-State Repair

The views expose a behavioral mismatch, but they do not specify a target answer. S2R resolves the mismatch inside the current forward computation: all branches share one temporary edit of visual-token states, optimized only for this request (Figure 2).

4.1 Shared hard-sparse edit

Let $H_{\text{vis}}^{\ell,b}$ be the visual-token states at editable layer ℓ for branch b. Every branch receives the same additive edit,

$$\widetilde{H}_{\text{vis}}^{\ell,b} = H_{\text{vis}}^{\ell,b} + \delta^{\ell}, \qquad \delta^{\ell} = M^{\ell} \odot V^{\ell}.$$

Sharing the edit prevents the branches from becoming individually consistent without yielding one state that works for the original request. To construct the binary masks M^{ℓ}, each scalar coordinate is scored by $\rho_t |\partial \bar{S}_0 / \partial H_{\text{vis},tj}^{\ell}|$, where $\bar{S}_0$ is the trajectory-averaged log probability of $\hat{y}_0$. The global top $K = 128$ coordinates across an eight-layer middle band are retained; only their fp32 values V^{ℓ} are optimized. The model parameters and all other activations remain fixed.

4.2 Selective-stability objective

Agreement is measured by symmetric KL and averaged across the available trajectory positions:

$$D_{\text{SKL}}(p,q) = \tfrac{1}{2}\,\text{KL}(p\|q) + \tfrac{1}{2}\,\text{KL}(q\|p), \qquad \bar{D}_{\delta}(c,c') = \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} D_{\text{SKL}}(\pi_{\delta}^{\tau}(\cdot \mid c), \pi_{\delta}^{\tau}(\cdot \mid c'))\,.$$

For matched image pair r, the evidence-attenuation gap is

$$d_r(\delta) = \frac{1}{|\mathcal{T}|}\sum_{\tau\in\mathcal{T}} \Big[\log \pi_{\delta}^{\tau}(\hat{y}_0 \mid \tilde{x}_{\text{irr}}^{(r)}, q) - \log \pi_{\delta}^{\tau}(\hat{y}_0 \mid \tilde{x}_{\text{rel}}^{(r)}, q)\Big]\,.$$

Because the pair differs primarily in location, a larger d_r means the anchor depends more on the proposed-evidence region than on the matched low-relevance region. S2R minimizes

$$\mathcal{L} = \mathcal{L}_{\text{eq}} + \lambda_1 \mathcal{L}_{\text{irr}} + \lambda_2 \mathcal{L}_{\text{rel}} + \lambda_3 \mathcal{L}_{\text{edit}}, \tag{1}$$

$$\mathcal{L}_{\text{eq}} = \frac{2}{m(m-1)} \sum_{i<j} \bar{D}_\delta \Big((x, q^{(i)}), (x, q^{(j)}) \Big), \tag{2}$$

$$\mathcal{L}_{\text{irr}} = \frac{1}{R} \sum_{r=1}^{R} \bar{D}_\delta \Big((x, q), (\tilde{x}_{\text{irr}}^{(r)}, q) \Big), \tag{3}$$

$$\mathcal{L}_{\text{rel}} = \frac{1}{R} \sum_{r=1}^{R} \max\{0, \kappa - d_r(\delta)\}, \tag{4}$$

$$\mathcal{L}_{\text{edit}} = \sum_{\ell} \left(\|\delta^\ell\|_1 + \gamma \|\delta^\ell\|_F^2 \right). \tag{5}$$

The preserving terms align distributions without forcing the original answer. The hinge requires an evidence-attenuation gap of at least κ and then becomes inactive. The final term limits the magnitude of the already sparse edit.

4.3 Optimization, regeneration, and gating

Default S2R caches states below the editable band and, at each update, samples two question variants, one matched image pair, and one trajectory position. A fixed checkpoint tuple is evaluated after steps 2, 4, and 6. Optimization stops when its residual improves by less than $\epsilon = 0.03$ and the top answer is unchanged; otherwise the six-update cap is used. S2R-Lite uses two image pairs and at most four updates. S2R-Full evaluates every view and readout for six updates without early stopping. The calibrated coefficients are $(\lambda_1, \lambda_2, \lambda_3, \gamma, \kappa) = (0.60, 0.80, 0.015, 0.20, 0.10)$ with step size 0.08.

After optimization, the base trace and text-side caches are discarded. The model rebuilds the multimodal context with the repaired visual states and generates a new reasoning trajectory and answer; the edit is then removed. This separation keeps the fixed trace as an optimization probe rather than part of the final prediction.

The online gate uses a cheaper version of the same diagnostic. It first standardizes disagreement with one question variant together with answer entropy. Clearly stable inputs keep the base answer, clear violations run Lite, and intermediate cases receive one low-relevance image check. Thresholds are selected under a 30% repair budget on calibration data and then frozen (Appendix C.4).

5 Experiments

The evaluation asks four questions: whether repair improves final answers beyond additional inference, which constraints and support choices matter, what changes on corrected and regressed examples, and how much repair can be avoided by gating.

5.1 Benchmarks and models

We use the complete public evaluation splits of six visual-reasoning benchmarks: MMMU validation (900), MathVista testmini (1,000), all five MathVerse testmini versions (3,940), DynaMath variants 1–10 (5,010), LogicVista test (448), and WeMath testmini (1,740), for 13,038 inputs (Yue et al., 2024; Lu et al., 2024; Zhang et al., 2024; Zou et al., 2025; Xiao et al., 2024; Qiao et al., 2025). MathVerse results include all versions of each underlying problem; DynaMath contains ten variants of each of 501 seed questions. *Avg* is the unweighted mean over the six benchmarks.

To check that the evaluation is not driven by text-solvable items, two annotators label 200 examples per benchmark for visual dependence. The rate ranges from 74% on MMMU to 96% on DynaMath, with Cohen's κ between 0.78 and 0.88; disagreements are adjudicated by a third annotator. Split details are in Appendix A, and the audit protocol is in Appendix A.5.

Qwen3-VL-8B-Thinking is the primary backbone for comparisons, ablations, runtime, and gating (Bai et al., 2025). Transfer experiments use Qwen3-VL-8B-Instruct and InternVL3.5-8B (Wang et al., 2025). All parameters remain frozen. The Qwen models edit layers 14–21 and InternVL edits layers 12–19; the objective, support size, view construction, and update budgets otherwise remain unchanged.

5.2 Protocol and comparisons

One global configuration per method is selected on a disjoint 1,200-example MMMU pool: 40 examples from each of 30 subjects, sampled with random seed 2026. Selection maximizes mean subject accuracy, breaking ties toward the smaller inference budget. No benchmark-specific tuning is performed. Reported stochastic results average random seeds 2027–2029.

All deterministic methods use greedy decoding with a 4,096-token generation cap. Self-consistency and Best-of-n draw eight samples; verifier reranking uses six candidates and one same-backbone verification pass; VisRef uses two refocusing rounds; Test-Time Consistency uses four branches; and TTAug uses 16 views. The same prompt, answer extraction, and normalization are shared across methods. Appendix A.3 gives the calibration grids and complete decoding settings.

Baselines and controls. Baselines cover candidate search (self-consistency, Best-of-n, and same-backbone reranking), visual refocusing (VisRef), and view-based test-time adaptation or aggregation (Test-Time Consistency and TTAug) (Wang et al., 2022; Ghosal et al., 2026; Chou et al., 2026; Kaya et al., 2025). VTI and ClearSight intervene directly on multimodal representations (Liu et al., 2025; Yin et al., 2025). A forward-only view ensemble averages logits over preserving views.

Two controls isolate the repair mechanism. The dense edit uses the full S2R objective but removes the sparse mask. The sparse-symmetric edit keeps S2R's support and optimization budget but treats proposed-evidence views as preserving, removing the directional hinge. Their implementations are given in Appendix A.4.

Measurement. We report normalized exact-match accuracy, correctness transitions, and the three selective-stability residuals. Bootstrap intervals cluster MathVerse versions by underlying problem and DynaMath variants by seed question. Latency and peak memory are measured on one A100-80GB GPU over 1,000 requests after 200 warm-ups, including the base decode, view construction, repair, and fresh regeneration. Appendix A.3 provides the stage-level measurement protocol.

6 Results

6.1 Overall accuracy and cost

Table 1 compares S2R with methods that spend additional inference on sampling, reranking, refocusing, or view aggregation. They improve the 72.4% base accuracy to at most 74.5%; S2R reaches 75.5%, and S2R-Full reaches 75.7%.

VisRef is the strongest external baseline at 74.5%. S2R exceeds it by 1.0 point (95% CI $[+0.6, +1.4]$, $p < 0.001$), and S2R-Full by 1.2 points ($[+0.8, +1.6]$, $p < 0.001$). Every benchmark improves over the base model, with positive example-clustered intervals (Appendix Table D.3). On DynaMath, the gain also holds under the stricter seed-question criterion: worst-case accuracy rises from 41.5% to 46.8% with S2R and 47.5% with S2R-Full.

Table 1: Accuracy (%) on Qwen3-VL-8B-Thinking. Avg is the unweighted benchmark mean. Bold marks the best result; underlining marks the strongest evaluated external baseline in each column. DynaMath reports average accuracy over 5,010 variants; seed-question worst-case results are in Appendix D.

Method	MMMU	MathVista	MathVerse	DynaMath	LogicVista	WeMath	Avg
Base	73.6	80.8	77.0	72.8	64.7	65.7	72.4
Self-consistency	74.1	81.1	77.6	73.5	64.6	66.2	72.8
Best-of-n	74.4	81.7	78.2	73.7	65.2	66.4	73.3
Verifier reranking	74.9	82.0	78.4	74.5	65.6	66.9	73.7
VisRef	<u>75.6</u>	<u>83.1</u>	<u>79.7</u>	<u>75.1</u>	66.2	67.3	<u>74.5</u>
VTI	75.1	82.5	79.1	74.8	65.9	67.4	74.1
ClearSight	75.3	82.8	79.3	75.0	66.0	<u>67.6</u>	74.3
View ensemble	75.0	82.4	79.1	74.7	66.1	67.0	74.0
TT-Consistency	75.4	82.6	79.0	74.9	<u>66.3</u>	67.1	74.2
TTAug	75.2	82.8	79.2	75.0	66.1	67.4	74.3
Dense edit	74.4	81.8	78.2	74.0	65.4	66.5	73.4
Sparse symmetric	75.0	82.3	78.8	74.6	65.8	66.9	73.9
S2R	76.7	83.6	80.2	76.4	68.0	68.1	75.5
S2R-Full	**76.9**	**83.8**	**80.4**	**76.8**	**68.4**	**68.2**	**75.7**

Table 2: Accuracy–cost comparison on Qwen3-VL-8B-Thinking. Latency and peak memory are measured on one A100-80GB GPU.

Method	Avg Acc	Latency (s)	Memory (GiB)
Base	72.4	12.8	22.6
VisRef	74.5	23.4	31.6
TT-Consistency	74.2	41.8	34.2
S2R-Lite	75.1	28.1	44.6
S2R	75.5	31.3	47.2
S2R-Full	75.7	42.4	49.6

Default S2R is faster and more accurate than TT-Consistency, while Lite reaches 75.1% in 28.1 s. For default S2R, the 31.3 s total divides into 12.8 s for the base decode, 1.4 s for view/readout construction, 7.0 s for repair, and 10.1 s for fresh generation. S2R-Full gains another 0.2 point but raises latency to 42.4 s. Editable activations are the main systems cost, increasing peak memory to 44.6–49.6 GiB.

6.2 Transfer and ablations

Table 3 shows that S2R-Full exceeds VisRef by 1.4 points on Qwen3-VL-8B-Instruct and 1.5 points on InternVL3.5-8B; Appendix Table D.1 gives the per-benchmark results. The ablations mirror the construction of the method. Adding low-relevance preservation to question agreement gives 0.9 point, the directional evidence term adds 0.6, and edit regularization supplies the final 0.4. Accuracy peaks with 128 coordinates in the middle band, three matched pairs, and six updates; larger edits or longer optimization do not help.

The forward ensemble, dense edit, and sparse-symmetric edit reach 74.0%, 73.4%, and 73.9%, respectively. Thus the views are not sufficient by themselves: their largest gain comes from constraining one localized edit and treating proposed-evidence attenuation directionally rather than as another view to match.

Support corruption produces a graded decline (Appendix Table D.2): a two-patch shift loses 1.3 points, replacing 75% of coordinates loses 1.9, and random support loses 2.6. Random matched regions reach 73.6%, while swapping proposed-evidence and low-relevance regions reaches 72.9%. The location and ordering of the pair, not merely its perturbation budget, carry the repair signal.

Table 3: Transfer and ablations. Panel (a) reports macro-average accuracy on three frozen 8B backbones. Panel (b) varies one component on Qwen3-VL-8B-Thinking; slash-separated entries follow the order in the row.

(a) Transfer across backbones			
Backbone	Base	VisRef	S2R-Full
Qwen3-VL-8B-Thinking	72.4	74.5	75.7
Qwen3-VL-8B-Instruct	68.5	70.4	71.8
InternVL3.5-8B	60.9	62.6	64.1
(b) Objective, support, and optimization ablations			
Variant	Avg		
$\mathcal{L}_{\text{eq}}$ / $+\mathcal{L}_{\text{irr}}$ / $+\mathcal{L}_{\text{rel}}$ / Full	73.8 / 74.7 / 75.3 / 75.7		
Matched pairs $R = 1/2/3$	74.9 / 75.4 / 75.7		
$K = 64/128/256$	75.2 / 75.7 / 75.4		
Low / middle / high band	74.9 / 75.7 / 75.2		
Steps $= 2/4/6/8$	74.6 / 75.3 / 75.7 / 75.5		
Lite / default / Full	75.1 / 75.5 / 75.7		

6.3 Behavioral effects of repair

Across pooled predictions, S2R-Full changes 5.29% of examples from wrong to correct and 1.80% from correct to wrong, a net gain of 3.49 points. Table 4 summarizes how the diagnostic and objective move during repair.

Table 4: Behavioral and optimization dynamics on Qwen3-VL-8B-Thinking. Correction and regression columns report mean within-group changes.

(a) Selective-stability residuals			
Metric	Overall before $\rightarrow$ after	Correction Δ	Regression Δ
Paraphrase disagreement	$.118 \rightarrow .045$	$-.097$	$-.028$
Low-relevance sensitivity	$.107 \rightarrow .039$	$-.083$	$-.022$
Evidence-attenuation gap	$.061 \rightarrow .142$	$+.109$	$+.021$
(b) Optimization depth			
Updates	2	4	6 / 8
Objective	.79	.54	$.41/.39$
Accuracy (%)	74.6	75.3	$75.7/75.5$

All three residuals move farther on corrections than regressions. Default S2R stops after 2/4/6 updates on 27/44/29% of examples.

6.4 Selective-repair effectiveness

Table 5: Selective repair on Qwen3-VL-8B-Thinking. Accuracy is the six-benchmark macro-average.

Policy	Repair fraction	Avg Acc
None	0%	72.4
Question threshold, Lite	24%	74.5
Online cascade, Lite	28%	74.7
All-input visual check, Lite	30%	74.8
Always-on Lite	100%	75.1

The online cascade repairs 28% of examples and recovers 2.3 of the 2.7 points gained by always-on Lite. Running the visual check on every request adds only 0.1 point, indicating that the question-side screen already captures most repairable failures.

A budget-matched ranking analysis in Appendix D.5 (Table D.4) shows that the richer score ranks requests more precisely: repairing its top 50% reaches 77.4% pooled accuracy, compared with 77.0% for always-on S2R-Full. The sweep uses pooled rather than benchmark macro-average accuracy, so it measures ranking quality rather than replacing Table 5 as the deployment result.

7 Discussion and Limitations

The experiments separate two effects that are often conflated at test time. Preserving-view agreement reduces incidental variation, but the sparse-symmetric control shows that agreement alone is not enough. The evidence-attenuation gap adds information about where the answer should cease to be stable. Its benefit, together with the monotonic loss under support corruption, is the central evidence that selective rather than universal consistency drives the repair.

The evidence proposal is also the main source of uncertainty. Gradient and attention scores can emphasize spurious regions (Adebayo et al., 2018); if that happens, S2R may suppress a useful cue or reinforce an irrelevant one. The method assumes that the frozen backbone already contains a usable visual representation and a partially informative answer distribution. It cannot recover perception or knowledge that is absent from the model.

The trajectory-conditioned distribution is an optimization proxy. Free-form tasks use a fixed short-answer set, and Thinking branches reuse a detached base trace, whereas evaluation regenerates without either restriction. The gap between objective value and final accuracy after six updates reflects this mismatch. Adaptive stopping limits over-optimization, but editable activations still make memory, rather than latency alone, a practical constraint.

Finally, the study covers three frozen 8B backbones on one A100-80GB GPU, with hyperparameters calibrated on MMMU. Larger, quantized, or domain-specific models may require different layer bands and repair budgets.

8 Conclusion

Selective stability asks a VLM to remain stable under answer-preserving changes and to become less confident when its proposed evidence is weakened. S2R enforces this relation at test time with a temporary sparse edit of visual states, then regenerates from the repaired context. Across three frozen 8B backbones, the resulting gains depend on both the preserving constraints and the directional evidence signal; a calibrated gate retains most of Lite's improvement while repairing 28% of requests.

Ethics Statement

This work studies test-time behavior of vision–language models on public benchmarks and collects no personal or sensitive information. S2R does not guarantee correct predictions, and an inaccurate model-derived relevance proposal may direct repair toward a spurious visual cue. Human oversight remains necessary for consequential uses.

Reproducibility Statement

The experimental protocol, benchmark splits, filtering rules, baseline configurations, hyperparameters, optimization procedure, randomization, hardware, and statistical analyses are documented in the main paper and Appendix A.

AI Use Statement

Generative AI tools were used to assist with language editing and LaTeX formatting. All AI-assisted changes were reviewed by the authors, who take responsibility for the final content of this work.

References

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197. Association for Computational Linguistics, 2020.

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020.

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*, 2025.

Chentao Cao, Zhun Zhong, Tongliang Liu, Yang Liu, Kun Zhang, Bo Han, et al. Noisy test-time adaptation in vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 59008–59039, 2025.

Shih-Han Chou, Shivam Chandhok, James J Little, and Leonid Sigal. Test-time consistency in vision language models. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 7789–7798. IEEE, 2026.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. Plug and play language models: A simple approach to controlled text generation. *arXiv preprint arXiv:1912.02164*, 2019.

Soumya Suvra Ghosal, Youngeun Kim, Zhuowei Li, Ritwick Chaudhry, Linghan Xu, Hongjing Zhang, Jakub Zablocki, Yifan Xing, and Qin Zhang. Visref: Visual refocusing while thinking improves test-time scaling in multi-modal large reasoning models. *arXiv preprint arXiv:2603.00207*, 2026.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.

Mehmet Onurcan Kaya, Desmond Elliott, and Dim P Papadopoulos. Efficient test-time scaling for small vision-language models. *arXiv preprint arXiv:2510.03574*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*, 2025.

Qing'an Liu, Juntong Feng, Yuhao Wang, Xinzhe Han, Yujie Cheng, Yue Zhu, Haiwen Diao, Yunzhi Zhuge, and Huchuan Lu. Vista-bench: Do vision-language models really understand visualized text as well as pure text? *arXiv preprint arXiv:2602.04802*, 2026.

Sheng Liu, Haotian Ye, and James Y Zou. Reducing hallucinations in large vision-language models via latent space steering. In *International Conference on Learning Representations*, volume 2025, pages 72402–72419, 2025.

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, pages 23439–23554, 2024.

Duy Nguyen, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. GrAInS: Gradient-based attribution for inference-time steering of LLMs and VLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 46523–46543. Association for Computational Linguistics, 2026.

Runqi Qiao, Qiuna Tan, Guanting Dong, MinhuiWu MinhuiWu, Chong Sun, Xiaoshuai Song, Jiapeng Wang, Zhuoma Gongque, Shanglin Lei, Yifan Zhang, et al. We-math: Does your large multimodal model achieve human-like mathematical reasoning? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20023–20070, 2025.

Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking robustness of multimodal image-text models under distribution shift. *arXiv preprint arXiv:2212.08044*, 2022.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 2020.

Binesh Sadanandan and Vahid Behzadan. PSF-Med: Measuring and explaining paraphrase sensitivity in medical vision language models. *arXiv preprint arXiv:2602.21428*, 2026.

Lijun Sheng, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. The illusion of progress? a critical look at test-time adaptation for vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 14274–14289, 2022.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

Cheng Tang, Junzhi Ning, Min Cen, Wei Li, Xinyi Zeng, Pinxian Zeng, Rongbin Li, Qiming Zhu, Yuqiang Li, Junjun He, Yirong Chen, and Ming Hu. SIVA-RL: Sensitivity-invariance visual alignment for multimodal reinforcement learning. *arXiv preprint arXiv:2607.13931*, 2026.

Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *International conference on machine learning*, pages 9561–9571. PMLR, 2020.

Shravan Venkatraman, Ritesh Thawkar, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Salman Khan, and Fahad Khan. Paying more attention to visual tokens in self-evolving large multimodal models. *arXiv preprint arXiv:2606.27373*, 2026.

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.

Xin Wang, Yixu Wang, Jiaming Zhang, Ruofan Wang, Jiaqi Yu, Kai Chen, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. TAME: Test-time adversarial prompt tuning via mixture-of-experts for vision-language models. *arXiv preprint arXiv:2605.17577*, 2026.

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.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. ControlMLLM: Training-free visual prompt learning for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 45206–45234, 2024.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *arXiv preprint arXiv:2407.04973*, 2024.

Jinlun Ye, Jiang Liao, Runhe Lai, Xinhua Lu, Jiaxin Zhuang, Zhiyong Gan, and Ruixuan Wang. Ttl: Test-time textual learning for ood detection with pretrained vision-language models. *arXiv preprint arXiv:2604.15756*, 2026.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14625–14634, 2025.

Davis Yoshida and Kevin Gimpel. Reconsidering the past: Optimizing hidden states in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4099–4105. Association for Computational Linguistics, 2021.

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*, pages 9556–9567, 2024.

Marvin Zhang, Sergey Levine, and Chelsea Finn. MEMO: Test time robustness via adaptation and augmentation. In *Advances in Neural Information Processing Systems*, volume 35, pages 38629–38642, 2022.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024.

Xingyu Zhu, Huanshen Wu, Shuo Wang, Beier Zhu, Jiannan Ge, Jiaheng Zhang, and Long Chen. Robustifying vision-language models via test-time prompt adaptation. In *Forty-third International Conference on Machine Learning*, 2026.

Yuhan Zhu, Guozhen Zhang, Chen Xu, Haocheng Shen, Xiaoxin Chen, Gangshan Wu, and Limin Wang. Efficient test-time prompt tuning for vision-language models. *arXiv preprint arXiv:2408.05775*, 2024.

Chengke Zou, Xingang Guo, Rui Yang, Junyu Zhang, Bin Hu, and Huan Zhang. Dynamath: A dynamic visual benchmark for evaluating mathematical reasoning robustness of vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

A Experimental Details

A.1 Data and calibration

We evaluate the full public splits: MMMU validation (900), MathVista testmini (1,000), MathVerse testmini (788 underlying problems in five versions; 3,940 examples), DynaMath sample variant1–10 (501 seed questions with 10 variants; 5,010 examples), LogicVista test (448), and WeMath testmini (1,740). Malformed-input filtering removes no examples. For calibration, we sample 40 examples from each of 30 subjects in the public MMMU test split, giving 1,200 examples with random seed 2026. The same pool is used for S2R and baseline selection. It has no exact overlap with any evaluation split and contains none of DynaMath's 501 underlying seed questions.

A.2 Trajectory-conditioned answer readout

For the Thinking backbone, let r be the reasoning trace produced once by the unedited model. The prefix $r_{\le\tau}$ is empty for $\tau = 0$, ends at the trace midpoint for $\tau = \frac{1}{2}$, and contains the full trace for $\tau = 1$. Every controlled view uses the same detached prefix during repair. Non-Thinking backbones use only the empty prefix. The readout then appends `Final:` before scoring an answer.

For a multiple-choice input, we score the valid option labels after this cue and renormalize their probabilities over the valid options. For a free-form input with fixed candidate set $\mathcal{A}$, the length-normalized sequence score and answer distribution are

$$s_\delta^\tau(a \mid c) = \frac{1}{|a|} \sum_{t=1}^{|a|} \log p_\delta(a_t \mid c, r_{\le\tau}, \texttt{Final:}, a_{<t}), \qquad \pi_\delta^\tau(a \mid c) = \mathrm{softmax}_{a\in\mathcal{A}}\, s_\delta^\tau(a \mid c).$$

The unedited model constructs the candidate set once. Starting with the normalized base answer $\hat{y}_0$, we greedily complete the two highest-probability answer starts at every applicable readout of the original and retained answer-preserving views. We apply the evaluation-time normalization, remove duplicates, retain at most eight candidates in source–readout–rank order, and ensure that $\hat{y}_0$ remains in the set. The result is fixed across repair steps and controlled views.

The scalar differentiated for relevance and support selection is

$$\bar{S}_\delta(c) = \frac{1}{|\mathcal{T}|} \sum_{\tau\in\mathcal{T}} \log \pi_\delta^\tau(\hat{y}_0 \mid c).$$

Thus g_t uses the local sensitivity of $\bar{S}_0((x,q))$, and each scalar support coordinate is ranked by $\rho_t \left|\partial \bar{S}_0((x,q))/\partial H^\ell_{\mathrm{vis},tj}\right|$. All gradients used to construct the relevance regions and hard mask are computed once at $\delta = 0$ and detached before repair.

A.3 Compute and decoding budgets

All methods use a 40,960-token context window and a 4,096-token generation cap. Base decoding, VisRef, the view ensemble, TT-Consistency, TTAug, and all S2R variants and controls use greedy decoding. Self-consistency and Best-of-n sample with temperature 0.7, top-p 0.95, and top-k 20.

Self-consistency and Best-of-n use eight sampled candidates. Verifier reranking generates six candidates and adds one same-backbone verification pass. VisRef uses two refocusing rounds. TT-Consistency uses four answer-preserving branches and at most four LM-head updates (2.7 on average), while TTAug aggregates 16 augmented views. Dense edit, sparse symmetric repair, and

S2R-Full optimize all controlled views and readouts for six updates. S2R-Lite uses stochastic five-branch updates, averaging 3.4 steps under a four-step maximum; default S2R averages 4.0 steps under a six-step maximum.

Latency is measured on one A100-80GB GPU after 200 warm-up requests, using 1,000 synchronized timed requests. Controlled views within an S2R request are batched when shapes permit; independently sampled candidates are decoded sequentially. Peak memory is measured on the same requests.

For S2R, end-to-end latency includes the initial unedited reasoning decode used to construct the trajectory readout, candidate/readout and relevance construction, repair optimization, and a fresh autoregressive regeneration from the repaired visual context. In that order, these stages take $12.8/1.4/3.7/10.2$ s for Lite, $12.8/1.4/7.0/10.1$ s for default S2R, and $12.8/1.4/18.4/9.8$ s for Full. The mean generated lengths before and after repair are 1,100 and 870 tokens, respectively.

For calibration, the candidate budgets are $\{4, 8\}$ for self-consistency and Best-of-n, $\{4, 6, 8\}$ for verifier reranking, $\{1, 2, 3\}$ refocusing rounds for VisRef, $\{2, 4\}$ maximum updates for TT-Consistency, and $\{8, 16\}$ views for TTAug. One configuration per method is selected by mean accuracy over the 30 MMMU calibration subjects, with the smaller inference budget used to break ties; no benchmark-specific tuning is performed. The selected configurations are fixed before evaluation.

A.4 Baseline and control implementation

Self-consistency returns the majority normalized answer across its sampled reasoning traces. Best-of-n ranks sampled candidates by length-normalized log-probability under the same backbone. Verifier reranking adds a deterministic same-backbone pass to select the most evidence-consistent candidate. VisRef iteratively refocuses the visual input.

The forward-only view ensemble averages first-answer-token logits across the original input, retained paraphrases, and low-relevance image views; proposed-evidence views are excluded. TT-Consistency updates only the language-model head using answer-preserving branches and selects a checkpoint by answer consistency. TTAug averages token-level distributions across deterministic text/image augmentations during generation.

The dense control uses the full S2R objective and editable layer band without the sparse mask. The sparse symmetric control keeps the top-K support, optimizer, six-step budget, and edit penalty of S2R-Full, but treats proposed-evidence views as preserving views. Let $\mathcal{C}_{\mathrm{ctrl}}$ contain every paraphrase, low-relevance-region, and proposed-evidence branch. The control minimizes

$$\mathcal{L}_{\mathrm{stab}} = \frac{1}{|\mathcal{C}_{\mathrm{ctrl}}|} \sum_{c \in \mathcal{C}_{\mathrm{ctrl}}} \bar{D}_{\delta}((x, q), c) + \lambda_3 \mathcal{L}_{\mathrm{edit}}. \tag{A.1}$$

This removes the directional proposed-evidence hinge while retaining the support and optimization budget of S2R-Full.

Among probability averaging, logit averaging, majority voting, and confidence weighting for the forward-only ensemble, logit averaging gives the highest calibration accuracy and is used in the main comparison.

All models use the same prompt template:

```
You are given one image and one question.  Solve carefully
using visual evidence.  Return only the final answer after
the tag 'Final:'.  Question:  {q}
```

The stop sequence, answer-extraction expression, and normalization code are also shared; only the tokenizer and chat wrapper vary by model family.

A.5 Example-level visual-dependence audit

Two annotators independently label 200 examples per benchmark for visual dependence. The visual-dependent rate ranges from 74% on MMMU to 96% on DynaMath, with Cohen's κ between 0.78

and 0.88. A third annotator resolves disagreements. This audit does not collect spatial localization labels.

B Controlled View Construction

B.1 Answer-preserving question views

For each (x, q), we generate up to four constrained paraphrases with a deterministic rule-based paraphraser using an ordered library of meaning-preserving edits:

- wh-form alternation, such as "which" $\leftrightarrow$ "what" or "how many" $\leftrightarrow$ "what number of";
- clause reordering that preserves logical scope;
- active–passive alternation for relational questions; and
- synonym replacement from a fixed, question-type-specific lexicon.

For ordinary questions, these operations rewrite the question text itself. For image-only examples with an empty external question field, the same deterministic procedure is applied to fixed answer-preserving task cues. Answer options are never rewritten. We reject a candidate if it changes or removes a number, comparison, negation, named entity, unit, option identifier, relation, or required answer type. For each input, the parser selects three applicable rewrites and retains a fourth when another rule is valid. Thus $m \in \{3, 4\}$, with the original text kept as a separate view. The rules and lexicon are fixed before evaluation and produce the same variants in every run.

B.2 Low-relevance and proposed-evidence image views

Low-relevance views are expected to preserve the answer; proposed-evidence views attenuate the high-relevance region associated with the current prediction. We construct $R = 3$ matched pairs per example by sampling three operator families without replacement.

Table B.1: Operator-strength levels used to construct matched image views.

Operator	Weak / medium / strong
Gaussian blur	$\sigma = 0.8/1.2/1.8$
Patch masking	$10/15/20\%$ region area
Occlusion block	side $= 8/12/16\%$ image width
Additive noise	std $= 0.02/0.04/0.06$
Local contrast	factor $= 0.75/0.55/0.35$
Patch shuffle	grid $= 2\times2/3\times3/4\times4$

The two members of each proposed-evidence–low-relevance pair share the operator and strength tier and have approximately matched perturbed area. All six benchmarks use the families in Table B.1; a keyed random seed selects three families for each example.

C Optimization Details

C.1 Calibrated configuration

Hyperparameters maximize mean accuracy over the 30 MMMU calibration subjects. The selected values are $K = 128$, $\eta = 0.08$, $\epsilon = 0.03$, and $(\lambda_1, \lambda_2, \lambda_3, \gamma, \kappa) = (0.60, 0.80, 0.015, 0.20, 0.10)$. The editable band spans eight middle layers, with global top-K support and top/bottom 25% relevance-mass regions. Default S2R, S2R-Lite, and S2R-Full use $R = 3/2/3$ and maximum update budgets of $6/4/6$, respectively.

Activations use bf16, repair variables use fp32, and experiments use random seeds 2027, 2028, and 2029.

C.2 SINGLE-EXAMPLE PROCEDURE

Algorithm C.1 gives the complete repair loop. The answer anchor and evidence proposal are computed by the unedited model and remain fixed during optimization.

Algorithm C.1 Default stochastic S2R inference for one example

Require: Frozen VLM f_θ, image x, question q, $N = 6$, $K = 128$, $R = 3$
Ensure: Final answer $\hat{a}$
Construct and validate answer-preserving question views $\{q^{(i)}\}_{i=1}^{m}$
Run the unedited model; fix its readout context, $\hat{y}_0$, and the answer candidate set
Compute and detach evidence scores and top-K masks $\{M^\ell\}$
Construct matched sets $\{\tilde{x}_{\text{irr}}^{(r)}, \tilde{x}_{\text{rel}}^{(r)}\}_{r=1}^{R}$
Fix the first paraphrase pair, matched image pair, and readout for checkpoint evaluation
Cache every view immediately below the editable band
Initialize fp32 repair variables $V^\ell \leftarrow 0$
for $n = 1$ to N **do**
  Sample two distinct paraphrases and one matched image pair using the keyed random seed
  Sample one trajectory readout $\tau \in \mathcal{T}$
  Inject $\delta^\ell = M^\ell \odot V^\ell$
  Run the cached suffixes for the sampled views and readout
  Evaluate the answer-level selective-stability objective on the sampled views and readout
  Backpropagate only to $\{V^\ell\}$ and apply the proximal update
  if $n \in \{2, 4, 6\}$ and the stopping criterion holds **then**
    Set $\delta \leftarrow \delta_n$ and **break**
  end if
end for
Discard the fixed readout context and text-side caches
Rebuild the multimodal context with the final repaired visual states
Autoregressively regenerate the reasoning trajectory and final answer $\hat{a}$
return $\hat{a}$

S2R-Lite changes only R to 2 and N to 4. S2R-Full evaluates all controlled views and all trajectory readouts at each update, runs six updates, and disables early stopping. Support construction, optimizer settings, decoding, and answer normalization are otherwise identical.

C.3 ADAPTIVE STOPPING RULE

Before optimization, we fix one checkpoint tuple: the first retained paraphrase pair, the first matched image pair, and the direct readout $\tau = 0$. At $n = 0$ and checkpoints $n \in \{2, 4, 6\}$, we evaluate this tuple independently of the views sampled at step n. Let $\mathcal{L}_{\text{eq}}^{\text{check}}(\delta_n)$, $\mathcal{L}_{\text{irr}}^{\text{check}}(\delta_n)$, and $\mathcal{L}_{\text{rel}}^{\text{check}}(\delta_n)$ denote the three residuals, and define

$$\chi_n = \mathcal{L}_{\text{eq}}^{\text{check}}(\delta_n) + \lambda_1 \mathcal{L}_{\text{irr}}^{\text{check}}(\delta_n) + \lambda_2 \mathcal{L}_{\text{rel}}^{\text{check}}(\delta_n).$$

We stop when

$$0 \le \frac{\chi_{n-2} - \chi_n}{\max(\chi_{n-2}, 10^{-6})} < \epsilon$$

and the top answer under the fixed readout is unchanged from the preceding checkpoint, with $\epsilon = 0.03$. S2R retains the current checkpoint δ_n. Lite checks steps 2 and 4; default S2R also checks step 6.

C.4 ONLINE GATE DETAILS

Gate thresholds are selected on the calibration pool. The first-stage score is

$$s_q = D_{\text{SKL}}\big(\pi_0^0(\cdot \mid x, q), \pi_0^0(\cdot \mid x, q_{\text{gate}})\big) + 0.25\,\text{Ent}\big[\pi_0^0(\cdot \mid x, q)\big].$$

where $q_{\text{gate}} = q^{(1)}$ is the first retained constrained paraphrase, and

$$\zeta_q = \frac{s_q - \mu_q}{\sigma_q}.$$

If $0 \leq \zeta_q \leq 1$, the gate forms one attention-only low-relevance view using the medium patch-masking setting in Table B.1 and computes

$$s_{\mathrm{irr}} = D_{\mathrm{SKL}}\big(\pi_0^0(\cdot \mid x, q), \pi_0^0(\cdot \mid \tilde{x}_{\mathrm{irr}}^{\mathrm{gate}}, q)\big),$$
$$\zeta_{\mathrm{irr}} = \frac{s_{\mathrm{irr}} - \mu_{\mathrm{irr}}}{\sigma_{\mathrm{irr}}}.$$

The calibration pool supplies (μ_q, σ_q) and $(\mu_{\mathrm{irr}}, \sigma_{\mathrm{irr}})$. The gate keeps the base answer when $\zeta_q < 0$, repairs when $\zeta_q > 1$, and otherwise repairs when $(\zeta_q + \zeta_{\mathrm{irr}})/2 > 0.5$. A grid search selects these thresholds to maximize mean accuracy over the 30 subject groups while repairing at most 30% of calibration inputs. The normalization statistics and thresholds are then fixed across benchmarks and random seeds.

C.5 Software, precision, and randomization

Experiments use PyTorch 2.3, CUDA 12.1, and FlashAttention-2 kernels on one NVIDIA A100-80GB GPU. All backbone parameters have `requires_grad=False`. Each stochastic choice is keyed by (random seed, benchmark name, item identifier), making perturbations, view sampling, and stopping repeatable independently of batch order.

C.6 Editable band by backbone

We edit layers 14–21 for both Qwen3-VL models and layers 12–19 for InternVL3.5-8B.

D Additional Results

D.1 Dataset-specific reporting

MathVerse `testmini` contains 788 underlying problems presented in Text Dominant, Text Lite, Vision Intensive, Vision Dominant, and Vision Only versions. The main table reports accuracy over all 3,940 examples.

On DynaMath, average variant / worst-case seed-question accuracy is 72.8 / 41.5 for Base, 75.1 / 44.7 for VisRef, 76.4 / 46.8 for S2R, and 76.8 / 47.5 for S2R-Full. Worst-case accuracy counts a seed question as correct only when all 10 generated variants are answered correctly.

D.2 Transfer across backbones

Table D.1: Per-benchmark transfer to two additional frozen 8B backbones. Values are accuracy (%); Avg is the unweighted benchmark mean.

Backbone	Method	MMMU	MathVista	MathVerse	DynaMath	LogicVista	WeMath	Avg
Qwen3-VL-8B-Instruct	Base	69.8	76.5	72.1	67.3	61.2	64.1	68.5
	VisRef	71.6	78.0	73.6	69.1	62.7	67.4	70.4
	S2R-Full	72.8	79.2	75.1	70.8	64.8	68.1	71.8
InternVL3.5-8B	Base	61.5	70.2	65.0	58.8	53.3	56.4	60.9
	VisRef	63.1	71.7	66.4	60.5	54.9	59.0	62.6
	S2R-Full	64.6	73.0	68.0	62.0	56.8	60.2	64.1

D.3 SUPPORT ROBUSTNESS

Table D.2: Support-quality stress test. Overlap@K is measured against the default hard mask.

Support variant	Avg	Δ vs. default	Overlap@K
Gradient + attention (0.6/0.4)	75.7	0.0	1.00
Gradient only	75.5	-0.2	0.71
Attention only	75.0	-0.7	0.58
Spatial shift: 1 patch	75.2	-0.5	0.49
Spatial shift: 2 patches	74.4	-1.3	0.32
Replace 25% coordinates	75.4	-0.3	0.75
Replace 50% coordinates	74.6	-1.1	0.50
Replace 75% coordinates	73.8	-1.9	0.25
Random support	73.1	-2.6	0.08

Accuracy declines with support overlap. Gradient-only support remains within 0.2 points of the default, while attention-only support loses 0.7; the latter still provides useful spatial context but a weaker coordinate-level signal. With the default hard mask fixed, random equal-area matched regions reach 73.6%, and swapping the proposed-evidence and low-relevance regions reaches 72.9%.

D.4 PER-BENCHMARK CLUSTERED CONFIDENCE INTERVALS

Table D.3: S2R-Full improvement over base with example-clustered confidence intervals. MathVerse clusters its five versions by underlying problem and DynaMath clusters its 10 variants by seed question.

Benchmark	Δ S2R-Full–Base	Example-clustered 95% CI
MMMU	+3.3	[+1.6, +5.3]
MathVista	+3.0	[+1.3, +4.9]
MathVerse	+3.4	[+1.6, +5.0]
DynaMath	+4.0	[+2.1, +5.8]
LogicVista	+3.7	[+0.9, +6.4]
WeMath	+2.5	[+1.3, +4.0]
Macro	+3.3	[+2.5, +4.3]

D.5 BUDGETED RANK-BASED GATING SWEEP

For the controlled budget sweep, examples are ranked in descending order of ζ_q. Let z_{eq}, z_{irr}, and z_{rel} denote the standardized pre-repair values of paraphrase disagreement, low-relevance sensitivity, and the evidence-attenuation gap. The full diagnostic ranks by

$$g_{\mathrm{full}} = \frac{z_{\mathrm{eq}} + z_{\mathrm{irr}} - z_{\mathrm{rel}}}{3},$$

using all retained paraphrases and matched image views. Normalization statistics are fixed on calibration data. This sweep measures ranking quality at matched repair budgets; Table 5 reports the deployable cascade.

Table D.4: Rank-based selective-repair sweep. Accuracy and transition rates are pooled over examples rather than macro-averaged over benchmarks.

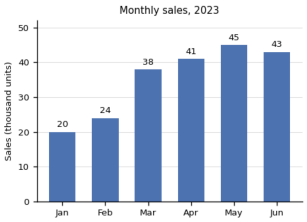
Gate	Repair fraction	Pooled Accuracy	$W \to C$	$C \to W$
None	0%	73.5	0.00%	0.00%
Lightweight	10%	74.9	1.85%	0.45%
Lightweight	25%	75.9	3.00%	0.60%
Lightweight	50%	76.8	4.25%	0.95%
Always-on Full	100%	77.0	5.29%	1.80%
Full diagnostic	10%	75.1	2.05%	0.45%
Full diagnostic	25%	76.2	3.35%	0.65%
Full diagnostic	50%	77.4	4.75%	0.85%

The 50% full-diagnostic policy reaches 77.4% pooled accuracy, compared with 77.0% for always-on S2R-Full. This sweep pools examples, whereas the main tables report macro-averages over benchmarks.

E WORKED EXAMPLE

We trace one evaluation example through the diagnostic, sparse repair, and fresh regeneration. The chart values are Jan–Jun $= 20, 24, 38, 41, 45, 43$; the gold answer is Mar, while the unedited model answers May. The trace is abbreviated.

(1) Input and unedited base run



Input image x (gold answer: Mar)

Prompt (shared by all backbones)

```
[image] You are given one image and one question.  Solve
carefully using visual evidence.  Return only the final answer
after the tag 'Final:'.  Question:  Which month had the largest
increase in sales over the previous month?
```

Base reasoning trace r (computed once and detached)

```
<think> The chart reports Jan 20, Feb 24, Mar 38, Apr 41, May
45, Jun 43.  τ=1/2  May reaches the highest value, 45, so the
largest increase occurs in May.  τ=1  </think> Final:  May
```

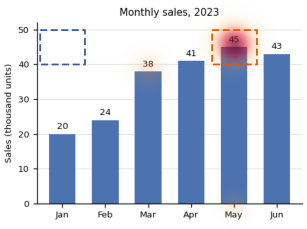
Anchor $\hat{y}_0 =$ May ✗ $\quad \mathcal{A} = \{\text{May}, \text{Mar}, \text{Apr}, \text{Jun}\}$

Online gate $\zeta_q = 0.62 \in [0, 1]$; $\zeta_{\text{irr}} = 1.05$; $(\zeta_q + \zeta_{\text{irr}})/2 = 0.84 > 0.5 \Rightarrow$ **repair**

(2) Answer-preserving question views ($m = 3$ retained)

View	Rule	Question text (edit highlighted)	Base	Repaired
q	original	Which month had the largest increase in sales over the previous month?	May ✗	Mar ✓
$q^{(1)}$	wh-form	What month had the largest increase in sales over the previous month?	Mar	Mar
$q^{(2)}$	clause order	Compared with the previous month, which month had the largest increase in sales?	Mar	Mar
$q^{(3)}$	synonym	Which month had the biggest increase in sales over the previous month?	May	Mar

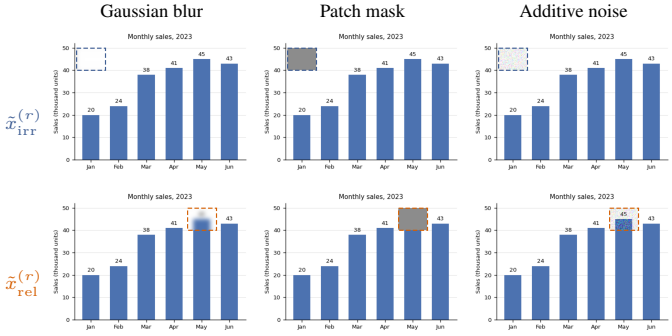
(3) Relevance proposal and matched image views ($R = 3$ pairs)



$\rho_t = 0.6g_t + 0.4a_t$ at $\delta = 0$.

proposed evidence

low relevance



Pair r (before → after)	$\pi(\hat{y}_0 \mid \tilde{x}^{(r)}_{\text{irr}})$	$\pi(\hat{y}_0 \mid \tilde{x}^{(r)}_{\text{rel}})$	d_r
Gaussian blur	.41 → .23	.40 → .20	.037 → .140
Patch mask	.55 → .26	.53 → .23	.037 → .123
Additive noise	.44 → .24	.42 → .21	.047 → .134

Figure E.1: **Worked example: constructing the diagnostic.** The fixed anchor is evaluated under answer-preserving question views and matched image perturbations before repair.

(4) Trajectory-conditioned answer readout ($\mathcal{T} = \{0, \frac{1}{2}, 1\}$)

```
τ=0    [image] ... c <think> </think> Final:  a
τ=½    [image] ... c <think> The chart reports ... Jun 43.  Final:  a
τ=1    [image] ... c <think> ... occurs in May.  </think> Final:  a
```

Each controlled view uses the same detached prefix $r_{\leq\tau}$; free-form answers are scored over the fixed set $\mathcal{A}$.

	Before repair ($\delta = 0$)			After repair (δ_6)		
$\pi_\delta^\tau(a \mid x, q)$	τ=0	τ=$\frac{1}{2}$	τ=1	τ=0	τ=$\frac{1}{2}$	τ=1
May ($\hat{y}_0$)	**.58**	**.63**	**.71**	.24	.27	.36
Mar (gold)	.31	.28	.22	**.69**	**.66**	**.57**
Apr	.07	.05	.04	.04	.04	.04
Jun	.04	.04	.03	.03	.03	.03

(5) Hard-sparse visual-state edit and optimization ($K = 128$, $N = 6$)

Support. Rank coordinates by $\rho_t |\partial \bar{S}_0 / \partial H^{\ell}_{\text{vis},tj}|$ and keep the global top K over layers 14–21. Only their fp32 repair variables are optimized.

Layer	14	15	16	17	18	19	20	21
# coords	9	12	17	21	24	19	15	11

81% of selected coordinates lie on visual tokens inside the proposed-evidence region.

Fixed checkpoint tuple: $(q^{(1)}, q^{(2)})$, blur pair, $\tau = 0$; $\chi_n = L_{\text{eq}} + 0.6L_{\text{irr}} + 0.8L_{\text{rel}}$.

Step	L_{eq}	L_{irr}	L_{rel}	χ_n	rel. drop	Top
0	.214	.183	.063	.374	–	May
2	.121	.102	.031	.207	.447	May
4	.068	.061	.000	.105	.493	Mar
6	.049	.042	.000	.074	.295	Mar

The stopping condition is not met, so S2R retains δ_6. The directional hinge is inactive from step 4 onward.

(6) Regeneration from the repaired visual context

```
<think> Changes from the previous month are Feb +4, Mar +14, Apr +3, May +4, and Jun -2.  The
largest increase is +14, from Feb to Mar.  </think>
Final:  Mar ✓
```

View	q	$q^{(1)}$	$q^{(2)}$	$q^{(3)}$	blur	mask	noise
Base	May	Mar	Mar	May	Mar	May	Mar
Repaired	Mar	Mar	Mar	Mar	Mar	Mar	Mar

Residual before → after	paraphrase disagreement	low-relevance sensitivity	evidence-attenuation gap
	.21 → .05	.18 → .04	.04 → .13

Figure E.1 (continued): Sparse repair and regeneration. The shared visual-state edit changes the answer under every trajectory readout; the final pass discards the base trace and regenerates Mar.