P-GVAR: SELECTIVE VISUAL RE-GROUNDING UNDER IMAGE–TEXT GROUNDING DRIFT

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) can follow misleading textual context even when the image supports a different answer. We study a simple recovery mechanism: querying the same frozen VLM again after removing potentially conflicting context. This fresh visual query repairs many context-induced errors, but unconditional context removal is not generally safe because reliable context can also improve prediction. We therefore formulate recovery as a *selective visual re-grounding* problem with two distinct components: constructing a fresh visual candidate and deciding whether it should replace the original response. We introduce Policy-Guided Visual Answer Refresh (P-GVAR), a training-free response-level policy that uses controlled cross-condition responses to make this retain-or-refresh decision without task labels or access to model internals. Across three open VLMs and four stress benchmarks, selective re-grounding preserves most of the gain of unconditional visual reset, while controlled context experiments show that arbitration becomes important when context can be either helpful or misleading. These findings motivate evaluating candidate quality and routing quality separately when studying recovery from image–text grounding drift.

1 INTRODUCTION

Vision–language models (VLMs) increasingly answer questions in the presence of conversation history, intermediate reasoning, retrieved passages, generated descriptions, and other supplementary text. Such context can be useful, but it can also conflict with the image. When this happens, a VLM may follow textual evidence despite contradictory visual information. Prior work documents hallucination propagation across conversational turns (Zhong et al., 2024), direct image–text conflict (Deng et al., 2025; Zhang et al., 2026a), and failures of apparent visual reconsideration (Shi et al., 2026). We refer to this response-level mismatch as *image–text grounding drift*, without assuming a particular internal causal mechanism.

Our motivation comes from a simple observation. A frozen VLM that answers incorrectly under misleading context can often answer the same question correctly when queried again with the same image after that context is removed. This suggests a simple repair primitive: restore a clean visual evidence state and obtain a fresh answer. Yet context is not always harmful. When it is correct and relevant, removing it can discard useful information. The central problem is therefore not simply whether a model should re-ground on the image, but *when* it should do so.

This leads to the key decomposition studied in this paper. Recovery from grounding drift contains two distinct operations: *candidate construction*, which asks whether a fresh visual query can produce a useful alternative answer, and *candidate selection*, which asks whether that alternative should replace the context-conditioned response. A method can obtain a large gain over the original response even when most of that gain comes from constructing a stronger candidate rather than deciding correctly when to intervene.

We study candidate selection using only completed responses from the same frozen model. Given a context-conditioned answer A and a fresh visual-reset answer B, we ask whether controlled changes to the available evidence can provide a label-free signal for deciding which answer should be retained. This leads to *Policy-Guided Visual Answer Refresh* (P-GVAR). P-GVARuses a text-only diagnostic to test whether the original answer can be reproduced without visual evidence. Its scored extension, P-GVAR-S, additionally uses an independently prompted visual response to corroborate the reset

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.

candidate. Neither method requires additional training, task labels during inference, or access to logits, hidden states, or model parameters.

Our experiments follow the scientific questions created by this formulation. How much recovery already comes from constructing a fresh visual candidate? When does selective intervention matter beyond unconditional reset? Can cross-condition responses identify beneficial versus harmful refreshes when candidate quality is fixed? And which measured gains arise from context removal, prompt wording, response validity, or additional inference compute?

Contributions. Our main contributions are:

- **Selective re-grounding formulation.** We formulate recovery from image–text grounding drift as two distinct problems: constructing a fresh visual candidate and selectively deciding whether that candidate should replace the context-conditioned response.
- **Training-free response-level arbitration.** P-GVAR uses text-only reproducibility as evidence about the context-conditioned answer, while P-GVAR-S adds independent visual corroboration of the reset candidate. Both operate on completed responses from the same frozen checkpoint and require no labels during selection.
- **A controlled characterization of selective recovery.** We show that selective re-grounding preserves most of the recovery obtained by unconditional visual reset under strong conflict, while controlled context, fixed-candidate routing, prompt-factorial, validity, and compute experiments characterize when selective arbitration matters.

2 Related Work

Language overriding visual evidence. MMHalSnowball studies hallucination propagation and introduces Residual Visual Decoding (RVD) to re-emphasize history-free visual evidence (Zhong et al., 2024). Words-or-Vision compares image-only, text-only, and mixed evidence under matching, irrelevant, and corrupted text (Deng et al., 2025), while ConText-VQA studies robustness to textual misinformation (Zhang et al., 2026a). Mechanistic analyses further distinguish whether visual evidence is represented from whether it influences the final response (Nooralahzadeh et al., 2026; Long et al., 2026). Our work instead asks whether completed responses under controlled evidence conditions can support an explicit retain-or-refresh decision.

Inference-time hallucination mitigation. Visual Contrastive Decoding (VCD) contrasts predictions under original and perturbed visual inputs (Leng et al., 2024); Instruction Contrastive Decoding (ICD) contrasts standard and disturbance instructions (Wang et al., 2024); and Residual Visual Decoding (RVD) contrasts history-conditioned and history-free visual evidence (Zhong et al., 2024). Other methods intervene through attention, intermediate representations, or multi-scale visual signals, including OPERA (Huang et al., 2024), DeCo (Wang et al., 2025), SECOND (Park et al., 2025), and more recent internal interventions (Huo et al., 2024; Li et al., 2026a;b; Zhu et al., 2026a; Sun et al., 2025). P-GVAR considers a different access regime: it requires only completed responses from a frozen VLM.

Response-level correction and evaluation. Woodpecker and LURE post-process multimodal outputs (Yin et al., 2024; Zhou et al., 2024); AVCR combines visual checking with contextual refinement (Zhu et al., 2026b); and iterative methods acquire additional visual evidence during reasoning (Zhang et al., 2026b). Recent work also shows that visual reconsideration can differ from verbal rechecking (Shi et al., 2026), and that apparent inference-time gains can reflect decoding artifacts (Yin et al., 2025). These observations motivate our separation of candidate construction from routing quality.

3 Why Selective Visual Re-Grounding?

Let a sample contain image I, question q, and supplementary context h. The ordinary context-conditioned response is

$$A = M_\theta(I, q, h), \tag{1}$$

while a fresh visual-reset response is obtained after removing supplementary context:

$$B = M_\theta(I, q). \tag{2}$$

Across the primary stress suite, A obtains 25.20% mean accuracy while B reaches 50.18%. Moreover, 1,593 of 6,000 evaluations are A-wrong/B-correct. Thus a large part of the apparent recovery is already present in candidate construction.

Observation 1: removing conflicting context is a strong repair primitive. A fresh visual query can recover many failures without changing model parameters or decoding internals. We therefore use unconditional reset as a reference that isolates the benefit of applying this repair primitive on every example.

Reset, however, is not always desirable. Under the released Words-or-Vision context conditions, Qwen2.5-VL reaches 69.82% with matching context compared with 58.15% for the context-free candidate, but the same context-conditioned branch falls to 26.76% under corrupted context. LLaVA-OneVision and InternVL3 show the same qualitative pattern.

Observation 2: context can either help or hurt. Removing supplementary text repairs corrupted-context failures but can also discard useful information. Recovery therefore separates naturally into

$$\underbrace{\text{construct a fresh visual answer}}_{\text{candidate construction}} \quad \neq \quad \underbrace{\text{decide whether to use it}}_{\text{candidate selection}}. \tag{3}$$

4 Policy-Guided Visual Answer Refresh

4.1 Design rationale

Suppose A disagrees with the fresh visual answer B. The disagreement alone cannot identify which candidate should be selected: A may have followed misleading text, but B may also have discarded useful context.

We therefore isolate different evidence sources. A text-only response asks whether the context-conditioned answer can be reproduced without seeing the image. If it reproduces A, then A is compatible with the textual evidence independently of the image. This is a routing signal, not proof that A is wrong. P-GVAR-S asks the complementary question: does an independently prompted visual response support the fresh visual candidate?

4.2 Controlled response branches

For a frozen VLM M_θ, we construct

$$A = M_\theta(\mathcal{P}_A(I, q, h)), \qquad B = M_\theta(\mathcal{P}_B(I, q)), \tag{4}$$

$$C = M_\theta(\mathcal{P}_C(q, h)), \qquad D = M_\theta(\mathcal{P}_D(I_D, q)). \tag{5}$$

A is the context-conditioned candidate, B the visual-reset candidate, C the text-only diagnostic, and D an independently prompted direct-visual response used for corroboration in P-GVAR-S and as a rare fallback when both primary candidates are invalid. All branches share the same frozen checkpoint and decoding configuration.

Let V_X indicate response validity, let T_D indicate that the direct-visual response addresses the intended target, and let $e(X, Y) \in \{0, 1\}$ denote normalized answer equivalence. The core P-GVAR signal is

$$A \not\equiv B \ \wedge \ C \equiv A. \tag{6}$$

Equation 6 identifies disagreements in which the context-conditioned answer remains reproducible from textual evidence alone; it is used as a routing signal rather than as a correctness test.

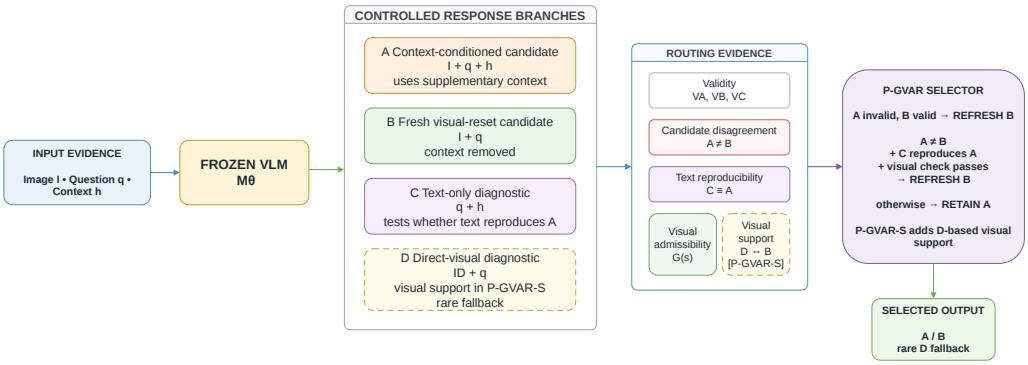

Figure 1: **P-GVAR architecture.** A frozen VLM generates controlled responses under different evidence conditions: the context-conditioned candidate A, visual-reset candidate B, text-only diagnostic C, and direct-visual diagnostic D. P-GVAR uses response validity, A–B disagreement, text-only reproducibility ($C \equiv A$), and visual admissibility to retain A or refresh to B. D provides visual corroboration in P-GVAR-S and serves as a fallback when both primary candidates are invalid. Ground-truth labels are used only for post-selection evaluation.

4.3 Retain-or-refresh policy

Let $G(s)$ denote a fixed visual-applicability check. The original refresh trigger is

$$F_0 = (\neg V_A \wedge V_B) \vee (V_A \wedge V_B \wedge A \not\equiv B \wedge V_C \wedge C \equiv A \wedge G(s)). \tag{7}$$

Under the decision rule in Equation 7, P-GVAR returns B when F_0 is satisfied and retains A otherwise. If both primary candidates are invalid, a valid same-target D can be used as a rare fallback. The complete P-GVAR architecture is summarized in Figure 1.

4.4 Scored visual corroboration

P-GVAR-S augments text-only reproducibility with direct-visual corroboration:

$$c(s) = \mathbf{1}[V_C]\{e(C, A) - e(C, B)\}, \tag{8}$$

$$v(s) = \mathbf{1}[V_D \wedge T_D]\{e(D, B) - e(D, A)\}, \tag{9}$$

$$S(s) = c(s) + v(s). \tag{10}$$

For valid non-equivalent candidates satisfying $G(s)$, P-GVAR-S refreshes when $S(s) > 0$ and otherwise retains A. Since $S \in \{-2, -1, 0, 1, 2\}$, we treat it as an ordinal routing score rather than a calibrated probability.

5 Experimental Setup

Models and benchmarks. We evaluate Qwen2.5-VL-7B (Bai et al., 2025), LLaVA-OneVision-7B (Li et al., 2024), and InternVL3-8B (Zhu et al., 2025). The primary stress suite contains MMHal-Snowball (Zhong et al., 2024), Words-or-Vision (Deng et al., 2025), RHBench, and VisualSwap, with 500 examples per benchmark and model (6,000 stress evaluations per method). MathVista (Lu et al., 2024) is used separately as a clean visual-reasoning control.

Baselines and reset control. We compare against Vanilla, Rule-GVAR, Visual Contrastive Decoding (VCD) (Leng et al., 2024), Instruction Contrastive Decoding (ICD) (Wang et al., 2024), Residual Visual Decoding (RVD) (Zhong et al., 2024), DeCo (Wang et al., 2025), and SECOND (Park et al., 2025). We additionally report unconditional visual reset (B-only) as a reference control that always returns the fresh visual candidate, isolating the effect of applying the reset intervention on every example. Published decoding baselines are common-backbone implementations of their equations rather than official-repository executions; details are in Appendix A.

Table 1: **Primary stress accuracy (%).** Each model–benchmark cell contains 500 examples. B-only always returns the fresh visual-reset candidate. Bold denotes the highest executed value including ties.

Model	Benchmark	Van.	B-only	VCD	ICD	RVD	DeCo	SECOND	Rule	P-GVAR
Qwen2.5-VL	MMHalSnowball	12.4	**73.4**	19.2	11.6	70.6	12.2	13.6	72.6	72.6
Qwen2.5-VL	RHBench	28.8	**46.0**	39.0	28.0	43.4	39.8	36.2	28.6	45.8
Qwen2.5-VL	VisualSwap	33.0	30.8	32.4	32.8	**34.2**	33.4	32.2	32.6	32.8
Qwen2.5-VL	Words-or-Vision	26.6	**57.8**	30.0	24.8	56.8	26.4	27.8	55.2	55.2
LLaVA-OV	MMHalSnowball	17.2	**77.6**	28.4	21.4	75.8	17.2	19.2	76.0	76.0
LLaVA-OV	RHBench	39.0	**39.4**	38.0	38.8	38.6	39.0	39.0	**39.4**	**39.4**
LLaVA-OV	VisualSwap	22.4	22.2	21.6	22.2	**22.6**	22.4	21.6	22.4	**22.6**
LLaVA-OV	Words-or-Vision	23.2	**44.0**	25.8	28.6	43.4	23.2	23.8	42.2	42.4
InternVL3	MMHalSnowball	11.6	**75.0**	17.8	10.8	67.4	11.4	14.2	73.8	73.8
InternVL3	RHBench	24.4	**45.8**	32.8	22.8	37.2	29.4	32.2	24.8	43.6
InternVL3	VisualSwap	25.6	**33.0**	22.8	21.0	25.8	24.6	24.8	25.4	32.0
InternVL3	Words-or-Vision	38.2	**57.2**	39.4	36.8	56.8	38.8	40.6	53.0	53.4
Mean stress		25.20	**50.18**	28.93	24.97	47.72	26.48	27.10	45.50	49.13

Metrics and decoding. We report task accuracy, recovery of A-wrong/B-correct cases, harmful switching on A-correct/B-wrong cases, and average precision (AP) for ranking beneficial refreshes among eligible disagreements, where beneficial refreshes are relatively sparse. Primary paired comparisons use bootstrap confidence intervals and exact McNemar tests. Primary decoding is greedy with a 64-token output limit.

6 Results

6.1 Fresh visual re-grounding explains most of the robustness gain

Table 1 reports the complete primary stress comparison. P-GVAR raises mean accuracy from 25.20% to 49.13%, an improvement of 23.93 percentage points over Vanilla, and improves 11 of 12 model–benchmark cells.

Among our common-backbone implementations of the published decoding baselines, RVD (Zhong et al., 2024) is strongest at 47.72% mean accuracy; P-GVAR exceeds it by 1.42 points on average. Across the 12 stress cells, P-GVAR records 11 wins and one loss against Vanilla; nine improvements remain significant after Holm correction and none is significantly negative. Against RVD, P-GVAR records seven wins, one tie, and four losses. These paired results indicate that the recovery effect is broad across model–benchmark combinations rather than being driven by a single benchmark.

Unconditional reset provides a deliberately aggressive reference: B-only applies the fresh visual query on every example and reaches 50.18%, compared with 49.13% for P-GVAR. The small gap is consistent with a stress suite dominated by misleading context, where removing that context is usually beneficial. Importantly, B-only does not solve the selection problem—it removes context regardless of whether that context is useful. This distinction supports the first central claim of the paper: *candidate construction must be separated from candidate selection*. A clean visual re-query establishes a strong recovery primitive, while selective re-grounding asks when that primitive should actually be used.

6.2 The primary stress suite strongly favors reset

The mean results hide an important benchmark-level difference: some datasets contain many opportunities for visual reset, while others contain almost none. Table 2 aggregates the original P-GVAR action diagnostics over the three backbones. MMHalSnowball is a high-recoverability regime: 62.3% of examples have A wrong and B correct, and P-GVAR realizes 60.9% of those cases. VisualSwap is the opposite regime, with only 4.4% recoverable examples. This explains why large gains concentrate on conflict-heavy benchmarks and why a small harmful-switch rate matters much more when recovery opportunities are scarce.

Table 2: **Where recovery is available.** Benchmark-level P-GVAR action diagnostics (%), averaged over the three backbones. "Recoverable" denotes A-wrong/B-correct cases.

Benchmark	Refresh	Recoverable	Recovered	Harmful	Invalid
MMHalSnowball	74.0	62.3	60.9	0.5	0.0
RHBench	30.9	13.7	12.5	0.3	1.0
VisualSwap	12.3	4.4	2.8	0.8	2.0
Words-or-Vision	39.2	25.8	21.3	0.3	4.2

We next pool all 6,000 candidate pairs to ask how much selection headroom remains after B has been constructed. There are 1,593 A-wrong/B-correct cases that can be repaired by reset, but only 94 A-correct/B-wrong cases in which retention is preferable. An oracle restricted to A/B can therefore improve B-only by at most $94/6000 = 1.57$ percentage points. Moreover, 48.25% of examples are wrong under both candidates and are unreachable by any A/B selector. Thus B-only is already close to the A/B accuracy ceiling on this conflict-heavy distribution. The relevant question for selective re-grounding is therefore how much of this reset benefit can be preserved while retaining the ability not to intervene when the original context is useful.

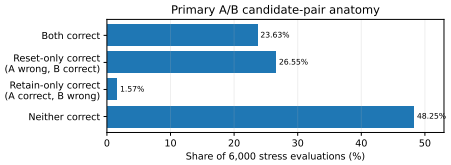

Figure 2: **Candidate-pair anatomy.** The stress suite contains far more reset-only than retain-only correct cases, placing unconditional reset close to the A/B accuracy ceiling.

Selector	Acc.	Δ_B	AP	Recovery	Harm
B-only	50.18	0.00	–	1593/1593	94/94
P-GVAR	49.13	-1.05	.677	1463/1593	29/94
C-margin	49.12	-1.07	.680	1462/1593	29/94
Visual-only	49.68	-0.50	.648	1523/1593	56/94
P-GVAR-S	49.68	-0.50	**.721**	1514/1593	47/94

Table 3: **Fixed-candidate routing.** AP is measured over eligible A/B disagreements; recovery and harm use their respective candidate-pair denominators.

The asymmetry in Figure 2 explains why final accuracy strongly favors unconditional reset. Every missed refresh can sacrifice one of the many A-wrong/B-correct opportunities, whereas retention can improve accuracy only in the much smaller A-correct/B-wrong region. Nevertheless, P-GVAR recovers 1,463 of the 1,593 reset-favorable cases while avoiding 65 of the 94 harmful resets incurred by B-only, leaving final accuracy only 1.05 points lower. Thus, on strongly corrupted distributions, recovery coverage and harmful intervention should be interpreted jointly with final accuracy. The value of arbitration becomes clearer when the evaluation contains enough useful context for retention to matter.

6.3 Cross-condition diagnostics improve routing discrimination

Holding the candidate pair fixed isolates routing quality from candidate construction. P-GVAR-S obtains the highest AP, 0.721, compared with 0.677 for P-GVAR, 0.680 for the text-only margin, and 0.648 for visual-only routing (Table 3). It recovers 1,514 of 1,593 available corrections while making 47 of 94 possible harmful switches.

P-GVAR-S reaches 49.68%, within 0.50 points of unconditional reset, while obtaining the highest routing AP. Relative to visual-only routing, P-GVAR-S avoids nine additional harmful switches while giving up nine recoveries, leaving their final accuracy tied. This illustrates why endpoint accuracy alone is insufficient for evaluating selection on a reset-dominated distribution: the combined signal improves discrimination even when the operating-point accuracy changes little.

6.4 CONTEXT CAN BE EITHER USEFUL OR HARMFUL

The primary stress suite is intentionally conflict-heavy. To test the regime in which retention can matter, we evaluate the matching, irrelevant, and corrupted context conditions released with Words-or-Vision. Matching context increases A to 66.8–69.8% across the three VLMs, whereas corrupted context sharply degrades the same branch. This is precisely the retain-or-refresh tension that unconditional reset cannot resolve.

Model	Context	A	B	P0	P1	P1−B (95% CI)
Qwen2.5-VL	Matching	69.82	58.15	58.95	60.16	+2.01 [0.00, 4.02]
	Irrelevant	57.75	58.15	57.75	57.95	-0.20 [-1.61, 1.21]
	Corrupted	26.76	58.15	55.53	56.94	-1.21 [-2.41, 0.00]
LLaVA-OV	Matching	67.40	52.72	54.12	54.12	+1.41 [-0.40, 3.22]
	Irrelevant	52.92	52.72	53.52	52.92	+0.20 [-1.21, 1.61]
	Corrupted	23.94	52.72	50.10	51.51	-1.21 [-2.41, -0.20]
InternVL3	Matching	66.80	57.55	56.74	56.94	-0.60 [-2.82, 1.61]
	Irrelevant	53.52	57.55	53.72	55.33	-2.21 [-4.43, 0.00]
	Corrupted	38.63	57.55	53.72	56.14	-1.41 [-3.22, 0.40]

Table 4: **Controlled context conditions.** Accuracy (%) on 497 complete Words-or-Vision triplets per model. A uses supplementary context and B removes it; P0 and P1 denote P-GVAR and P-GVAR-S.

The controlled conflict-load experiment provides the complementary direction. As the number of misleading context slots rises from $k = 0$ to $k = 4$, A deteriorates sharply while the visually re-grounded branches remain comparatively stable. To keep the two trends directly comparable without consuming a full page, we place them side by side.

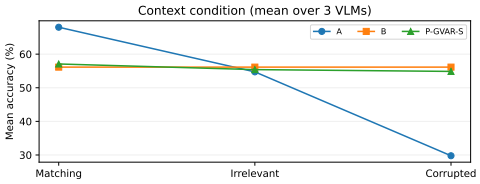

Figure 3: **Context reliability.** Matching context helps A, whereas corrupted context strongly degrades it.

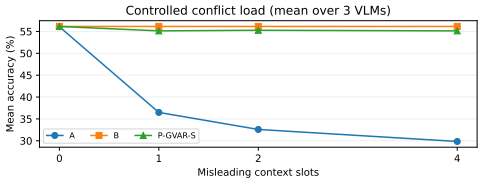

Figure 4: **Repeated conflict.** Increasing misleading-context load progressively degrades A while re-grounded responses remain stable.

Together, Table 4 and Figures 3–4 show why selective re-grounding is a genuine decision problem. The reset candidate is intentionally invariant to h, so its stability across context conditions is expected; the informative quantity is how sharply the context-conditioned branch moves around that reference. Matching context can place A well above B, whereas corrupted or repeated misleading context drives it far below B. Thus "always retain" and "always reset" are fixed policies whose relative advantage reverses with context reliability, and a useful selector must distinguish these regimes without access to ground truth.

6.5 WHAT ACTUALLY DRIVES THE IMPROVEMENT?

A response-level intervention can appear effective for reasons unrelated to the intended routing signal. We therefore isolate prompt wording, response validity, and alternative allocations of additional inference.

Prompt versus context removal. A matched 2×2 factorial independently varies context presence and the reset instruction (Table 5). A, A^+, B_0, and B obtain 32.0%, 45.0%, 54.0%, and 57.0%,

Table 7: **Independent MathVista clean accuracy (%).** The effect of re-grounding is model-dependent rather than uniformly beneficial.

Model	Vanilla	B-only	Rule	P-GVAR	Oracle A/B
Qwen2.5-VL	63.2	63.6	63.2	64.4	65.2
LLaVA-OV	51.2	51.8	50.8	50.8	53.0
InternVL3	50.0	59.6	50.2	60.4	61.8

respectively. Averaging across prompt conditions, context removal contributes +17.0 points; averaging across context conditions, the reset directive contributes +8.0 points. Removing conflicting context is therefore the larger factor.

Validity handling. Matched 256/512-token RHBench runs reduce truncation substantially for Qwen and InternVL. P-GVAR-S differs from the validity-only policy by only -0.4 to $+0.4$ points across the six tested cells, showing that performance under matched output budgets is largely explained by validity handling. The complete matched-budget table is moved to Appendix D.

Alternative test-time computation. At a common 256-token cap, P-GVAR-S reaches 60.0%, compared with 58.0% for one-call B-only, 59.0% for four-sample reset voting, and 59.5% for direct A/B verification (Table 6). One-call reset remains substantially cheaper, but under comparable additional inference P-GVAR-S reaches the highest accuracy among the tested allocations. This suggests that the diagnostic calls provide information beyond simple resampling of the reset candidate.

Candidate	Context	Reset prompt	Acc.
A	yes	no	32.0
A^+	yes	yes	45.0
B_0	no	no	54.0
B	no	yes	57.0

Table 5: **Prompt-factorial candidate construction.** Context removal contributes more than reset-prompt wording in this matched control.

Method	Acc.	Calls	Sec.
B-only	58.0	1.00	0.80
Vote-B-4	59.0	4.00	3.35
Verify-AB cond.	59.5	2.50	2.05
P-GVAR cond.	57.5	2.50	2.02
P-GVAR-S cond.	**60.0**	2.98	2.40

Table 6: **Alternative inference allocations.** P-GVAR-S attains the highest accuracy in this matched control, while one-call B-only remains substantially cheaper.

6.6 Clean behavior is model-dependent

MathVista provides a complementary clean visual-reasoning control in which the central concern is unnecessary intervention rather than recovery from conflicting text. Table 7 shows that the effect is model-dependent: P-GVAR changes Qwen2.5-VL from 63.2% to 64.4%, LLaVA-OneVision from 51.2% to 50.8%, and InternVL3 from 50.0% to 60.4%. B-only shows a similarly model-dependent pattern. We therefore do not claim uniform clean retention; instead, this control reinforces the paper's broader point that the usefulness of refresh depends on the candidate regime.

7 Discussion

Fresh visual reset is the dominant recovery mechanism under strong conflict: the primary suite contains 1,593 reset-only recoveries but only 94 cases in which retaining A is uniquely beneficial. This asymmetry places unconditional reset close to the A/B accuracy ceiling, making it an aggressive reference for candidate construction rather than the primary objective of selective routing.

Selection addresses a complementary failure mode. The controlled context experiments show that correct context can materially improve prediction while corrupted context can sharply degrade it. P-GVAR-S consequently should be interpreted as a response-level arbitration mechanism whose value depends on the mixture of helpful and harmful context.

Routing AP and final accuracy should also be kept conceptually separate. P-GVAR-S provides the strongest tested routing discrimination while remaining within 0.50 points of unconditional reset on

the conflict-heavy stress distribution. Because this distribution offers little retention headroom, small endpoint differences can mask meaningful changes in which interventions are beneficial or harmful. Finally, nearly half of the stress cases are wrong under both A and B, so improving candidate construction remains a larger bottleneck than selection alone.

8 Limitations and Future Work

Benchmark balance and candidate coverage. The primary suite is conflict-heavy, placing unconditional reset near the A/B ceiling and therefore offering limited sensitivity to improvements in retention. Moreover, 48.25% of examples are wrong under both A and B. Future evaluation should include realistic mixtures of helpful, irrelevant, stale, and adversarial context, while future methods should generate richer grounded candidate sets.

Calibration and response processing. P-GVAR-S uses an ordinal, uncalibrated routing score and depends on response validity, normalization, and answer equivalence. Held-out calibration and stronger semantic equivalence models are natural extensions. The controlled matching-context results also leave substantial room for improving retention when supplementary context is genuinely informative.

Scope and cost. We evaluate three open 7–8B VLMs and tasks that remain interpretable after supplementary context removal. Long multimodal dialogue, context-dependent anaphora, larger proprietary models, and agentic settings remain untested. P-GVAR also requires more inference than one-call reset, motivating adaptive execution that invokes diagnostics only when arbitration is necessary.

No internal causal claim. Text-only reproducibility and visual corroboration are response-level signals, not proofs of internal causal grounding. Connecting selective re-grounding to activation-level or causal analyses remains future work.

9 Conclusion

We studied image–text grounding drift as a selective visual re-grounding problem. A fresh visual query after removing supplementary context can recover many context-induced errors, making candidate construction the dominant source of improvement under strong conflict. However, correct context can also improve prediction, so unconditional reset is a strong recovery reference rather than a universally desirable policy.

P-GVAR makes this distinction explicit by separating fresh candidate construction from the decision to retain or replace the original response. Its scored extension combines text-only reproducibility with direct visual corroboration to improve routing discrimination when candidate quality is fixed. The resulting picture is therefore not simply that visual reset is better than contextual answering: visual reset is a strong repair primitive, while selective arbitration determines when that primitive should be applied.

Reproducibility Statement

The evaluation uses fixed sample identifiers and deterministic primary decoding. The appendix provides the complete routing procedure together with action diagnostics, controlled-conflict results, matched output-budget controls, conditional execution details, and representative decision traces.

AI Use Statement

Generative AI tools were used for language editing, code debugging, and manuscript preparation. All scientific claims, experimental choices, implementations, evaluations, and reported conclusions were reviewed by the authors. Generative AI tools were not used to construct ground-truth labels.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876, 2025.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024. URL https://arxiv.org/abs/2408.02032.

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*, pp. 13872–13882, 2024. URL https://arxiv.org/abs/2311.16922.

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

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, Ming Li, Xiaonan Luo, and Guanbin Li. Cross-modal attention calibration for LVLM hallucination mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 40186–40196, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Cross-Modal_Attention_Calibration_for_LVLM_Hallucination_Mitigation_CVPR_2026_paper.html.

Xinwei Li, Li Lin, Hui Jiao, Li Yao, Tien-Tsin Wong, and Hanqian Wu. DiVE: Decoupling intra-layer visual evidence for mitigating hallucinations in large vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37552–37568, San Diego, California, United States, 2026b. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1742. URL https://aclanthology.org/2026.acl-long.1742/.

Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. Understanding language prior of LVLMs by contrasting chain-of-embedding. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/98b89a6e3e53eb97960fbd8d3bff37be-Abstract-Conference.html.

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*, 2024.

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. URL https://arxiv.org/abs/2604.09364.

Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. SECOND: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 48027–48040. PMLR, 2025. URL https://proceedings.mlr.press/v267/park25c.html.

Chufan Shi, Cheng Yang, Yaokang Wu, Linhao Jin, Bo Shui, Taylor Berg-Kirkpatrick, and Xuezhe Ma. Are VLMs seeing or just saying? uncovering the illusion of visual re-examination. *arXiv preprint arXiv:2605.15864*, 2026. URL https://arxiv.org/abs/2605.15864.

Nan Sun, Zhenyu Zhang, Xixun Lin, Kun Wang, Yanmin Shang, Naibin Gu, Shuohuan Wang, Yu Sun, Hua Wu, Haifeng Wang, and Yanan Cao. V-ITI: Mitigating hallucinations in multimodal large language models via visual inference-time intervention. *arXiv preprint arXiv:2512.03542*, 2025. URL https://arxiv.org/abs/2512.03542.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4z3IguA4Zg.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024. URL https://arxiv.org/abs/2403.18715.

Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in MLLMs? In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2f89a23a19d1617e7fb16d4f7a049ce2-Abstract-Conference.html.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2024. URL https://arxiv.org/abs/2310.16045.

Chi Zhang, Wenxuan Ding, Jiale Liu, Mingrui Wu, Qingyun Wu, and Ray Mooney. Do images speak louder than words? investigating the effect of textual misinformation in VLMs. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6872–6895, Rabat, Morocco, 2026a. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.323. URL https://aclanthology.org/2026.eacl-long.323/.

Yongchang Zhang, Oliver Ma, Tianyi Liu, Guangquan Zhou, and Yang Chen. See it, say it, sorted: An iterative training-free framework for visually-grounded multimodal reasoning in LVLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11933–11942, 2026b. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_See_It_Say_It_Sorted_An_Iterative_Training-Free_Framework_for_CVPR_2026_paper.html.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11991–12011, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.648. URL https://aclanthology.org/2024.acl-long.648/.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2024. URL https://arxiv.org/abs/2310.00754.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. In *International Conference on Learning Representations*, 2026a. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/83749a89d03df30c66af88020e4c7892-Abstract-Conference.html.

Zhihao Zhu, Jiafeng Liang, Shixin Jiang, Jinlan Fu, Ming Liu, Guanglu Sun, See-Kiong Ng, and Bing Qin. Analyzing reasoning consistency in large multimodal models under cross-modal conflicts. *arXiv preprint arXiv:2601.04073*, 2026b. URL https://arxiv.org/abs/2601.04073.

A Implementation Details and Exact Routing Procedure

All response branches use the same frozen checkpoint, image preprocessing, answer options when present, output format, and primary deterministic decoding. A receives (I, q, h); B receives (I, q) with the visual-reset instruction; C receives (q, h) without an image; and D receives the target image and question with an independent direct-visual instruction. Ground truth is unavailable during generation, parsing, validity checks, answer-equivalence tests, and policy selection.

Ordered P-GVAR procedure.

```
Generate A and B

if not valid(A) and valid(B):
    return B

if valid(A) and not valid(B):
    return A

if not valid(A) and not valid(B):
    generate/check D
    return D if valid(D) and same_target(D)
    else INVALID

if equivalent(A, B):
    return A

generate/check C

if valid(C)
   and equivalent(C, A)
   and visual_admissible(sample):
    return B

return A
```

P-GVAR-S replaces the final binary routing condition on admissible disagreements with $S(s) = c(s) + v(s) > 0$.

B Complete Action Diagnostics

Table 8 reports the full model–benchmark action breakdown.

Table 8: Complete P-GVAR action diagnostics on the four stress benchmarks. Rates are percentages of 500 examples per row.

Model	Benchmark	Refresh	Recoverable	Recovered	Harmful	Invalid
Qwen	MMHalSnowball	76.4	62.4	61.4	1.2	0.0
Qwen	RHBench	42.2	17.6	17.4	0.4	0.8
Qwen	VisualSwap	6.8	2.2	1.0	1.2	0.8
Qwen	Words-or-Vision	44.2	32.0	28.6	0.0	3.0
LLaVA	MMHalSnowball	70.2	61.0	59.0	0.2	0.0
LLaVA	RHBench	2.6	1.2	0.6	0.2	0.4
LLaVA	VisualSwap	3.4	1.2	0.4	0.4	3.4
LLaVA	Words-or-Vision	42.2	24.0	19.6	0.4	3.8
InternVL	MMHalSnowball	75.4	63.6	62.4	0.2	0.0
InternVL	RHBench	47.8	22.2	19.4	0.2	1.8
InternVL	VisualSwap	26.6	9.8	7.0	0.8	1.8
InternVL	Words-or-Vision	31.2	21.4	15.8	0.6	5.8

Across the 12 stress cells, P-GVAR exceeds Vanilla by 23.93 percentage points on average, with 11 wins and one loss. Nine of the twelve comparisons against Vanilla are Holm-significant positive and

none is Holm-significant negative. Relative to RVD (Zhong et al., 2024), P-GVAR records seven wins, one tie, and four losses.

C Complete Controlled-Conflict Results

Table 9 gives the complete model-wise conflict-load results underlying Figures 3–4.

Table 9: Complete controlled-conflict load results. Accuracy is in percent.

Model / k	A	B	P0	P1	P1−B
Qwen / 0	58.55	58.15	58.55	58.55	+0.40
Qwen / 1	38.83	58.15	56.14	56.74	-1.41
Qwen / 2	35.41	58.15	56.54	56.94	-1.21
Qwen / 4	32.80	58.15	55.73	57.34	-0.80
LLaVA / 0	53.32	52.72	53.52	52.31	-0.40
LLaVA / 1	30.78	52.72	51.71	52.31	-0.40
LLaVA / 2	25.55	52.72	50.30	51.71	-1.01
LLaVA / 4	23.34	52.72	51.71	52.11	-0.60
InternVL / 0	56.34	57.55	56.34	57.55	0.00
InternVL / 1	39.84	57.55	53.12	56.34	-1.21
InternVL / 2	36.82	57.55	53.32	57.14	-0.40
InternVL / 4	33.40	57.55	52.92	55.94	-1.61

D Matched Output-Budget Controls

Table 10 reports the complete matched output-budget control discussed in Section 6.5.

Table 10: RHBench matched-budget control. Accuracy and truncation are percentages. Δ_V denotes P1 minus the validity-only policy.

Model/cap	A	B	Valid.	P0	P1	A tr.	B tr.	Δ_V
Qwen/256	35.8	40.6	42.8	42.2	42.8	25.6	16.8	0.0
Qwen/512	41.4	42.8	43.6	43.2	43.8	10.4	7.6	+0.2
LLaVA/256	42.6	41.8	42.6	43.0	42.2	0.4	0.2	-0.4
LLaVA/512	42.6	41.8	42.6	43.0	42.2	0.0	0.2	-0.4
InternVL/256	41.4	42.8	44.0	43.4	44.4	16.8	13.4	+0.4
InternVL/512	44.4	45.0	45.4	44.2	45.6	6.6	4.4	+0.2

E Inference Cost and Conditional Execution

Conditional P-GVAR-S first generates A/B and requests C/D only when the diagnostics can affect the decision. Ignoring retries,

$$\mathbb{E}[N] = 2 + 2q + b, \tag{11}$$

where q is the admissible-disagreement rate and

$$b = \Pr(\neg V_A \wedge \neg V_B). \tag{12}$$

Four-sample reset voting uses temperature 0.7 followed by majority voting over normalized answers. Direct verification receives the image, supplementary context, and order-balanced A/B candidates and must return one of the two candidates. Generation count is not treated as equivalent computational cost; latency is measured separately.

F Representative Decision Traces

Table 11 gives representative successful, harmful, retained, and fallback decisions.

These examples reinforce that $C \equiv A$ indicates textual reproducibility rather than proof that A is incorrect. Likewise, agreement between D and B provides visual corroboration rather than a correctness guarantee.

Table 11: Representative stored decision traces. Gold labels are shown only after routing.

Case	A	B	C	D	Action	Gold	Outcome
RHBench counting	invalid	2	0	2	refresh B	2	correct
RHBench backlight	B	A	B	B	refresh B	B	harmful
MathVista clean	8	8	0	8	retain A	8	correct
VisualSwap 2619	invalid	invalid	invalid	C	fallback D	C	correct