arXiv:2608.01207v1 [cs.CV] 2 Aug 2026

It's the Decoding Format, Not the Perturbation: Auditing Consistency-Based Selection for Vision-Language Test-Time Scaling

Puzhuo Zheng, Hasan Kurban

Hamad Bin Khalifa University
{puzh90198, hkurban}@hbku.edu.qa

Abstract

Test-time scaling lifts large language model reasoning by sampling many candidate solutions and selecting among them, yet the same recipe transfers poorly to vision-language models (VLMs): recent work shows that simple majority voting beats selection methods built on the model's own self-verification, apparently because at the selection layer an image-grounded answer and a confident guess from the language prior look the same. A natural fix is to make the selection signal one that cannot be computed without the image. We study Perturbation-Grounded Selection (PGS), a label-free, training-free rule that scores each candidate by whether the model re-derives it under label-preserving perturbations of the input (cropping, background masking, mild photometric or geometric jitter); PGS recovers majority voting when the perturbation set is empty. The decisive question is not whether PGS beats chain-of-thought-only majority voting, but whether the perturbation term adds anything once decoding format and budget are controlled. We therefore introduce a *format-matched* control (MatchedCtrl): the same short, no-CoT draws spent on the *original* image. Across TextVQA, MATH-Vision, MMMU, and ViLP, with a Qwen headline (three-seed means) and LLaVA-OneVision coverage in matched-budget selector tables, PGS appears to beat plain majority voting by up to $+31.8$ points on TextVQA (Qwen), but MatchedCtrl tracks or exceeds PGS within noise on every benchmark, including the vision-required ViLP; no Qwen category shows a significant gain over this control. The preserve/destroy stability gap is real and image-dependent (up to $+0.48$), yet does not predict per-instance wins. The result is negative and diagnostic: perturbation consistency is at best a partial diagnostic of visual dependence and, on its own, not a usable selection signal once format is controlled; gains reported against CoT-only majority voting overstate such methods. We release code and audit tooling.

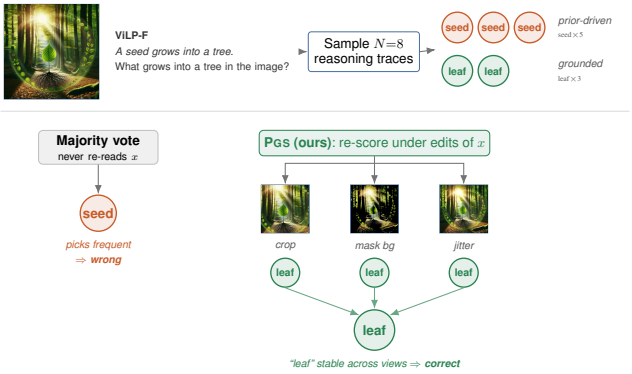

Figure 1: PGS on Real ViLP-F example (Qwen2.5-VL-7B). A VLM samples $N=8$ traces for an image–question pair whose prompt plants a language prior (*"A seed grows into a tree"*) while the image-grounded answer is "leaf". The prior guess "seed" is most *frequent* (5 vs. 3), so **majority voting picks it and is wrong**—it never re-reads the image. PGS (ours) instead re-scores each candidate by adding re-rendering x under label-preserving edits (center crop, background mask, photometric jitter): "leaf" stays modal across views. So PGS might selects the stable, image-grounded answer. Majority voting is the empty-perturbation special case of PGS (Section 3).

1 Introduction

Sampling many candidate solutions at inference and selecting a good one (test-time scaling) has become one of the most reliable ways to improve reasoning in large language models (LLMs) (Wei et al. 2022; Wang et al. 2023; Snell et al. 2024; Brown et al. 2024). The recipe assumes a usable *selection signal*: either many samples agree on the right answer (self-consistency / majority voting) or the model can verify its own candidates well enough to pick the best one (best-of-N with self-verification).

This recipe transfers poorly to vision-language models, and recent work converges on one cause. For RL-tuned VLMs on visual math, majority voting beats verification-centric selection and the self-correction "aha moment" yields no reliable gain (Wu et al. 2026a). VLMs often attend to the correct evidence yet still answer wrongly ("seeing but not believing" (Liu et al. 2025)) and can hold a fixed answer while

their internal representation drifts under label-preserving edits (Wani et al. 2026), so grounded and ungrounded answers are indistinguishable at the output level and output stability is not evidence of grounding. Self-consistency assumes correctness correlates with answer frequency, which fails when spurious paths dominate (Chung et al. 2025); and across seven VLMs, single-model majority voting yields only modest, chain-of-thought-dependent gains that vanish once outputs are correlated (Tong et al. 2026). The common thread: the selection layer cannot tell an image-grounded answer from a confident guess, because the usual signals (frequency, verbalized confidence) never test whether the answer depends on the pixels. The selection layer, not the generation layer, is where vision-language test-time scaling leaks.

The field's responses have largely gone in two directions, both of which leave the core problem untouched for a practitioner with modest resources. One trains the deficiency away with reinforcement learning or tool-use curricula so the model learns when to re-examine visual evidence (Wu et al. 2026b; Zhu et al. 2026; Marsili and Gkioxari 2025); this needs training compute and data many groups do not have. The other intervenes inside a *single* generation, for example by masking deep-layer attention toward evidence regions (Liu et al. 2025) or adding an external verifier or process reward model that scores candidates (Marsili and Gkioxari 2025); these need access to model internals, a second trained network, or supervision. Neither gives a label-free, training-free fix at the *selection* layer that runs on a single consumer GPU, which is precisely the regime in which test-time scaling is most attractive.

We turn this diagnosis into a design constraint: *if the usual selection signals fail because they never test whether an answer depends on the image, then a corrective signal must be one that cannot be computed without the image.* A natural candidate, which we call Perturbation-Grounded Selection (PGS), scores each of the N sampled answers by re-rendering the visual input under a set of *label-preserving perturbations*, transformations that should not change the correct answer to a genuinely image-grounded question (cropping toward question-relevant regions, masking plausibly irrelevant background, mild photometric or geometric jitter), and measuring how stable the model's support for that answer is across these views. The intuition is that an answer re-derived from many perturbed views is grounded in the pixels, whereas an answer that survives only on the original view is the fingerprint of a language-prior guess. Majority voting is the empty-perturbation special case of PGS, so this is a strict generalization of the standard rule rather than a competing heuristic (Section 3).

PGS inherits none of the costs that make trained fixes inaccessible (no training, labels, reward model, or second network, at a constant multiple of ordinary best-of-N cost), so if a purely inference-time grounding signal can improve selection, PGS is where it should show. We therefore ask a sharper question than "does PGS beat majority voting?": *does the perturbation signal add anything once the comparison is fair?*

The comparison is not fair by default, and this is the crux. PGS's perturbation-side draws are *short, answer-only* samples with no chain of thought, whereas plain majority voting aggregates only the N long CoT samples, so any gain of PGS over MV conflates the perturbation reweighting we care about with a switch in decoding format (CoT→short) that we do not, and that format is not incidental for visual tasks (Sammani, Chamiti, and Deligiannis 2026; Kaya, Elliott, and Papadopoulos 2026). We isolate the two with a *format-matched control* (MatchedCtrl): reuse the N CoT answers and add the same number of short, no-CoT answers, but drawn from the *original* image with no perturbation. MatchedCtrl spends PGS's exact extra budget in its exact decoding style and changes only whether those short answers pass through perturbed views, so comparing PGS to MatchedCtrl rather than to CoT-only MV tests grounding rather than format.

Our measurements across four automatically scored benchmarks—with Qwen as the headline model and LLaVA-OneVision in the selector tables—return a negative result. Against plain MV, PGS appears to help substantially (up to $+31.8$ points on TextVQA; three-seed mean), but against MatchedCtrl it shows no reliable advantage on any benchmark, including the vision-required ViLP, and no category yields a significant gain. The preserve/destroy stability gap the method is built on is real and image-dependent (up to $+0.48$), yet its per-instance value does not predict when PGS wins. Perturbation consistency is thus at best a partial diagnostic of visual dependence, not a usable label-free selection signal once decoding format is held fixed.

Contributions. (1) A decoding-format confound and its control: comparing a perturbation- or consistency-based selection rule against chain-of-thought-only majority voting conflates the selection mechanism with a decoding-format change, and our format-matched control (MatchedCtrl), the same short, no-CoT budget spent on the unperturbed image, isolates it, naming a cause distinct from the diversity (Tong et al. 2026) and internal-structure (Cai, Kulik, and Choudhury 2026) accounts of weak VLM selection. **(2) A negative result under that control:** across four benchmarks (Qwen headline; LLaVA-OneVision in matched-budget selectors), PGS, a label-free, training-free rule that strictly generalizes majority voting, gives no advantage over MatchedCtrl on any Qwen headline benchmark, including the vision-required ViLP, and a real, image-dependent stability gap (up to $+0.48$) does not predict when it wins. **(3) An experiment-facing isolation protocol:** we formalize the information and budget difference among MV, PGS, and MatchedCtrl, and state what would count as evidence that perturbation reweighting supplies visual grounding at the selection layer (Section 4). **(4) Released measurement and audit tooling** (the PGS score, stability gap, paired bootstrap CIs), so the format-matched comparison can be reused to audit perturbation- and consistency-based selection claims for VLM test-time scaling.

2 Related Work

Test-time scaling and self-verification. LLM test-time scaling selects among many samples via majority vote or best-of-N verification (Wang et al. 2023; Brown et al. 2024; Snell et al. 2024). The same recipe is attractive for VLMs,

yet the selection layer is fragile: for RL-tuned visual math models, majority voting beats verification-centric selection and self-correction gains are unreliable (Wu et al. 2026a), while self-refinement often *degrades* open VLMs (Ahmadpour et al. 2025). Adaptive compute allocation and trained multi-view policies (Jeddi et al. 2026; Avogaro et al. 2026; Yang et al. 2025; Zhu et al. 2026) improve *generation*—when to look again, which crop to take—rather than how to aggregate already-drawn candidates. PGS is orthogonal: it leaves the generator unchanged and only reweights answers at selection time, with no training and no second network.

Perturbation signals and confounds. Input perturbations already appear as decoding interventions (e.g., contrastive decoding (Leng et al. 2024; Wan et al. 2025)) and as confidence or consistency signals (Kim and Chelikavada 2026; Chou et al. 2026). Closest neighbors differ in what they require (attention access, trained modules) or in what they claim (generation-time grounding vs. selection-time reweighting); a compact design-properties comparison is in the Supplementary Material (Miyazawa and Okuno 2026; Liu et al. 2025). Parallel accounts of weak VLM selection emphasize sample diversity (Tong et al. 2026), internal commitment versus correctness (Cai, Kulik, and Choudhury 2026), and the decoding substrate (Baxevanakis and Yang 2026; Sammani, Chamiti, and Deligiannis 2026; Kaya, Elliott, and Papadopoulos 2026) (short answers often dominate verbose CoT on visual tasks). We add a *format-isolation* account: when a perturbation-based selector spends its extra budget as short, no-CoT draws, a gain over CoT-only majority voting can be almost entirely a CoT→short effect that a format-matched control (MatchedCtrl) removes. Under that control, preserve/destroy stability remains useful as an uncertainty *diagnostic* (Zhang, Zhang, and Zheng 2024), but not as a selector.

3 Method

3.1 Setup and notation

A VLM defines a distribution $p_\theta(a \mid x, q)$ over answers a given an image $x \in \mathcal{X}$ and a question q. Test-time scaling draws N candidates $a_1, \dots, a_N \sim p_\theta(\cdot \mid x, q)$ (with chain-of-thought, then extracting the final answer) and applies a selection function $S(\{a_i\}, x, q) \to \hat{a}$. Let $\mathrm{ans}(\cdot)$ map a sampled generation to its normalized final answer and let $\mathcal{A} = \{\mathrm{ans}(a_i)\}$ be the set of distinct candidate answers. Throughout, the image x and question q are fixed for an instance; what changes across selectors is which additional draws are taken and how they enter S.

Majority voting. The standard rule is $S_{\mathrm{MV}} = \arg\max_{c \in \mathcal{A}} \sum_i \mathbf{1}[\mathrm{ans}(a_i) = c]$. Each sample a_i is generated conditioned on the image x, so MV *does* use vision at generation time; the aggregation step itself only counts extracted answers and does not re-read x. That is a limitation of the *selection* layer, not a claim that MV ignores the image. Separately, Wu et al. (2026a) show that self-verification-based selection fails to integrate visual evidence effectively and is outperformed by MV—a generation–verification gap that motivates seeking selection signals beyond unverified self-checks, while still leaving open whether a perturbation-based reweighting can beat a format-matched control.

3.2 Label-preserving perturbations

Definition 1 (Label-preserving perturbation set)**.** *A perturbation set $\mathcal{T} = \{t_1, \dots, t_M\}$ is a finite collection of maps $t_m : \mathcal{X} \to \mathcal{X}$ such that for the true answer $a^\star$ to (x, q), $a^\star$ remains the correct answer to $(t_m(x), q)$ for all m. Examples used here: (i) crops toward question-relevant regions proposed by a cheap saliency heuristic; (ii) masking of background regions unlikely to contain the answer; (iii) mild photometric jitter (brightness/contrast) and small-angle rotation or rescaling.*

Perturbations are *label-preserving by construction*, not by assumption about the model: a crop that still contains the evidence, a mask over background, or a brightness change does not alter the ground-truth answer to a well-posed visual question. We make this concrete and auditable in Section 5 via the label-*destroying* control, which deliberately removes the evidence region and must reduce PGS's support for the correct answer if the signal is genuinely grounded.

3.3 The selection rule

For candidate answer $c \in \mathcal{A}$, define its *grounded support*

$$g(c) \;=\; \underbrace{\textstyle\sum_i \mathbf{1}[\mathrm{ans}(a_i) = c]}_{\text{original-view votes}} \;+\; \lambda \textstyle\sum_{m=1}^{M} w_m \, \rho\big(c \mid t_m(x), q\big), \tag{1}$$

where $\rho(c \mid t_m(x), q)$ is the model's *re-derivation strength* for answer c under perturbed view $t_m(x)$, $w_m \geq 0$ weights perturbation m, and $\lambda \geq 0$ trades off original-view agreement against perturbation stability. We instantiate ρ as the consistency score

$$\rho(c \mid t_m(x), q) = \tfrac{1}{K} \textstyle\sum_{k=1}^{K} \mathbf{1}\big[\mathrm{ans}(a^{(m,k)}) = c\big], \tag{2}$$

with $a^{(m,k)} \sim p_\theta(\cdot \mid t_m(x), q)$: we draw K short samples per perturbed view and count how often they re-derive c. The selected answer is $\hat{a} = \arg\max_{c \in \mathcal{A}} g(c)$.

Algorithm 1 Perturbation-Grounded Selection (PGS)

Require: image x, question q, model p_θ, perturbations $\mathcal{T}$, counts N, K, weight λ

sample $a_1, \dots, a_N \sim p_\theta(\cdot \mid x, q)$; $\mathcal{A} \leftarrow \{\mathrm{ans}(a_i)\}$
for each perturbation $t_m \in \mathcal{T}$ **do**
  sample $a^{(m,1)}, \dots, a^{(m,K)} \sim p_\theta(\cdot \mid t_m(x), q)$
  record $\rho(c \mid t_m(x), q)$ for every $c \in \mathcal{A}$
end for
compute $g(c)$ by Eq. (1) for every $c \in \mathcal{A}$
return $\hat{a} = \arg\max_c g(c)$

Proposition 1 (Majority voting is a special case)**.** *If $\mathcal{T} = \emptyset$ (equivalently $\lambda = 0$), then $g(c)$ reduces to the original-view vote count and $\hat{a} = S_{\mathrm{MV}}$.*

This is immediate from Eq. (1). PGS is therefore a strict generalization of majority voting rather than a competing heuristic: the original-view vote count is always present in $g(c)$, and the perturbation term only re-weights candidates. We do not claim PGS can never underperform majority voting (a large λ can in principle override a correct original-view majority), so the choice of λ matters. Unless noted otherwise, all primary results use a fixed operating point λ=2 (chosen once for the protocol, not computed per example). Separately, we also evaluate a *label-free* selection rule that does not look at ground truth: split the K samples of each perturbed view into two halves, run selection on each half, and among the λ values whose two halves agree on the chosen answer at least 80% of the time, take the *largest*. Choosing the smallest such λ would trivially return λ=0, since majority voting always agrees with itself. That rule is reported in SensAblations alongside the full λ sweep and a label-aware oracle; it is *not* the λ used to produce Table 1.

3.4 Cost

The full procedure is Algorithm 1. The generation cost is $N+MK$ forward samples versus N for plain best-of-N / CoT-only majority voting. Throughout the experiments we hold the total number of generations fixed at a matched budget (default $N+MK$=32; Section 5), so differences among PGS, and MatchedCtrl are about *how* those draws are spent rather than about spending more. Perturbed-view samples are short (answer-only), so the MK term stays cheap relative to N long CoT traces; MatchedCtrl spends that same short-answer mass on the original image.

4 What Would Count as a Grounding Gain?

The empirical claim is not that PGS can never help relative to *some* baseline, but that a widely used comparison is misleading, and that under the comparison that isolates the intended mechanism the gain disappears. This section fixes that comparison before the numbers appear.

4.1 Three selectors, three information budgets

Fix an instance (x, q) and a generation budget B=$N+MK$. Write $\mathrm{CoT}_N(x)$ for N chain-of-thought samples on the original image and $\mathrm{Short}_K(y)$ for K short, no-CoT samples on an image y.

- **MV** uses only $\mathrm{CoT}_N(x)$ and returns a majority vote. Cost N.
- **PGS** uses $\mathrm{CoT}_N(x)$ together with $\{\mathrm{Short}_K(t_m(x))\}_{m=1}^{M}$ and returns $\arg\max_c g(c)$ (Eq. (1)). Cost $N+MK$.
- **MatchedCtrl** uses $\mathrm{CoT}_N(x)$ together with $\mathrm{Short}_{MK}(x)$—the *same* short-answer mass, still on the original image—and returns a majority vote on the pooled answers. Cost $N+MK$.

MV and PGS differ in *two* ways at once: decoding format of the extra draws (CoT vs. short) and whether those draws see perturbed pixels. MatchedCtrl matches PGS on budget and short-answer format and differs *only* in whether the short draws are routed through $\mathcal{T}$. Therefore:

A gain of PGS over MV is not evidence of perturbation grounding.
A gain of PGS over MatchedCtrl would be.

Conversely, if $\mathrm{acc}(\text{PGS}) \approx \mathrm{acc}(\mathrm{MatchedCtrl})$, the perturbation term is not buying selection accuracy beyond spending the same short-answer budget on the original view. That is the decision-relevant null for contribution (2).

4.2 Diagnostic vs. routing; operating point

Even under a null against MatchedCtrl, the perturbation channel may still track visual content. We separate two roles:

- **Diagnostic (StabilityGap/BlankAblation).** The preserve/destroy stability gap, and the collapse of PGS when perturbation inputs are blanked, test whether ρ depends on the image. A large gap can coexist with a selection null.
- **Routing (RoutingTest).** For the same instances, does a larger gap predict a larger per-example gain $\mathbf{1}[\text{PGS}] - \mathbf{1}[\mathrm{MatchedCtrl}]$? If not, the gap is not a usable switch for when to trust perturbation weighting over MatchedCtrl.

A *grounding gain at the selection layer* would require (i) a reliably positive $\Delta(\text{PGS}-\mathrm{MatchedCtrl})$ overall or in a prespecified slice, and/or (ii) a monotone routing relationship in RoutingTest. Section 5 tests both and finds neither. Primary tables fix λ=2, pool `union`, and vote weight 1, holding total generations at $N+MK$=32 unless noted. Sensitivity to λ, M, K, and perturbation family is reported in SensAblations; those ablations do not overturn the MatchedCtrl null.

5 Experiments

We evaluate the isolation protocol of Section 4: whether PGS improves over the format-matched control MatchedCtrl, and whether the stability gap routes that comparison—not merely whether PGS beats CoT-only majority voting. The design therefore reports MV (for the familiar but confounded contrast), MatchedCtrl (for the decision-relevant null), and the StabilityGap/BlankAblation/RoutingTest diagnostics that separate "the signal sees the image" from "the signal improves selection." We evaluate on four benchmarks that vary in how much the answer depends on the image: TextVQA (Singh et al. 2019), MATH-Vision (Wang et al. 2024), MMMU (Yue et al. 2024), and ViLP (Luo et al. 2025), the last of which pairs a language-prior-aligned answer with a vision-required answer for each question and is the natural stress test for a grounding-at-selection claim. Coverage uses **two** open VLMs—**Qwen2.5-VL-7B-Instruct** (Qwen Team 2025) and **LLaVA-OneVision-7B** (Li et al. 2024). The *headline* mechanism table (Table 1) and the StabilityGap/ BlankAblation/ SensAblations/RoutingTest analyses report Qwen three-seed means for a single readable story; matched-budget selector comparisons and reproducibility tables report *both* models (Table 2, Table 3; and compute tables in the Supplementary Material). Unless stated otherwise, PGS is recomputed offline with candidate pool `union`, vote weight vw=1, and λ=2 on saved runs. Accuracy is hard-match rate in percent (TextVQA: any-annotator match).

Table 1: **Main results on Qwen2.5-VL-7B-Instruct**. Mean hard-match accuracy over seeds $\{0, 23, 42\}$. PGS appears to beat plain majority voting (MV) by up to $+31.8$ pp (column Δ), yet the format-matched control MatchedCtrl (shaded)—the same short, no-CoT budget on the *original* image—tracks PGS within seed noise on every benchmark (TextVQA: MatchedCtrl still ahead). Accuracy is hard-match rate (%; TextVQA any-annotator match); PGS uses `union`, vote weight 1, λ=2.

		Accuracy (%)				
Bench.	N	**MV**	**PGS**	**MatchedCtrl**[a]	Δ[b]	**Gap**[c]
TextVQA	300	54.4	86.2	87.6	+31.8	+0.478
MATH-V	299	25.6	26.0	25.3	+0.4	+0.025
MMMU	900	49.5	50.3	50.1	+0.7	+0.071
ViLP	600	53.2	52.3	53.7	−0.9	+0.445

[a] **MatchedCtrl**, format-matched control: reuse the N CoT answers + $M{\cdot}K$ short no-CoT answers on the original image ($N{+}MK{=}32$).
[b] $\Delta = \mathrm{acc}(\mathrm{PGS}) - \mathrm{acc}(\mathrm{MV})$, from seed-mean accuracies.
[c] Mean Gap $\rho_{\text{preserve}} - \rho_{\text{destroy}}$ over seeds; ViLP gaps are computed on non-prior (grounded) slots in the dump.

5.1 Setup and controls

Protocols. For each example we draw N original-view answers with chain-of-thought (CoT) and, for PGS, K *short, no-CoT* answers on each of M label-preserving perturbations. PGS scores candidates by votes plus λ times re-derivation support ρ estimated from the perturbation views. We report:

- **MV:** majority vote over the N original-view *CoT* samples only.
- **PGS:** ρ-weighted selection as above.
- **MVLift:** paired difference PGS$-$MV (percentage points).
- **MatchedCtrl:** format-matched control that reuses the N CoT answers and adds $M{\cdot}K$ short no-CoT answers on the *original* image ($N{+}MK{=}32$), matching PGS's extra budget and decoding style with no perturbation.
- **StabilityGap:** mean paired stability gap $\rho_{\text{preserve}} - \rho_{\text{destroy}}$ on the MV-selected candidate under label-destroying crops.

Comparing PGS to MV confounds perturbation weighting with the CoT↔short format change; comparing PGS to MatchedCtrl isolates whether routing those short answers through perturbed views and ρ helps beyond keeping them on the original image.

Baselines. Our comparison isolates the perturbation mechanism against the format-matched control (MatchedCtrl), the decisive test for our claim; it is not a survey of selectors. Still, under the same matched-budget protocol on each VLM we evaluate standard label-free alternatives—self-certainty (SC), SC with Borda voting (Kang, Zhao, and Song 2026), confidence-weighted self-consistency (Taubenfeld et al. 2025), and mean token-entropy selection—against MV, MatchedCtrl, and PGS at $N{+}MK \in \{16, 32\}$ (Table 2, Table 3). Both LLaVA-OneVision-7B and Qwen appear in these selector tables, so the MatchedCtrl null is not an artifact of a single checkpoint family. Confidence selectors use the matched $N{+}MK$ decode pool after teacher-force rescoring; budget 16 is an offline subsample (N=4, K=2) of the budget-32 dumps. Consistent with the intro diagnosis, these selectors do not systematically beat MatchedCtrl: SC and entropy often fall below MatchedCtrl, while SC+Borda and CISC typically track MatchedCtrl within noise. Where PGS also fails to separate from MatchedCtrl, the null remains diversity / format rather than grounding; where absolute numbers differ across models, the gap is still relative to a control that already absorbs the strongest confidence-based aggregators.

Reproducibility and benchmarks. We evaluate TextVQA, MATH-Vision, MMMU, and ViLP (Score / ViLP-P / ViLP-F splits as noted per table; ViLP StabilityGap gaps use non-prior grounded slots in the dumps, N=600 per seed). Decode settings, exact (N, M, K), compute footprint for *both* VLMs, dataset versions/splits/licenses, and checkpoint revisions appear in the Supplementary Material. Default generation budget is N=8, M=6, K=4 ($N{+}MK$=32) on a single NVIDIA RTX 4090.

5.2 Main results

Table 1 is the headline comparison (three-seed means on Qwen dumps). Against plain MV, PGS gains substantially on TextVQA ($+31.8$ pp) and is near-flat on MATH-V, MMMU, and ViLP. That comparison is unfair in exactly the sense of Section 4: MV aggregates only the N *CoT* samples, whereas both PGS and MatchedCtrl add $M{\cdot}K$ *short, no-CoT* answers. The all-CoT MV pool is therefore the weaker aggregator at a smaller effective budget in short-answer mass: accuracy rises at matched $N{+}MK$ even when the extra draws never leave the original image. MatchedCtrl spends those same short answers on the *original* image and, under this format-matched control, tracks PGS within seed noise on every benchmark (TextVQA 87.6 vs. 86.2; MATH-V 25.3 vs. 26.0; MMMU 50.1 vs. 50.3; ViLP 53.7 vs. 52.3), with MatchedCtrl still ahead on TextVQA. The MV→PGS lift is therefore the effect of adding short no-CoT mass, not evidence that re-derivation support ρ supplies visual grounding at the selection layer; isolating the perturbation piece leaves no reliable advantage. The StabilityGap reinforces this reading rather than rescuing it. Large mean preserve/destroy gaps on TextVQA and ViLP ($+0.478$ / $+0.445$) coexist with no win over MatchedCtrl, while smaller gaps on MATH-V and MMMU ($+0.025$ / $+0.071$) likewise fail to separate PGS from MatchedCtrl. Preserve/destroy asymmetry is real and image-dependent, but its magnitude does not identify when perturbation-weighted selection beats format-matched short-answer MV. In other words, the diagnostic channel can fire without a routing gain over MatchedCtrl—the distinction Section 4 insists on.

5.3 Ablations: the signal is visual but does not route selection

Blanking the perturbation inputs (BlankAblation, Table 4) collapses PGS's accuracy where the signal is strong (TextVQA $87.7 \to 7.9$, ViLP-F $46.2 \to 0.4$, MMMU $50.2 \to 23.3$) while symbolic MATH-V is unaffected, confirming the

Table 2: Label-free selectors vs. MV/MatchedCtrl/PGS at matched budget $N+MK=32$ (rescored confidence metrics; mean±std over seeds).

Benchmark	Model	MV	MatchedCtrl	SC	Entropy	SC+Borda	CISC	PGS
MATH-Vision	LLaVA-OV-7B	18.8±1.1	20.7±0.9	15.8±1.1	17.8±0.7	19.2±2.3	19.5±0.8	20.3±0.2
MATH-Vision	Qwen2.5-VL-7B	25.6±1.6	25.3±1.0	17.9±0.5	15.4±0.6	25.1±1.0	25.4±1.0	26.0±0.7
MMMU	LLaVA-OV-7B	47.4±0.5	47.6±0.6	43.9±0.3	45.7±0.3	47.5±0.7	47.9±0.7	48.3±0.7
MMMU	Qwen2.5-VL-7B	49.5±0.3	50.1±0.4	44.7±0.5	44.7±0.7	50.1±0.5	50.4±0.4	50.3±0.7
TextVQA	LLaVA-OV-7B	78.8±0.2	81.3±0.3	78.8±0.8	80.4±0.8	82.2±0.5	81.6±0.5	81.7±0.3
TextVQA	Qwen2.5-VL-7B	54.4±2.0	87.6±0.4	47.1±2.1	84.0±4.1	88.1±0.5	88.0±0.3	86.2±0.7
ViLP	LLaVA-OV-7B	49.9±0.3	49.4±0.3	50.4±0.5	49.1±0.1	50.4±0.9	50.1±0.7	49.5±0.7
ViLP	Qwen2.5-VL-7B	53.2±0.7	53.7±0.4	50.9±0.6	48.6±0.6	54.0±0.2	54.0±0.4	52.3±0.7

Table 3: Label-free selectors vs. MV/MatchedCtrl/PGS at matched budget $N+MK=16$ (offline subsample from budget-32 dumps; mean±std over seeds).

Benchmark	Model	MV	MatchedCtrl	SC	Entropy	SC+Borda	CISC	PGS
MATH-Vision	LLaVA-OV-7B	20.9±2.2	20.7±1.2	17.4±1.4	19.0±0.7	20.4±0.6	20.3±1.2	22.9±0.8
MATH-Vision	Qwen2.5-VL-7B	23.3±1.6	26.8±2.6	23.0±1.9	21.6±2.2	26.9±2.3	26.3±2.6	26.8±1.1
MMMU	LLaVA-OV-7B	48.2±1.2	48.3±0.6	44.8±0.1	47.1±0.4	47.6±0.4	48.2±0.9	48.8±1.1
MMMU	Qwen2.5-VL-7B	49.3±0.8	51.0±0.5	45.4±0.6	45.7±0.6	51.3±0.1	51.2±0.1	50.6±0.3
TextVQA	LLaVA-OV-7B	75.7±0.3	81.8±0.2	79.3±0.7	81.3±0.3	81.8±0.4	82.0±0.7	81.6±0.2
TextVQA	Qwen2.5-VL-7B	57.0±2.0	87.9±0.4	49.9±2.9	84.7±2.7	88.4±0.7	88.4±0.2	88.7±0.7
ViLP	LLaVA-OV-7B	49.8±0.3	49.5±0.4	50.5±0.2	49.4±0.2	50.1±0.8	50.1±0.6	48.9±0.8
ViLP	Qwen2.5-VL-7B	52.9±0.8	53.3±0.3	51.8±1.2	49.7±0.3	53.7±0.4	53.6±0.2	51.2±0.7

Table 4: **BlankAblation** (earlier single-seed Qwen dual-PGS run; not present in the multi-seed dumps used for Table 1). Holding the original-view CoT pool fixed and blanking the perturbation inputs collapses accuracy where the signal is visual (TextVQA, ViLP-F, MMMU) but not on symbolic MATH-V. Absolutes are *not* aligned with the multi-seed main table; only the within-row comparison is meaningful.

Benchmark	N	**MV**	**PGS**	**BlankAblation**
TextVQA	302	55.0	87.7	7.9
MATH-V	304	22.4	26.3	23.4
MMMU	900	49.9	50.2	23.3
ViLP-F	266	48.9	46.2	0.4

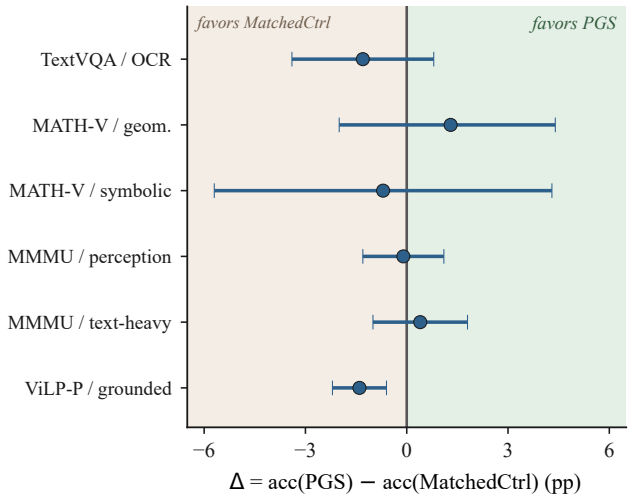

Figure 2: **No category shows a significant gain of PGS over the format-matched control MatchedCtrl.** Per-category difference $\Delta = \mathrm{acc}(\mathrm{PGS}) - \mathrm{acc}(\mathrm{MatchedCtrl})$ with 95% bootstrap confidence intervals (Supplementary Material), pooled over seeds $\{0, 23, 42\}$ on Qwen dumps. No interval lies entirely in the "favors PGS" half-plane (shaded); the ViLP interval lies entirely below zero. Points are Δ; whiskers are the 95% CI.

score is genuinely image-dependent rather than a format artifact of short decoding alone. Yet image-dependence and the stability gap decouple (MMMU drops sharply while its multi-seed mean gap is only $+0.071$), so the gap is only a partial measure of visual dependence: BlankAblation can fail even when StabilityGap looks mild. Ablations over perturbation family, λ, M, and K (SensAblations, Supplementary Material: perturbation-family / λ-sweep / $\hat{\lambda}$ tables; recomputed offline from the same multi-seed Qwen dumps as Table 1) show PGS is active almost only on OCR-heavy TextVQA (crop/geometry-sensitive) while other benchmarks are flat, consistent with a format-sensitive rather than a universal grounding effect. Neither ablation overturns MatchedCtrl: where the signal is visually live, matched short answers on the original image still absorb the lift.

5.4 RoutingTest: where the gap predicts help (it does not)

We break results down by perception-heavy vs. text-heavy categories and correlate per-instance gain ($\mathbf{1}[\mathrm{PGS}] - \mathbf{1}[\mathrm{MatchedCtrl}]$) with the StabilityGap; on ViLP we additionally report prior→grounded flip rates relative to

MatchedCtrl. These analyses ask the routing question of Section 4 directly: even if the mean $\Delta(\text{PGS}-\text{MatchedCtrl})$ is near zero, a pre-specified slice or a monotone gap→gain relationship could still salvage a conditional use of ρ. The Supplementary Material category table and Figure 2 close the isolation protocol under the decision-relevant baseline: no category yields a significant positive $\Delta(\text{PGS}-\text{MatchedCtrl})$ (every 95% bootstrap CI overlaps zero). The stability gap also fails to *route* the selector: instance-level correlations between gap and gain are near zero (Supplementary Material), gap quartiles are non-monotone, and PGS's recoveries of MatchedCtrl errors on ViLP are rare (Supplementary Material). A large measured re-derivation advantage therefore does not forecast when perturbation-weighted selection beats format-matched short-answer MV. The gap remains a useful *diagnostic* of preserve/destroy asymmetry; it is not a reliable *routing* signal for the selector studied here.

6 Discussion

The controls converted a plausible mechanism into a measured null. StabilityGap and BlankAblation show that the re-derivation score tracks visual content when the pixels are removed or destroyed; once the same short draws sit on the original image, the perturbation term buys nothing at selection, and the per-instance gap does not route the selector (RoutingTest). Diagnostic image dependence is therefore not a routing gain over a format-matched control (Section 4). Gains against CoT-only majority voting can be almost entirely a decoding-format effect—here as large as $+31.8$ pp on TextVQA—and should not be read as evidence that perturbation consistency supplies visual grounding. ViLP sharpens the point: even where language priors and image answers are split, MatchedCtrl still meets or beats PGS, so a large preserve/destroy gap coexists with a selection null rather than rescuing it. This isolates a format piece that coexisting 2026 accounts do not foreground in the same way: diversity-limited self-consistency (Tong et al. 2026) and commitment-versus-correctness (Cai, Kulik, and Choudhury 2026) explain other failure modes of VLM selection, while MatchedCtrl shows that a CoT↔short confound can inflate apparent perturbation gains even when sample diversity and internal stability are held aside. Holding the *number* of draws fixed is not enough if the extra draws change answer format; the control must also hold format fixed on the original view. Progress beyond MatchedCtrl will likely need an information source that answer-level consistency alone does not supply—denser visual probes, process-level rewards, or cross-model disagreement. We do *not* show that a trained verifier, or multi-model ensemble beats MatchedCtrl; whether such a signal recovers the gap remains open, and any such claim must still clear a format-matched control.

Practical takeaways. Any selection rule that spends extra draws in a different decoding format than the MV baseline needs a format-matched control; without one, a format effect is easily misread as a grounding gain. Short-answer aggregation on the original image is a strong, cheap baseline under matched budget, and in our grid it is the hurdle that confidence-style and perturbation-weighted selectors alike must clear. Perturbation consistency remains worth computing as a partial *diagnostic* of visual dependence (StabilityGap/BlankAblation)—useful for auditing whether a score still sees the image—but not as a drop-in *selector* once format is controlled. Reporting MV lifts without a MatchedCtrl-style arm should be treated as incomplete for grounding-at-selection claims.

Limitations

Our conclusion is bounded, and stating the bounds is part of reading it honestly. The evidence covers two open VLMs (LLaVA-OneVision-7B and Qwen2.5-VL-7B) on four automatically scored benchmarks. A stronger family or larger decode budget could in principle show a benefit ours do not; we do not claim universality beyond the measured grid. Main results are aggregated over three decoding seeds ($\{0, 23, 42\}$) as mean±std (Supplementary Material); paired PGS-vs-MatchedCtrl comparisons carry instance-level bootstrap CIs (RoutingTest), the right uncertainty for the claim that no category shows a significant gain over MatchedCtrl. Three seeds cannot rule out rarer decoding regimes. The perturbation families remain hand-designed heuristics (crop, mask, mild photometric and geometric jitter); a learned set could carry more signal, but would still need a MatchedCtrl-style control before any gain is attributed to grounding, and our SensAblations sweeps already show that family choice does not overturn the null under matched format. We do not include a stronger-signal reference point (trained verifier or multi-model ensemble) that beats MatchedCtrl, so recoverability under training is not demonstrated. And the negative finding is about *answer selection*: the stability gap remains a partial diagnostic of visual dependence—none of which we test here for abstention or routing.

Ethics Statement

PGS is inference-only and uses public benchmarks and open models, raising no annotation- or participant-related concerns. Favoring image-grounded answers may reduce confidently-wrong outputs that ignore the pixels, but a stability signal can be satisfied by inputs that are stable yet wrong, so PGS is not a correctness or safety guarantee and does not replace verification in safety-critical use; it also inherits the underlying model's biases and failure modes. Deploying perturbation probes still incurs extra decode cost and can amplify whatever stereotypes or dataset artifacts the base VLM already exhibits under short-answer sampling. Because our headline finding is negative under MatchedCtrl, the main ethical risk of overstating PGS as a grounding selector is mitigated by the release itself: we release the analysis code, perturbation specifications, and evaluation tooling with paired bootstrap CIs, so the null against format-matched short-answer aggregation is reproducible. We encourage downstream work that claims a grounding gain at selection to publish the same MatchedCtrl-style contrast rather than MV-only lifts. The authors used AI-based tools for assistance with language editing and software debugging during the development of this work. All AI-generated suggestions were reviewed, verified, and modified by the authors.

References

Ahmadpour, M.; Meighani, A.; Taebi, P.; Ghahroodi, O.; Izadi, A.; and Soleymani Baghshah, M. 2025. Limits and Gains of Test-Time Scaling in Vision-Language Reasoning. *arXiv preprint arXiv:2512.11109*.

Avogaro, N.; Debnath, N.; Mi, L.; Frick, T.; Wang, J.; He, Z.; Hua, H.; Schindler, K.; and Rigotti, M. 2026. SPARC: Separating Perception And Reasoning Circuits for Test-time Scaling of VLMs. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*.

Baxevanakis, S.; and Yang, P.-J. 2026. Test-Time Scaling for Small VLMs on Multilingual Visual MCQ. *arXiv preprint arXiv:2607.09438*.

Brown, B.; Juravsky, J.; Ehrlich, R.; Clark, R.; Le, Q. V.; Ré, C.; and Mirhoseini, A. 2024. Large Language Monkeys: Scaling Inference Compute with Repeated Sampling. *arXiv preprint arXiv:2407.21787*.

Cai, M.; Kulik, L.; and Choudhury, F. 2026. ARBITER: Reasoning Trajectory Basins and Majority Vote Failures in Test-Time Sampling. *arXiv preprint arXiv:2605.26172*.

Chou, S.-H.; Chandhok, S.; Little, J. J.; and Sigal, L. 2026. Test-Time Consistency in Vision Language Models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

Chung, H.-L.; Hsiao, T.-Y.; Huang, H.-Y.; Cho, C.; Lin, J.-R.; Zhang, Z.; Hsieh, J.; and Chen, Y.-N. 2025. Revisiting Test-Time Scaling: A Survey and a Diversity-Aware Method for Efficient Reasoning. *arXiv preprint arXiv:2506.04611*.

Jeddi, A.; Le, M. N.; Kazerouni, A.; Karaimer, H. C.; Nguyen, H.; Mohomed, I.; Brudno, M.; Levinshtein, A.; Derpanis, K. G.; Taati, B.; and Grzeszczuk, R. 2026. AVIS: Adaptive Test-Time Scaling for Vision–Language Models. *arXiv preprint arXiv:2606.11576*.

Kang, Z.; Zhao, X.; and Song, D. 2026. Scalable best-of-n selection for large language models via self-certainty. *Advances in neural information processing systems*.

Kaya, M. O.; Elliott, D.; and Papadopoulos, D. P. 2026. Efficient Test-Time Scaling for Small Vision-Language Models. In *International Conference on Learning Representations (ICLR)*.

Kim, K.; and Chelikavada, K. 2026. Zoom Consistency: A Free Confidence Signal in Multi-Step Visual Grounding Pipelines. *arXiv preprint arXiv:2604.15376*.

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 2024. 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 (CVPR)*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Liu, Z.; et al. 2025. Seeing but Not Believing: Probing the Disconnect Between Visual Attention and Answer Correctness in VLMs. *arXiv preprint arXiv:2510.17771*.

Luo, T.; Cao, A.; Lee, G.; Johnson, J.; and Lee, H. 2025. Probing Visual Language Priors in VLMs (ViLP). In *International Conference on Machine Learning (ICML)*.

Marsili, D.; and Gkioxari, G. 2025. No Labels, No Problem: Training Visual Reasoners with Multimodal Verifiers. *arXiv preprint arXiv:2512.08889*.

Miyazawa, T.; and Okuno, H. 2026. Answer Self-Consistency with Margin-Triggered Question Re-Arbitration for the CVPR 2026 VidLLMs Challenge. *arXiv preprint arXiv:2606.04323*.

Qwen Team. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Sammani, F.; Chamiti, T.; and Deligiannis, N. 2026. On Test-Time Scaling for Vision-Language Models. In *European Conference on Computer Vision (ECCV)*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Snell, C.; Lee, J.; Xu, K.; and Kumar, A. 2024. Scaling LLM Test-Time Compute Optimally Can Be More Effective Than Scaling Model Parameters. *arXiv preprint arXiv:2408.03314*.

Taubenfeld, A.; Sheffer, T.; Ofek, E.; Feder, A.; Goldstein, A.; Gekhman, Z.; and Yona, G. 2025. Confidence improves self-consistency in llms. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Tong, Y.; Hou, Y.; Cui, S.; Bosselut, A.; and Sachan, M. 2026. Diversity Matters: Revisiting Test-Time Compute in Vision-Language Models. In *International Conference on Machine Learning (ICML)*.

Wan, D.; Cho, J.; Stengel-Eskin, E.; and Bansal, M. 2025. Contrastive Region Guidance: Improving Grounding in Vision-Language Models Without Training. In *European Conference on Computer Vision (ECCV)*.

Wang, K.; Pan, J.; Shi, W.; Lu, Z.; Ren, H.; Zhou, A.; Zhan, M.; and Li, H. 2024. Measuring Multimodal Mathematical Reasoning with MATH-Vision Dataset. In *Advances in Neural Information Processing Systems*.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*.

Wani, F. A.; Suglia, A.; Saxena, R.; Gema, A. P.; Kwan, W.-C.; Barez, F.; Bucarelli, M. S.; Silvestri, F.; and Minervini, P. 2026. Same Answer, Different Representations: Hidden Instability in VLMs. *arXiv preprint arXiv:2602.06652*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Wu, M.; Li, M.; Yang, J.; Jiang, J.; Yan, K.; Li, Z.; Yu, H.; Zhang, M.; and Nahrstedt, K. 2026a. Aha Moment Revisited: Are VLMs Truly Capable of Self Verification in Inference-time Scaling? *arXiv preprint arXiv:2506.17417*.

Wu, M.; Yang, J.; Jiang, J.; Li, M.; et al. 2026b. VTool-R1: VLMs Learn to Think with Images via Reinforcement Learning on Multimodal Tool Use. In *International Conference on Learning Representations (ICLR)*.

Yang, Y.; Liu, J.; Zhang, Z.; Zhou, S.; Tan, R.; Yang, J.; Du, Y.; and Gan, C. 2025. MindJourney: Test-Time Scaling with World Models for Spatial Reasoning. In *Advances in Neural Information Processing Systems*.

Yue, X.; et al. 2024. 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 (CVPR)*.

Zhang, R.; Zhang, H.; and Zheng, Z. 2024. VL-Uncertainty: Detecting Hallucination in Large Vision-Language Models via Uncertainty Estimation. *arXiv preprint arXiv:2411.11919*.

Zhu, C.; Lin, J.; Long, Y.; Cao, P.; Wang, T.; Pang, J.; and Liu, X. 2026. Thinking with Imagination: Agentic Visual Spatial Reasoning with World Simulators. *arXiv preprint arXiv:2606.06476*.