UNDERSTANDING AND IMPROVING ADVERSARIAL ANSWER RECOVERY VIA ACTIVATION STEERING IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Activation steering provides a controlled way to modify internal representations without updating backbone parameters. We study its use for *behavioral recovery*: restoring pre-attack responses after adversarial image perturbations while preserving responses to pre-attack inputs. Controlled analyses first examine where and how to intervene. We find that attack status is decodable at the answer-prediction state, yet detector and recovery directions are weakly aligned, and input-specific answer-margin directions recover substantially more responses under matched interventions. These findings motivate AM-GARD, which separates intervention selection from answer correction using a learned intervention gate and input-specific answer-margin updates. Across multiple LVLMs, tasks, and answer formats, AM-GARD improves recovery while maintaining high pre-attack response preservation. Further analyses show that broader selection trades preservation for recovery opportunity, while successful correction depends on identifying a useful answer boundary rather than merely changing the output. Together, these results show that behavioral recovery depends on separating *whether* to intervene from *how* to correct the answer, providing both a diagnostic account of recovery behavior and an effective inference-time intervention. [1]

1 INTRODUCTION

Visual adversarial perturbations can alter the responses of large vision–language models (LVLMs) even when the accompanying question is unchanged (Carlini et al., 2023; Schlarmann & Hein, 2023; Cui et al., 2024). Recent defenses have explored several intervention points, including input transformation, adversarial training, image purification, and generation-time guidance (Mirza et al., 2026; Wang et al., 2025b; Zhang et al., 2025; Zhou et al., 2025). These approaches improve robustness in different settings, but leave a related question less directly examined: when an attack changes a model's response, what kind of intervention can restore its own pre-attack behavior while avoiding unnecessary changes to responses on pre-attack inputs?

We study this problem as *behavioral recovery*: restoring the model's pre-attack answers on attacked inputs while preserving its responses to pre-attack inputs. Recovery goes beyond detecting an attack or merely changing the attacked output: if an attack changes answer A to B, producing C does not recover the original behavior (Figure 1a). Recovering attacked answers while preserving pre-attack responses motivates separating intervention selection from answer correction.

Activation steering lets us study this distinction by intervening in internal representations while keeping the observed input and backbone fixed, as illustrated in Figure 1b. Prior work has used activation-level interventions to analyze and control model behavior, including representation editing, behavioral steering, and multimodal alignment and safety (Belrose et al., 2023; Li et al., 2023a; Rimsky et al., 2024; Qiu et al., 2024; Sivakumar et al., 2025; Wang et al., 2025a; Lu et al., 2026; Wu et al., 2025). However, these studies do not directly characterize which activation interventions restore pre-attack responses rather than merely alter attacked outputs.

[1] Code and the demo are available (anonymized) at `https://anonymous.4open.science/r/AM-GARD-A8FB`.

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.

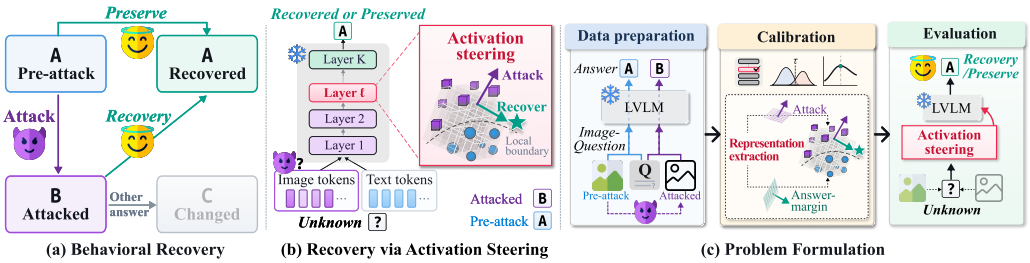


Figure 1: Overview of behavioral recovery, activation-based intervention, and the evaluation setup.

We study recovery through two questions: *where* to intervene and *which direction* to use. Our analyses identify the answer-prediction state as a useful intervention site, where attack status is decodable after visual and textual information are integrated. However, this does not imply that the corresponding detector direction is suitable for recovery. Under matched interventions, an input-specific answer-margin direction recovers substantially more pre-attack answers, while the two directions exhibit weak alignment.

Motivated by this separation, we introduce **AM-GARD** (**A**nswer-**M**argin-**G**uided **A**ctivation for **R**obust **D**efense), decoupling intervention selection from answer correction. AM-GARD uses detector scores to determine *whether* to intervene and an input-specific answer-margin signal to determine *how* to intervene, enabling targeted recovery while preserving pre-attack responses. The backbone remains frozen, and inference requires neither the paired pre-attack input nor the pre-attack response.

Our contributions are threefold: **(1)** we characterize behavioral recovery through *answer-conditioned interventions*, showing that attack detection and recovery can share the same state while requiring different intervention directions; **(2)** building on this finding, we introduce AM-GARD, which separates *whether to intervene* from *how to recover* through detector-based selection and input-specific answer-margin correction; and **(3)** we evaluate this principle across multiple LVLMs, tasks, and answer formats, jointly examining recovery and pre-attack preservation to better understand what governs behavioral recovery and how that understanding can be used to improve it.

2 Problem Formulation

Attack and recovery objective. For a frozen LVLM f_θ, let $Y_i^0 = f_\theta(x_i, q_i)$ be its pre-attack response to an image–question pair. Behavioral recovery seeks agreement with Y_i^0 after attacking the image while preserving responses to pre-attack inputs, irrespective of whether Y_i^0 matches the dataset ground truth. The question and decoding rule remain fixed. Our primary white-box attack reduces the first-response margin of the pre-attack leading token a_i^0:

$$m_i^{\mathrm{atk}}(\tilde{x}) = z_{a_i^0}(\tilde{x}, q_i) - \max_{v \neq a_i^0} z_v(\tilde{x}, q_i), \qquad L_{\mathrm{atk}} = m_i^{\mathrm{atk}}(\tilde{x}) + \lambda_p [d_{\mathrm{LPIPS}}(x_i, \tilde{x}) - \rho]_+. \tag{1}$$

Image-domain gradient updates and a budget search seek a small perceptual perturbation that changes the leading token (Madry et al., 2018; Zhang et al., 2018). This preparation-time criterion defines the attacked decision; final recovery is evaluated from generated answers. The full test set retains both successful and unsuccessful attack attempts. Auxiliary visual analyses use separately specified attacks.

Intervention and information access. Let u denote the observed image: $u = x_i$ for pre-attack evaluation and $u = \tilde{x}_i$ for attacked evaluation. At the last input position predicting the first response token, we consider

$$h'_\ell(u, q) = h_\ell(u, q) + D(u, q)\Delta h_\ell(u, q),$$

where $D(u, q) \in \{0, 1\}$ determines whether to intervene and $\Delta h_\ell(u, q)$ specifies the update. The update is computed from the answer-prediction state and reused during autoregressive generation, as detailed in Section 4. At inference, deployed methods access only the observed image, question, and frozen model's activations and gradients, without access to the paired pre-attack image, pre-attack response, or ground-truth answer. Reference-based masks or targets are confined to identified diagnostic experiments.

Evaluation protocol. We prepare paired pre-attack and attacked inputs and record the pre-attack responses as recovery references. Pairs are partitioned into disjoint estimation, calibration, and test splits. Learned components are fitted on the estimation split, and intervention settings are selected on the calibration split. With these choices fixed, the test split measures recovery gain over no-op and response preservation on pre-attack inputs. Metric definitions and scoring details are provided in Section 5. Figure 1c summarizes the problem formulation.

3 Understanding What Governs Answer Recovery

We use controlled activation interventions to study which internal changes recover an attacked answer. We first examine the intervention state, then test whether a direction trained for attack detection also supports answer recovery.

3.1 Where Should Intervention Be Conditioned?

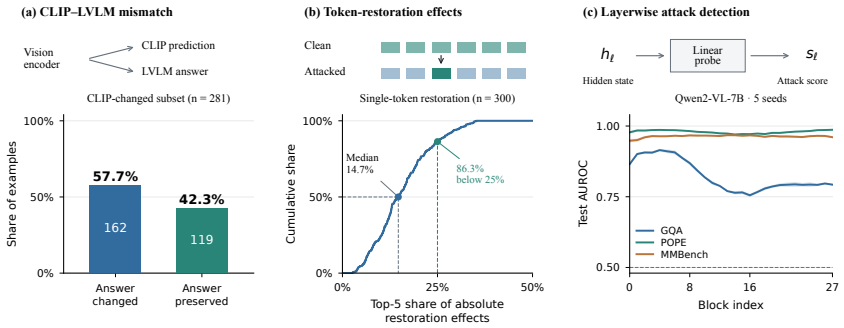


Figure 2: Analyses motivating the answer-prediction state as the intervention site.

Recovery Is Not Determined by Visual Changes Alone. Although CLIP (Radford et al., 2021) and LLaVA-1.5-7B (Liu et al., 2024a) share the same vision encoder, a CLIP prediction change does not reliably imply an LVLM answer change (Figure 2a). Full experimental settings are in Appendix D.1. Likewise, restoring individual visual tokens to their pre-attack representations on attacked CLEVR inputs (Johnson et al., 2017) produces effects that are broadly distributed across positions rather than concentrated in a small subset (Figure 2b). A localized object-deletion control, in contrast, places the dominant restoration effect near the modified region, confirming that the analysis can recover spatially localized effects when they are present. These observations make individual visual decisions or a small set of visual tokens insufficient as a general basis for conditioning recovery.

Attack detection at the answer-prediction state. We therefore examine the downstream state used for answer prediction. At the final input position, linear probes trained to distinguish pre-attack from attacked representations achieve above-chance test AUROC on disjoint data across Qwen2-VL-7B blocks and tasks (Wang et al., 2024b) (Figure 2c). The answer-prediction state provides a suitable point for conditioning an intervention, while leaving open which direction should be used to recover the answer. Probe training and evaluation details are provided in Appendix D.2.

3.2 Which Direction Should Recover the Answer?

Matched-norm direction comparison. We compare the detector direction from Section 3.1 with an input-specific answer-margin direction. We select attacked examples whose current no-op answers differ from their saved pre-attack answers, rather than using the flipped subset defined during attack preparation. This selection requires saved pre-attack answers and is therefore a diagnostic control, not a deployable gate. To isolate the effect of intervention direction, both interventions use the same examples at layer 24 with matched per-example L_2 update norms, holding input selection and update magnitude fixed. The detector direction yields little recovery, whereas the answer-margin direction recovers substantially more pre-attack answers (Figure 3a). Thus, a direction learned for attack detection need not be effective for answer recovery. The controlled evaluation protocol is detailed in Appendix D.3.

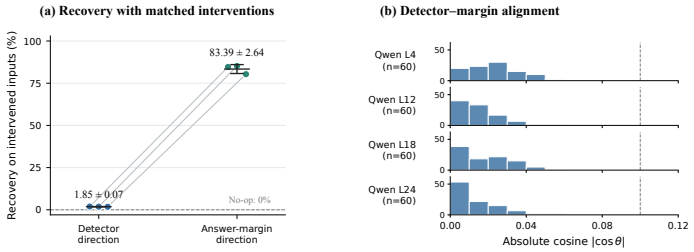


Figure 3: **Intervention direction and answer recovery.** (a) Recovery on the same attacked examples whose current no-op answers differ from their pre-attack answers, with matched per-example L_2 update norms. (b) Absolute cosine similarity between detector directions and answer-margin gradients across Qwen2-VL-7B layers.

Linearized answer-boundary analysis. We analyze the displacement required to reach a selected token-pair boundary using local linearization (Moosavi-Dezfooli et al., 2016). For an attacked answer-prediction state h, let a be the current leading token and c a fixed competitor. Define $g(h) = z_a(h) - z_c(h) > 0$ and $n = \nabla_h g(h) \neq 0$, with token indices held fixed during differentiation. The linearized margin is $g(h + \delta) \approx g + n^\top \delta$.

Proposition 3.1 (Minimum-norm boundary correction)**.** *The unique minimum-L_2-norm displacement satisfying $g + n^\top \delta \leq 0$ is*

$$\delta^* = -\frac{g}{\|n\|_2^2} n, \qquad \|\delta^*\|_2 = \frac{g}{\|n\|_2}. \tag{2}$$

Proposition 3.2 (Alignment determines correction cost)**.** *For a unit direction v, consider displacements $\delta = -tv$ with $t \geq 0$. If $\cos\theta = n^\top v / \|n\|_2 > 0$, the minimum step length required to reach the same linearized boundary is*

$$t_v = \frac{g}{\|n\|_2 \cos\theta}, \qquad \frac{t_v}{\|\delta^*\|_2} = \frac{1}{\cos\theta}. \tag{3}$$

If $\cos\theta \leq 0$, no displacement along $-v$ reaches this boundary under the local linearization.

The local answer-margin direction $-n$ therefore minimizes the displacement to the selected boundary, whereas weak alignment requires a larger update norm. These statements concern a linearized token-pair boundary; reaching it does not identify the pre-attack answer or guarantee recovery of the generated response. Proofs and additional geometric details are provided in Appendix C.

Detector–margin alignment. We measure absolute cosine similarity between detector directions and answer-margin gradients across Qwen2-VL-7B layers. Similarities concentrate near zero (Figure 3b). Since absolute cosine is sign-invariant, reversing the detector direction does not resolve the weak alignment. Proposition 3.2 associates this alignment with a larger displacement to the selected linearized boundary, consistent with the limited recovery observed under matched update norms. Appendix C further visualizes this geometry using projections of the stored intervention vectors.

AM-GARD uses these results to separate the two operations: detector scores select inputs for intervention, while input-specific answer-margin gradients determine the correction. The following sections evaluate whether this separation improves answer recovery.

4 AM-GARD: Testing the Recovery Principle through Intervention

Following Section 3, AM-GARD tests whether separating intervention selection from answer correction improves recovery. Detector scores select inputs for intervention, while input-specific answer-margin gradients determine the correction.

4.1 Intervention Gate

We fit linear logistic probes on estimation-split pre-attack ($t = 0$) and attacked ($t = 1$) representations, minimizing binary cross-entropy with an L_2 penalty on the probe weights. We divide each

coordinate by its estimation-set standard deviation to obtain $\bar{h}$. The detector and gate are

$$S_\ell(h) = \sigma(w_\ell^\top \bar{h} + b_\ell), \qquad D(u,q) = \mathbb{1}\{S_{\ell_d}(h_{\ell_d}(u,q)) > \tau\}. \tag{4}$$

On calibration data, ℓ_d is selected by detection rate at a specified false-positive rate on pre-attack inputs, and τ is set from pre-attack detector scores. The probes receive no answer-token or correction-vector supervision. Inputs with $D(u,q) = 0$ retain the original generation procedure.

4.2 Iterative Answer-Margin Correction

For a selected input, let $h_t = h_{\ell_s}(u,q) + v_t$ with $v_0 = 0$. Let $z(h_t)$ denote the first-response logits, $a_t = \arg\max_a z_a(h_t)$ the current leader, $a^{\text{obs}} = a_0$ the initial observed-input leader, and $\mathcal{C}_k(h_t)$ the highest-logit k tokens excluding a_t. For each competitor c, evaluate the margin and gradient from Section 3.2 at h_t for the fixed token pair (a_t, c), denoting them by $g_{t,c}$ and $n_{t,c}$. We select

$$d_{t,c} = \frac{g_{t,c}}{\|n_{t,c}\|_2}, \qquad c_t^* = \arg\min_{c \in \mathcal{C}_k(h_t)} d_{t,c}. \tag{5}$$

This selects the nearest linearized boundary among the candidates, adapting DeepFool's local linearization (Moosavi-Dezfooli et al., 2016) to hidden states.

Write (g_t, n_t, d_t) for the quantities associated with c_t^*. Following Proposition 3.1, apply

$$\delta_t = -\frac{g_t + \epsilon_r |g_t|}{\|n_t\|_2^2} n_t, \qquad v_{t+1} = v_t + \delta_t, \tag{6}$$

where $\epsilon_r \geq 0$ supplies relative margin slack beyond the linearized boundary. Gradient-norm denominators are numerically stabilized. After each update, we recompute the logits and stop when the leading token differs from the initial observed-input leader a^{obs} or after T updates. Otherwise, candidates and gradients are recomputed at the updated state.

An optional distance limit r returns the unmodified output if $d_0 = d_{0,c_0^*} > r$. Before any update, if the proposed step length $\alpha_t = (g_t + \epsilon_r |g_t|)/\|n_t\|_2$ exceeds r, iteration stops without executing that step and retains the displacement accumulated by earlier updates. Setting $r = \infty$ disables screening.

Answer generation. After T' executed updates, set $\Delta h = \beta v_{T'}$, where $\beta \geq 0$ controls generation-time strength separately from the search slack ϵ_r. This fixed vector is added at layer ℓ_s to the last processed position at every generation step. Boundary search is not repeated per token, and greedy decoding remains unchanged. The selected competitor is not forced as output; recovery is assessed from generated answers under Section 5, not from the stopping condition. Full inference pseudocode is provided in the Appendix B.

5 Experimental Setup

Models and comparisons. We evaluate Qwen2-VL-7B and LLaVA-1.5-7B with frozen backbones (Wang et al., 2024b; Liu et al., 2024a). Our primary comparisons use activation-based methods (Rimsky et al., 2024; Li et al., 2023a; Qiu et al., 2024; Belrose et al., 2023; Wang et al., 2025a; Lu et al., 2026); F3 purifies adversarial images, while DPS guides generation using partial-image responses (Zhang et al., 2025; Zhou et al., 2025) (Table 1). F3 operates through image-space purification, which differs from AM-GARD's gated representation-level intervention. Detailed implementations and adaptations of the comparison methods are provided in Appendix E.1.

AM-GARD variants. All variants selectively apply the same answer-margin correction to the hidden state, differing only in intervention selection. Outlier-Gate is fitted on pre-attack representations only, and Probe-Gate on both pre-attack and attacked representations. Probe-Gate is the primary deployed configuration; neither requires paired references at inference. Gold-Gate uses pre-attack answers for diagnostic intervention selection. Detailed definitions of the AM-GARD variants are provided in Appendix E.2.

Table 1: **Comparison methods and evaluation datasets.** Left: what decides *whether* to intervene, what sets *how* to correct, and where the correction acts. Right: abbreviations and datasets by answer format. Full settings are in Appendix E.

Comparison methods

	Whether					*How*			*Locus*		
			Det.								
Method	None	Impl.	Pre.	Atk.	Gold	Fix.	Inp.	Ans.	Img.	Rep.	Gen.
Activation steering											
CAA / ITI-head	✓					✓				✓	
SEA / LEACE	✓					✓				✓	
ASTRA		✓				✓				✓	
ARGUS				✓		✓				✓	
Visual adversarial defense											
F3	✓						✓		✓		
DPS		✓						✓			✓
AM-GARD											
+ Outlier-Gate			✓					✓		✓	
+ **Probe-Gate**				✓				✓		✓	
+ Gold-Gate					✓			✓		✓	

Abbreviations

Whether	None: every input; Impl.: set by the correction signal; Det.: detector on pre-attack samples (Pre.), pre-attack and attacked samples (Atk.), or gold labels (Gold)
How	Fix.: input-independent; Inp.: input-specific, answer-agnostic; Ans.: conditioned on the model's answer
Locus	Img.: image; Rep.: hidden state; Gen.: prompt or decoding

Evaluation datasets

Format	Datasets	Output
Open	GQA, VQAv2, A-OKVQA	one word
Binary	POPE, VSR, NaturalBench	yes / no
MC	MMBench, MMMU-Pro, MathVision	letter
Numeric	MathVista	value

Attack, splits, and datasets. We follow the attack and data-split protocol in Section 2. Main test results are averaged over five random seeds, with methods sharing inputs and splits within each seed and all settings fixed before testing. The datasets in Table 1 cover open-ended (Hudson & Manning, 2019; Goyal et al., 2017; Schwenk et al., 2022), binary (Li et al., 2023b; Liu et al., 2023; Li et al., 2024), multiple-choice (Liu et al., 2024b; Yue et al., 2025; Wang et al., 2024a), and numeric answers (Lu et al., 2024).

Evaluation. For method output $Y_{i,m}$, answer-agreement score s, and $p \in \{\text{all}, \text{flip}\}$, define

$$R^p_{m,s} = \frac{1}{|\mathcal{T}_p|} \sum_{i \in \mathcal{T}_p} s\big(Y_{i,m}(\tilde{x}_i, q_i), Y_i^0\big)\,,$$
$$C_{m,s} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} s\big(Y_{i,m}(x_i, q_i), Y_i^0\big)\,. \tag{7}$$

Here $\mathcal{T}_{\text{all}} = \mathcal{T}$ and $\mathcal{C}$ denote the attacked and pre-attack test sets, respectively. The flipped subset $\mathcal{T}_{\text{flip}} \subseteq \mathcal{T}$ is fixed across methods using preparation-time answer records. Scores use exact match (EM) of extracted, normalized answers or semantic agreement, where semantically equivalent answer forms are counted as matches under a fixed VQA-grading criterion; full scoring rules and the judging procedure are provided in Appendix E.5. No-op serves as the baseline, generating answers from identical evaluation inputs without intervention.

6 Experimental Results

We examine *(1)* whether AM-GARD improves recovery while preserving pre-attack responses, *(2)* how selection affects recovery, *(3)* when correction recovers rather than merely changes an answer, and *(4)* how far recovery extends beyond the first response token.

6.1 Recovery Requires Both Correction and Preservation

Selective answer-margin correction improves recovery. Table 2 shows improved recovery over no-op for Probe-Gate across most evaluated tasks. Across ten datasets, mean improvements are $8.1pp$ and $17.3pp$ for Qwen and LLaVA, respectively, with mean pre-attack response preservation of 96.3% and 94.1%. Preservation is lower on NaturalBench for both models.

Defense-oriented steering has variable recovery effects. The ASTRA and ARGUS adaptations improve recovery in some settings but often remain near or below no-op, whereas Probe-Gate improves recovery across a broader range of tasks. Their defensive objectives therefore do not directly translate into consistent recovery of pre-attack answers. This motivates evaluating the recovery effect of an intervention separately from its original defensive purpose.

Table 2: **Recovery and pre-attack response preservation** ($R^{\text{all}}_{\text{EM}}/C_{\text{EM}}$, %). Cells show recovery on all attacked inputs and preservation on pre-attack inputs (small red type), averaged over five seeds (42, 1337, 2026, 123, 7); no-op is the regeneration baseline. **Bold** and underline mark the best and second-best recovery among representation-level interventions with $C_{\text{EM}} \geq 90$, excluding Gold-Gate. F3 and DPS (gray) are shown as complementary image- and generation-level defenses.

	Open-ended			Binary			Multiple-choice			Numeric
Method	GQA	VQAv2	AOK	POPE	VSR	NatB	MMB	MMMU	MVsn	MVsta
Qwen2-VL-7B										
No-op	42.4 /100.0	63.0 /100.0	61.9 /100.0	84.9 /100.0	80.0 /100.0	69.5 /100.0	37.2 /100.0	33.1 /100.0	41.3 /100.0	47.7 /100.0
Visual adversarial defense										
F3	83.9 83.5	77.8 81.0	69.9 73.3	90.8 91.1	76.8 80.9	82.2 80.0	90.0 92.8	58.7 65.7	62.5 64.0	65.8 70.3
DPS	68.1 72.6	64.6 69.5	59.7 61.7	87.9 93.1	81.8 94.5	69.1 73.6	48.5 98.0	41.5 59.1	49.0 75.4	49.0 75.8
Activation steering										
CAA	42.2 /99.1	67.3 /99.0	61.1 /95.4	84.7 /99.6	81.8 /98.2	68.9 /98.4	37.4 /99.7	35.4 /94.6	41.6 /91.2	48.7 /97.3
ITI-head	46.0 /98.1	66.0 /98.6	62.8 /94.3	84.7 /99.6	80.4 /98.2	67.6 /99.7	40.3 /99.0	32.2 /85.2	43.0 /89.3	47.7 /96.7
SEA	36.8 /59.1	60.9 /82.9	59.8 /80.6	61.4 /60.5	65.0 /70.0	67.3 /84.1	43.7 /95.8	37.2 /76.8	34.3 /72.8	41.6 /86.8
LEACE	43.4 /96.2	0.0 /0.1	0.2 /0.3	20.7 /20.0	22.7 /23.2	13.7 /13.0	2.1 /4.4	4.3 /4.3	5.1 /7.5	1.3 /1.3
ASTRA	45.2 /97.4	64.7 /98.5	61.0 /95.6	53.0 /64.8	60.5 /80.5	60.3 /93.7	39.6 /81.2	35.7 /86.3	42.8 /86.2	45.5 /96.5
ARGUS	42.6 /99.7	63.7 /99.9	60.3 /99.6	84.7 /100.0	**82.3** /99.1	**70.5** /100.0	36.7 /99.9	32.8 /99.4	42.2 /99.1	45.8 /99.7
AM-GARD										
+ Outlier-Gate	45.7 /91.0	64.0 /94.3	62.8 /97.7	74.2 /99.8	60.0 /98.6	68.9 /95.6	38.9 /91.3	39.4 /94.6	43.5 /96.1	46.8 /88.1
+ **Probe-Gate**	**66.6** /96.5	**70.3** /95.7	**64.7** /99.1	**87.8** /98.5	**82.3** /98.6	72.4 /87.0	**57.0** /96.2	**39.8** /97.0	**49.2** /97.4	**51.9** /96.8
+ Gold-Gate	89.2 /100.0	89.2 /100.0	85.6 /100.0	96.7 /100.0	87.3 /100.0	93.7 /100.0	89.9 /100.0	70.0 /100.0	79.7 /100.0	89.0 /100.0
LLaVA-1.5-7B										
No-op	17.4 /100.0	17.3 /100.0	14.8 /100.0	28.6 /100.0	35.5 /100.0	23.5 /100.0	31.0 /100.0	23.6 /100.0	53.6 /100.0	41.2 /100.0
Visual adversarial defense										
F3	90.2 91.2	85.9 90.0	86.0 86.0	93.5 94.6	88.0 89.3	86.9 89.8	93.3 94.0	62.1 63.6	84.5 86.0	58.8 76.2
DPS	72.0 78.8	72.0 79.8	65.8 72.6	76.5 78.2	55.7 61.1	70.5 74.9	51.7 91.7	26.4 55.5	58.5 67.3	38.8 52.5
Activation steering										
CAA	17.0 /98.5	17.4 /98.4	15.7 /97.3	24.2 /98.9	70.9 /81.0	**46.2** /90.9	34.2 /97.5	24.6 /70.9	57.4 /89.5	51.2 /80.0
ITI-head	34.7 /94.8	**30.4** /96.8	23.1 /96.7	21.1 /99.2	72.6 /77.2	60.5 /79.7	36.8 /96.6	25.2 /74.6	53.3 /87.6	46.3 /86.2
SEA	13.6 /59.9	15.9 /98.0	25.7 /82.5	45.8 /98.0	25.4 /96.3	31.4 /94.7	35.5 /92.4	25.5 /73.9	35.3 /77.0	40.0 /67.5
LEACE	18.2 /98.6	7.6 /12.9	8.4 /20.8	26.3 /98.7	12.3 /11.7	36.1 /35.7	9.4 /16.2	32.1 /71.8	21.5 /22.1	3.8 /2.5
ASTRA	20.8 /97.9	18.1 /98.0	15.4 /97.2	24.4 /98.9	60.4 /82.0	34.2 /92.6	35.1 /96.4	24.6 /69.1	49.0 /88.0	40.0 /82.5
ARGUS	17.5 /99.8	16.3 /99.8	15.2 /99.6	28.3 /99.8	41.8 /98.8	24.0 /99.8	31.5 /99.3	20.3 /98.8	**52.7** /98.9	**45.0** /98.8
AM-GARD										
+ Outlier-Gate	24.2 /91.4	19.3 /93.0	23.7 /92.2	30.4 /98.9	**42.6** /93.9	27.0 /94.5	30.2 /94.9	27.6 /98.5	47.5 /97.9	36.2 /91.2
+ **Probe-Gate**	**45.3** /96.6	30.0 /96.4	**39.2** /94.5	**74.0** /90.5	39.6 /94.2	54.6 /87.0	**59.1** /94.5	**29.1** /90.6	51.6 /97.9	37.5 /98.8
+ Gold-Gate	84.1 /100.0	65.3 /100.0	77.2 /100.0	94.2 /100.0	83.1 /100.0	87.0 /100.0	89.8 /100.0	68.2 /100.0	85.8 /100.0	47.5 /100.0

Recovery gains vary across models and answer formats. Figure 4 shows the relative benefit of input-specific and fixed-direction steering is task-dependent. Probe-Gate improves recovery across open-ended and most multiple-choice tasks, while fixed-direction methods can achieve higher recovery on some LLaVA binary tasks. Answer format alone does not explain this variation: on the same mathematical tasks, MathVision and MathVista, Probe-Gate improves recovery for Qwen but not LLaVA. Thus, the intervention effectiveness varies across models even within the same task.

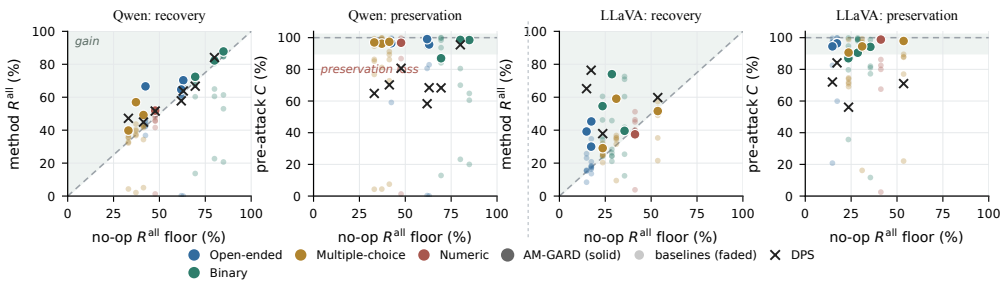


Figure 4: **Recovery beyond no-op versus pre-attack response preservation.** Points are method–dataset pairs; crosses are exploratory DPS runs.

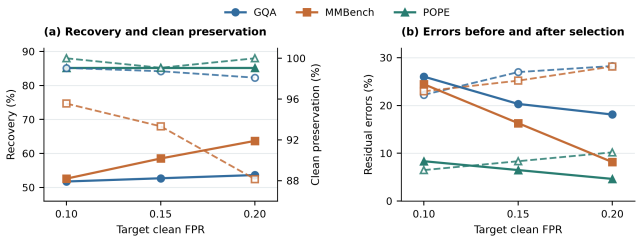


Figure 5: **Gate operating point and failure-stage handoff.** Qwen2-VL-7B on GQA, MMBench, and POPE. (a) Recovery (solid, left axis) and pre-attack response preservation (dashed, right axis). (b) Missed-intervention (solid) and post-intervention (dashed) errors across target false positive rate (FPR).

6.2 Selection Determines Whether Correction Helps

Selection controls the recovery–preservation tradeoff. We vary the target FPR on pre-attack inputs while keeping the answer-margin corrector fixed (Figure 5a). Broader selection improves recovery most clearly on MMBench, but also reduces preservation. GQA improves modestly, while POPE changes little. Thus, expanding intervention coverage does not uniformly improve recovery; its benefit and preservation cost depend on the task.

Error sources cross as selection broadens. Figure 5b separates missed-intervention errors from errors remaining after intervention. Across all three tasks, broader selection reduces missed interventions, while post-intervention errors persist or increase. The two error curves cross, shifting the larger contribution to residual error from selection to correction. Beyond this crossover, further recovery improvement depends increasingly on *how* selected answers are corrected, rather than simply expanding intervention coverage.

6.3 Correction Must Identify a Useful Answer Boundary

We examine how competitor selection affects answer recovery while holding the gate and remaining correction settings fixed.

Competitor selection affects recovery. We compare normalized boundary-distance selection with *highest-logit competitor selection*, which selects the highest-logit token excluding the current leading token. The former uses both the token margin and its gradient norm; the latter uses only logit ranking.

Highest-logit competitor selection changes more answers but yields lower recovery and more off-target answer changes (Table 3). Normalized boundary-distance selection improves recovery despite fewer answer changes. Thus, competitor selection affects whether correction recovers the pre-attack answer rather than merely changes the output.

Table 3: **Competitor selection and answer recovery.** Qwen GQA; three-seed mean ± sample standard deviation (%), with identical Probe-Gate decisions. Off-target answer change denotes arrival at a competitor other than the pre-attack answer.

Metric	Normalized boundary distance	Highest-logit competitor
$R_{\mathrm{EM}}^{\mathrm{all}}$	60.7 ± 3.7	57.6 ± 3.4
Answer change	49.8 ± 3.1	56.0 ± 4.1
Off-target answer change	17.3 ± 2.4	23.6 ± 2.7

Recovery depends on identifying a useful boundary. Stronger steering produces saturating recovery, while expanding the candidate set has little effect (Table 14). In a separate Gold-Gate diagnostic, supplying the saved pre-attack first token as the correction target further improves recovery. These results show that correction strength and candidate coverage alone do not determine recovery: reaching a nearby boundary does not necessarily recover the pre-attack answer.

Remaining recovery gaps reflect both selection and correction. Table 2 shows that large Gold–Probe gaps indicate missed recoverable cases, whereas low Gold-Gate recovery points to limitations after selection. Recovery can also decrease under unnecessary intervention. Thus, end-to-end recovery depends on both *when* to intervene and *how* to correct the answer. Qualitative examples are in Appendix G.2.

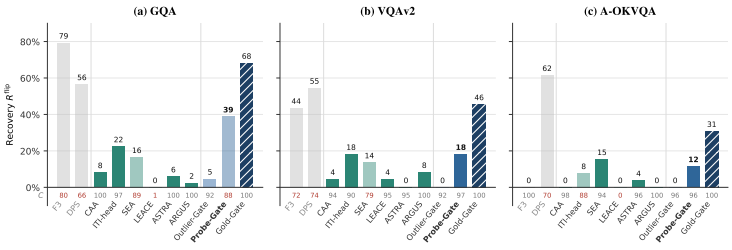


Figure 6: **Sentence-form answer-content recovery on Qwen.** Recovery on the flipped subset and pre-attack response preservation across GQA, VQAv2, and A-OKVQA. All AM-GARD variants use oracle answer-position localization from saved pre-attack responses.

6.4 Beyond First-Response-Token Recovery

Our main formulation computes correction at the first response-token prediction state, but sentence-form answer content may appear later. We use oracle, reference-based localization to identify the answer-bearing position from the saved pre-attack response and apply the same answer-margin correction at that position. All variants share this oracle localization; Gold-Gate additionally uses pre-attack answers for intervention selection. Full protocols are provided in Appendix J. Figure 6 shows that Probe-Gate recovers answer content across the three datasets while largely preserving pre-attack responses. Gold-Gate yields higher recovery under the same oracle localization setting. These results support applying answer-margin correction beyond the first response token when the answer position is known.

7 Related Work

7.1 Defenses against Visual Adversarial Attacks

Recent LVLM defenses intervene at different stages of inference. F3 uses cross-modal attention for image purification (Zhang et al., 2025), while DPS uses partial-image responses to guide generation on the original image (Zhou et al., 2025). Both provide inference-time defenses against visual adversarial perturbations.

7.2 Activation Steering for Behavioral Control

Activation steering modifies internal representations to control model behavior. CAA and ITI apply activation interventions in language models (Rimsky et al., 2024; Li et al., 2023a), while SteerVLM extends this setting to VLMs (Sivakumar et al., 2025). ASTRA and ARGUS further use activation steering for multimodal safety and defense (Wang et al., 2025a; Lu et al., 2026). Our experiments adapt these approaches as comparison methods for adversarial answer recovery.

8 Conclusion

We used controlled activation interventions to understand which internal changes support behavioral recovery in LVLMs. At the answer-prediction state, input-specific answer-margin directions recover more pre-attack answers than detector directions under matched input selection and update norms. This distinguishes directions useful for attack detection from those effective for answer recovery.

Building on this understanding, AM-GARD combines detector-based selection with answer-margin correction, improving recovery over no-op on most Qwen2-VL and LLaVA tasks while largely preserving pre-attack responses. Our experiments further clarify the conditions for effective correction: broader selection reduces missed interventions but leaves post-intervention failures, while boundary-distance selection improves recovery despite fewer answer changes. Thus, recovery depends not simply on increasing intervention coverage or changing outputs, but on which inputs are selected and which competitors guide correction. Together, these results connect an understanding of internal intervention effects with practical improvements in behavioral recovery.

Reproducibility Statement

The experimental setup, including the attack setting, intervention formulation, information-access assumptions, and data splits, is specified in Section 2. AM-GARD is fully specified in Section 4, with inference pseudocode provided in Appendix B. Proofs of Propositions 3.1 and 3.2 are provided in Appendix C. Models, datasets, and baseline methods are summarized in Section 5, while Appendix E details data-split construction, baseline adaptations, answer-scoring procedures, subset definitions, and calibration procedures. Appendix G provides additional selection and correction diagnostics, including sensitivity and controlled sweep analyses. The number of random seeds for each result is stated in the corresponding table or text. An anonymized implementation covering attack generation, intervention, and evaluation is included in the supplementary material.

Use of Generative AI

Generative AI tools were used to assist with English translation, language refinement, and editorial feedback on the presentation and organization of the manuscript. They were also used for limited writing assistance, including improving readability, suggesting organizational revisions, formatting references, and assisting with code for figures and plotting. These tools were not used to formulate mathematical claims, develop or write proofs, propose or refine research hypotheses, design the proposed method, implement the method, preprocess or reformat datasets, or scientifically interpret experimental results. Separately, Claude Opus was used as a fixed semantic-equivalence judge for the auxiliary semantic-agreement evaluation described in Appendix E.5, under a predefined prompt and grading criterion. All AI-assisted material was independently reviewed by the authors, who take full responsibility for the accuracy, validity, and final content of the manuscript.

References

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 274–283, 2018.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Nicholas Carlini, Milad Nasr, Christopher A. Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei Koh, Daphne Ippolito, Florian Tramèr, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24625–24634, 2024.

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*, pp. 6904–6913, 2017.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. NaturalBench: Evaluating

vision-language models on natural adversarial samples. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Kenneth Li, Oam Patel, Fernanda Vi'egas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023b.

Fangyu Liu, Guy Emerson, and Nigel Collier. Visual spatial reasoning. *Transactions of the Association for Computational Linguistics*, 11:635–651, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, et al. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, 2024b.

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.

Weikai Lu, Ziqian Zeng, Kehua Zhang, Haoran Li, Huiping Zhuang, Ruidong Wang, Cen Chen, and Hao Peng. Argus: Defending against multimodal indirect prompt injection via steering instruction-following behavior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31–40, 2026.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Chengzhi Mao, Scott Geng, Junfeng Yang, Xin Wang, and Carl Vondrick. Understanding zero-shot adversarial robustness for large-scale models. In *International Conference on Learning Representations*, 2023.

Mujtaba Hussain Mirza, Antonio D'Orazio, Odelia Melamed, and Iacopo Masi. A provable energy-guided test-time defense boosting adversarial robustness of large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2574–2582, 2016.

Yifu Qiu, Zheng Zhao, Yftah Ziser, Anna Korhonen, Edoardo M. Ponti, and Shay B. Cohen. Spectral editing of activations for large language model alignment. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering Llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024. doi: 10.18653/v1/2024.acl-long.828.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 3677–3685, 2023.

Christian Schlarmann, Naman Deep Singh, Francesco Croce, and Matthias Hein. Robust CLIP: Unsupervised adversarial fine-tuning of vision embeddings for robust large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 43685–43704, 2024.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, 2022.

Anushka Sivakumar, Andrew Zhang, Zaber Ibn Abdul Hakim, and Chris Thomas. SteerVLM: Robust model control through lightweight activation steering for vision language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23640–23665, 2025. doi: 10.18653/v1/2025.findings-emnlp.1285.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248*, 2023.

Han Wang, Gang Wang, and Huan Zhang. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29947–29957, June 2025a.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. Measuring multimodal mathematical reasoning with the MATH-Vision dataset. In *Advances in Neural Information Processing Systems*, volume 37, 2024a.

Peng Wang, Shuai Bai, Sinan Tan, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Zeyu Wang, Cihang Xie, Brian Bartoldson, and Bhavya Kailkhura. Double visual defense: Adversarial pre-training and instruction tuning for improving vision-language model robustness. *arXiv preprint arXiv:2501.09446*, 2025b.

Lyucheng Wu, Mengru Wang, Ziwen Xu, Tri Cao, Nay Oo, Bryan Hooi, and Shumin Deng. Automating steering for safe multimodal large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 792–814, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.41.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. MMMU-Pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186, 2025.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Yudong Zhang, Ruobing Xie, Yiqing Huang, Jiansheng Chen, Xingwu Sun, Zhanhui Kang, Di Wang, and Yu Wang. Fighting fire with fire (f3): A training-free and efficient visual adversarial example purification method in lvlms. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12266–12275, 2025.

Qi Zhou, Dongxia Wang, Tianlin Li, Yun Lin, Yang Liu, Jin Song Dong, and Qing Guo. Defending LVLMs against vision attacks through partial-perception supervision. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79254–79280, 2025.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A Extended Related Work

Visual adversarial robustness. Visual attacks can substantially alter LVLM responses, motivating defenses at different stages of the multimodal pipeline (Carlini et al., 2023; Schlarmann & Hein, 2023; Cui et al., 2024). TeCoA and FARE improve adversarial robustness at the vision-encoder level (Mao et al., 2023; Schlarmann et al., 2024), whereas F3 purifies adversarial images and DPS guides generation using partial visual evidence (Zhang et al., 2025; Zhou et al., 2025). These approaches establish effective intervention points from visual representations to generation. Our focus is different: visual robustness does not by itself characterize whether the downstream LVLM returns to its own pre-attack response, motivating our analysis at the answer-prediction state.

Representation-level intervention. Representation Engineering and Activation Addition establish that internal representations can be directly manipulated to control model behavior (Zou et al., 2023; Turner et al., 2023). Subsequent methods including CAA, ITI, SteerVLM, and ASTRA develop activation-level interventions for language and vision–language models (Rimsky et al., 2024; Li et al., 2023a; Sivakumar et al., 2025; Wang et al., 2025a), while VTI steers multimodal latent representations to reduce LVLM hallucination (Liu et al., 2025). These works demonstrate that representation-level interventions can change behavior, but behavioral change does not necessarily imply recovery: an attacked response must be moved toward its instance-specific pre-attack behavior rather than merely altered.

Detection, selective intervention, and adaptive attacks. Selective defenses introduce a further distinction between deciding *whether* to intervene and determining *how* to intervene; ARGUS, for example, combines an intervention decision with representation-level steering (Lu et al., 2026). Adaptive-attack literature further shows that detection and preprocessing mechanisms can themselves become part of the attack surface (Athalye et al., 2018; Tramèr et al., 2020). These works motivate treating intervention selection and correction as distinct components, while evaluating selection under defense-aware attacks. Our controlled analyses further show that a signal useful for identifying an attacked state need not provide an effective direction for restoring pre-attack behavior, motivating AM-GARD's separation of intervention selection from answer-conditioned correction.

B AM-GARD Pseudocode

Algorithms 1 and 2 summarize the estimation/calibration and inference procedures corresponding to Section 4. Learned components are fitted on the estimation split, while the detection and steering layers, gate threshold, and correction settings are selected on the calibration split. All settings are fixed before test evaluation.

C Geometry of Local Answer Boundaries

We prove the two propositions in Section 3.2. Fix the token pair (a, c) and let $g = g(h) > 0$ and $n = \nabla_h g(h) \neq 0$. The boundary considered here is the linearized hyperplane $g + n^\top \delta = 0$.

C.1 Proof of Proposition 3.1

Every feasible displacement satisfies $-n^\top \delta \geq g$. By Cauchy–Schwarz,

$$g \leq -n^\top \delta \leq \|n\|_2 \|\delta\|_2, \tag{8}$$

Algorithm 1 AM-GARD estimation and calibration

Require: Estimation split $\mathcal{E}$, calibration split $\mathcal{D}_{\text{cal}}$, candidate detection layers $\mathcal{L}_d$, correction grid $\mathcal{G}$, target pre-attack FPR η
Ensure: Fitted gate $(S_{\ell_d}, \ell_d, \tau)$ and correction settings $(\ell_s, k, T, \epsilon_r, \beta, r)$
// Estimation: fit attack-detection probes
for all $\ell \in \mathcal{L}_d$ **do**
  $\sigma_\ell \leftarrow$ coordinate-wise standard deviation of $\{h_\ell(u, \eta)\}$ over $\mathcal{E}$
  Fit S_ℓ by logistic regression on $\{(h_\ell(u, \eta)/\sigma_\ell, t)\}_{\mathcal{E}}$ ▷ pre-attack t=0, attacked t=1
end for
// Calibration: select the gate operating point
for all $\ell \in \mathcal{L}_d$ **do**
  Set τ_ℓ from pre-attack scores on $\mathcal{D}_{\text{cal}}$ to satisfy target FPR η
  $\text{TPR}_\ell \leftarrow \text{mean}[S_\ell(\text{attacked}) > \tau_\ell]$ on $\mathcal{D}_{\text{cal}}$
end for
$\ell_d \leftarrow \arg\max_\ell \text{TPR}_\ell$
$\tau \leftarrow \tau_{\ell_d}$
$D(u, \eta) \leftarrow \mathbf{1}[S_{\ell_d}(h_{\ell_d}(u, \eta)) > \tau]$
// Calibration: select correction settings
$(\ell_s, k, T, \epsilon_r, \beta, r) \leftarrow \arg\max_{\mathcal{G}}$ recovery on $\mathcal{D}_{\text{cal}}$ subject to the preservation criterion
return $(S_{\ell_d}, \ell_d, \tau, \ell_s, k, T, \epsilon_r, \beta, r)$

Algorithm 2 AM-GARD inference

Require: Observed image u, question q, frozen model, fitted gate, settings $(\ell_s, k, T, \epsilon_r, \beta, r)$
Ensure: Generated response
if $D(u, q) = 0$ **then**
  return the unmodified response
end if
$h \leftarrow h_{\ell_s}(u, q)$; $v \leftarrow 0$; $a^{\text{obs}} \leftarrow \arg\max_a z_a(h)$
for $t = 0, \ldots, T-1$ **do**
  $h_t \leftarrow h + v$; $a_t \leftarrow \arg\max_a z_a(h_t)$
  if $a_t \neq a^{\text{obs}}$ **then**
    break
  end if
  Form $\mathcal{C}_k(h_t)$ from the top-k tokens excluding a_t
  for all $c \in \mathcal{C}_k(h_t)$ **do**
    $g_{t,c} \leftarrow z_{a_t}(h_t) - z_c(h_t)$
    $n_{t,c} \leftarrow \nabla_{h_t} g_{t,c}$
    $d_{t,c} \leftarrow g_{t,c}/\|n_{t,c}\|_2$
  end for
  $c_t^* \leftarrow \arg\min_{c \in \mathcal{C}_k(h_t)} d_{t,c}$
  $(g_t, n_t, d_t) \leftarrow (g_{t,c_t^*}, n_{t,c_t^*}, d_{t,c_t^*})$
  if $t = 0$ **and** $d_t > r$ **then**
    return the unmodified response
  end if
  $\alpha_t \leftarrow (g_t + \epsilon_r|g_t|)/\|n_t\|_2$
  if $\alpha_t > r$ **then**
    break ▷ Retain the accumulated displacement
  end if
  $v \leftarrow v - \alpha_t n_t/\|n_t\|_2$
end for
$\Delta h \leftarrow \beta v$
return response generated with Δh at layer ℓ_s

so $\|\delta\|_2 \geq g/\|n\|_2$. The displacement $\delta^* = -gn/\|n\|_2^2$ lies on the boundary and attains this lower bound. Equality requires δ to be a positive multiple of $-n$, and the boundary condition uniquely determines that multiple. Hence δ^* is the unique minimum-norm solution.

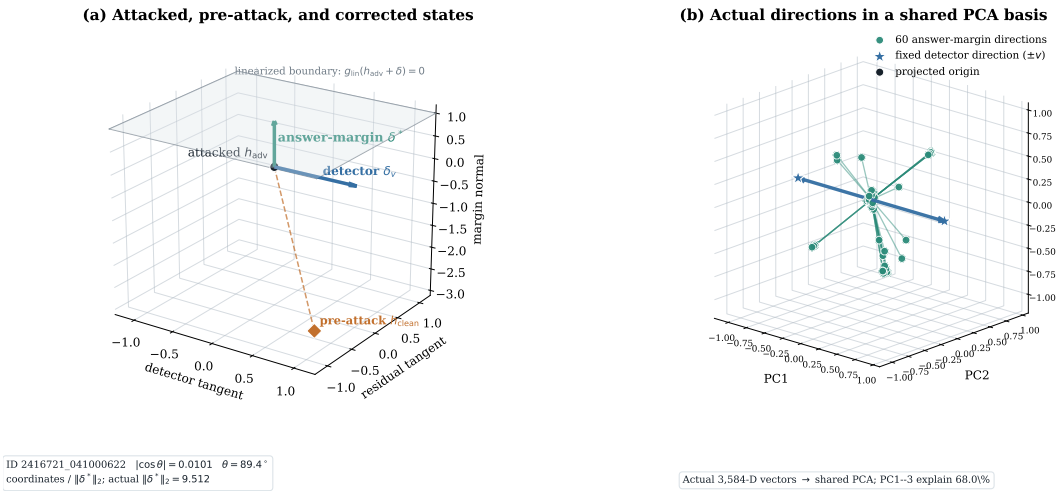


Figure 7: **Three-dimensional projection of the raw answer-correction vectors.** (a) The example at the median absolute detector–margin cosine, projected onto an orthonormal basis containing both observed displacements. The black point and orange diamond denote the attacked and paired pre-attack states; the green and blue arrows denote the answer-margin and matched-norm detector corrections. The gray plane is the local linearized answer boundary $g_i + n_i^\top \delta = 0$. Dividing all coordinates by $\|\delta^*\|_2$ preserves relative lengths and angles. (b) A common PCA projection fitted directly to the 60 unit answer-margin displacements and their detector-direction counterparts. Because the detector is fixed, the blue arrows and stars overlap at at most two sign-selected locations. The panel reports the variance explained by the first three components.

C.2 Proof of Proposition 3.2

For a unit direction v, write $\delta = -tv$ with $t \geq 0$. The linearized margin is

$$g - tn^\top v = g - t\|n\|_2 \cos\theta. \tag{9}$$

If $\cos\theta > 0$, the first boundary intersection occurs at $t_v = g/(\|n\|_2 \cos\theta)$. Dividing by the minimum distance $t_* = g/\|n\|_2$ from Proposition 3.1 gives $t_v/t_* = 1/\cos\theta$. If $\cos\theta \leq 0$, the margin stays positive for every $t \geq 0$, so the ray does not reach the linearized boundary.

The empirical alignment analysis uses $|\cos\theta|$. This corresponds to the alignment available after choosing the better of the two signs, without assuming that the implemented correction uses that favorable sign.

Both propositions concern a linearized boundary for a fixed token pair. They do not guarantee the shortest distance to the nonlinear model's actual boundary, convergence of repeated updates, or agreement between the competitor and the pre-attack token. We use them as geometric reference points for direction comparisons and measure recovery from regenerated outputs separately.

C.3 Three-Dimensional Projection of the Raw Vectors

Figure 7 projects the original 3,584-dimensional vectors saved for Qwen GQA at layer 24. For each example, we use the paired pre-attack and attacked answer-prediction hidden states, the answer-margin gradient for the nearest competing token, and the linear detector direction learned on the estimation split. We compute the answer-margin displacement as the minimum-norm update to the linearized boundary. The detector displacement has the same norm and uses the sign that decreases the selected margin. The pre-attack state is shown only as a geometric reference and is not used to construct either correction direction. Thus, the plotted points and arrows are orthogonal projections of stored vectors rather than coordinates reconstructed from cosine summaries.

D CONTROLLED-ANALYSIS DETAILS

D.1 EXPERIMENTAL SETTINGS FOR VISUAL PREDICTION AND TOKEN RESTORATION

Matched CLIP–LVLM Comparison We compare `openai/clip-vit-large-patch14-336` with `llava-hf/llava-1.5-7b-hf`. The shared vision-tower weights were checked in float16: all 391 parameter tensors agree exactly. The coarse-label experiment contains 495 images drawn from 99 ImageNet classes, grouped into ten categories. Three readouts are evaluated: CLIP closed-set classification, LVLM scoring of candidate first tokens, and LVLM free generation. CLIP class prompts use "a photo of a {category}." The LVLM prompt is "What is the main object in this image? Answer with one word." Generation uses at most eight new tokens. Free responses are mapped to the candidate categories by substring matching, with nearest CLIP text-embedding matching as a fallback. Thus, this analysis measures category agreement rather than unrestricted response-string EM.

Attacks are applied to the 281 examples correctly classified by all three readouts on clean inputs. The attack minimizes the CLIP ground-truth category logit using 30 sign-gradient steps with an L_∞ budget of $16/255$ and step size $4/255$, starting from zero perturbation. Pixel values are clipped to their valid range. This CLIP-specific attack differs from the LPIPS-budget answer-token attack in the main recovery experiments. All 281 examples change their CLIP prediction; 162 also change their mapped LVLM answer.

Input purification in this run uses Gaussian blur with $\sigma = 1.5$ and a 7×7 kernel. Among the 162 LVLM-flipped examples, 138 recover the pre-attack category after purification. Among the 119 examples whose LVLM category survives the attack, 117 remain unchanged and two change after purification. Figure 2(a) reports the CLIP and LVLM prediction changes on the same 281 examples. The $138/162$ recovery result uses the LVLM-flipped population. The separate statistic $232/242$ conditions on successful CLIP purification and therefore uses a different denominator.

Single-Token Restoration on CLEVR Counting The restoration analysis uses `Qwen/Qwen2-VL-7B-Instruct` on CLEVR counting (Johnson et al., 2017), with the prompt "How many objects are in this image? Answer with only the number." Labels in `clip-benchmark/wds_vtab-clevr_count_all` are converted to object counts by adding three. Decoding is greedy; inference uses bfloat16, with logits cast to float32 when computing effects. Examples must have a correct pre-attack answer and a changed generated answer and first answer token after attack. The retained set contains 300 examples.

This experiment uses a separate pixel-space L_∞ attack. It runs 15 PGD steps per budget on the pre-attack answer-token margin, with budgets tried in the order $6/255$, $12/255$, and $20/255$. Of the retained examples, 296 flip at $6/255$ and four at $12/255$. LPIPS was not recorded for this experiment, so these settings are reported separately from the main LPIPS-budget protocol.

The intervention unit is one visual-token vector at the visual injection interface. Let a_i^0 and a_i^{adv} denote the pre-attack and attacked first answer tokens, held fixed during the replacement analysis, and define the contrastive logit

$$d_i = z_{a_i^{\mathrm{adv}}} - z_{a_i^0}. \tag{10}$$

For token position j, we replace its attacked vector with the paired pre-attack vector at the same position, keeping the remaining attacked vectors unchanged. If the resulting contrastive logit is $d_i^{(j)}$, the restoration effect is $e_{ij} = d_i^{(j)} - d_i^{\mathrm{adv}}$. Its concentration in the five most influential positions is

$$M_{5,i} = \frac{\sum_{j \in \mathrm{Top5}(|e_i|)} |e_{ij}|}{\sum_j |e_{ij}|}. \tag{11}$$

The reported strata are $M_{5,i} < 0.25$, $0.25 \le M_{5,i} < 0.50$, and $M_{5,i} \ge 0.50$, containing 259, 41, and zero examples, respectively. The median is 0.147. These are normalized absolute single-token effects and do not assume that joint replacements have additive effects.

A separate positive control deletes one annotated object by replacing its region with the background color. On 30 such examples, the token with the largest restoration effect lies in the ground-truth object region in all cases. This control checks whether localized image changes can produce localized restoration effects under the same measurement procedure.

D.2 Layerwise Probes for Attack Detection

Data and representations. We use the existing Qwen2-VL-7B full caches: 4,533 pre-attack–attacked pairs for GQA, 3,878 for POPE, and 3,719 for MMBench. Both representations are available for every pair. We retain attack attempts regardless of success and label representations by input provenance (pre-attack: 0; attacked: 1), rather than answer correctness or recoverability. The representation is the answer-prediction state, i.e., the language-model block output at the last input position predicting the first response token. Cached block indices range from 0 to 27, with 3,584 features per block; an embedding-only layer is not included. No attack or VLM forward pass is rerun for this analysis.

Training and splits. We reproduce the AM-GARD resplitting procedure with seeds 42, 1337, 2026, 7, and 123. For N pairs, we shuffle their original row indices using NumPy's default random generator and allocate $\lfloor 0.4N \rfloor$ pairs to estimation, $\lfloor 0.3N \rfloor$ to calibration, and the remainder to test. Both members of a pair stay in the same split. The grouping unit is the cached pair ID; no additional grouping by image is introduced. For each layer and seed, we subtract the estimation-set feature mean and divide by its standard deviation plus 10^{-8}. The original gate divides by the standard deviation without centering. With an unpenalized intercept, centering is an intercept reparameterization of the same logistic-regression objective; we use it here to improve numerical convergence and refit the probes rather than reuse historical scores. We fit logistic regression with $C = 0.5$, an intercept, the lbfgs solver, a maximum of 2,000 iterations, and no class weighting, matching the gate implementation. The same fitted probe is evaluated on calibration and test examples.

Reporting. The main figure reports test AUROC at every layer as the five-seed mean with sample standard deviation. The shading is not a confidence interval: test examples can overlap across seed resplits. No layer is chosen using test performance. Calibration AUROCs are stored in the original outputs under the development field and are distinguished from the test curves. We check pair coverage, split disjointness by pair ID, and solver convergence, and retain source-row hashes, software versions, split IDs, per-layer scores, and iteration counts with the analysis outputs.

D.3 Evaluation Protocol for the Controlled Direction Comparison

Each Qwen GQA layer-24 calibration run contains 105 examples. Saved pre-attack answer information selects 52, 54, and 56 examples for intervention in seeds 42, 1337, and 2026, respectively. We report both the full evaluation population and the actual intervention subset.

Table 4: **Controlled direction comparison on Qwen GQA at layer 24.** Values are pre-attack answer agreement (%), reported as the mean and standard deviation across three seeds for the intervention conditions. Each seed contains 105 evaluation examples, of which 52/54/56 are selected for intervention. Both directions use the same diagnostic intervention mask defined from saved pre-attack answers and the same per-example update norm. The nonzero full-set no-op value comes from examples outside the intervention mask.

Direction	All examples	Intervened only
No-op	48.6	0.00
Detector direction	49.5 ± 1.9	1.85 ± 0.07
Answer-margin direction	$\mathbf{91.4} \pm 1.6$	$\mathbf{83.39} \pm 2.64$

The detector direction recovers one example in each seed. The answer-margin direction recovers 44/52, 46/54, and 45/56 examples, respectively. The conditional denominator is the intervention mask at evaluation, not the recorded-flip subset used during attack preparation.

E Experimental Setup Details

This section provides additional details on comparison-method adaptations, AM-GARD variants, data splits, and the evaluation and scoring procedures used in the main experiments.

E.1 Models and Comparison Methods

We evaluate Qwen2-VL-7B and LLaVA-1.5-7B with frozen backbones (Wang et al., 2024b; Liu et al., 2024a). All activation-based comparison methods are adapted to the same pre-attack–attacked evaluation setting used by AM-GARD.

Activation-based comparisons. CAA (Rimsky et al., 2024) and ITI-head (Li et al., 2023a) apply fixed activation directions during generation. SEA (Qiu et al., 2024) and LEACE (Belrose et al., 2023) modify representations through fixed editing or erasure operators. ASTRA (Wang et al., 2025a) uses an activation-dependent steering rule, while ARGUS (Lu et al., 2026) combines an explicit intervention decision with a fixed steering direction. We use these methods to compare different choices of *whether* to intervene and *how* to modify the internal representation.

Visual adversarial defenses. F3 (Zhang et al., 2025) and DPS (Zhou et al., 2025) are included as complementary inference-time defenses against visual adversarial perturbations. F3 intervenes at the image level through purification, whereas DPS modifies generation using responses to partial image views. These methods differ from the activation-based comparisons in both intervention locus and correction mechanism. Additional input-level and adaptive-attack comparisons are provided in Appendix H and I.

E.2 AM-GARD Variants

All AM-GARD variants share the same input-specific answer-margin correction and differ only in intervention selection.

Outlier-Gate. Outlier-Gate is fitted using pre-attack representations only and selects inputs whose representations deviate from the pre-attack distribution.

Probe-Gate. Probe-Gate is trained to distinguish pre-attack from attacked representations and is the primary deployed configuration. At inference, it uses only the observed input and model representation; the paired pre-attack input and pre-attack response are not available.

Gold-Gate. Gold-Gate is a diagnostic condition that uses saved pre-attack answers to determine whether correction is needed. It is used to separate selection errors from errors that remain after intervention and is not a deployable configuration.

E.3 Attack and Data Splits

We use the white-box image-attack protocol defined in Section 2, keeping the question and decoding rule fixed; the maximum LPIPS search budget is 0.05. The full attacked population retains both successful and unsuccessful attack attempts.

Pre-attack–attacked pairs are partitioned into disjoint estimation, calibration, and test splits of size $4{:}3{:}3$. Both members of each pair remain in the same split. Learned components and feature statistics are fitted using the estimation split, while intervention layers, thresholds, and other operating settings are selected on the calibration split. All settings are fixed before test evaluation. Main results are averaged over five random seeds (42, 1337, 2026, 123, and 7), with methods evaluated on the same inputs and splits within each seed.

E.4 Evaluation Sets and Metrics

We evaluate recovery on two attacked test sets. The full attacked set, $\mathcal{T}_{\text{all}}$, includes all attacked test inputs, regardless of whether the attack changes the recorded response. The recorded-flip subset, $\mathcal{T}_{\text{flip}} \subseteq \mathcal{T}_{\text{all}}$, contains examples whose preparation-time attacked response differs from the corresponding pre-attack response. This subset is fixed before method evaluation and shared across methods.

Pre-attack test inputs, denoted by $\mathcal{C}$, are used to measure response preservation. For method m, answer-agreement function s, and $p \in \{\text{all}, \text{flip}\}$, we define

$$R^p_{m,s} = \frac{1}{|\mathcal{T}_p|} \sum_{i \in \mathcal{T}_p} s\big(Y_{i,m}(\tilde{x}_i, q_i), Y_i^0\big)\,,$$
$$C_{m,s} = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} s\big(Y_{i,m}(x_i, q_i), Y_i^0\big)\,, \tag{12}$$

where Y_i^0 is the saved pre-attack response. No-op regeneration on the same evaluation inputs serves as the recovery baseline.

E.5 Answer Scoring

We report LLM-judged semantic agreement alongside EM. For semantic evaluation, Claude Opus, invoked in headless command-line mode, compares one-word answers extracted from the evaluated and saved pre-attack responses without rule-based normalization. The judge returns a strict-JSON equivalence label under a fixed, liberal VQA-grading standard (Prompt E.5). Pairs absent from the judgment set are counted as non-equivalent. We use `claude-opus-4-8` as the semantic-equivalence judge with a fixed grading prompt.

Answer formats. Evaluation covers open-ended answers (GQA, VQAv2, A-OKVQA) (Hudson & Manning, 2019; Goyal et al., 2017; Schwenk et al., 2022), binary answers (POPE, VSR, NaturalBench) (Li et al., 2023b; Liu et al., 2023; Li et al., 2024), multiple-choice answers (MMBench, MMMU-Pro, MathVision) (Liu et al., 2024b; Yue et al., 2025; Wang et al., 2024a), and numeric answers (MathVista) (Lu et al., 2024). The corresponding output formats are short answers, yes/no responses, option letters, and numeric values.

LLM Judge Prompt for Semantic Agreement

You are a visual question answering (VQA) answer-equivalence judge. Given two extracted one-word answers to the same question, determine whether a human VQA grader would accept them as the same answer. Apply the following liberal grading standard.

Equivalent answers
Mark the pair as equivalent in the following cases:

- Synonyms or near-synonyms: *couch / sofa*, *lady / woman*.
- A specific and a general term naming the same referent: *beagle / dog*, *jet / airplane*.
- The same place or time at different granularity: *london / england*.
- A role and a generic description of the same person: *driver / man*.
- A plausible first-word truncation of a common compound naming the same object.
- Adjacent shades within the same color family: *gray / silver*.

Non-equivalent answers
Mark the pair as not equivalent when the answers would be distinguished by a VQA grader, including *left / right*, *yes / no*, *open / closed*, and *large / small*.

Input
`{answer_a} | {answer_b}`

Output
Return exactly one of the following JSON objects:
`{"equivalent": true}`
`{"equivalent": false}`
Do not include explanations, Markdown formatting, or additional fields.

E.6 Information Access

At test time, deployed AM-GARD variants access only the observed image, question, and frozen model's internal activations and gradients. They do not access the paired pre-attack image, saved pre-attack response, or dataset ground-truth answer. Any condition using saved pre-attack answers, including Gold-Gate and reference-target diagnostics, is reported explicitly as diagnostic only.

E.7 Baseline Evaluation Protocol

All comparison methods use the same estimation, calibration, and test splits as AM-GARD. Method-specific directions, subspaces, probes, or heads are fitted only on the estimation split, and intervention layers and strengths are selected on calibration data and fixed before test evaluation. Unless intervention selection is part of the method itself, the learned operator is applied to all evaluation inputs.

All methods are evaluated on the same test inputs using $R_{\mathrm{EM}}^{\mathrm{all}}$ and $R_{\mathrm{sem}}^{\mathrm{all}}$ for recovery against the saved pre-attack response, together with C_{EM} for pre-attack response preservation. Results use the same five random seeds and data splits. In the controlled direction comparison of Section 3.2, detector and answer-margin directions additionally share the same intervention mask and per-example L_2 update norm.

E.8 Baseline Implementation

Table 5 summarizes the implementations used in our evaluation. Fixed activation operators are applied at the same intervention site as AM-GARD where applicable. The comparison methods are adapted to the same LVLMs, prompts, and behavioral-recovery evaluation used for AM-GARD. We preserve each method's intervention mechanism while fitting or calibrating its components on our estimation and calibration splits. Accordingly, reported results should be interpreted as task-adapted comparisons under a common evaluation protocol rather than reproductions of the original papers' end-to-end results.

Table 5: **Baseline implementation parameters.** Intervention forms and principal settings used in our evaluation. F3 and DPS are included for context.

Method	Intervention	Key parameters
Activation steering		
CAA	Fixed contrastive direction	$\alpha = 1.0$; L24/L16
ITI-head	Selected-head steering	Top-$K = 48$ heads by probe AUROC; $\alpha = 5.0$; layers $\{12, 18, 24\}$
SEA	Subspace editing	Explained-variance threshold $k = 0.5$; L24/L16
LEACE	Linear concept erasure	Shrinkage$= 0.05$; L24/L16
ASTRA	Adaptive direction suppression	Gain $= 1.0$; L24
ARGUS	Detector-based adaptive steering with post-intervention filtering	Detector $C = 0.5$; gate FPR $= 0.10$; utility-decouple rank $= 8$; mid-layer steering with a late filter
Visual and generation defenses (for context)		
F3	Image-space cross-modal-attention purification	Single gradient step; $\alpha = 16$, $\beta = 32$ on the 0–255 pixel scale; random-perturbation attention reference; applied to both pre-attack and attacked inputs
DPS	Partial-perception supervision during generation	Three crops: center (0.5), random ($1/4$–$1/2$), and adaptive (object bounding box); partial-view answers supplied to a fixed correction prompt; black-box access
AM-GARD	Detector-based selection and iterative answer-margin correction	Gate FPR $\eta = 0.20$; $k = 4$; ϵ_r, T, β; layer selected on calibration data

F RECOVERY FULL RESULTS

Table 6: **Recovery and pre-attack preservation, all attacked inputs, exact match** ($R^{\text{all}}_{\text{EM}}/C_{\text{EM}}$, %). Each cell shows recovery followed in smaller type by pre-attack preservation; subscripts give the sample SD over five seeds (42, 1337, 2026, 123, 7). **Bold** and underlined values mark the best and second-best recovery per dataset among representation-level interventions with $C_{\text{EM}} \geq 90$, excluding Gold-Gate. F3 and DPS (gray) are shown as complementary image- and generation-level defenses.

Method	Open-ended			Binary			Multiple-choice			Numeric
	GQA	VQAv2	AOK	POPE	VSR	NatB	MMB	MMMU	MVsn	MVsta
Qwen2-VL-7B										
No-op	42.4$_{\pm 0.5}$ /100.0$_{\pm 0.0}$	63.0$_{\pm 3.6}$ /100.0$_{\pm 0.0}$	61.9$_{\pm 2.9}$ /100.0$_{\pm 0.0}$	84.9$_{\pm 0.4}$ /100.0$_{\pm 0.0}$	80.0$_{\pm 5.9}$ /100.0$_{\pm 0.0}$	69.5$_{\pm 3.8}$ /100.0$_{\pm 0.0}$	37.2$_{\pm 0.8}$ /100.0$_{\pm 0.0}$	33.1$_{\pm 5.6}$ /100.0$_{\pm 0.0}$	41.3$_{\pm 3.7}$ /100.0$_{\pm 0.0}$	47.7$_{\pm 4.4}$ /100.0$_{\pm 0.0}$
Visual adversarial defense (reference)										
F3	83.9$_{\pm 0.6}$ 83.5$_{\pm 0.7}$	77.8$_{\pm 1.9}$ 81.0$_{\pm 3.8}$	69.9$_{\pm 4.1}$ 73.3$_{\pm 4.2}$	90.8$_{\pm 0.7}$ 91.1$_{\pm 1.0}$	76.8$_{\pm 2.5}$ 80.9$_{\pm 2.6}$	82.2$_{\pm 4.7}$ 80.0$_{\pm 7.3}$	90.0$_{\pm 0.7}$ 92.8$_{\pm 0.8}$	58.7$_{\pm 4.7}$ 65.7$_{\pm 7.8}$	62.5$_{\pm 3.8}$ 64.0$_{\pm 3.3}$	65.8$_{\pm 3.8}$ 70.3$_{\pm 7.4}$
DPS	68.1$_{\pm 1.5}$ 72.6$_{\pm 1.5}$	64.6$_{\pm 2.9}$ 69.5$_{\pm 4.9}$	59.7$_{\pm 1.3}$ 61.7$_{\pm 2.7}$	87.9$_{\pm 0.9}$ 93.1$_{\pm 0.3}$	81.8$_{\pm 6.4}$ 94.5$_{\pm 2.6}$	69.1$_{\pm 7.9}$ 73.6$_{\pm 4.0}$	48.5$_{\pm 1.4}$ 98.0$_{\pm 0.3}$	41.5$_{\pm 1.5}$ 59.1$_{\pm 4.1}$	49.0$_{\pm 4.1}$ 75.4$_{\pm 3.3}$	49.0$_{\pm 8.6}$ 75.8$_{\pm 7.2}$
Activation steering										
CAA	42.2$_{\pm 1.0}$ /99.1$_{\pm 0.1}$	67.3$_{\pm 3.6}$ /99.0$_{\pm 0.8}$	61.1$_{\pm 3.3}$ /95.4$_{\pm 1.2}$	84.7$_{\pm 0.4}$ /99.6$_{\pm 0.2}$	81.8$_{\pm 7.4}$ /98.2$_{\pm 2.5}$	68.9$_{\pm 3.3}$ /98.4$_{\pm 1.6}$	37.4$_{\pm 0.8}$ /99.7$_{\pm 0.1}$	35.4$_{\pm 4.9}$ /94.6$_{\pm 1.4}$	41.6$_{\pm 3.8}$ /91.2$_{\pm 2.3}$	48.7$_{\pm 4.8}$ /97.3$_{\pm 2.5}$
ITI-head	46.0$_{\pm 0.8}$ /98.1$_{\pm 0.3}$	66.0$_{\pm 3.0}$ /98.6$_{\pm 0.6}$	62.8$_{\pm 3.7}$ /94.3$_{\pm 1.5}$	84.7$_{\pm 0.7}$ /99.6$_{\pm 0.2}$	80.4$_{\pm 5.5}$ /98.2$_{\pm 1.9}$	67.6$_{\pm 3.1}$ /99.7$_{\pm 0.7}$	40.3$_{\pm 1.3}$ /99.0$_{\pm 0.2}$	32.2$_{\pm 4.9}$ /85.2$_{\pm 6.8}$	43.0$_{\pm 3.0}$ /89.3$_{\pm 2.4}$	47.7$_{\pm 6.9}$ /96.7$_{\pm 3.5}$
SEA	36.8$_{\pm 0.9}$ /59.1$_{\pm 1.3}$	60.9$_{\pm 2.1}$ /82.9$_{\pm 1.6}$	59.8$_{\pm 5.6}$ /80.6$_{\pm 1.6}$	61.4$_{\pm 1.3}$ /60.5$_{\pm 1.2}$	65.0$_{\pm 8.0}$ /70.0$_{\pm 8.7}$	67.3$_{\pm 1.8}$ /84.1$_{\pm 9.2}$	43.7$_{\pm 0.6}$ /95.8$_{\pm 0.5}$	37.2$_{\pm 6.5}$ /76.8$_{\pm 4.5}$	34.3$_{\pm 3.2}$ /72.8$_{\pm 2.2}$	41.6$_{\pm 2.6}$ /86.8$_{\pm 3.9}$
LEACE	43.4$_{\pm 1.4}$ /96.2$_{\pm 1.2}$	0.0$_{\pm 0.0}$ /0.1$_{\pm 0.2}$	0.2$_{\pm 0.5}$ /0.3$_{\pm 0.3}$	20.7$_{\pm 15.7}$ /20.0$_{\pm 15.6}$	22.7$_{\pm 16.3}$ /23.2$_{\pm 14.1}$	13.7$_{\pm 13.6}$ /13.0$_{\pm 15.0}$	2.1$_{\pm 0.5}$ /4.4$_{\pm 0.8}$	4.3$_{\pm 3.5}$ /4.3$_{\pm 3.2}$	5.1$_{\pm 3.1}$ /7.5$_{\pm 3.5}$	1.3$_{\pm 0.7}$ /1.3$_{\pm 1.4}$
ASTRA	45.2$_{\pm 2.4}$ /97.4$_{\pm 0.6}$	64.7$_{\pm 3.2}$ /98.5$_{\pm 1.4}$	61.0$_{\pm 3.1}$ /95.6$_{\pm 1.4}$	53.0$_{\pm 21.1}$ /64.8$_{\pm 23.7}$	60.5$_{\pm 17.5}$ /80.5$_{\pm 17.1}$	60.3$_{\pm 4.6}$ /93.7$_{\pm 8.2}$	39.6$_{\pm 6.2}$ /81.2$_{\pm 13.6}$	35.7$_{\pm 3.7}$ /86.3$_{\pm 6.0}$	42.8$_{\pm 3.7}$ /86.2$_{\pm 2.0}$	45.5$_{\pm 3.5}$ /96.5$_{\pm 2.1}$
ARGUS	42.6$_{\pm 0.8}$ /99.7$_{\pm 0.1}$	63.7$_{\pm 3.0}$ /99.9$_{\pm 0.2}$	60.3$_{\pm 3.9}$ /99.6$_{\pm 0.5}$	84.7$_{\pm 0.7}$ /100.0$_{\pm 0.0}$	**82.3**$_{\pm 5.7}$ /99.1$_{\pm 1.2}$	**70.5**$_{\pm 4.1}$ /100.0$_{\pm 0.0}$	36.7$_{\pm 0.7}$ /99.9$_{\pm 0.1}$	32.8$_{\pm 5.6}$ /99.4$_{\pm 0.8}$	42.2$_{\pm 3.0}$ /99.1$_{\pm 0.6}$	45.8$_{\pm 3.7}$ /99.7$_{\pm 0.7}$
AM-GARD										
+ Outlier-Gate	45.7$_{\pm 0.7}$ /91.0$_{\pm 0.9}$	64.0$_{\pm 4.4}$ /94.3$_{\pm 0.8}$	62.8$_{\pm 3.0}$ /97.7$_{\pm 0.7}$	74.2$_{\pm 0.4}$ /99.8$_{\pm 0.1}$	60.0$_{\pm 2.0}$ /98.6$_{\pm 1.2}$	68.9$_{\pm 8.2}$ /95.6$_{\pm 3.4}$	38.9$_{\pm 1.6}$ /91.3$_{\pm 0.9}$	39.4$_{\pm 5.1}$ /94.6$_{\pm 3.8}$	43.5$_{\pm 4.0}$ /96.1$_{\pm 1.8}$	46.8$_{\pm 6.5}$ /88.1$_{\pm 5.4}$
+ **Probe-Gate**	**66.6**$_{\pm 1.4}$ /96.5$_{\pm 0.7}$	**70.3**$_{\pm 3.5}$ /95.7$_{\pm 1.2}$	**64.7**$_{\pm 4.1}$ /99.1$_{\pm 0.8}$	**87.8**$_{\pm 1.1}$ /98.5$_{\pm 0.5}$	**82.3**$_{\pm 4.7}$ /98.6$_{\pm 1.2}$	72.4$_{\pm 3.6}$ /87.0$_{\pm 6.8}$	**57.0**$_{\pm 4.7}$ /96.2$_{\pm 0.6}$	**39.8**$_{\pm 3.5}$ /97.0$_{\pm 2.0}$	**49.2**$_{\pm 5.4}$ /97.4$_{\pm 0.9}$	**51.9**$_{\pm 3.1}$ /96.8$_{\pm 3.2}$
+ Gold-Gate	89.2$_{\pm 1.1}$ /100.0$_{\pm 0.0}$	89.2$_{\pm 2.2}$ /100.0$_{\pm 0.0}$	85.6$_{\pm 2.9}$ /100.0$_{\pm 0.0}$	96.7$_{\pm 0.6}$ /100.0$_{\pm 0.0}$	87.3$_{\pm 2.0}$ /100.0$_{\pm 0.0}$	93.7$_{\pm 1.1}$ /100.0$_{\pm 0.0}$	89.9$_{\pm 0.6}$ /100.0$_{\pm 0.0}$	70.0$_{\pm 5.0}$ /100.0$_{\pm 0.0}$	79.7$_{\pm 4.4}$ /100.0$_{\pm 0.0}$	89.0$_{\pm 3.5}$ /100.0$_{\pm 0.0}$
LLaVA-1.5-7B										
No-op	17.4$_{\pm 0.4}$ /100.0$_{\pm 0.0}$	17.3$_{\pm 0.8}$ /100.0$_{\pm 0.0}$	14.8$_{\pm 2.9}$ /100.0$_{\pm 0.0}$	28.6$_{\pm 1.1}$ /100.0$_{\pm 0.0}$	35.5$_{\pm 3.2}$ /100.0$_{\pm 0.0}$	23.5$_{\pm 3.4}$ /100.0$_{\pm 0.0}$	31.0$_{\pm 0.9}$ /100.0$_{\pm 0.0}$	23.6$_{\pm 4.4}$ /100.0$_{\pm 0.0}$	53.6$_{\pm 4.4}$ /100.0$_{\pm 0.0}$	41.2$_{\pm 15.7}$ /100.0$_{\pm 0.0}$
Visual adversarial defense (reference)										
F3	90.2$_{\pm 0.3}$ 91.2$_{\pm 0.5}$	85.9$_{\pm 2.1}$ 90.0$_{\pm 2.4}$	86.0$_{\pm 4.9}$ 86.0$_{\pm 3.1}$	93.5$_{\pm 0.4}$ 94.6$_{\pm 0.5}$	88.0$_{\pm 1.9}$ 89.3$_{\pm 0.7}$	86.9$_{\pm 3.0}$ 89.8$_{\pm 2.1}$	93.3$_{\pm 0.6}$ 94.0$_{\pm 0.4}$	62.1$_{\pm 3.0}$ 63.6$_{\pm 3.0}$	84.5$_{\pm 2.2}$ 86.0$_{\pm 2.1}$	58.8$_{\pm 18.0}$ 76.2$_{\pm 12.0}$
DPS	72.0$_{\pm 0.3}$ 78.8$_{\pm 0.6}$	72.0$_{\pm 2.6}$ 79.8$_{\pm 2.3}$	65.8$_{\pm 2.6}$ 72.6$_{\pm 1.7}$	76.5$_{\pm 1.0}$ 78.2$_{\pm 1.1}$	55.7$_{\pm 1.3}$ 61.1$_{\pm 2.2}$	70.5$_{\pm 3.3}$ 74.9$_{\pm 1.4}$	51.7$_{\pm 0.5}$ 91.7$_{\pm 0.7}$	26.4$_{\pm 5.1}$ 55.5$_{\pm 6.2}$	58.5$_{\pm 2.0}$ 67.3$_{\pm 3.9}$	38.8$_{\pm 6.8}$ 52.5$_{\pm 9.5}$
Activation steering										
CAA	17.0$_{\pm 0.4}$ /98.5$_{\pm 0.2}$	17.4$_{\pm 1.5}$ /98.4$_{\pm 0.5}$	15.7$_{\pm 2.6}$ /97.3$_{\pm 0.9}$	24.2$_{\pm 1.2}$ /98.9$_{\pm 0.2}$	70.9$_{\pm 3.0}$ /81.0$_{\pm 2.4}$	**46.2**$_{\pm 3.4}$ /90.9$_{\pm 1.7}$	34.2$_{\pm 1.4}$ /97.5$_{\pm 0.6}$	24.6$_{\pm 4.3}$ /70.9$_{\pm 3.9}$	57.4$_{\pm 4.7}$ /89.5$_{\pm 2.2}$	51.2$_{\pm 20.4}$ /80.0$_{\pm 6.9}$
ITI-head	34.7$_{\pm 2.2}$ /94.8$_{\pm 1.1}$	**30.4**$_{\pm 2.4}$ /96.8$_{\pm 1.1}$	23.1$_{\pm 4.8}$ /96.7$_{\pm 0.9}$	21.1$_{\pm 0.9}$ /99.2$_{\pm 0.2}$	72.6$_{\pm 2.7}$ /77.2$_{\pm 1.8}$	60.5$_{\pm 4.3}$ /79.7$_{\pm 2.9}$	36.8$_{\pm 1.1}$ /96.6$_{\pm 0.6}$	25.2$_{\pm 3.6}$ /74.6$_{\pm 4.7}$	53.3$_{\pm 4.9}$ /87.6$_{\pm 1.8}$	46.3$_{\pm 19.1}$ /86.2$_{\pm 11.2}$
SEA	13.6$_{\pm 0.5}$ /59.9$_{\pm 1.0}$	15.9$_{\pm 1.0}$ /98.0$_{\pm 0.7}$	25.7$_{\pm 2.8}$ /82.5$_{\pm 3.4}$	45.8$_{\pm 1.0}$ /98.0$_{\pm 0.2}$	25.4$_{\pm 3.9}$ /96.3$_{\pm 1.3}$	31.4$_{\pm 1.3}$ /94.7$_{\pm 2.0}$	35.5$_{\pm 0.9}$ /92.4$_{\pm 0.6}$	25.5$_{\pm 2.9}$ /73.9$_{\pm 3.4}$	35.3$_{\pm 2.2}$ /77.0$_{\pm 2.0}$	40.0$_{\pm 14.4}$ /67.5$_{\pm 2.8}$
LEACE	18.2$_{\pm 0.4}$ /98.6$_{\pm 0.1}$	7.6$_{\pm 3.8}$ /12.9$_{\pm 4.9}$	8.4$_{\pm 3.6}$ /20.8$_{\pm 9.0}$	26.3$_{\pm 2.4}$ /98.7$_{\pm 0.1}$	12.3$_{\pm 22.4}$ /11.7$_{\pm 19.5}$	36.1$_{\pm 19.3}$ /35.7$_{\pm 17.4}$	9.4$_{\pm 4.8}$ /16.2$_{\pm 7.6}$	32.1$_{\pm 8.1}$ /71.8$_{\pm 8.0}$	21.5$_{\pm 8.4}$ /22.1$_{\pm 9.7}$	3.8$_{\pm 5.6}$ /2.5$_{\pm 5.6}$
ASTRA	20.8$_{\pm 3.6}$ /97.9$_{\pm 0.9}$	18.1$_{\pm 1.4}$ /98.0$_{\pm 1.3}$	15.4$_{\pm 3.4}$ /97.2$_{\pm 1.3}$	24.4$_{\pm 0.9}$ /98.9$_{\pm 0.2}$	60.4$_{\pm 15.4}$ /82.0$_{\pm 8.3}$	34.2$_{\pm 10.4}$ /92.6$_{\pm 4.5}$	35.1$_{\pm 2.4}$ /96.4$_{\pm 0.8}$	24.6$_{\pm 5.0}$ /69.1$_{\pm 10.1}$	49.0$_{\pm 5.9}$ /88.0$_{\pm 1.2}$	40.0$_{\pm 16.3}$ /82.5$_{\pm 5.2}$
ARGUS	17.5$_{\pm 0.9}$ /99.8$_{\pm 0.0}$	16.3$_{\pm 0.6}$ /99.8$_{\pm 0.3}$	15.2$_{\pm 2.6}$ /99.6$_{\pm 0.3}$	28.3$_{\pm 0.9}$ /99.8$_{\pm 0.1}$	41.8$_{\pm 3.1}$ /98.8$_{\pm 0.8}$	24.0$_{\pm 2.8}$ /99.8$_{\pm 0.3}$	31.5$_{\pm 1.6}$ /99.3$_{\pm 0.4}$	20.3$_{\pm 4.6}$ /98.8$_{\pm 1.3}$	**52.7**$_{\pm 4.9}$ /98.9$_{\pm 1.2}$	**45.0**$_{\pm 18.4}$ /98.8$_{\pm 2.8}$
AM-GARD										
+ Outlier-Gate	24.2$_{\pm 0.5}$ /91.4$_{\pm 0.6}$	19.3$_{\pm 1.4}$ /93.0$_{\pm 2.3}$	23.7$_{\pm 2.8}$ /92.2$_{\pm 1.0}$	30.4$_{\pm 1.0}$ /98.9$_{\pm 0.4}$	**42.6**$_{\pm 2.6}$ /93.9$_{\pm 0.8}$	27.0$_{\pm 3.9}$ /94.5$_{\pm 1.5}$	30.2$_{\pm 1.1}$ /94.9$_{\pm 0.5}$	27.6$_{\pm 6.3}$ /98.5$_{\pm 2.6}$	47.5$_{\pm 3.9}$ /97.9$_{\pm 0.8}$	36.2$_{\pm 17.9}$ /91.2$_{\pm 10.5}$
+ **Probe-Gate**	**45.3**$_{\pm 5.2}$ /96.6$_{\pm 1.4}$	30.0$_{\pm 5.9}$ /96.4$_{\pm 1.8}$	**39.2**$_{\pm 4.5}$ /94.5$_{\pm 1.4}$	**74.0**$_{\pm 0.9}$ /90.5$_{\pm 0.9}$	39.6$_{\pm 2.6}$ /94.2$_{\pm 1.6}$	54.6$_{\pm 4.5}$ /87.0$_{\pm 2.8}$	**59.1**$_{\pm 1.5}$ /94.5$_{\pm 2.7}$	**29.1**$_{\pm 5.4}$ /90.6$_{\pm 3.3}$	51.6$_{\pm 1.8}$ /97.9$_{\pm 1.8}$	37.5$_{\pm 15.9}$ /98.8$_{\pm 2.8}$
+ Gold-Gate	84.1$_{\pm 0.6}$ /100.0$_{\pm 0.0}$	65.3$_{\pm 1.3}$ /100.0$_{\pm 0.0}$	77.2$_{\pm 3.9}$ /100.0$_{\pm 0.0}$	94.2$_{\pm 0.4}$ /100.0$_{\pm 0.0}$	83.1$_{\pm 1.0}$ /100.0$_{\pm 0.0}$	87.0$_{\pm 2.4}$ /100.0$_{\pm 0.0}$	89.8$_{\pm 0.9}$ /100.0$_{\pm 0.0}$	68.2$_{\pm 4.5}$ /100.0$_{\pm 0.0}$	85.8$_{\pm 1.2}$ /100.0$_{\pm 0.0}$	47.5$_{\pm 19.6}$ /100.0$_{\pm 0.0}$

Table 7: **Recovery and pre-attack preservation, all attacked inputs, semantic agreement** ($R^{\text{all}}_{\text{sem}}/C_{\text{EM}}$, %). Each cell shows recovery followed in smaller type by pre-attack preservation; subscripts give the sample SD over five seeds (42, 1337, 2026, 123, 7). **Bold** and underlined values mark the best and second-best recovery per dataset among representation-level interventions with $C_{\text{EM}} \geq 90$, excluding Gold-Gate. F3 and DPS (gray) are shown as complementary image- and generation-level defenses.

Method	Open-ended			Binary			Multiple-choice			Numeric
	GQA	VQAv2	AOK	POPE	VSR	NatB	MMB	MMMU	MVsn	MVsta
Qwen2-VL-7B										
No-op	47.5$_{\pm 0.9}$ /100.0$_{\pm 0.0}$	69.2$_{\pm 3.7}$ /100.0$_{\pm 0.0}$	70.1$_{\pm 4.2}$ /100.0$_{\pm 0.0}$	84.9$_{\pm 0.4}$ /100.0$_{\pm 0.0}$	80.0$_{\pm 5.9}$ /100.0$_{\pm 0.0}$	69.5$_{\pm 3.8}$ /100.0$_{\pm 0.0}$	37.2$_{\pm 0.8}$ /100.0$_{\pm 0.0}$	33.1$_{\pm 5.6}$ /100.0$_{\pm 0.0}$	41.3$_{\pm 3.7}$ /100.0$_{\pm 0.0}$	47.7$_{\pm 4.4}$ /100.0$_{\pm 0.0}$
Visual adversarial defense (reference)										
F3	85.3$_{\pm 0.7}$ 84.8$_{\pm 0.8}$	81.4$_{\pm 2.1}$ 83.5$_{\pm 3.3}$	75.7$_{\pm 4.0}$ 78.7$_{\pm 2.7}$	90.8$_{\pm 0.7}$ 91.1$_{\pm 1.0}$	76.8$_{\pm 2.5}$ 80.9$_{\pm 2.6}$	82.2$_{\pm 4.7}$ 80.0$_{\pm 7.3}$	90.0$_{\pm 0.7}$ 92.8$_{\pm 0.8}$	58.7$_{\pm 4.7}$ 65.7$_{\pm 7.8}$	62.8$_{\pm 3.9}$ 64.4$_{\pm 3.3}$	65.8$_{\pm 3.9}$ 70.3$_{\pm 7.4}$
DPS	72.4$_{\pm 1.0}$ 76.3$_{\pm 1.2}$	73.7$_{\pm 3.9}$ 78.2$_{\pm 4.4}$	72.3$_{\pm 1.3}$ 73.1$_{\pm 2.7}$	87.9$_{\pm 0.9}$ 93.1$_{\pm 0.3}$	81.8$_{\pm 6.4}$ 94.5$_{\pm 2.6}$	69.1$_{\pm 7.9}$ 73.6$_{\pm 4.0}$	48.5$_{\pm 1.4}$ 98.0$_{\pm 0.3}$	41.7$_{\pm 1.9}$ 59.4$_{\pm 3.8}$	49.0$_{\pm 4.1}$ 75.4$_{\pm 3.3}$	49.0$_{\pm 8.6}$ 75.8$_{\pm 7.2}$
Activation steering										
CAA	47.2$_{\pm 1.4}$ /99.4$_{\pm 0.2}$	71.5$_{\pm 4.3}$ /99.2$_{\pm 0.9}$	69.8$_{\pm 4.3}$ /96.6$_{\pm 0.9}$	84.7$_{\pm 0.4}$ /99.6$_{\pm 0.2}$	81.8$_{\pm 7.4}$ /98.2$_{\pm 2.5}$	68.9$_{\pm 3.3}$ /98.4$_{\pm 1.6}$	37.4$_{\pm 0.8}$ /99.7$_{\pm 0.1}$	35.4$_{\pm 4.9}$ /94.6$_{\pm 1.4}$	41.6$_{\pm 3.8}$ /91.2$_{\pm 2.3}$	48.7$_{\pm 4.8}$ /97.3$_{\pm 2.5}$
ITI-head	51.1$_{\pm 0.3}$ /98.4$_{\pm 0.3}$	70.7$_{\pm 3.5}$ /99.1$_{\pm 0.4}$	70.7$_{\pm 4.4}$ /95.7$_{\pm 1.7}$	84.7$_{\pm 0.7}$ /99.6$_{\pm 0.2}$	80.5$_{\pm 5.5}$ /98.2$_{\pm 1.9}$	67.6$_{\pm 3.1}$ /99.7$_{\pm 0.7}$	40.3$_{\pm 1.3}$ /99.0$_{\pm 0.2}$	32.2$_{\pm 4.9}$ /85.2$_{\pm 6.8}$	43.0$_{\pm 3.0}$ /89.3$_{\pm 2.4}$	47.7$_{\pm 6.9}$ /96.7$_{\pm 3.5}$
SEA	38.4$_{\pm 1.1}$ /59.6$_{\pm 1.4}$	68.6$_{\pm 1.9}$ /85.3$_{\pm 1.3}$	68.4$_{\pm 4.8}$ /84.6$_{\pm 1.9}$	61.4$_{\pm 1.3}$ /60.5$_{\pm 1.2}$	65.0$_{\pm 8.0}$ /70.0$_{\pm 8.7}$	67.3$_{\pm 1.8}$ /84.1$_{\pm 9.2}$	43.7$_{\pm 0.6}$ /95.8$_{\pm 0.5}$	37.2$_{\pm 6.5}$ /76.8$_{\pm 4.5}$	34.3$_{\pm 3.2}$ /72.8$_{\pm 2.2}$	41.6$_{\pm 2.6}$ /86.8$_{\pm 3.9}$
LEACE	47.6$_{\pm 1.8}$ /96.8$_{\pm 1.4}$	0.0$_{\pm 0.0}$ /0.2$_{\pm 0.5}$	0.3$_{\pm 0.5}$ /0.3$_{\pm 0.3}$	27.8$_{\pm 15.2}$ /26.3$_{\pm 15.5}$	23.2$_{\pm 16.3}$ /23.2$_{\pm 14.1}$	19.4$_{\pm 11.0}$ /14.9$_{\pm 14.0}$	2.1$_{\pm 0.5}$ /4.4$_{\pm 0.8}$	4.3$_{\pm 3.5}$ /4.3$_{\pm 3.2}$	5.1$_{\pm 3.1}$ /7.5$_{\pm 3.5}$	6.8$_{\pm 9.4}$ /6.7$_{\pm 9.8}$
ASTRA	50.7$_{\pm 2.3}$ /97.7$_{\pm 0.6}$	70.8$_{\pm 3.3}$ /98.9$_{\pm 1.5}$	**70.9**$_{\pm 2.9}$ /97.2$_{\pm 1.3}$	53.0$_{\pm 21.1}$ /64.8$_{\pm 23.7}$	60.5$_{\pm 17.5}$ /80.5$_{\pm 17.1}$	60.3$_{\pm 4.6}$ /93.7$_{\pm 8.2}$	39.6$_{\pm 6.2}$ /81.2$_{\pm 13.6}$	35.7$_{\pm 3.7}$ /86.3$_{\pm 6.0}$	42.8$_{\pm 3.7}$ /86.2$_{\pm 2.0}$	45.5$_{\pm 3.5}$ /96.5$_{\pm 2.1}$
ARGUS	47.9$_{\pm 1.3}$ /100.0$_{\pm 0.0}$	69.8$_{\pm 3.2}$ /100.0$_{\pm 0.0}$	70.4$_{\pm 3.5}$ /99.6$_{\pm 0.5}$	84.7$_{\pm 0.7}$ /99.9$_{\pm 0.1}$	**82.3**$_{\pm 5.7}$ /100.0$_{\pm 0.0}$	**70.5**$_{\pm 4.1}$ /100.0$_{\pm 0.0}$	36.7$_{\pm 0.7}$ /99.9$_{\pm 0.1}$	32.8$_{\pm 5.6}$ /99.6$_{\pm 0.5}$	42.2$_{\pm 3.0}$ /99.5$_{\pm 0.7}$	45.8$_{\pm 3.7}$ /99.4$_{\pm 1.4}$
AM-GARD										
+ Outlier-Gate	45.7$_{\pm 0.7}$ /91.0$_{\pm 0.9}$	64.0$_{\pm 4.4}$ /94.3$_{\pm 0.8}$	62.8$_{\pm 3.0}$ /97.7$_{\pm 0.7}$	74.2$_{\pm 0.4}$ /99.8$_{\pm 0.1}$	60.0$_{\pm 2.0}$ /98.6$_{\pm 1.2}$	68.9$_{\pm 8.2}$ /95.6$_{\pm 3.4}$	38.9$_{\pm 1.6}$ /91.3$_{\pm 0.9}$	39.4$_{\pm 5.1}$ /94.6$_{\pm 3.8}$	43.5$_{\pm 4.0}$ /96.1$_{\pm 1.8}$	46.8$_{\pm 6.5}$ /88.1$_{\pm 5.4}$
+ **Probe-Gate**	**67.9**$_{\pm 1.4}$ /96.5$_{\pm 0.7}$	**71.8**$_{\pm 4.3}$ /95.7$_{\pm 1.2}$	66.7$_{\pm 3.6}$ /99.1$_{\pm 0.8}$	**87.8**$_{\pm 1.1}$ /98.5$_{\pm 0.5}$	**82.3**$_{\pm 4.7}$ /98.6$_{\pm 1.2}$	72.4$_{\pm 3.6}$ /87.0$_{\pm 6.8}$	**57.0**$_{\pm 4.7}$ /96.2$_{\pm 0.6}$	**39.8**$_{\pm 3.5}$ /97.0$_{\pm 2.0}$	**49.2**$_{\pm 5.4}$ /97.4$_{\pm 0.9}$	**51.9**$_{\pm 3.1}$ /96.8$_{\pm 3.2}$
+ Gold-Gate	89.2$_{\pm 1.1}$ /100.0$_{\pm 0.0}$	89.2$_{\pm 2.2}$ /100.0$_{\pm 0.0}$	85.6$_{\pm 2.9}$ /100.0$_{\pm 0.0}$	96.7$_{\pm 0.6}$ /100.0$_{\pm 0.0}$	87.3$_{\pm 2.0}$ /100.0$_{\pm 0.0}$	93.7$_{\pm 1.1}$ /100.0$_{\pm 0.0}$	89.9$_{\pm 0.6}$ /100.0$_{\pm 0.0}$	70.0$_{\pm 5.0}$ /100.0$_{\pm 0.0}$	79.7$_{\pm 4.4}$ /100.0$_{\pm 0.0}$	89.0$_{\pm 3.5}$ /100.0$_{\pm 0.0}$
LLaVA-1.5-7B										
No-op	24.7$_{\pm 0.4}$ /100.0$_{\pm 0.0}$	25.6$_{\pm 1.1}$ /100.0$_{\pm 0.0}$	28.9$_{\pm 4.4}$ /100.0$_{\pm 0.0}$	28.6$_{\pm 1.1}$ /100.0$_{\pm 0.0}$	35.5$_{\pm 3.2}$ /100.0$_{\pm 0.0}$	23.5$_{\pm 3.4}$ /100.0$_{\pm 0.0}$	31.0$_{\pm 0.9}$ /100.0$_{\pm 0.0}$	23.6$_{\pm 4.4}$ /100.0$_{\pm 0.0}$	53.6$_{\pm 4.4}$ /100.0$_{\pm 0.0}$	41.2$_{\pm 15.7}$ /100.0$_{\pm 0.0}$
Visual adversarial defense (reference)										
F3	90.8$_{\pm 0.3}$ 91.7$_{\pm 0.5}$	87.9$_{\pm 2.3}$ 90.9$_{\pm 2.3}$	88.0$_{\pm 4.4}$ 87.7$_{\pm 3.1}$	93.5$_{\pm 0.4}$ 94.6$_{\pm 0.5}$	88.0$_{\pm 1.9}$ 89.3$_{\pm 0.7}$	86.9$_{\pm 3.0}$ 89.8$_{\pm 2.1}$	93.3$_{\pm 0.6}$ 94.0$_{\pm 0.4}$	62.1$_{\pm 3.0}$ 63.6$_{\pm 3.0}$	84.5$_{\pm 2.2}$ 86.0$_{\pm 2.1}$	58.8$_{\pm 18.0}$ 76.2$_{\pm 12.0}$
DPS	74.2$_{\pm 0.3}$ 80.3$_{\pm 0.6}$	76.8$_{\pm 2.8}$ 84.1$_{\pm 2.5}$	73.2$_{\pm 2.1}$ 78.2$_{\pm 2.7}$	76.5$_{\pm 1.0}$ 78.2$_{\pm 1.1}$	55.7$_{\pm 1.3}$ 61.1$_{\pm 2.2}$	70.5$_{\pm 3.3}$ 74.9$_{\pm 1.4}$	51.7$_{\pm 0.5}$ 91.7$_{\pm 0.7}$	26.4$_{\pm 5.1}$ 55.5$_{\pm 6.2}$	58.5$_{\pm 2.0}$ 67.3$_{\pm 3.9}$	38.8$_{\pm 6.8}$ 52.5$_{\pm 9.5}$
Activation steering										
CAA	24.1$_{\pm 0.5}$ /98.8$_{\pm 0.2}$	29.7$_{\pm 2.5}$ /98.4$_{\pm 0.5}$	30.4$_{\pm 3.8}$ /97.9$_{\pm 0.5}$	24.2$_{\pm 1.2}$ /98.9$_{\pm 0.2}$	70.9$_{\pm 3.0}$ /81.0$_{\pm 2.4}$	**46.2**$_{\pm 3.4}$ /90.9$_{\pm 1.7}$	34.2$_{\pm 1.4}$ /97.5$_{\pm 0.6}$	24.6$_{\pm 4.3}$ /70.9$_{\pm 3.9}$	57.4$_{\pm 4.7}$ /89.5$_{\pm 2.2}$	51.2$_{\pm 20.4}$ /80.0$_{\pm 6.9}$
ITI-head	40.7$_{\pm 2.2}$ /95.3$_{\pm 1.1}$	**35.6**$_{\pm 2.1}$ /98.5$_{\pm 0.8}$	35.0$_{\pm 5.1}$ /97.4$_{\pm 0.5}$	21.1$_{\pm 0.9}$ /99.2$_{\pm 0.2}$	72.6$_{\pm 2.7}$ /77.2$_{\pm 1.8}$	60.5$_{\pm 4.3}$ /79.7$_{\pm 2.9}$	36.8$_{\pm 1.1}$ /96.6$_{\pm 0.6}$	25.1$_{\pm 3.6}$ /74.6$_{\pm 4.7}$	53.3$_{\pm 4.9}$ /87.6$_{\pm 1.8}$	46.2$_{\pm 19.1}$ /86.2$_{\pm 11.2}$
SEA	15.4$_{\pm 0.6}$ /60.3$_{\pm 1.0}$	29.3$_{\pm 2.0}$ /98.0$_{\pm 0.7}$	37.8$_{\pm 2.9}$ /84.8$_{\pm 2.5}$	45.8$_{\pm 1.0}$ /98.0$_{\pm 0.2}$	25.4$_{\pm 3.9}$ /96.3$_{\pm 1.3}$	31.4$_{\pm 1.3}$ /94.7

Table 8: **Recovery and pre-attack preservation, flipped inputs, exact match** ($R^{\text{flip}}_{\text{EM}}/C_{\text{EM}}$, %). Each cell shows recovery followed in smaller type by pre-attack preservation; subscripts give the sample SD over five seeds (42, 1337, 2026, 123, 7). **Bold** and underlined values mark the best and second-best recovery per dataset among representation-level interventions with $C_{\text{EM}} \geq 90$, excluding Gold-Gate. F3 and DPS (gray) are shown as complementary image- and generation-level defenses.

	Open-ended			Binary			Multiple-choice			Numeric
Method	GQA	VQAv2	AOK	POPE	VSR	NatB	MMB	MMMU	MVsn	MVsta
Qwen2-VL-7B										
No-op	16.3$_{\pm 0.6}$ /100.0$_{\pm 0.0}$	14.7$_{\pm 4.8}$ /100.0$_{\pm 0.0}$	16.4$_{\pm 3.5}$ /100.0$_{\pm 0.0}$	19.8$_{\pm 0.7}$ /100.0$_{\pm 0.0}$	19.1$_{\pm 12.7}$ /100.0$_{\pm 0.0}$	27.1$_{\pm 7.2}$ /100.0$_{\pm 0.0}$	14.7$_{\pm 0.9}$ /100.0$_{\pm 0.0}$	25.8$_{\pm 5.0}$ /100.0$_{\pm 0.0}$	29.5$_{\pm 3.7}$ /100.0$_{\pm 0.0}$	38.9$_{\pm 5.4}$ /100.0$_{\pm 0.0}$
Visual adversarial defense (reference)										
F3	79.1$_{\pm 0.9}$ 83.5$_{\pm 0.7}$	65.2$_{\pm 5.6}$ 81.0$_{\pm 3.8}$	58.8$_{\pm 5.2}$ 73.3$_{\pm 4.2}$	74.8$_{\pm 3.5}$ 91.1$_{\pm 1.0}$	55.2$_{\pm 5.9}$ 80.9$_{\pm 2.6}$	69.1$_{\pm 7.5}$ 80.0$_{\pm 7.3}$	87.0$_{\pm 0.8}$ 92.8$_{\pm 0.8}$	56.2$_{\pm 5.2}$ 65.7$_{\pm 7.8}$	58.6$_{\pm 4.4}$ 64.0$_{\pm 3.3}$	64.7$_{\pm 3.9}$ 70.3$_{\pm 7.4}$
DPS	59.2$_{\pm 1.8}$ 72.6$_{\pm 1.5}$	39.7$_{\pm 4.2}$ 69.5$_{\pm 4.9}$	46.3$_{\pm 3.1}$ 61.7$_{\pm 2.7}$	53.2$_{\pm 3.8}$ 93.1$_{\pm 0.3}$	23.5$_{\pm 12.1}$ 94.5$_{\pm 2.6}$	65.2$_{\pm 8.6}$ 73.6$_{\pm 4.0}$	30.8$_{\pm 1.4}$ 98.0$_{\pm 0.3}$	39.1$_{\pm 2.3}$ 59.1$_{\pm 4.1}$	41.0$_{\pm 4.3}$ 75.4$_{\pm 3.3}$	43.2$_{\pm 7.8}$ 75.8$_{\pm 7.2}$
Activation steering										
CAA	16.2$_{\pm 1.3}$ /99.1$_{\pm 0.1}$	23.8$_{\pm 4.0}$ /99.0$_{\pm 0.8}$	15.4$_{\pm 2.9}$ /95.4$_{\pm 1.2}$	19.8$_{\pm 1.0}$ /99.6$_{\pm 0.2}$	**22.3**$_{\pm 18.2}$ /98.2$_{\pm 2.5}$	29.7$_{\pm 9.2}$ /98.4$_{\pm 1.6}$	14.8$_{\pm 1.4}$ /99.7$_{\pm 0.1}$	28.2$_{\pm 4.6}$ /94.6$_{\pm 1.4}$	28.5$_{\pm 2.9}$ /91.2$_{\pm 2.3}$	40.1$_{\pm 5.8}$ /97.3$_{\pm 2.5}$
ITI-head	21.9$_{\pm 1.5}$ /98.1$_{\pm 0.3}$	20.8$_{\pm 4.3}$ /98.6$_{\pm 0.6}$	19.3$_{\pm 5.1}$ /94.3$_{\pm 1.5}$	18.3$_{\pm 1.9}$ /99.6$_{\pm 0.2}$	20.7$_{\pm 13.7}$ /98.2$_{\pm 1.9}$	23.6$_{\pm 8.4}$ /99.7$_{\pm 0.7}$	19.3$_{\pm 2.3}$ /99.0$_{\pm 0.2}$	24.9$_{\pm 4.5}$ /85.2$_{\pm 6.8}$	30.9$_{\pm 2.3}$ /89.3$_{\pm 2.4}$	39.6$_{\pm 9.2}$ /96.7$_{\pm 3.5}$
SEA	23.2$_{\pm 1.3}$ /59.1$_{\pm 1.3}$	22.2$_{\pm 3.0}$ /82.9$_{\pm 1.6}$	27.2$_{\pm 7.3}$ /80.6$_{\pm 1.6}$	56.8$_{\pm 3.0}$ /60.5$_{\pm 1.2}$	29.8$_{\pm 22.4}$ /70.0$_{\pm 8.7}$	34.5$_{\pm 4.0}$ /84.1$_{\pm 9.2}$	26.5$_{\pm 1.1}$ /95.8$_{\pm 0.5}$	33.0$_{\pm 6.6}$ /76.8$_{\pm 4.5}$	25.6$_{\pm 3.5}$ /72.8$_{\pm 2.2}$	32.8$_{\pm 4.6}$ /86.8$_{\pm 3.9}$
LEACE	19.4$_{\pm 1.7}$ /96.2$_{\pm 1.2}$	0.0$_{\pm 0.0}$ /0.1$_{\pm 0.2}$	0.0$_{\pm 0.0}$ /0.3$_{\pm 0.3}$	19.3$_{\pm 14.5}$ /20.0$_{\pm 15.6}$	31.4$_{\pm 37.8}$ /23.2$_{\pm 14.1}$	8.2$_{\pm 11.5}$ /13.0$_{\pm 15.0}$	1.4$_{\pm 0.4}$ /4.4$_{\pm 0.8}$	4.3$_{\pm 3.4}$ /4.3$_{\pm 3.2}$	5.3$_{\pm 3.7}$ /7.5$_{\pm 3.8}$	1.6$_{\pm 0.9}$ /1.3$_{\pm 1.4}$
ASTRA	21.2$_{\pm 3.6}$ /97.4$_{\pm 0.6}$	17.7$_{\pm 1.4}$ /98.5$_{\pm 1.4}$	15.0$_{\pm 4.8}$ /95.6$_{\pm 1.4}$	34.4$_{\pm 15.3}$ /64.8$_{\pm 23.7}$	16.2$_{\pm 6.2}$ /80.5$_{\pm 17.1}$	22.7$_{\pm 12.4}$ /93.7$_{\pm 8.2}$	26.4$_{\pm 4.3}$ /81.2$_{\pm 13.6}$	29.0$_{\pm 4.3}$ /86.3$_{\pm 6.0}$	30.8$_{\pm 2.7}$ /86.2$_{\pm 2.0}$	37.2$_{\pm 3.9}$ /96.5$_{\pm 2.1}$
ARGUS	16.6$_{\pm 1.1}$ /99.7$_{\pm 0.1}$	15.4$_{\pm 3.3}$ /99.9$_{\pm 0.2}$	12.8$_{\pm 4.9}$ /99.6$_{\pm 0.5}$	18.2$_{\pm 1.8}$ /100.0$_{\pm 0.0}$	19.1$_{\pm 12.7}$ /99.1$_{\pm 1.2}$	26.3$_{\pm 7.3}$ /100.0$_{\pm 0.0}$	14.0$_{\pm 0.5}$ /99.9$_{\pm 0.1}$	25.6$_{\pm 4.9}$ /99.4$_{\pm 0.8}$	30.0$_{\pm 2.4}$ /99.1$_{\pm 0.6}$	36.5$_{\pm 4.7}$ /99.7$_{\pm 0.7}$
AM-GARD										
+ Outlier-Gate	22.5$_{\pm 0.6}$ /91.0$_{\pm 0.9}$	24.3$_{\pm 5.9}$ /94.3$_{\pm 0.8}$	**31.2**$_{\pm 5.5}$ /97.7$_{\pm 0.7}$	24.9$_{\pm 1.7}$ /99.8$_{\pm 0.1}$	19.1$_{\pm 8.1}$ /98.6$_{\pm 1.2}$	**36.3**$_{\pm 16.9}$ /95.6$_{\pm 3.4}$	24.4$_{\pm 1.7}$ /91.3$_{\pm 0.9}$	33.4$_{\pm 5.0}$ /94.6$_{\pm 3.8}$	32.6$_{\pm 3.8}$ /96.1$_{\pm 1.8}$	39.3$_{\pm 6.6}$ /98.1$_{\pm 5.4}$
+ **Probe-Gate**	**56.2**$_{\pm 1.9}$ /96.5$_{\pm 0.7}$	**39.0**$_{\pm 3.5}$ /95.7$_{\pm 1.2}$	26.9$_{\pm 5.9}$ /99.1$_{\pm 0.8}$	**38.3**$_{\pm 8.9}$ /98.5$_{\pm 0.5}$	18.5$_{\pm 9.8}$ /98.6$_{\pm 1.2}$	44.7$_{\pm 13.3}$ /87.0$_{\pm 6.8}$	**45.5**$_{\pm 7.2}$ /96.2$_{\pm 0.6}$	**33.9**$_{\pm 5.3}$ /97.0$_{\pm 2.0}$	**42.4**$_{\pm 7.6}$ /97.4$_{\pm 0.9}$	**43.6**$_{\pm 3.2}$ /96.8$_{\pm 3.2}$
+ Gold-Gate	86.0$_{\pm 1.7}$ /100.0$_{\pm 0.0}$	75.6$_{\pm 5.5}$ /100.0$_{\pm 0.0}$	67.5$_{\pm 6.8}$ /100.0$_{\pm 0.0}$	82.5$_{\pm 3.3}$ /100.0$_{\pm 0.0}$	40.2$_{\pm 9.5}$ /100.0$_{\pm 0.0}$	86.5$_{\pm 3.2}$ /100.0$_{\pm 0.0}$	86.4$_{\pm 0.7}$ /100.0$_{\pm 0.0}$	66.6$_{\pm 5.9}$ /100.0$_{\pm 0.0}$	75.1$_{\pm 4.3}$ /100.0$_{\pm 0.0}$	87.0$_{\pm 3.9}$ /100.0$_{\pm 0.0}$
LLaVA-1.5-7B										
No-op	14.8$_{\pm 0.4}$ /100.0$_{\pm 0.0}$	13.7$_{\pm 0.8}$ /100.0$_{\pm 0.0}$	13.1$_{\pm 2.6}$ /100.0$_{\pm 0.0}$	26.3$_{\pm 0.8}$ /100.0$_{\pm 0.0}$	37.2$_{\pm 3.1}$ /100.0$_{\pm 0.0}$	21.2$_{\pm 3.5}$ /100.0$_{\pm 0.0}$	24.5$_{\pm 1.4}$ /100.0$_{\pm 0.0}$	24.6$_{\pm 4.1}$ /100.0$_{\pm 0.0}$	46.4$_{\pm 5.5}$ /100.0$_{\pm 0.0}$	20.6$_{\pm 10.6}$ /100.0$_{\pm 0.0}$
Visual adversarial defense (reference)										
F3	89.8$_{\pm 0.4}$ 91.2$_{\pm 0.5}$	86.0$_{\pm 2.7}$ 90.0$_{\pm 2.4}$	85.8$_{\pm 5.2}$ 86.0$_{\pm 3.1}$	93.3$_{\pm 0.4}$ 94.6$_{\pm 0.5}$	89.1$_{\pm 1.9}$ 89.3$_{\pm 0.7}$	86.7$_{\pm 3.2}$ 89.8$_{\pm 2.1}$	92.8$_{\pm 0.5}$ 94.0$_{\pm 0.4}$	62.0$_{\pm 2.7}$ 63.6$_{\pm 3.0}$	84.7$_{\pm 3.2}$ 86.0$_{\pm 2.1}$	49.6$_{\pm 13.9}$ 76.2$_{\pm 12.0}$
DPS	71.7$_{\pm 0.4}$ 78.8$_{\pm 0.6}$	72.1$_{\pm 2.3}$ 79.8$_{\pm 2.3}$	65.1$_{\pm 2.5}$ 72.6$_{\pm 1.7}$	76.9$_{\pm 0.7}$ 78.2$_{\pm 1.1}$	54.1$_{\pm 1.7}$ 61.1$_{\pm 2.2}$	70.3$_{\pm 3.3}$ 74.9$_{\pm 1.4}$	47.7$_{\pm 0.7}$ 91.7$_{\pm 0.7}$	25.0$_{\pm 4.4}$ 55.5$_{\pm 6.2}$	54.8$_{\pm 3.9}$ 67.3$_{\pm 3.9}$	30.1$_{\pm 6.6}$ 52.5$_{\pm 9.5}$
Activation steering										
CAA	14.9$_{\pm 0.3}$ /98.5$_{\pm 0.2}$	15.0$_{\pm 1.8}$ /98.4$_{\pm 0.5}$	13.5$_{\pm 1.9}$ /97.3$_{\pm 0.9}$	22.2$_{\pm 1.2}$ /98.9$_{\pm 0.2}$	74.4$_{\pm 2.8}$ /81.0$_{\pm 2.4}$	**47.8**$_{\pm 3.9}$ /90.9$_{\pm 1.7}$	28.3$_{\pm 1.4}$ /97.5$_{\pm 0.6}$	25.3$_{\pm 4.1}$ /70.9$_{\pm 3.9}$	49.9$_{\pm 4.9}$ /89.5$_{\pm 2.2}$	31.8$_{\pm 14.9}$ /80.0$_{\pm 6.9}$
ITI-head	34.2$_{\pm 2.3}$ /94.8$_{\pm 1.1}$	**27.2**$_{\pm 1.6}$ /96.8$_{\pm 1.1}$	21.3$_{\pm 4.5}$ /96.7$_{\pm 0.9}$	19.7$_{\pm 0.8}$ /99.2$_{\pm 0.2}$	76.2$_{\pm 2.4}$ /77.2$_{\pm 1.8}$	62.9$_{\pm 4.1}$ /79.7$_{\pm 2.9}$	31.9$_{\pm 1.4}$ /96.6$_{\pm 0.6}$	25.9$_{\pm 3.9}$ /74.6$_{\pm 4.7}$	46.6$_{\pm 6.4}$ /87.6$_{\pm 1.8}$	27.3$_{\pm 12.1}$ /86.2$_{\pm 11.2}$
SEA	12.6$_{\pm 0.5}$ /59.9$_{\pm 1.0}$	14.1$_{\pm 1.0}$ /98.0$_{\pm 0.7}$	25.2$_{\pm 3.2}$ /82.5$_{\pm 3.4}$	43.9$_{\pm 0.8}$ /98.0$_{\pm 0.2}$	23.7$_{\pm 4.4}$ /96.3$_{\pm 1.3}$	28.9$_{\pm 1.7}$ /94.7$_{\pm 2.0}$	30.4$_{\pm 1.3}$ /92.4$_{\pm 0.6}$	25.9$_{\pm 3.0}$ /73.9$_{\pm 5.4}$	32.1$_{\pm 4.0}$ /77.0$_{\pm 2.0}$	29.2$_{\pm 13.2}$ /67.5$_{\pm 2.8}$
LEACE	15.9$_{\pm 0.6}$ /98.6$_{\pm 0.1}$	7.7$_{\pm 3.9}$ /12.9$_{\pm 4.9}$	7.8$_{\pm 3.2}$ /20.8$_{\pm 9.0}$	24.3$_{\pm 2.1}$ /98.7$_{\pm 0.1}$	12.5$_{\pm 22.8}$ /11.7$_{\pm 19.5}$	37.0$_{\pm 20.2}$ /35.7$_{\pm 17.4}$	9.0$_{\pm 4.3}$ /16.2$_{\pm 7.6}$	32.1$_{\pm 8.3}$ /71.8$_{\pm 8.0}$	22.8$_{\pm 9.9}$ /22.1$_{\pm 9.7}$	4.2$_{\pm 5.9}$ /2.5$_{\pm 5.6}$
ASTRA	18.6$_{\pm 4.3}$ /97.9$_{\pm 0.9}$	15.9$_{\pm 1.3}$ /98.0$_{\pm 1.3}$	13.3$_{\pm 2.9}$ /97.2$_{\pm 1.3}$	22.3$_{\pm 0.8}$ /98.9$_{\pm 0.2}$	63.4$_{\pm 16.3}$ /82.0$_{\pm 8.3}$	33.6$_{\pm 12.3}$ /92.6$_{\pm 4.5}$	29.5$_{\pm 2.5}$ /96.4$_{\pm 0.8}$	25.6$_{\pm 4.5}$ /69.1$_{\pm 10.1}$	40.8$_{\pm 7.2}$ /88.0$_{\pm 1.2}$	20.6$_{\pm 10.6}$ /82.5$_{\pm 5.2}$
ARGUS	15.0$_{\pm 0.8}$ /99.8$_{\pm 0.0}$	14.1$_{\pm 0.6}$ /99.8$_{\pm 0.3}$	13.5$_{\pm 2.1}$ /99.6$_{\pm 0.3}$	26.0$_{\pm 0.7}$ /99.8$_{\pm 0.1}$	43.6$_{\pm 3.0}$ /98.8$_{\pm 0.8}$	22.2$_{\pm 3.4}$ /99.8$_{\pm 0.3}$	25.0$_{\pm 1.9}$ /99.3$_{\pm 0.4}$	21.1$_{\pm 4.4}$ /98.8$_{\pm 1.3}$	45.2$_{\pm 4.6}$ /98.9$_{\pm 1.2}$	**23.1**$_{\pm 9.6}$ /98.8$_{\pm 2.8}$
AM-GARD										
+ Outlier-Gate	22.1$_{\pm 0.6}$ /91.4$_{\pm 0.6}$	5.6$_{\pm 1.2}$ /93.0$_{\pm 2.3}$	22.1$_{\pm 2.5}$ /92.2$_{\pm 1.0}$	28.3$_{\pm 0.8}$ /98.9$_{\pm 0.4}$	**44.3**$_{\pm 2.3}$ /93.9$_{\pm 0.8}$	24.8$_{\pm 4.1}$ /94.5$_{\pm 1.5}$	25.9$_{\pm 1.6}$ /94.9$_{\pm 0.5}$	28.4$_{\pm 6.2}$ /98.5$_{\pm 2.6}$	41.6$_{\pm 4.6}$ /97.9$_{\pm 0.8}$	15.0$_{\pm 14.9}$ /91.2$_{\pm 10.5}$
+ **Probe-Gate**	**44.4**$_{\pm 5.6}$ /96.6$_{\pm 1.4}$	19.7$_{\pm 5.5}$ /96.4$_{\pm 1.8}$	**39.0**$_{\pm 4.8}$ /94.5$_{\pm 1.4}$	**75.4**$_{\pm 0.9}$ /90.5$_{\pm 0.9}$	39.2$_{\pm 2.7}$ /94.2$_{\pm 1.6}$	53.9$_{\pm 4.9}$ /87.0$_{\pm 2.8}$	**60.0**$_{\pm 1.7}$ /94.5$_{\pm 2.7}$	**29.1**$_{\pm 5.9}$ /90.6$_{\pm 3.3}$	**46.5**$_{\pm 2.0}$ /97.9$_{\pm 1.8}$	17.8$_{\pm 10.9}$ /98.8$_{\pm 2.8}$
+ Gold-Gate	83.7$_{\pm 0.6}$ /100.0$_{\pm 0.0}$	58.7$_{\pm 1.7}$ /100.0$_{\pm 0.0}$	76.6$_{\pm 4.0}$ /100.0$_{\pm 0.0}$	94.1$_{\pm 0.4}$ /100.0$_{\pm 0.0}$	82.5$_{\pm 1.1}$ /100.0$_{\pm 0.0}$	86.4$_{\pm 2.5}$ /100.0$_{\pm 0.0}$	88.9$_{\pm 1.0}$ /100.0$_{\pm 0.0}$	67.5$_{\pm 4.3}$ /100.0$_{\pm 0.0}$	82.7$_{\pm 2.3}$ /100.0$_{\pm 0.0}$	31.1$_{\pm 14.2}$ /100.0$_{\pm 0.0}$

Table 9: **Recovery and pre-attack preservation, flipped inputs, semantic agreement** ($R^{\text{flip}}_{\text{sem}}/C_{\text{EM}}$, %). Each cell shows recovery followed in smaller type by pre-attack preservation; subscripts give the sample SD over five seeds (42, 1337, 2026, 123, 7). **Bold** and underlined values mark the best and second-best recovery per dataset among representation-level interventions with $C \geq 90$, excluding Gold-Gate, whose gate uses flip labels; visual adversarial defenses are shown in gray for reference. F3 and DPS (gray) are shown as complementary image- and generation-level defenses.

	Open-ended			Binary			Multiple-choice			Numeric
Method	GQA	VQAv2	AOK	POPE	VSR	NatB	MMB	MMMU	MVsn	MVsta
Qwen2-VL-7B										
No-op	23.8$_{\pm 1.2}$ /100.0$_{\pm 0.0}$	29.7$_{\pm 5.8}$ /100.0$_{\pm 0.0}$	34.4$_{\pm 5.9}$ /100.0$_{\pm 0.0}$	19.8$_{\pm 0.7}$ /100.0$_{\pm 0.0}$	19.1$_{\pm 12.7}$ /100.0$_{\pm 0.0}$	27.1$_{\pm 7.2}$ /100.0$_{\pm 0.0}$	14.7$_{\pm 0.9}$ /100.0$_{\pm 0.0}$	25.8$_{\pm 5.0}$ /100.0$_{\pm 0.0}$	29.5$_{\pm 3.7}$ /100.0$_{\pm 0.0}$	38.9$_{\pm 5.4}$ /100.0$_{\pm 0.0}$
Visual adversarial defense (reference)										
F3	80.7$_{\pm 1.0}$ 84.8$_{\pm 0.8}$	71.3$_{\pm 5.1}$ 83.5$_{\pm 3.3}$	68.0$_{\pm 4.2}$ 78.7$_{\pm 2.7}$	74.8$_{\pm 3.5}$ 91.1$_{\pm 1.0}$	55.2$_{\pm 5.9}$ 80.9$_{\pm 2.6}$	69.1$_{\pm 7.5}$ 80.0$_{\pm 7.3}$	87.0$_{\pm 0.8}$ 92.8$_{\pm 0.8}$	56.2$_{\pm 5.2}$ 65.7$_{\pm 7.8}$	59.0$_{\pm 4.4}$ 64.4$_{\pm 3.3}$	64.7$_{\pm 3.9}$ 70.3$_{\pm 7.4}$
DPS	64.6$_{\pm 1.1}$ 76.3$_{\pm 1.2}$	54.2$_{\pm 4.8}$ 78.2$_{\pm 4.4}$	62.0$_{\pm 1.7}$ 73.1$_{\pm 2.7}$	53.2$_{\pm 3.8}$ 93.1$_{\pm 0.3}$	23.5$_{\pm 12.1}$ 94.5$_{\pm 2.6}$	65.2$_{\pm 8.6}$ 73.6$_{\pm 4.0}$	30.8$_{\pm 1.4}$ 98.0$_{\pm 0.3}$	39.1$_{\pm 2.3}$ 59.4$_{\pm 3.8}$	41.0$_{\pm 4.3}$ 75.4$_{\pm 3.3}$	43.2$_{\pm 7.8}$ 75.8$_{\pm 7.2}$
Activation steering										
CAA	23.5$_{\pm 1.9}$ /99.4$_{\pm 0.2}$	34.0$_{\pm 6.3}$ /99.2$_{\pm 0.9}$	34.4$_{\pm 5.1}$ /96.6$_{\pm 0.9}$	19.8$_{\pm 1.0}$ /99.6$_{\pm 0.2}$	**22.3**$_{\pm 18.2}$ /98.2$_{\pm 2.5}$	29.7$_{\pm 9.2}$ /98.4$_{\pm 1.6}$	14.8$_{\pm 1.4}$ /99.7$_{\pm 0.1}$	28.2$_{\pm 4.6}$ /94.6$_{\pm 1.4}$	28.5$_{\pm 2.9}$ /91.2$_{\pm 2.3}$	40.1$_{\pm 5.8}$ /97.3$_{\pm 2.5}$
ITI-head	29.3$_{\pm 1.0}$ /98.4$_{\pm 0.3}$	32.1$_{\pm 5.9}$ /99.1$_{\pm 0.4}$	36.8$_{\pm 6.6}$ /95.7$_{\pm 1.7}$	18.3$_{\pm 1.9}$ /99.6$_{\pm 0.2}$	20.7$_{\pm 13.7}$ /98.2$_{\pm 1.9}$	23.6$_{\pm 8.4}$ /99.7$_{\pm 0.7}$	19.3$_{\pm 2.3}$ /99.0$_{\pm 0.2}$	24.9$_{\pm 4.5}$ /85.2$_{\pm 6.8}$	30.9$_{\pm 2.3}$ /89.3$_{\pm 2.4}$	39.6$_{\pm 9.2}$ /96.7$_{\pm 3.5}$
SEA	25.3$_{\pm 1.5}$ /59.6$_{\pm 1.4}$	35.7$_{\pm 2.9}$ /85.3$_{\pm 1.3}$	42.2$_{\pm 5.4}$ /84.6$_{\pm 1.9}$	56.8$_{\pm 3.0}$ /60.5$_{\pm 1.2}$	29.8$_{\pm 22.4}$ /70.0$_{\pm 8.7}$	34.5$_{\pm 4.0}$ /84.1$_{\pm 9.2}$	26.5$_{\pm 1.1}$ /95.8$_{\pm 0.5}$	33.0$_{\pm 6.6}$ /76.8$_{\pm 4.5}$	25.6$_{\pm 3.5}$ /72.8$_{\pm 2.2}$	32.8$_{\pm 4.6}$ /86.8$_{\pm 3.9}$
LEACE	25.4$_{\pm 2.3}$ /96.8$_{\pm 1.4}$	0.0$_{\pm 0.0}$ /0.2$_{\pm 0.5}$	0.3$_{\pm 0.6}$ /0.3$_{\pm 0.3}$	26.6$_{\pm 13.6}$ /26.3$_{\pm 15.3}$	31.4$_{\pm 37.8}$ /23.2$_{\pm 14.1}$	18.0$_{\pm 14.2}$ /14.9$_{\pm 14.6}$	1.4$_{\pm 0.4}$ /4.4$_{\pm 0.8}$	4.3$_{\pm 3.4}$ /4.3$_{\pm 3.2}$	5.3$_{\pm 3.7}$ /7.5$_{\pm 3.8}$	7.2$_{\pm 9.3}$ /6.7$_{\pm 9.8}$
ASTRA	28.8$_{\pm 3.4}$ /97.7$_{\pm 0.6}$	31.4$_{\pm 4.1}$ /98.9$_{\pm 1.5}$	34.9$_{\pm 4.8}$ /97.2$_{\pm 1.3}$	34.4$_{\pm 15.3}$ /64.8$_{\pm 23.7}$	16.2$_{\pm 6.2}$ /80.5$_{\pm 17.1}$	22.7$_{\pm 12.4}$ /93.7$_{\pm 8.2}$	26.4$_{\pm 4.3}$ /81.2$_{\pm 13.6}$	29.0$_{\pm 4.3}$ /86.3$_{\pm 6.0}$	30.8$_{\pm 2.7}$ /86.2$_{\pm 2.0}$	37.2$_{\pm 3.9}$ /96.5$_{\pm 2.1}$
ARGUS	24.1$_{\pm 1.7}$ /100.0$_{\pm 0.0}$	29.2$_{\pm 5.2}$ /100.0$_{\pm 0.0}$	33.1$_{\pm 6.2}$ /99.6$_{\pm 0.5}$	18.2$_{\pm 1.8}$ /99.9$_{\pm 0.1}$	19.1$_{\pm 12.7}$ /100.0$_{\pm 0.0}$	26.3$_{\pm 7.3}$ /100.0$_{\pm 0.0}$	14.0$_{\pm 0.5}$ /99.9$_{\pm 0.1}$	25.6$_{\pm 4.9}$ /99.6$_{\pm 0.5}$	30.0$_{\pm 2.4}$ /99.5$_{\pm 0.7}$	36.5$_{\pm 4.7}$ /99.4$_{\pm 1.4}$
AM-GARD										
+ Outlier-Gate	27.2$_{\pm 0.9}$ /91.0$_{\pm 0.9}$	37.0$_{\pm 6.3}$ /94.3$_{\pm 0.8}$	**43.4**$_{\pm 5.5}$ /97.7$_{\pm 0.7}$	24.9$_{\pm 1.7}$ /99.8$_{\pm 0.1}$	19.1$_{\pm 8.1}$ /98.6$_{\pm 1.2}$	**36.3**$_{\pm 16.9}$ /95.6$_{\pm 3.4}$	24.4$_{\pm 1.7}$ /91.3$_{\pm 0.9}$	33.4$_{\pm 5.0}$ /94.6$_{\pm 3.8}$	32.6$_{\pm 3.8}$ /96.1$_{\pm 1.8}$	39.3$_{\pm 6.6}$ /98.1$_{\pm 5.4}$
+ **Probe-Gate**	**58.6**$_{\pm 1.7}$ /96.5$_{\pm 0.7}$	**44.8**$_{\pm 1.9}$ /95.7$_{\pm 1.2}$	39.0$_{\pm 8.0}$ /99.1$_{\pm 0.8}$	**38.3**$_{\pm 8.9}$ /98.5$_{\pm 0.5}$	18.5$_{\pm 9.8}$ /98.6$_{\pm 1.2}$	44.7$_{\pm 13.3}$ /87.0$_{\pm 6.8}$	**45.5**$_{\pm 7.2}$ /96.2$_{\pm 0.6}$	**33.9**$_{\pm 5.3}$ /97.0$_{\pm 2.0}$	**42.4**$_{\pm 7.6}$ /97.4$_{\pm 0.9}$	**43.6**$_{\pm 3.2}$ /96.8$_{\pm 3.2}$
+ Gold-Gate	86.0$_{\pm 1.7}$ /100.0$_{\pm 0.0}$	75.6$_{\pm 5.5}$ /100.0$_{\pm 0.0}$	67.5$_{\pm 6.8}$ /100.0$_{\pm 0.0}$	82.5$_{\pm 3.3}$ /100.0$_{\pm 0.0}$	40.2$_{\pm 9.5}$ /100.0$_{\pm 0.0}$	86.5$_{\pm 3.2}$ /100.0$_{\pm 0.0}$	86.4$_{\pm 0.7}$ /100.0$_{\pm 0.0}$	66.6$_{\pm 5.9}$ /100.0$_{\pm 0.0}$	75.1$_{\pm 4.3}$ /100.0$_{\pm 0.0}$	87.0$_{\pm 3.9}$ /100.0$_{\pm 0.0}$
LLaVA-1.5-7B										
No-op	22.2$_{\pm 0.5}$ /100.0$_{\pm 0.0}$	26.7$_{\pm 1.8}$ /100.0$_{\pm 0.0}$	28.1$_{\pm 3.9}$ /100.0$_{\pm 0.0}$	26.3$_{\pm 0.8}$ /100.0$_{\pm 0.0}$	37.2$_{\pm 3.1}$ /100.0$_{\pm 0.0}$	21.2$_{\pm 3.5}$ /100.0$_{\pm 0.0}$	24.5$_{\pm 1.4}$ /100.0$_{\pm 0.0}$	24.6$_{\pm 4.1}$ /100.0$_{\pm 0.0}$	46.4$_{\pm 5.9}$ /100.0$_{\pm 0.0}$	20.6$_{\pm 10.6}$ /100.0$_{\pm 0.0}$
Visual adversarial defense (reference)										
F3	90.5$_{\pm 0.4}$ 91.7$_{\pm 0.5}$	88.1$_{\pm 2.7}$ 90.9$_{\pm 2.3}$	88.0$_{\pm 4.7}$ 87.7$_{\pm 3.1}$	93.3$_{\pm 0.4}$ 94.6$_{\pm 0.5}$	89.1$_{\pm 1.9}$ 89.3$_{\pm 0.7}$	86.7$_{\pm 3.2}$ 89.8$_{\pm 2.1}$	92.8$_{\pm 0.5}$ 94.0$_{\pm 0.4}$	62.0$_{\pm 2.7}$ 63.6$_{\pm 3.0}$	84.7$_{\pm 3.2}$ 86.0$_{\pm 2.1}$	49.6$_{\pm 13.9}$ 76.2$_{\pm 12.0}$
DPS	73.9$_{\pm 0.4}$ 80.3$_{\pm 0.6}$	76.8$_{\pm 2.2}$ 84.1$_{\pm 2.5}$	72.6$_{\pm 2.1}$ 78.2$_{\pm 2.7}$	76.9$_{\pm 0.7}$ 78.2$_{\pm 1.1}$	54.1$_{\pm 1.7}$ 61.1$_{\pm 2.2}$	70.3$_{\pm 3.3}$ 74.9$_{\pm 1.4}$	47.7$_{\pm 0.7}$ 91.7$_{\pm 0.7}$	25.0$_{\pm 4.4}$ 55.5$_{\pm 6.2}$	54.8$_{\pm 3.9}$ 67.3$_{\pm 3.9}$	30.1$_{\pm 6.6}$ 52.5$_{\pm 9.5}$
Activation steering										
CAA	22.1$_{\pm 0.4}$ /98.8$_{\pm 0.2}$	27.0$_{\pm 2.5}$ /98.4$_{\pm 0.5}$	28.9$_{\pm 3.5}$ /97.9$_{\pm 0.5}$	22.2$_{\pm 1.2}$ /98.9$_{\pm 0.2}$	74.4$_{\pm 2.8}$ /81.0$_{\pm 2.4}$	**47.8**$_{\pm 3.9}$ /90.9$_{\pm 1.7}$	28.3$_{\pm 1.4}$ /97.5$_{\pm 0.6}$	25.3$_{\pm 4.1}$ /70.9$_{\pm 3.9}$	49.9$_{\pm 4.9}$ /89.5$_{\pm 2.2}$	31.8$_{\pm 14.9}$ /80.0$_{\pm 6.9}$
ITI-head	40.3$_{\pm 2.2}$ /95.3$_{\pm 1.1}$	**36.2**$_{\pm 2.1}$ /98.5$_{\pm 0.8}$	34.0$_{\pm 4.9}$ /97.4$_{\pm 0.5}$	19.7$_{\pm 0.8}$ /99.2$_{\pm 0.2}$	76.2$_{\pm 2.4}$ /77.2$_{\pm 1.8}$	62.9$_{\pm 4.1}$ /79.7$_{\pm 2.9}$	31.9$_{\pm 1.4}$ /96.6$_{\pm 0.6}$	25.9$_{\pm 3.9}$ /74.6$_{\pm 4.7}$	46.6$_{\pm 6.4}$ /87.6$_{\pm 1.8}$	27.3$_{\pm 12.1}$ /86.2$_{\pm 11.2}$
SEA	14.4$_{\pm 0.6}$ /60.3$_{\pm 1.0}$	27.1$_{\pm 2.1}$ /98.0$_{\pm 0.7}$	37.9$_{\pm 3.5}$ /84.8$_{\pm 2.5}$	43.9$_{\pm 0.8}$ /98.0$_{\pm 0.2}$	23.7$_{\pm 4.4}$ /96.3$_{\pm 1.3}$	28.9$_{\pm 1.7}$ /94

Table 10: **Pre-attack preservation on all pre-attack inputs versus on flipped samples** (C_{EM} and $C_{\mathrm{EM}}^{\mathrm{flip}}$, %). Supplementary analysis: C^{flip} is measured only on pre-attack inputs whose attacked counterpart flipped the no-op answer. Only baselines with per-sample pre-attack logs are reported, and AM-GARD is not included. Subscripts give the sample SD over five seeds (42, 1337, 2026, 123, 7).

	Open-ended			Binary		Multiple-choice			Numeric
Method	GQA	VQAv2	AOK	VSR	NatB	MMB	MMMU	MVsn	MVsta
Qwen2-VL-7B									
Visual adversarial defense									
F3, C_{EM}	$83.5_{\pm 0.7}$	$81.0_{\pm 3.8}$	$73.3_{\pm 4.2}$	$80.9_{\pm 2.6}$	$80.0_{\pm 7.3}$	$92.8_{\pm 0.8}$	$65.7_{\pm 7.8}$	$64.0_{\pm 3.3}$	$70.3_{\pm 7.4}$
F3, C^{flip}	$78.2_{\pm 0.9}$	$67.6_{\pm 7.8}$	$60.4_{\pm 4.4}$	$53.2_{\pm 7.6}$	$69.7_{\pm 10.6}$	$90.5_{\pm 1.0}$	$62.3_{\pm 8.7}$	$60.5_{\pm 4.4}$	$68.2_{\pm 6.7}$
DPS, C_{EM}	$72.6_{\pm 1.5}$	$69.5_{\pm 4.9}$	$61.7_{\pm 2.7}$	$94.5_{\pm 2.6}$	$73.6_{\pm 4.0}$	$98.0_{\pm 0.3}$	$59.1_{\pm 4.1}$	$75.4_{\pm 3.3}$	$75.8_{\pm 7.2}$
DPS, C^{flip}	$66.4_{\pm 1.9}$	$49.7_{\pm 4.8}$	$49.4_{\pm 3.4}$	$87.0_{\pm 8.7}$	$74.3_{\pm 7.9}$	$97.2_{\pm 0.4}$	$57.0_{\pm 2.9}$	$72.5_{\pm 3.2}$	$75.7_{\pm 7.4}$
Activation steering									
ASTRA, C_{EM}	$97.4_{\pm 0.6}$	$98.5_{\pm 1.4}$	$95.6_{\pm 1.4}$	$80.5_{\pm 17.1}$	$93.7_{\pm 8.2}$	$81.2_{\pm 13.6}$	$86.3_{\pm 6.0}$	$86.2_{\pm 2.0}$	$96.5_{\pm 2.1}$
ASTRA, C^{flip}	$97.4_{\pm 0.6}$	$98.5_{\pm 1.4}$	$95.5_{\pm 1.5}$	$80.5_{\pm 17.1}$	$93.7_{\pm 8.2}$	$79.8_{\pm 13.8}$	$85.7_{\pm 6.2}$	$83.7_{\pm 2.7}$	$95.8_{\pm 2.3}$
ARGUS, C_{EM}	$99.7_{\pm 0.1}$	$99.9_{\pm 0.2}$	$99.6_{\pm 0.5}$	$99.1_{\pm 1.2}$	$100.0_{\pm 0.0}$	$99.9_{\pm 0.1}$	$99.4_{\pm 0.8}$	$99.1_{\pm 0.6}$	$99.7_{\pm 0.7}$
ARGUS, C^{flip}	$99.7_{\pm 0.1}$	$99.9_{\pm 0.2}$	$99.5_{\pm 0.5}$	$99.1_{\pm 1.2}$	$100.0_{\pm 0.0}$	$99.9_{\pm 0.1}$	$99.4_{\pm 0.9}$	$98.9_{\pm 0.8}$	$99.6_{\pm 0.9}$
LLaVA-1.5-7B									
Visual adversarial defense									
F3, C_{EM}	$91.2_{\pm 0.5}$	$90.0_{\pm 2.4}$	$86.0_{\pm 3.1}$	$89.3_{\pm 0.7}$	$89.8_{\pm 2.1}$	$94.0_{\pm 0.4}$	$63.6_{\pm 3.0}$	$86.0_{\pm 2.1}$	$76.2_{\pm 12.0}$
F3, C^{flip}	$90.9_{\pm 0.6}$	$90.0_{\pm 2.3}$	$85.8_{\pm 3.3}$	$90.3_{\pm 0.8}$	$89.3_{\pm 2.2}$	$93.5_{\pm 0.4}$	$62.9_{\pm 3.3}$	$86.0_{\pm 1.9}$	$68.1_{\pm 13.1}$
DPS, C_{EM}	$78.8_{\pm 0.6}$	$79.8_{\pm 2.3}$	$72.6_{\pm 1.7}$	$61.1_{\pm 2.2}$	$74.9_{\pm 1.4}$	$91.7_{\pm 0.7}$	$55.5_{\pm 6.2}$	$67.3_{\pm 3.9}$	$52.5_{\pm 9.5}$
DPS, C^{flip}	$78.7_{\pm 0.7}$	$79.6_{\pm 2.1}$	$71.8_{\pm 1.6}$	$59.4_{\pm 2.6}$	$75.6_{\pm 1.1}$	$91.1_{\pm 0.9}$	$55.4_{\pm 6.1}$	$64.9_{\pm 4.9}$	$41.1_{\pm 12.8}$
Activation steering									
ASTRA, C_{EM}	$97.9_{\pm 0.9}$	$98.0_{\pm 1.3}$	$97.2_{\pm 1.3}$	$82.0_{\pm 8.3}$	$92.6_{\pm 4.5}$	$96.4_{\pm 0.8}$	$69.1_{\pm 10.1}$	$88.0_{\pm 1.2}$	$82.5_{\pm 5.2}$
ASTRA, C^{flip}	$97.9_{\pm 0.8}$	$97.9_{\pm 1.3}$	$97.1_{\pm 1.3}$	$84.7_{\pm 7.3}$	$93.0_{\pm 3.9}$	$96.2_{\pm 0.9}$	$68.7_{\pm 9.4}$	$88.9_{\pm 0.6}$	$74.3_{\pm 4.8}$
ARGUS, C_{EM}	$99.8_{\pm 0.0}$	$99.8_{\pm 0.3}$	$99.6_{\pm 0.3}$	$98.8_{\pm 0.8}$	$99.8_{\pm 0.3}$	$99.3_{\pm 0.4}$	$98.8_{\pm 1.3}$	$98.9_{\pm 1.2}$	$98.8_{\pm 2.8}$
ARGUS, C^{flip}	$99.8_{\pm 0.0}$	$99.8_{\pm 0.3}$	$99.6_{\pm 0.3}$	$99.1_{\pm 0.8}$	$99.8_{\pm 0.3}$	$99.3_{\pm 0.4}$	$98.7_{\pm 1.3}$	$98.7_{\pm 1.3}$	$100.0_{\pm 0.0}$

agreement (sem). Every table uses the same methods, grouping, and ranking rule as Table 2, and reports the sample standard deviation over the five seeds wherever per-seed results are available. Pre-attack preservation C is always measured on the full pre-attack set; Table 10 additionally restricts it to the pre-attack counterparts of flipped samples for the baselines with per-sample pre-attack logs.

All attacked inputs, exact match. Table 6 adds seed variability to Table 2. The seed standard deviation of Probe-Gate recovery stays within a few points on most datasets, so the ranking in the main text is not driven by a single seed. The exception is LLaVA MathVista, whose test set contains only 16 items; every method, including no-op, varies widely across splits there, so conclusions on this dataset should be read as tentative.

Semantic agreement confirms recovery beyond exact-match scoring. Table 7 re-evaluates the same model outputs using semantic agreement, which credits equivalent answer forms that exact match treats as different. Semantic scoring raises absolute recovery on open-ended datasets and leaves binary and multiple-choice datasets essentially unchanged, since their answers are single labels. Because the no-op floor rises as well, gains over no-op shrink slightly, but the main trends under EM persist: Probe-Gate remains the strongest representation-level method on most datasets, and the relative behavior of the comparison methods is largely unchanged. The recovery gains are therefore not specific to exact string matching.

Flipped inputs. Tables 8 and 9 restrict recovery to attacked inputs whose answer the attack actually changed, removing inputs whose answer survived the perturbation. No-op recovery on this subset is lower than on all attacked inputs but not zero, since regeneration alone restores some answers. Probe-Gate's gain over no-op is larger on this subset, because inputs that did not need correction no longer contribute to R, and it remains the strongest representation-level method on most datasets.

The subset also makes the cost of imperfect selection more visible: every item requires intervention, so on Qwen the gap between Gold-Gate and Probe-Gate is wider than on all attacked inputs.

Pre-attack preservation on flipped samples. Table 10 compares pre-attack preservation on all pre-attack inputs with preservation on the pre-attack counterparts of flipped samples, for the four baselines. On Qwen, both visual defenses change the pre-attack response of flippable samples considerably more often than the full-set average suggests, most strongly on VQAv2, A-OKVQA, and VSR. Samples that an attack can flip thus tend to sit near an answer boundary, where input purification also perturbs the pre-attack decision. The two representation-level baselines show no such gap: ASTRA's preservation loss on Qwen VSR is already visible in C_{EM} and does not concentrate on flippable samples, and ARGUS stays near 100% on both sets. On LLaVA, the two measures are close for all four baselines except on MathVista.

G Selection and Correction Diagnostics

G.1 Selection

Recovery Estimation from the Development Set We further ask whether test recovery can be estimated from a small development set before full test evaluation. We compare a predictor based only on detection quality (TPR-only) with a fuller diagnostic that also incorporates correction reach. The detection-only predictor reflects primarily *whether* inputs requiring intervention are identified, whereas the full diagnostic additionally captures whether the selected inputs can be successfully corrected.

Table 11 reports two measures of predictive quality. Mean absolute error (MAE) is the average absolute difference between predicted and observed test recovery, so lower values indicate more accurate estimates. Spearman correlation measures how well the predicted ordering of configurations agrees with their observed test ordering, so higher values indicate better ranking consistency.

Including correction reach consistently improves both measures. With only 16 development examples, MAE decreases from 0.118 to 0.077, while Spearman correlation increases from 0.594 to 0.838. The same pattern holds as the development set grows. These results indicate that test recovery is better estimated when development-set diagnostics capture both intervention selection and subsequent correction, rather than detection quality alone.

Table 11: **Test-recovery estimation from the development set.** Detection quality alone (TPR-only) is compared with a fuller diagnostic that additionally accounts for correction reach. MAE measures the average absolute error between predicted and observed test recovery (lower is better), while Spearman correlation measures agreement in the ranking of configurations (higher is better).

Dev n	MAE (TPR-only)	MAE (full)	Spearman (TPR-only)	Spearman (full)
16	0.118	**0.077**	0.594	**0.838**
32	0.117	**0.072**	0.593	**0.849**
64	0.114	**0.068**	0.612	**0.837**

G.1.1 Which Stage Limits Recovery

We separate residual recovery error into cases where the gate remains off (*unselected error*) and those unresolved after the gate fires (*selected error*). As the gate broadens, unselected error decreases while selected error increases across all three datasets (Figure 8). Thus, stricter settings are more selection-limited, whereas broader settings increasingly expose limitations in correction. This decomposition does not assume that every unselected error would be recoverable if selected.

G.2 Correction

Candidate set size and target information. Table 12 extends the correction analysis in Section 6.3. Under Gold-Gate, enlarging the competitor set has little effect on recovery, whereas supplying the saved pre-attack first token as the correction target improves recovery. Thus, expanding the candidate set alone does not resolve target identification.

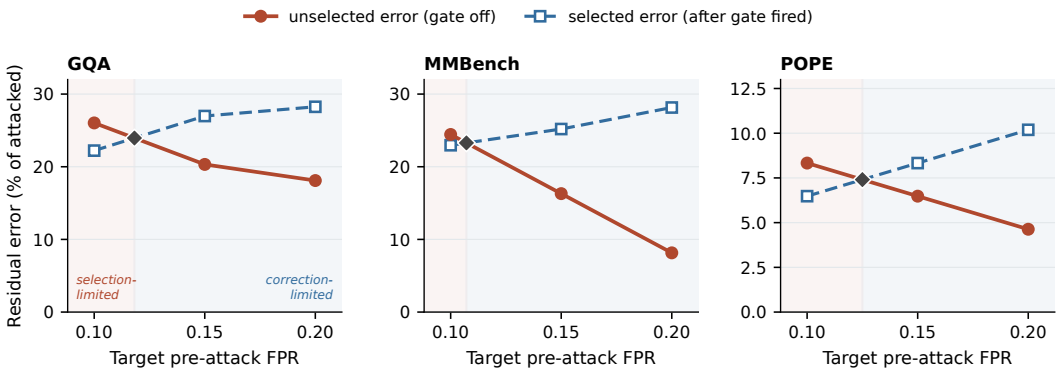


Figure 8: **Which stage limits recovery across gate settings.** Broadening the gate reduces unselected error while increasing error among selected inputs. Diamonds mark where the two error contributions cross. Points show three-seed means.

Table 12: **Candidate set size and target information.** Qwen GQA with Gold-Gate; three seeds.

Condition	Candidate count k	Recovery R^{diag} (%)
Nearest-boundary target	1	91.4
Nearest-boundary target	2	90.5
Nearest-boundary target	4 (base)	91.4
Nearest-boundary target	6	91.4
Pre-attack first token as target	4	97.8

Table 13: **Qualitative examples of competitor selection on Qwen GQA.** Normalized boundary-distance and highest-logit competitor selection use the same attacked image and gate decision in each example. The first case shows recovery after an attack-induced answer change, while the second shows preservation when correction is unnecessary. The first case is seed 2026 and the second seed 42. Capitalization is shown as generated; exact-match scoring normalizes it. Image perturbations are not visually apparent at this scale.

Question	Pre-attack image	Attacked image	Answers	Outcome
What is the material of the fence behind the horse?			Pre-attack: wood No-op: metal Normalized: wood Highest-logit: Metal	**Successful recovery.** Normalized boundary-distance selection recovers the pre-attack answer, whereas highest-logit selection does not.
On which side of the image is the curtain?			Pre-attack: left No-op: Left Normalized: left Highest-logit: Right	**Preservation under unnecessary intervention.** The no-op response already matches the pre-attack answer. Normalized selection preserves it, whereas highest-logit selection changes the answer.

Qualitative Examples of Competitor Selection The first example illustrates successful recovery after an attack-induced answer change. Although the pre-attack token *wood* has a larger raw logit gap than the highest-logit competitor (1.50 versus 1.25), its gradient-normalized boundary distance

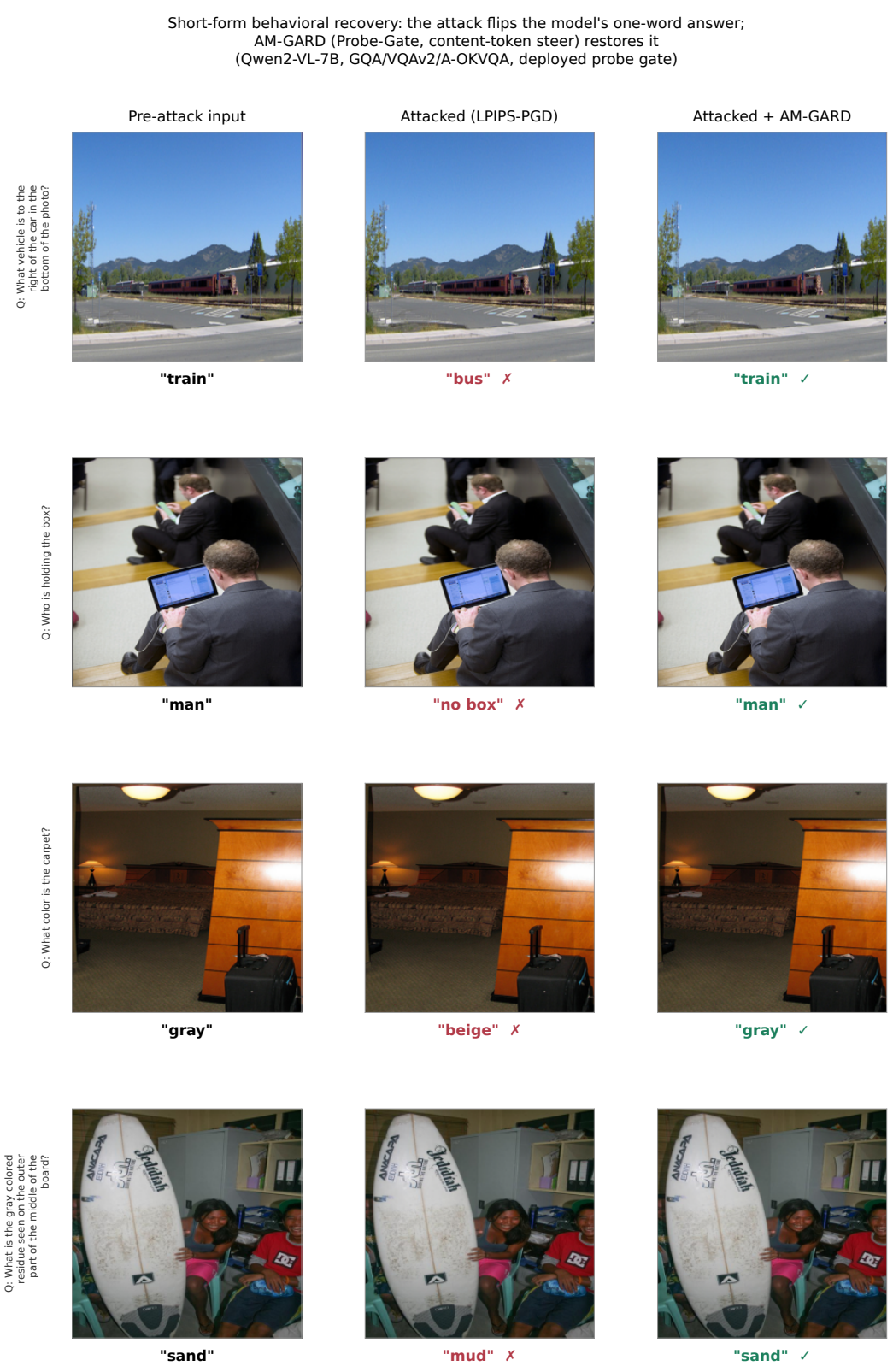


Figure 9: **Short-form behavioral recovery examples.** Representative Qwen2-VL-7B cases from GQA, VQAv2, and A-OKVQA. Each row shows the pre-attack response, the response changed by the LPIPS-PGD attack, and the response after AM-GARD with Probe-Gate. AM-GARD restores the pre-attack response in each illustrated case.

is smaller (3.57 versus 4.24). Normalized boundary-distance selection therefore recovers *wood*, whereas highest-logit competitor selection produces *Metal*.

The second example illustrates preservation when intervention is unnecessary. The attacked no-op response already agrees with the pre-attack answer after normalization. Normalized boundary-distance selection preserves *left*, whereas highest-logit competitor selection changes the response to *Right*. Together, these examples illustrate why changing an answer or selecting the highest-logit competitor does not necessarily lead to recovery.

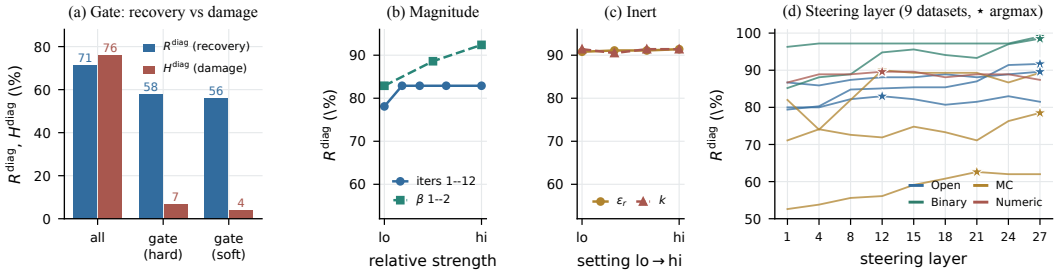


Figure 10: **Selection and correction sensitivity on the development set.** (a) Selective intervention substantially reduces response harm relative to applying correction to every input. (b) Recovery improves with correction strength, with diminishing returns from additional relinearization. (c) Recovery changes little across the evaluated values of ϵ_r and candidate-set size k. (d) The recovery-maximizing steering layer varies across the nine evaluated datasets; stars mark the best layer for each dataset. Each panel is a separate controlled sweep and is interpreted only within its own setting.

Selection and Correction Sensitivity Figure 10 examines how the main design choices affect selection and correction. Because the panels use different controlled development-set settings, absolute values are interpreted only within each sweep.

Selective intervention primarily improves preservation (Figure 10a). Applying correction to every input yields high recovery but also substantially changes responses that do not require correction. Introducing a gate sharply reduces this harm, and score-scaled gating further improves preservation with only a small change in recovery. This supports separating *whether* to intervene from the correction itself.

Correction strength has a clearer effect on recovery (Figure 10b). Additional relinearization improves recovery initially and then saturates, while increasing the final scale β produces a larger gain over the evaluated range. Recovery therefore benefits from a sufficiently strong correction, but repeated updates show diminishing returns.

By contrast, recovery is comparatively stable across the evaluated values of ϵ_r and candidate-set size k (Figure 10c). In particular, increasing the number of candidate competitors provides little additional recovery, consistent with Section 6.3: broader candidate coverage alone does not ensure that correction reaches the pre-attack answer. Finally, the preferred steering layer varies across datasets (Figure 10d), motivating layer selection on development data rather than fixing a single layer globally.

Table 14 summarizes the settings and corresponding development-set results.

Why Subspace Editing Can Hurt Pre-Attack Preservation We examine why subspace-based methods can improve recovery on some tasks while reducing pre-attack response preservation. If attack-induced representation changes were concentrated in a small low-rank subspace, a few singular directions would capture most of the displacement and could be removed with limited effect on the remaining representation.

Table 15 shows that real attacked–pre-attack differences are substantially less concentrated than a synthetic low-rank corruption. In several datasets, the fitted subspace also removes a substantial fraction of variance from pre-attack representations. Thus, subspace editing can suppress not only attack-related variation but also representation components that support the original response, helping explain the preservation cost observed for methods such as SEA and LEACE.

Table 14: **Selection and correction sweep results.** R^{dev} denotes recovery and H^{dev} response harm on the development set. Each row corresponds to a separate controlled sweep and should be compared only within that row.

Component	Sweep	Result (%)
Selection		
Intervention gate	apply-to-all → hard gate	$R^{\text{dev}}: 71.4 \to 57.8$; $H^{\text{dev}}: 76.2 \to 6.7$
Gate threshold	target FPR .10 → .15 → .20	$R^{\text{dev}}: 56.2 \to 62.9 \to 63.8$; $H^{\text{dev}}: 4.8 \to 5.7 \to 11.4$
Gate application	hard → score-scaled soft	$R^{\text{dev}}: 57.8 \to 56.2$; $H^{\text{dev}}: 6.7 \to 4.1$
Correction		
Relinearization	iterations $1 \to 3/5/8$	$R^{\text{dev}}: 78.1 \to 82.9$
Final scale β	$\beta: 1 \to 1.5$	$R^{\text{dev}}: 82.9 \to 88.6$
Final scale β	$\beta: 1 \to 2$ (seed 42)	$R^{\text{dev}}: 88.6 \to 92.4$
Margin parameter ϵ_r	.01 → .4	$R^{\text{dev}}: 90.8 \to 91.4$
Candidate set k	$1 \to 6$	$R^{\text{dev}}: 91.4 \to 91.4$

Table 15: **Low-rank concentration and pre-attack representation overlap.** κ_r is the fraction of attacked–pre-attack difference energy captured by the top r singular directions, and $\hat{\gamma}$ measures the fraction of pre-attack representation variance removed by projection onto the fitted subspace.

Dataset (Qwen2-VL-7B, L24)	κ_1	κ_4	κ_{16}	$\hat{\gamma}$
GQA	0.23	0.39	0.50	0.12
VQAv2	0.18	0.27	0.45	0.05
A-OKVQA	0.20	0.29	0.46	0.04
POPE	0.38	0.63	0.74	0.67
VSR	0.33	0.64	0.85	0.62
NaturalBench	0.40	0.67	0.82	0.49
MMBench	0.23	0.73	0.82	0.64
MMMU-Pro	0.10	0.31	0.64	0.14
MathVision	0.12	0.36	0.65	0.12
MathVista	0.10	0.33	0.72	0.11
Synthetic low-rank corruption	—	≈ 0.99	—	—

G.3 Qualitative Analysis

Figure 9 shows representative short-form recovery cases from GQA, VQAv2, and A-OKVQA. In each case, the attack changes the pre-attack answer, while AM-GARD with Probe-Gate restores the original response. The examples cover different answer types, including objects, persons, colors, and materials, and qualitatively illustrate recovery beyond a single question type.

H Robustness, Transfer, and Generalization

This section examines whether the selection–correction separation remains informative under adaptive attack objectives, alternative representation structure, and a qualitatively different attack. These analyses are diagnostic extensions of the main evaluation rather than additional end-to-end benchmarks.

H.1 Adaptive Attacks on the Gate

Gate-Aware Perturbation We first test an adaptive attacker that explicitly accounts for the deployed gate. The attack augments the answer-changing objective with a term that suppresses the probe score, while retaining the same perceptual budget as the primary attack in Section 2. Joint success therefore requires both changing the answer and avoiding intervention.

Across 40 Qwen examples per dataset, joint success is 67.5% on GQA, 57.5% on POPE, and 25.0% on MMBench. These joint-success cases are end-to-end recovery failures because gate evasion prevents correction. The criterion does not count additional failures after intervention. The adaptive

perturbations also remain close to the original perceptual budget: the median LPIPS ratio between the adaptive and original attacks is approximately one across the three datasets.

The lower joint success on MMBench coincides with its deeper selected gate layer, but gate depth and dataset vary together in this comparison. We therefore treat this as a dataset-level difference rather than evidence that deeper gates are intrinsically harder to evade.

Likelihood-Based Attack We next replace the first-token margin objective with a sequence negative-log-likelihood (seq-NLL) objective. This provides a complementary attack whose optimization target is not defined by the local first-token boundary used by AM-GARD.

When the attacker does not optimize against the gate, detection evasion is low on GQA and POPE and higher on MMBench (Table 16). Explicitly optimizing for gate evasion raises the evasion rate on all three datasets. The resulting tradeoff, however, is not uniform: answer-attack success changes little on GQA and POPE, whereas on MMBench it drops substantially from 0.700 to 0.275. Adaptive optimization also requires markedly larger perceptual changes, with median LPIPS increasing by approximately 2.6–$7\times$ and the 90th percentile by 9–$32\times$.

These results show that gate evasion is possible under an alternative attack objective, but doing so changes the attack tradeoff rather than simply bypassing selection at no additional cost. More importantly, the selection behavior is not specific to the first-token-margin attack used in the main experiments.

Table 16: **Adaptive robustness under the seq-NLL attack.** Gate-unaware and gate-aware attacks are compared by detection evasion and answer-attack success rate (ASR). Optimizing against the gate increases evasion, with the largest reduction in ASR observed on MMBench.

Dataset	Gate-unaware evasion	Gate-aware evasion	ASR
GQA	0.025	0.800	$0.925 \rightarrow 0.900$
POPE	0.000	0.450	$0.325 \rightarrow 0.300$
MMBench	0.375	0.550	$\mathbf{0.700 \rightarrow 0.275}$

H.2 Generalization to a Semantic-Swap Attack

Finally, we test whether the selection–correction decomposition remains useful under a qualitatively different perturbation. The semantic-swap attack inserts a distractor object into the scene rather than optimizing a small perceptual perturbation. We evaluate this setting on 36 Qwen GQA examples as a diagnostic rather than a full benchmark.

The same separation between selection and correction remains visible, but the dominant source of error changes. A matched-norm fixed direction provides essentially no improvement over no-op, again indicating that a single global direction does not reliably recover instance-specific answers. The selection signal also transfers weakly to this attack ($\mathrm{TPR} = 0.14$).

More importantly, supplying idealized intervention selection produces only a small additional recovery gain under the semantic-swap attack. This differs from the primary PGD setting, where improving selection leaves substantial recovery headroom. The semantic-swap diagnostic is therefore more correction-limited: once an affected input is selected, the remaining difficulty lies in moving the answer-prediction state toward the pre-attack answer.

Although the sample is small, the result supports the same conceptual decomposition under a different attack mechanism while showing that the relative importance of *whether* and *how* can change with the attack.

I Input-Space Defense Analysis

We separately examine input-space transformations to characterize how their recovery behavior differs from selective representation-level intervention.

Recovery under fixed attacks. On Qwen GQA, resizing the attacked image to $0.5\times$ yields $R = 0.886$, while JPEG-$q75$ compression yields $R = 0.857$ on the same evaluation set. Under the fixed attack used to generate these examples, both transformations recover a substantial fraction of the pre-attack answers. These results show that simple preprocessing can be effective when it attenuates the applied image perturbation. These results are measured on fixed attacked inputs and do not constitute an adaptive-attack evaluation.

Intervention coverage. Input-space transformations such as resize and JPEG are applied to every image, whereas AM-GARD uses detector-based selection to decide whether representation-level correction is applied. The two approaches therefore operate at different intervention coverage: input preprocessing modifies all inputs, while AM-GARD can leave selected pre-attack responses untouched. Their recovery and preservation results should be interpreted with this difference in mind.

J SENTENCE-FORM RECOVERY

We extend the recovery evaluation to sentence-form outputs, where the answer content may occur after the first generated token. For each example, the answer-bearing position is identified from the saved pre-attack response, and recovery is evaluated by whether the generated sentence contains the original answer content. All AM-GARD variants use this position localization; Gold-Gate additionally uses saved pre-attack answers for intervention selection.

Table 17: **Sentence-form recovery and pre-attack response preservation** (%). R^{all} is measured on all attacked inputs, R^{flip} on a separately constructed flipped subset, and C on all pre-attack inputs. Recovery is scored by the original answer content in the generated sentence. **Bold** and underlined values denote the best and second-best values among representation-level interventions with $C \geq 90$, excluding Gold-Gate. Visual adversarial defenses are shown for reference. F3 and DPS (gray) are shown as complementary image- and generation-level defenses (Table 1). All AM-GARD variants use oracle answer-position localization from saved pre-attack responses.

	Qwen2-VL-7B									LLaVA-1.5-7B								
	GQA			VQAv2			A-OKVQA			GQA			VQAv2			A-OKVQA		
Method	R^{all}	R^{flip}	C	R^{all}	R^{flip}	C	R^{all}	R^{flip}	C	R^{all}	R^{flip}	C	R^{all}	R^{flip}	C	R^{all}	R^{flip}	C
No-op	45.9	0.0	100.0	64.5	0.0	100.0	48.0	0.0	100.0	32.1	0.0	100.0	43.1	0.0	100.0	53.2	0.0	100.0
Visual adversarial defense (reference)																		
F3	77.8	79.3	80.4	63.8	43.5	72.4	46.9	0.0	100.0	89.2	84.9	94.2	92.3	92.4	94.0	86.0	77.3	89.2
DPS	59.9	56.5	65.6	62.9	54.5	74.2	72.0	61.5	70.0	70.7	70.5	83.6	78.4	77.3	86.2	67.0	72.7	75.5
Activation steering																		
CAA	50.3	8.2	**100.0**	66.1	4.5	93.5	48.0	0.0	98.0	33.6	2.1	99.3	47.0	4.6	99.1	55.3	4.5	98.9
ITI-head	**58.0**	**22.4**	96.8	**71.0**	**18.2**	90.3	52.0	7.7	88.0	**47.9**	**23.2**	96.4	54.7	18.5	97.4	**59.6**	**13.6**	96.8
SEA	54.8	16.5	89.2	69.4	13.6	79.0	**56.0**	**15.4**	94.0	40.0	11.6	86.4	**56.4**	**21.5**	92.3	68.1	31.8	86.2
LEACE	45.9	0.0	0.6	66.1	4.5	95.2	48.0	0.0	0.0	34.3	3.2	6.4	44.4	0.0	12.0	59.6	13.6	45.7
ASTRA	49.0	5.9	**100.0**	61.3	0.0	95.2	51.0	3.8	96.1	34.3	3.2	99.3	44.4	0.0	**100.0**	54.3	2.3	98.9
ARGUS	47.1	2.4	**100.0**	64.5	8.3	**100.0**	49.0	0.0	**100.0**	32.1	0.0	**100.0**	44.4	0.0	**100.0**	54.3	2.3	**100.0**
AM-GARD																		
+ Outlier-Gate	48.4	4.7	91.7	64.5	0.0	91.9	49.0	0.0	96.1	33.6	2.1	95.0	43.7	2.9	94.1	54.3	2.3	93.6
+ **Probe-Gate**	66.9	38.8	88.5	**71.0**	**18.2**	96.8	54.0	11.5	96.0	39.3	10.5	98.6	54.6	21.7	87.4	58.5	11.4	92.5
+ Gold-Gate	82.8	68.2	100.0	80.6	45.5	100.0	64.0	30.8	100.0	77.9	67.4	100.0	76.5	59.4	100.0	79.8	56.8	100.0

Results. Sentence-form recovery remains possible when correction is applied at the answer-bearing position rather than the first response token. Probe-Gate improves R^{flip} over no-op across both models and all three datasets, while Gold-Gate provides a higher diagnostic recovery level under the same localization setting. Fixed-direction steering is more variable across datasets and models, and high recovery can coincide with lower preservation. Overall, the results support extending answer-margin correction beyond the first-response-token setting while preserving the same distinction between intervention selection and correction.

Qualitative examples. Figure 11 illustrates representative sentence-form recovery cases. Each example shows how the attacked response differs from the saved pre-attack response and how answer-margin correction changes the generated sentence.



Figure 11: **Sentence-form recovery examples.** Representative Qwen GQA cases showing the question, pre-attack response, attacked response, and response after answer-margin correction.

Overall, the results support extending answer-margin correction beyond the first-response-token setting while preserving the same distinction between intervention selection and correction. Representative recovery cases are shown in Figure 11.