LOOKS THE SAME, ANSWERS DIFFERENTLY: FLIP-DIRECTION STEERING FOR ROBUST VISION-LANGUAGE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) achieve strong visual reasoning performance, yet subtle changes from routine image capture and processing can alter their reasoning trajectories even when images appear nearly identical. In long-horizon generation, the resulting activation shifts may accumulate across decoding steps, progressively altering reasoning tokens and ultimately changing the final answer, a phenomenon referred to as *answer flips*. To address this instability, we propose FLIPDIR (Flip-Direction Steering), a training-free inference-time method that estimates a low-rank flip-inducing activation subspace from contrastive pairs of original and answer-flipping inputs and selectively steers hidden states during decoding. A margin-based gate limits subspace attenuation to uncertain decoding steps, recovering original predictions while preserving stable ones. To evaluate robustness beyond accuracy or consistency on fixed test sets, we introduce VISFLIP, a benchmark framework that constructs evaluation groups for a target model and visual variation setting to separately assess recovery of original predictions and preservation of stable ones. VISFLIP spans nine dataset–variation combinations across scientific reasoning, robot-scene understanding, and medical VQA, covering subtle visual variations common in each domain. Experiments across 18 settings demonstrate that FLIPDIR consistently outperforms existing methods on the combined recovery and preservation metric. We will make our code publicly available.

1 INTRODUCTION

Recent vision-language models (VLMs) achieve strong performance across scientific reasoning, robot-scene understanding, and medical visual question answering (Zhang et al., 2024b; Chen et al., 2025; Li et al., 2023). Despite these advances, generating reliable and consistent responses to diverse visual inputs in real-world scenarios remains challenging (Rosenfeld et al., 2025). This requires models to preserve task-relevant visual information across image variations and integrate it consistently with the textual context throughout generation. These challenges have motivated research on several aspects of VLM reliability, including adversarial robustness (Xie et al., 2025), multimodal safety (Wang et al., 2025b), and visual hallucination (Liu et al., 2025). Recent studies further show that, even without deliberate attacks, benign corruptions such as additive noise can degrade VLM performance (Qiu et al., 2025).

However, common real-world scenarios involve benign inputs with subtle visual variations rather than deliberate attacks, unsafe requests, or pronounced image degradation. Although safety and resistance to adversarial prompts receive considerable attention in recent post-training (Gemma Team et al., 2025; Meta, 2025), the effects of subtle visual variations, such as slight exposure shifts or JPEG compression, on long-horizon visual reasoning remain underexplored. As shown in Table 1, we observe that subtle exposure shifts can change final predictions even in frontier models, a phenomenon we call *answer flips*. On M3CoT (Chen et al., 2024b), a benchmark for multi-step multimodal reasoning, flip rates are 9.6% for GPT-5.5 and 21.7% for Claude Opus 4.8 (OpenAI, 2026; Anthropic, 2026) (details in Appendix B.2). Figure 1 shows an *answer flip* under a subtle exposure shift, where initially identical generations first differ at *a new task* versus *another task* and later reach different action predictions. Moreover, these early token changes can remain locally plausible, making it difficult to identify when intervention is needed.

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

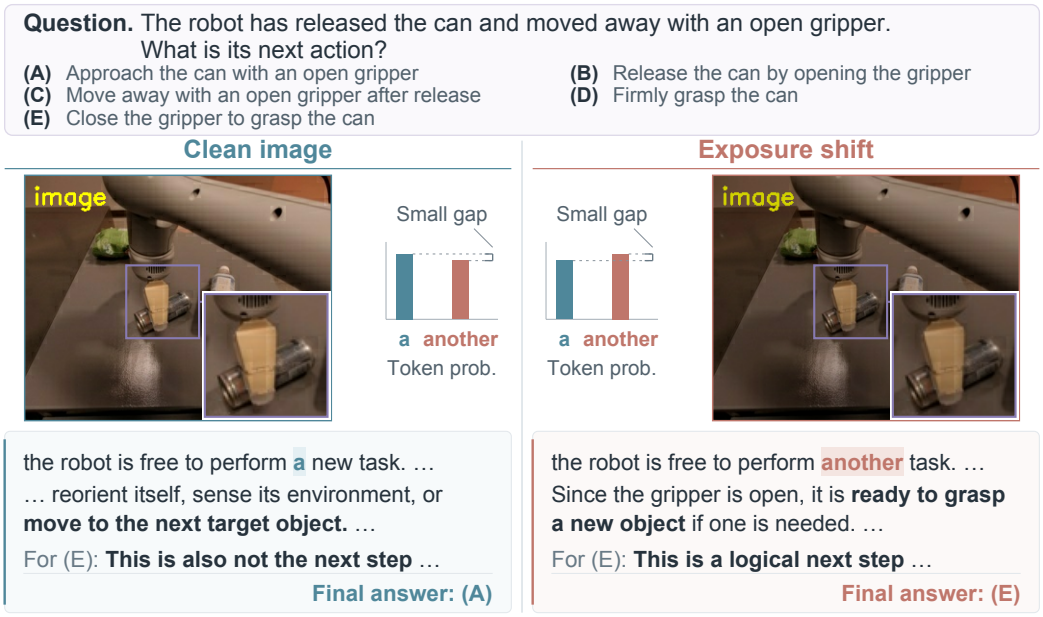


Figure 1: **A subtle exposure shift can flip the final answer.** For nearly identical Robo2VLM images, Qwen3-VL-8B produces identical text before diverging at *a* versus *another* and reaching different action predictions. Token probabilities are schematic.

Table 1: ***Answer flips* on recent models.**

	GPT-5.5	Claude Opus 4.8	Qwen3-VL-8B	Gemma-3-12B
Clean accuracy (%)	93.2	91.8	68.3	47.8
Answer flip (%)	**9.6**	**21.7**	**47.6**	**62.4**

To address this challenge, we propose FLIPDIR (Flip-Direction Steering), a training-free inference-time method that estimates a low-rank flip-inducing activation subspace and uses it to selectively steer hidden states during decoding. We estimate this subspace offline by applying singular value decomposition (SVD) to activation differences from a set of contrastive pairs of original and answer-flipping inputs. During inference on new inputs, a margin-based gate uses the gap between the two largest next-token probabilities to select uncertain decoding steps for intervention (Figure 5). When the gate opens, we attenuate the part of the current hidden state that lies in the estimated subspace; otherwise, we retain the unedited next-token distribution. Calibration requires no gradient updates or auxiliary model training, and inference uses the stored basis without a corresponding original image.

Our evaluation targets robustness to subtle visual variations rather than improvements in task accuracy. Existing benchmarks typically measure accuracy or consistency on fixed evaluation sets (Qiu et al., 2025; Zhang et al., 2024a). However, accuracy can increase when a visual change happens to correct an originally wrong prediction, even though the model fails to maintain its original decision. High consistency may result from perturbations that barely affect the model, while increasing their severity can make the evaluation less representative of real-world scenarios. To overcome these limitations, we introduce VISFLIP, a benchmark framework that constructs evaluation groups for a chosen model, dataset, and visual variation setting based on whether the original predictions change. This model-specific construction enables separate evaluation of recovery and preservation in the setting of interest. We instantiate the framework across nine dataset–variation combinations spanning scientific reasoning, robot-scene understanding, and medical VQA (Section 4; Appendix A.2).

Across 18 settings, FLIPDIR improves the harmonic mean of these metrics by 2.8 points on average and up to 6.3 points over the strongest baseline in each setting. Ablation studies show that subspace estimation from contrastive pairs and margin-based gating both improve the balance between recovery and preservation. We will publicly release the complete benchmark, a reproducible construction and evaluation framework for additional models and perturbation settings, and all experimental code.

Our contributions are threefold:

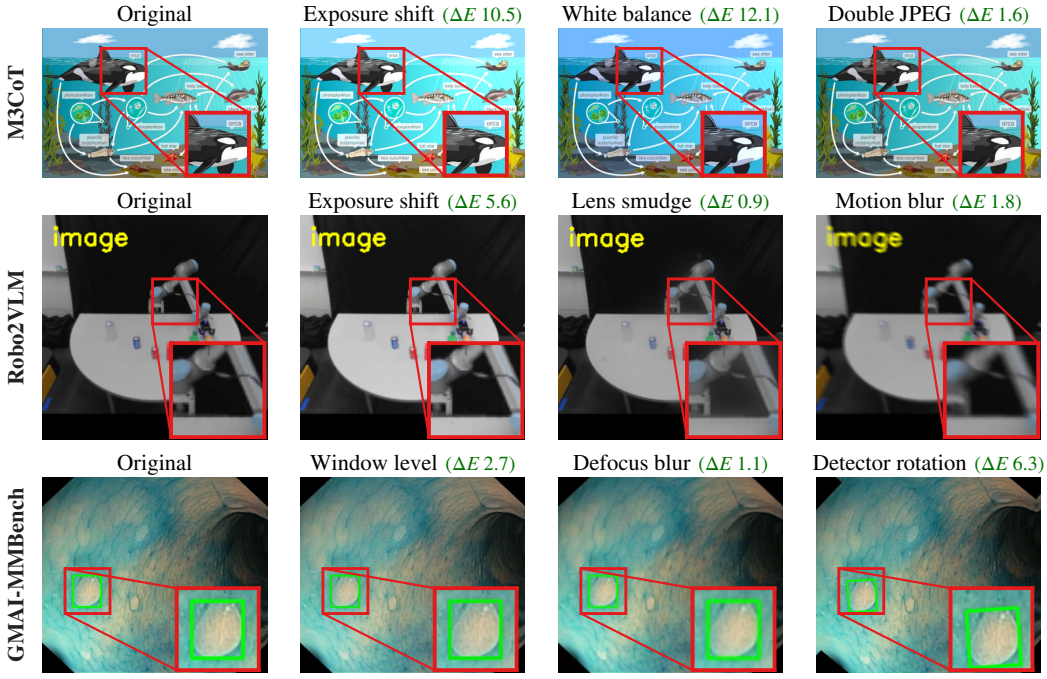


Figure 2: **Visual variations in VISFLIP.** Rows show subtle variations common to each domain. ΔE denotes the mean per-pixel color difference from the original image (parameter ranges in Appendix A.2).

- We analyze *answer flips* under subtle real-world visual variations, examining the associated activation shifts and decoding divergence.
- We propose FLIPDIR, a training-free inference-time method that estimates a low-rank flip-inducing activation subspace and uses margin-based gating to selectively steer hidden states during decoding.
- We introduce VISFLIP, a benchmark framework for model-specific evaluation of prediction recovery and preservation. It covers scientific reasoning, robot-scene understanding, and medical VQA under subtle visual variations, including lens contamination and medical image window-level changes.

2 RELATED WORK

Robustness to real-world visual perturbations. Prior work studies natural corruptions, acquisition shifts, and semantics-preserving transformations (Hendrycks & Dietterich, 2019; Di Salvo et al., 2024). Recent benchmarks evaluate task accuracy under image corruptions (Qiu et al., 2025) or aggregate output consistency over a fixed set of corrupted inputs (Zhang et al., 2024a). These fixed sets are constructed independently of the target model's responses. Accuracy can reward a prediction change that happens to correct an original error, whereas high consistency may simply indicate that the corruption barely affects the evaluated model. Increasing corruption severity can make the evaluation less representative of real-world visual variations. As a result, these scores do not directly measure robustness relative to a given model's original predictions. They combine cases where the visual change alters the original prediction with cases where the prediction remains unchanged.

Jailbreaking and safety steering. Refusal can be mediated by a single activation direction (Arditi et al., 2024), while multimodal defenses use activation differences associated with adversarial visual features to suppress harmful responses (Wang et al., 2025a). Such methods target an identifiable behavior.

Hallucination mitigation. Decoding methods improve visual grounding through visual contrast, token dependencies, or uncertainty (Leng et al., 2024; Huang et al., 2024; Xu et al., 2026); activation-based methods steer representations toward grounded behavior (Liu et al., 2025).

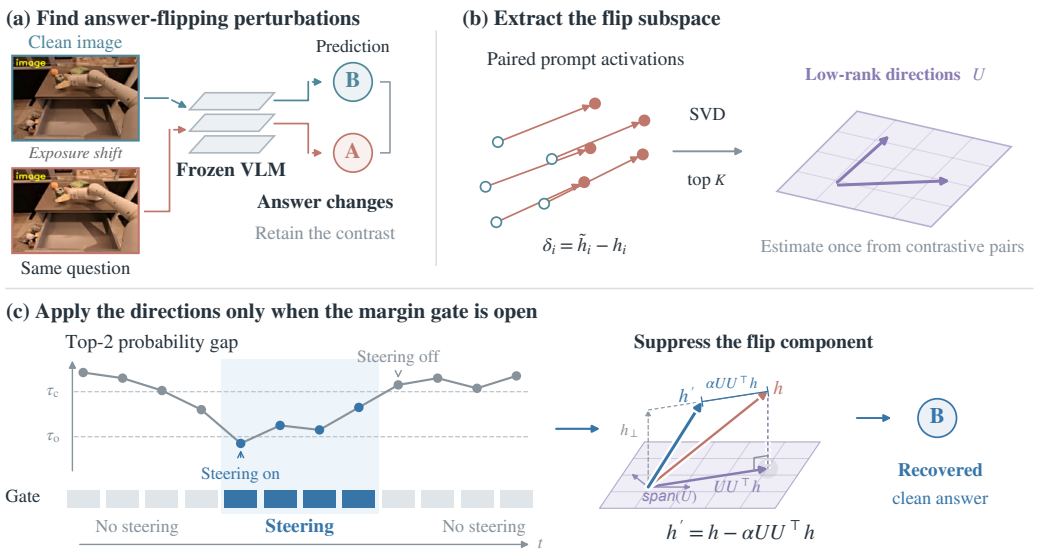


Figure 3: **FLIPDIR overview.** (a) Identify *answer flips*. (b) Estimate U via SVD of paired activation differences. (c) Attenuate the subspace projection at margin-gated decoding steps. The plane denotes $\mathrm{span}(U)$, and increasing α moves h' toward $h_\perp = (I - UU^\top)h$.

3 FLIPDIR: SELECTIVE STEERING FOR ANSWER-FLIP RECOVERY

FLIPDIR estimates a low-rank flip-inducing activation subspace from activation differences between original inputs and their answer-flipping variants. During inference on new inputs, a margin-based gate selects decoding steps at which to attenuate the part of the hidden state that lies in this subspace (Figure 3). Subspace estimation is performed offline using a set of contrastive pairs; inference reuses the resulting basis without gradient updates or auxiliary model training.

3.1 PROBLEM SETUP

We study answer instability under mild visual changes that leave images nearly identical, such as exposure shifts, blur, and JPEG recompression. The objective is to recover the model's original answer when a perturbation changes it, while preserving answers on unaffected inputs. Let f be a VLM with original answer $y = f(I, T)$ for image I and prompt T. Let $\mathcal{P}$ denote the sampling distribution over transformations in a predefined perturbation family. For $\tau \sim \mathcal{P}$, let $\tilde{I} = \tau(I)$. A perturbation is *flip-inducing* when

$$\mathrm{norm}(f(\tilde{I}, T)) \neq \mathrm{norm}(f(I, T)), \tag{1}$$

where $\mathrm{norm}(\cdot)$ maps the model output to the task-specific answer space. Flip labels are model-specific because they reference the model's clean prediction. Calibration is performed separately for each model, dataset, and known perturbation family; the resulting intervention is reused on new inputs from that setting without a clean reference or advance knowledge of their flip status.

3.2 ESTIMATING THE FLIP-INDUCING SUBSPACE

For each model and perturbation setting, let $\mathcal{D}_{\mathrm{flip}} = \{(I_i, \tilde{I}_i, T_i)\}_{i=1}^{n}$ be a set of contrastive pairs satisfying Eq. 1, with at most one pair per original input and $n = |\mathcal{D}_{\mathrm{flip}}|$. This selection isolates answer-changing perturbations within families that can also preserve predictions.

Paired activation differences. Let $h_\ell(I, T) \in \mathbb{R}^d$ be the hidden state at layer ℓ for the final prompt token after multimodal prefill. We extract this state from unedited passes on the clean and flipped inputs and compute

$$\delta_i = h_\ell(\tilde{I}_i, T_i) - h_\ell(I_i, T_i), \qquad \tilde{I}_i = \tau_i(I_i). \tag{2}$$

Thus, δ_i measures the paired activation shift at the final prompt-token position. Estimating directions from contrastive examples and reusing them during generation follows prior activation-steering work (Arditi et al., 2024; Rimsky et al., 2024; Liu et al., 2025). We use the final prompt token as a common comparison point before generated continuations diverge. We stack the differences from n pairs into $\Delta = [\delta_1, \ldots, \delta_n]^\top \in \mathbb{R}^{n \times d}$.

Subspace estimation. We approximate these activation differences with a k-dimensional subspace. An orthonormal basis U that minimizes the reconstruction error is obtained by

$$\min_{U^\top U = I_k} \|\Delta - \Delta U U^\top\|_F^2, \qquad U = V_{:,1:k} \in \mathbb{R}^{d \times k}, \tag{3}$$

where V contains the right singular vectors of $\Delta = P\Sigma V^\top$, ordered by decreasing singular value. Thus, U retains k singular directions, with $1 \le k \le \mathrm{rank}(\Delta) \le \min(n, d)$. We use uncentered SVD so that the mean difference, which may contain a shared clean-to-flip shift, contributes to the estimated subspace. The basis is computed once from these contrastive pairs and fixed during inference. We call $\mathcal{S}_{\text{flip}} = \mathrm{span}(U) \subseteq \mathbb{R}^d$ the estimated flip-inducing subspace. It summarizes activation shifts under answer-flipping perturbations and may also include other perturbation-related variation.

Spectral concentration. We measure the fraction of squared activation differences captured by the top k directions as $E(k) = \sum_{j=1}^{k} \sigma_j^2 / \sum_j \sigma_j^2$, where σ_j are the singular values. For 255 Qwen3-VL-8B/Robo2VLM lens-smudge pairs ($d = 4096$), Figure 4 compares layers 12 and 36 with a Gaussian null; layer 4 is shown in Appendix D.1 (Figure 10(a)). The leading direction alone captures 15.6% to 21.2% of the energy, compared with 0.6% for the null. The first eight directions account for only 3.1% of the maximum rank but retain 48.6–53.7% of the energy, suggesting that a compact subspace captures much of the observed variation. Prior work similarly finds that a single-neuron intervention at one layer can suppress refusal across several LLMs (Kazemi et al., 2026). Paired differences may also contain image- and prompt-specific variation, so retaining more directions need not improve steering on new inputs. Consistent with this interpretation, our rank ablation across five settings shows no systematic gain beyond $k = 8$ (Section 5.4), which we adopt throughout.

3.3 Margin-gated steering

At inference, we use the estimated basis U to attenuate the hidden-state component within the subspace. We apply this update selectively, using the next-token margin to determine when to intervene.

Hidden-state steering. Let $h \in \mathbb{R}^d$ be the hidden state of the current token at the selected layer ℓ. We steer this state by attenuating its projection onto the estimated subspace:

$$h' = h - \alpha U U^\top h = (I - UU^\top)h + (1 - \alpha)UU^\top h. \tag{4}$$

The update scales the projection $UU^\top h$ by $1 - \alpha$, leaving its orthogonal component unchanged. Thus, $\alpha = 1$ removes the projection, while $0 < \alpha < 1$ partially attenuates it. At each step, an unedited forward pass first determines the margin and gate state. When the gate opens, we apply the hidden-state update and recompute the next-token distribution; otherwise, we retain the distribution from the unedited pass. We edit only decoding states; prompt prefill is unchanged. During inference, steering uses only the current hidden state and stored basis, without a corresponding clean image, activation, or answer. Section 5.1 and Appendix C.3 detail the experimental settings and validation-based selection of ℓ and α.

Next-token margin. Applying this update at every decoding step can also alter stable predictions. We therefore use the model's next-token uncertainty to select when to intervene. At decoding step t, we first run an unedited forward pass conditioned on the tokens generated so far. Let $p_t^{(1)}$ and $p_t^{(2)}$ be the two largest next-token probabilities. Their difference, $m_t = p_t^{(1)} - p_t^{(2)}$, is small when the model assigns similar probabilities to competing tokens. To test its association with *answer flips*, we analyze stored unedited generation traces. For flipped cases, we label differing next-token predictions under a shared prefix as *divergence tokens*, realigning the prefix after each mismatch; matching predictions are *non-divergence tokens*. Across 7 Gemma and 8 Qwen settings (2,775 divergence and 88,071 non-divergence tokens), divergence tokens have lower median margins in every setting. Pooled over tokens, 93.0% of divergence tokens have $m_t < 0.3$, compared with 11.7% of non-divergence tokens (Figure 5 for Qwen; Appendix D.2, Figure 10b for Gemma). This association motivates using a low margin to trigger steering. The original input is used only to label tokens for this analysis; the gate does not require it at inference.

Margin gate. With $\tau_o < \tau_c$, the stateful gate opens when $m_t < \tau_o$, closes when $m_t > \tau_c$, and retains its previous state otherwise. Gate decisions always use the unedited probabilities, so they do not depend on confidence introduced by the current edit. When the gate is open, we apply Eq. 4; otherwise, we use the original next-token distribution.

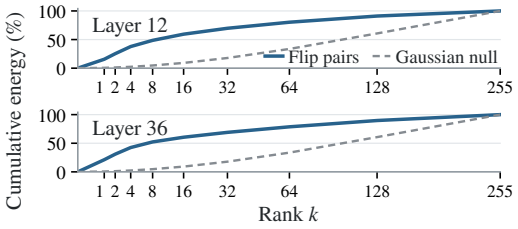


Figure 4: **Low-rank concentration.** Leading directions capture more energy in comparison with a Gaussian null.

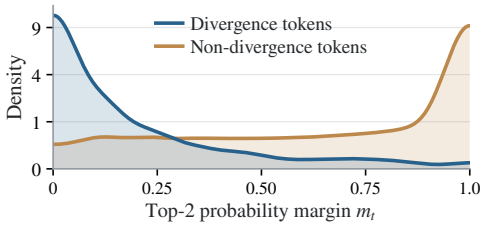


Figure 5: **Next-token margins.** Divergence tokens tend to have lower margins in Qwen3-VL-8B.

4 VISFLIP

VISFLIP is a benchmark framework for model-specific evaluation of prediction recovery and preservation under subtle visual variations. For a chosen model, dataset, and variation setting, it groups variants of each image–instruction pair according to whether they change the model's original prediction. Together with the original inputs, these flip and non-flip groups separately measure recovery of original predictions and preservation of stable ones, evaluating robustness rather than improvements in task accuracy (Section 5.1). We instantiate the framework across 18 settings spanning scientific reasoning, robot-scene understanding, and medical VQA, with subtle visual variations common in each domain's real-world settings (Figure 2).

Science reasoning. We use the physics subset of M3CoT (Chen et al., 2024b), which requires multi-step reasoning over scientific diagrams and scenes. We consider exposure shift, white-balance shift, and double JPEG compression, covering changes in illumination, camera color processing, and image recompression during capture and transmission (Hendrycks & Dietterich, 2019; Afifi et al., 2019).

Robot scenes. We use Robo2VLM (Chen et al., 2025), which covers spatial relations, robot state, and action outcomes in manipulation scenes. We consider exposure shift, lens contamination, and motion blur, reflecting lighting changes, contamination of robot cameras, and camera or platform motion (Hendrycks & Dietterich, 2019; Park et al., 2026).

Medical VQA. We use GMAI-MMBench (Chen et al., 2024a), which spans clinical VQA tasks, perceptual granularities, and imaging modalities. Motivated by medical image formation and display (Bushberg et al., 2020; Goodman, 2005; Mannan & Langer, 2016), we use synthetic operators for window-level variation, modality-specific blur, and small in-plane rotation (Appendix A.5).

Table 2 lists the number of parent examples in each split and their representative tasks or modalities. GMAI-MMBench splits are stratified by clinical task and perceptual granularity, preserving these distributions and modality proportions. Benchmark construction requires 12.8 ± 6.4 VLM calls per search on average, including unsuccessful searches (Table 8). Appendices A.1 through A.4 detail construction and filtering, and Appendix A.5 covers medical preprocessing. We will release the constructed evaluation sets and the framework for reproducible variant generation, model-specific flip labeling, and recovery–preservation evaluation.

Table 2: Composition of VISFLIP across three domains.

Domain	Source	Train	Val	Test	Coverage
Science reasoning	M3CoT	1,582	221	450	Forces & motion, energy, heat, magnetism
Robot scenes	Robo2VLM	1,500	200	567	Robot state, reachability, depth, next action
Medical VQA	GMAI-MMBench	300	200	200	X-ray, CT, endoscopy, fundus

5 EXPERIMENTS

We evaluate recovery, preservation, transfer, and computational cost on VISFLIP using two open-weight VLMs and inference-time baselines.

Table 3: Main results on **Qwen3-VL-8B** (%): clean preservation (C), non-flip preservation (N), flip recovery (F), and their harmonic mean H. Best H is bold; green differences indicate **FLIPDIR**'s gain in H over the strongest baseline.

M3CoT	Exposure				White balance				JPEG			
Method	C	N	F	H	C	N	F	H	C	N	F	H
Base	100	100	0.0	0.0	100	100	0.0	0.0	100	100	0.0	0.0
VTI	54.1	50.5	33.9	44.2	54.1	56.3	37.9	47.9	54.1	57.7	38.5	48.6
VCD	81.6	84.6	60.7	74.0	83.0	82.7	62.6	74.8	82.3	82.5	55.6	71.0
LEAD	87.6	89.0	62.8	77.8	87.6	87.6	56.3	73.9	87.6	87.4	56.1	73.8
FLIPDIR	87.4	87.9	64.5	**78.3** (+0.5)	84.6	84.1	64.2	**76.4** (+1.6)	87.2	87.1	66.3	**78.9** (+5.1)
Robo2VLM	Exposure				Lens smudge				Motion blur			
Base	100	100	0.0	0.0	100	100	0.0	0.0	100	100	0.0	0.0
VTI	90.2	90.0	61.7	78.1	90.3	89.9	54.7	74.1	90.2	90.7	55.3	74.6
VCD	91.1	90.5	62.6	79.0	91.9	93.0	66.3	81.7	91.1	89.4	63.8	79.3
LEAD	96.1	94.3	60.7	80.1	96.4	97.1	58.1	79.2	96.1	94.1	47.9	71.6
FLIPDIR	92.5	93.2	67.3	**82.4** (+2.3)	94.1	94.0	66.3	**82.5** (+0.8)	93.2	93.0	67.0	**82.4** (+3.1)
GMAI-MMBench	Window level				Defocus blur				Rotation			
Base	100	100	0.0	0.0	100	100	0.0	0.0	100	100	0.0	0.0
VTI	45.8	47.4	39.3	43.9	45.8	43.4	32.9	39.9	45.8	42.7	37.3	41.6
VCD	71.4	78.9	59.8	69.1	76.6	72.5	45.9	61.7	72.4	76.0	50.0	63.9
LEAD	85.4	81.1	45.5	65.2	85.4	78.8	40.0	60.7	85.4	84.4	44.5	65.2
FLIPDIR	79.7	84.2	54.5	**70.1** (+1.0)	80.2	81.5	44.7	**63.7** (+2.0)	80.2	81.8	56.4	**70.7** (+5.5)

Table 4: Main results on **Gemma-3-12B**; notation as in Table 3.

M3CoT	Exposure				White balance				JPEG			
Method	C	N	F	H	C	N	F	H	C	N	F	H
Base	100	100	0.0	0.0	100	100	0.0	0.0	100	100	0.0	0.0
VTI	39.2	34.9	31.7	35.0	39.2	38.8	29.0	35.0	39.2	35.4	33.9	36.0
VCD	66.4	66.2	42.4	55.8	66.1	66.3	48.2	58.9	66.1	67.3	44.3	57.1
LEAD	63.0	66.4	47.8	57.9	63.0	64.7	51.8	59.3	61.0	64.9	56.0	60.4
FLIPDIR	70.2	69.1	50.0	**61.6** (+3.7)	74.4	72.5	45.2	**60.8** (+1.5)	70.6	70.9	51.7	**63.0** (+2.6)
Robo2VLM	Exposure				Lens smudge				Motion blur			
Base	100	100	0.0	0.0	100	100	0.0	0.0	100	100	0.0	0.0
VTI	48.3	46.6	32.9	41.4	48.3	48.9	32.7	41.8	48.3	49.1	33.2	42.1
VCD	80.1	79.5	41.6	61.1	80.6	79.7	47.1	64.9	78.8	80.7	44.9	63.4
LEAD	84.5	85.3	56.4	72.7	83.1	83.6	52.9	69.9	84.5	83.0	44.4	64.6
FLIPDIR	88.2	87.8	55.0	**73.4** (+0.7)	87.8	86.4	56.2	**73.6** (+3.7)	85.7	82.8	50.8	**69.1** (+4.5)
GMAI-MMBench	Window level				Defocus blur				Rotation			
Base	100	100	0.0	0.0	100	100	0.0	0.0	100	100	0.0	0.0
VTI	39.5	41.7	34.3	38.3	39.5	42.9	33.8	38.4	39.5	41.6	35.7	38.8
VCD	82.0	81.9	45.1	64.4	78.5	81.2	42.9	62.0	79.0	83.2	42.1	61.9
LEAD	80.5	76.9	51.0	66.6	80.5	75.9	48.1	64.6	80.5	75.1	44.4	62.2
FLIPDIR	88.5	89.4	49.0	**70.0** (+3.4)	82.5	84.3	54.5	**70.9** (+6.3)	84.5	83.2	43.7	**64.2** (+2.0)

5.1 EXPERIMENTAL SETUP

Models. Since FLIPDIR requires access to intermediate hidden states, we evaluate it on two open-weight VLMs, Gemma-3-12B-IT (Gemma Team et al., 2025) and Qwen3-VL-8B-Instruct (Bai et al., 2025a).

Baselines. We compare against the unmodified model (Base), two decoding-time methods, VCD (Leng et al., 2024) and LEAD (Xu et al., 2026), and the activation-steering method VTI (Liu et al., 2025). Implementation details appear in Appendix C.4.

Metrics. We evaluate each model on its own original, flip, and non-flip groups: preservation of original predictions on clean (C) and non-flip (N) inputs, and recovery of those predictions on flipped inputs (F). Our primary metric is their harmonic mean,

$$H(\mathrm{C},\mathrm{N},\mathrm{F}) = \left[\frac{1}{3}\left(\frac{1}{\mathrm{C}}+\frac{1}{\mathrm{N}}+\frac{1}{\mathrm{F}}\right)\right]^{-1}, \tag{5}$$

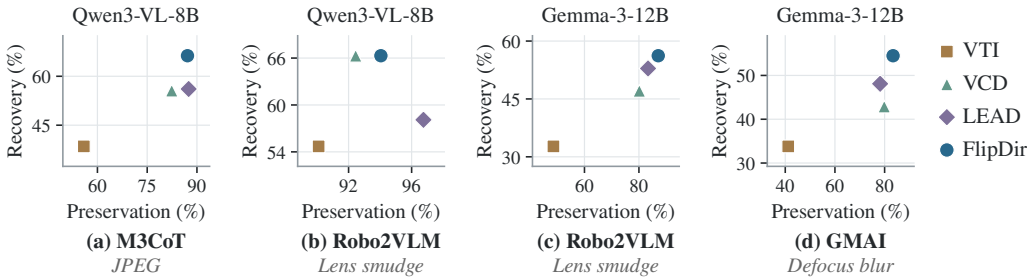


Figure 6: **Recovery–preservation trade-offs in selected settings.** Each point is a reported configuration, not a tuning sweep. All 18 settings, including Base, appear in Appendix E.1.

with $H(\mathrm{C}, \mathrm{N}, \mathrm{F}) = 0$ if any component is zero. All three metrics use the model's original prediction as the reference, measuring robustness to visual variations rather than task accuracy. Recovery measures restoration of that prediction, including when it is incorrect. Reporting F separately from C and N prevents stable cases from masking poor recovery.

Implementation details. We use rank $k = 8$ and shared gate thresholds $(\tau_{\mathrm{o}}, \tau_{\mathrm{c}}) = (0.3, 0.7)$ across all settings. For each model and perturbation setting, we estimate the flip subspace from the training split and select the steering layer ℓ and strength α on the validation split. The test split is used only for final evaluation. We use deterministic decoding where available for reproducibility. Implementation and selection details appear in Appendix C.

5.2 Main Results

FlipDir achieves the highest $H(\mathrm{C}, \mathrm{N}, \mathrm{F})$ in all 18 settings, exceeding the strongest baseline by 2.8 points on average (Tables 3 and 4). Mean gains are 2.5, 2.5, and 3.4 points for science, robotics, and medical VQA, and 2.4 and 3.2 points for Qwen3-VL-8B and Gemma-3-12B, respectively. These gains reflect a better balance between recovering original predictions and preserving stable ones, rather than higher task accuracy. On Gemma/Robo2VLM exposure, FlipDir slightly reduces recovery relative to LEAD (55.0 versus 56.4) but improves C/N from 84.5/85.3 to 88.2/87.8, raising H from 72.7 to 73.4. On Qwen/M3CoT JPEG, it instead raises recovery from the strongest baseline's 56.1 to 66.3 while keeping C/N above 87%, improving H by 5.1 points.

Figure 6 compares flip recovery (F) with average clean and non-flip preservation, $P = (\mathrm{C}{+}\mathrm{N})/2$, in four settings. Across all 18 settings, FlipDir is Pareto-nondominated among the reported operating points, with some baselines also on the frontier (Appendix E.1).

5.3 Transfer Across Datasets and Perturbation Combinations

We transfer only U across datasets and perturbation combinations, retaining target-specific layers and strengths (Figure 7). Across the twelve off-diagonal pairs, H decreases by 0.1 to 6.8 points relative to the target-matched subspace. The largest relative decrease occurs for Robo2VLM lens smudge → M3CoT exposure on Qwen (71.5 versus 78.3, retaining 91% of target-matched H). This probes shared subspace structure under target-specific calibration, rather than calibration-free deployment; dataset and perturbation effects are not isolated. Appendix E.2 gives full results.

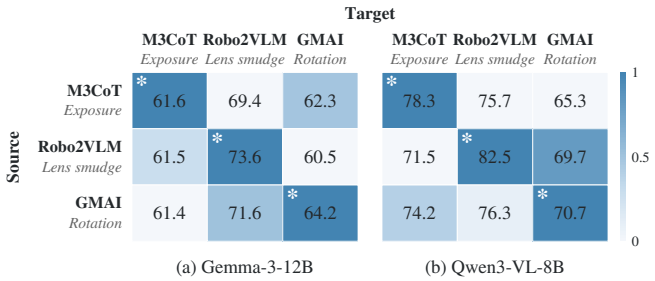


Figure 7: **Flip-subspace transfer** (H, %). Column-wise min–max colors compare sources for each target. White asterisks mark matched source–target pairs.

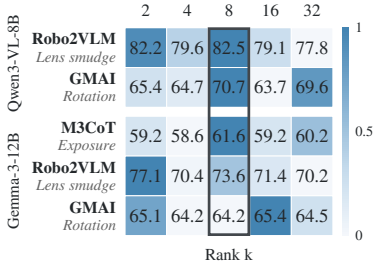


Figure 8: **Rank sensitivity** (H, %). Row-normalized colors; boxed $k = 8$. GMAI denotes GMAI-MMBench.

5.4 Ablation Studies

Representation difference. Flip-inducing and non-flipping perturbations yield similar mean activation-shift norms (15.2 versus 15.6) but differ in direction, suggesting that *answer flips* are associated with the direction rather than the magnitude of the representation change. On Robo2VLM exposure and GMAI window level, we re-estimate U for Gemma-3-12B from three representation differences while holding all other components fixed (Table 5). *Flip − Clean* achieves the highest H in both evaluations. These results suggest that conditioning on *answer flips* enriches perturbation-induced representation shifts for directions useful for recovery.

Table 5: Representation ablation (H, %).

	Gemma-3-12B	
Representation	Robo2VLM	GMAI
Flip − Clean	**73.4**	**70.0**
Non-flip − Clean	61.5	67.9
Flip − Non-flip	68.8	64.6

Table 6: Margin-gate ablation (H, %); Δ = On − Off.

	Gemma-3-12B			Qwen3-VL-8B		
Setting	Off	On	Δ	Off	On	Δ
M3CoT exp.	54.8	**61.6**	+6.8	75.6	**78.3**	+2.7
Robo2VLM lens	68.8	**73.6**	+4.8	73.5	**82.5**	+9.0
GMAI rot.	58.3	**64.2**	+5.9	69.2	**70.7**	+1.5

Margin gate. Table 6 compares margin-gated steering with steering at every decoding step, holding other components fixed within each setting. The gate improves H in all six comparisons. Figure 13 in Appendix F.2 shows gains in C, N, F, and H; Figure 9(a) varies thresholds on Qwen3-VL-8B/M3CoT exposure, with the H axis starting at 74%.

Steering strength. Figure 9(b) varies α on Gemma-3-12B/Robo2VLM exposure. Partial attenuation at $\alpha = 0.5$ achieves higher H than either weaker steering at $\alpha = 0.1$ or full projection removal at $\alpha = 1$. Each sweep evaluates one setting.

Subspace rank. Across the five settings in Figure 8, increasing k beyond 8 yields no consistent gain, although the best rank varies by setting. Additional ablations appear in Appendix F.

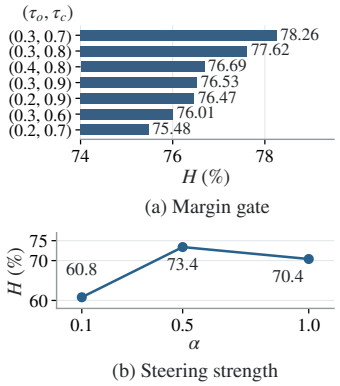


Figure 9: **Gate and strength sensitivity.** Other components are fixed.

Computational efficiency. FlipDir extracts paired activations and computes U by SVD once for reuse across inputs. At each decoding step, an unedited forward supplies the margin for gating. An additional forward pass applies steering only when the gate is open. With lower-layer activations reused, estimated decoding FLOPs increase by 8% on average at matched sequence lengths (Table 7). Appendices G.1 and G.3 detail offline costs, estimation assumptions, and measured runtimes.

Table 7: Offline preparation and mean estimated decoding FLOPs relative to Base.

	Base	LEAD	VTI	VCD	FlipDir
Offline directions	No	No	Yes	No	Yes
Decode FLOPs	1.00×	≈ 1.08×	≈ 1.00×	≈ 2.00×	≈ 1.08×

6 Conclusion

We study *answer flips* under subtle visual changes common in real-world image capture, processing, and transmission. Observations of activation spectra and token margins motivate FlipDir, which selectively attenuates a learned flip subspace during decoding without weight updates or a clean reference. VisFlip evaluates robustness relative to each model's original predictions, separating recovery from preservation across scientific reasoning, robot-scene understanding, and medical VQA. Across 18 settings, FlipDir improves the harmonic mean of preservation and recovery by 2.8 points on average over the strongest baseline.

Limitations and future work. FlipDir does not yet identify a single subspace that generalizes across perturbation types and supports steering throughout decoding without disrupting stable predictions. Future work will investigate whether such a shared subspace can remove the need for setting-specific calibration and margin gating.

AI USE STATEMENT

We used LLMs or other AI tools to aid or polish writing, for research ideation or execution, to draft sections of the paper, and for the implementation of experimental code. The authors guided and reviewed all AI-assisted work and take full responsibility for the final manuscript and associated artifacts.

ETHICS STATEMENT

Our benchmark construction procedure searches for subtle, model-specific visual variations that change a VLM's prediction. Although intended for robustness evaluation and mitigation, the same procedure could be misused to identify failure-inducing inputs for deployed models. This risk should be considered when applying the released code to safety-critical systems.

REPRODUCIBILITY STATEMENT

We will publicly release the complete VISFLIP dataset, the code used to construct it, and all experimental code, including implementations of FLIPDIR and the evaluated baselines, evaluation scripts, and analysis and figure-generation scripts. Section 3 describes the method, and Section 4 and Appendix A document benchmark construction, perturbation settings, and data splits. Appendix B specifies the evaluation pools and budget-matching protocol. Appendix C provides implementation details, hyperparameter selection, baseline adaptations, and inference settings. The release will also include versioned configurations, random seeds and provenance metadata, and the extracted flip subspaces, as described in Appendix C.5. Computational cost accounting is provided in Appendix G.1.

REFERENCES

Mahmoud Afifi, Brian Price, Scott Cohen, and Michael S. Brown. When color constancy goes wrong: Correcting improperly white-balanced images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1535–1544, 2019.

Anthropic. Introducing Claude Opus 4.8. `https://www.anthropic.com/news/claude-opus-4-8`, 2026. Model announcement, May 28, 2026. System card: `https://www.anthropic.com/claude-opus-4-8-system-card`.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Jerrold T. Bushberg, J. Anthony Seibert, Edwin M. Leidholdt, Jr., and John M. Boone. *The Essential Physics of Medical Imaging*. Wolters Kluwer, 4th edition, 2020.

Kaiyuan Eric Chen, Shuangyu Xie, Zehan Ma, Pannag Sanketi, and Ken Goldberg. Robo2VLM: Improving visual question answering using large-scale robot manipulation data. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL `https://papers.nips.cc/paper_files/paper/2025/`

hash/1f467c3e37abf9f86c78f44c6a27ee7c-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track. arXiv:2505.15517.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, Benyou Wang, Shaoting Zhang, Bin Fu, Jianfei Cai, Bohan Zhuang, Eric J Seibel, Yu Qiao, and Junjun He. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical AI. In *Advances in Neural Information Processing Systems*, volume 37, 2024a. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/ab7e02fd60e47e2a379d567f6b54f04e-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track. arXiv:2408.03361.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. M^3CoT: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8199–8221, Bangkok, Thailand, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.446. URL https://aclanthology.org/2024.acl-long.446/.

Francesco Di Salvo, Sebastian Doerrich, and Christian Ledig. MedMNIST-C: Comprehensive benchmark and improved classifier robustness by simulating realistic image corruptions. In *MICCAI Workshop on Advancing Data Solutions in Medical Imaging AI (ADSMI)*, pp. 1–11, 2024. URL https://arxiv.org/abs/2406.17536.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. VLMEvalKit: An open-source ToolKit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 11198–11201, 2024. doi: 10.1145/3664647.3685520.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Joseph W. Goodman. *Introduction to Fourier Optics*. Roberts & Company, 3rd edition, 2005.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13418–13427, 2024.

Hamid Kazemi, Atoosa Chegini, and Maria Safi. A single neuron is sufficient to bypass safety alignment in large language models. *arXiv preprint arXiv:2605.08513*, 2026.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.nips.cc/paper/2023/hash/5abcdf8ecdcacba028c6662789194572-Abstract-Datasets_and_Benchmarks.html. Datasets and Benchmarks Track.

Sheng Liu, Haotian Ye, and James Y. Zou. Reducing Hallucinations in Large Vision-Language Models via Latent Space Steering. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b4008025c2182bfe16fcc8566ee14d64-Abstract-Conference.html.

Fahim Mannan and Michael S. Langer. What is a good model for depth from defocus? In *13th Conference on Computer and Robot Vision (CRV)*, pp. 273–280, 2016. doi: 10.1109/CRV.2016.61.

Meta. Llama 4 model card. https://github.com/meta-llama/llama-models/blob/main/models/llama4/MODEL_CARD.md, 2025.

OpenAI. GPT-5.5 system card. https://openai.com/index/gpt-5-5-system-card/, April 2026.

Sungyong Park, Sooyoung Choi, Hyunsuh Koh, Youngjae Choi, and Heewon Kim. CLP: A real-world dataset of contaminated lens protectors for robust semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3794–3804, 2026.

Xinkuan Qiu, Meina Kan, Yongbin Zhou, and Shiguang Shan. Benchmarking multimodal large language models against image corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9014–9023, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Qiu_Benchmarking_Multimodal_Large_Language_Models_Against_Image_Corruptions_ICCV_2025_paper.html.

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. URL https://aclanthology.org/2024.acl-long.828/.

Amir Rosenfeld, Neta Glazer, and Ethan Fetaya. Questioning the stability of visual question answering. *arXiv preprint arXiv:2511.11206*, 2025. URL https://arxiv.org/abs/2511.11206.

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, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Wang_Steering_Away_from_Harm_An_Adaptive_Approach_to_Defending_Vision_CVPR_2025_paper.html.

Yu Wang, Xiaofei Zhou, Yichen Wang, Geyuan Zhang, and Tianxing He. Jailbreak large vision-language models through multi-modal linkage. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1466–1494, 2025b. doi: 10.18653/v1/2025.acl-long.74. URL https://aclanthology.org/2025.acl-long.74/.

Peng Xie, Yequan Bie, Jianda Mao, Yangqiu Song, Yang Wang, Hao Chen, and Kani Chen. Chain of attack: On the robustness of vision-language models against transfer-based adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14679–14689, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Xie_Chain_of_Attack_On_the_Robustness_of_Vision-Language_Models_Against_CVPR_2025_paper.html.

Zhongxing Xu, Zhonghua Wang, Zhe Qian, Dachuan Shi, Feilong Tang, Ming Hu, Shiyan Su, Xiaocheng Zou, Wei Feng, Dwarikanath Mahapatra, Yifan Peng, Minquan Lin, and Zongyuan Ge. Thinking in uncertainty: Mitigating hallucinations in MLRMs with latent entropy-aware decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11064–11075, 2026.

Jiawei Zhang, Tianyu Pang, Chao Du, Yi Ren, Bo Li, and Min Lin. Benchmarking large multimodal models against common corruptions. *arXiv preprint arXiv:2401.11943*, 2024a.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *Transactions on Machine Learning Research*, 2024b.

A Benchmark Construction and Data

A.1 Benchmark construction

Algorithm 1 instantiates the VisFlip framework for a specified dataset, model, and variation family. Applying it to another model requires obtaining that model's original predictions and flip labels, rather than reusing labels from another model. We first obtain a greedy clean-image answer and exclude parents whose answer cannot be normalized to the option space. Perturbations are then sampled from the fixed ranges in Table 9 using a per-parent seed chain and evaluated with the same prompt. Each child is labeled relative to its parent's clean prediction using Eq. 1. The full-budget procedure evaluates all B draws. When only the existence of both child types is required, sampling may stop after finding one flip and one non-flip child. The released field `stable` denotes non-flip children. Seeds, parameters, prompts, outputs, and parent identifiers are retained for reconstruction.

Algorithm 1 Answer-flip benchmark construction with a fixed draw budget

Require: Base dataset $\mathcal{D}$, VLM f, perturbation family $\mathcal{P}$, budget B

for all $(I, T, y^\star) \in \mathcal{D}$ **do**
  $y_{\text{clean}} \leftarrow f(I, T)$
  if y_{clean} cannot be normalized **then**
    continue
  end if
  for $b = 1$ to B **do**
    Sample $\tau_b \sim \mathcal{P}$ using a deterministic seed chain
    $\tilde{I}_b \leftarrow \tau_b(I)$; $\tilde{y}_b \leftarrow f(\tilde{I}_b, T)$
    $z_b \leftarrow \mathbb{1}[\text{norm}(\tilde{y}_b) \neq \text{norm}(y_{\text{clean}})]$
    Store image, seed, variation parameters, raw output, normalized answer, and label z_b
  end for
end for

Construction cost. Sampling stops when both child types are found, after 15 draws without a flip, or after at most 50 draws. This stopping rule differs from the full-budget procedure in Algorithm 1. Table 8 reports the number of VLM calls under this stopping rule, including the clean call and subsequent perturbed-image calls. Each search is defined by a parent, a perturbation family, and a backbone; the same source parent may therefore contribute to multiple searches. Across 27,478 searches, construction uses 352,970 calls: 87,384 for searches that obtain both child types and 265,586 for those that do not. Counts include all calls made within each search, not only the retained children.

Table 8: **VLM calls for benchmark construction.** Calls per search are reported as mean ± standard deviation. A search succeeds when it obtains both a flip and a non-flip child.

	Pair found			Pair not found			
Domain	Searches	Calls/search	Calls	Searches	Calls/search	Calls	All calls
Gemma-3-12B							
M3CoT	3,817	6.25 ± 5.08	23,858	2,276	16.82 ± 5.28	38,271	62,129
Robo2VLM	1,485	6.99 ± 5.09	10,387	4,716	16.10 ± 1.90	75,946	86,333
GMAI-MMBench	781	6.43 ± 5.00	5,021	719	17.02 ± 5.90	12,239	17,260
Qwen3-VL-8B							
M3CoT	2,971	9.71 ± 7.14	28,847	3,062	16.45 ± 3.93	50,357	79,204
Robo2VLM	1,312	10.02 ± 7.46	13,148	4,863	16.11 ± 1.94	78,333	91,481
GMAI-MMBench	876	6.99 ± 5.22	6,123	600	17.40 ± 6.86	10,440	16,563
All	11,242	7.77 ± 6.21	87,384	16,236	16.36 ± 3.52	265,586	**352,970**

A.2 Perturbation definitions and ranges

Perturbation operators. Exposure adjusts image intensity in EV units; white balance applies a global illuminant shift (Afifi et al., 2019); double JPEG recompresses the image twice at distinct quality levels. Motivated by real lens contamination (Park et al., 2026), our synthetic lens-smudge operator composites one to two Bezier wipe smears (opacity 0.28 to 0.45), two to four droplets (0.30 to 0.50), and two to six dust specks, with local blur ($\sigma = 4.5$), 22% haze, and an opacity cap

Table 9: **Perturbation ranges.** Each instance uses a deterministic seed chain. Appendix A.5 specifies modality-dependent medical operators.

Domain	Variation	Sampling band
M3CoT	Exposure shift	$\|\text{EV}\| \in [0.45, 0.65]$
M3CoT	White-balance shift	1000 to 3000 K
M3CoT	Double JPEG	$q_1 \in [85, 95]$, $q_2 \in [72, 85]$, $q_2 \neq q_1$
Robo2VLM	Exposure shift	$\|\text{EV}\| \in [0.45, 0.65]$
Robo2VLM	Lens smudge	wipe smears, droplets, dust; opacity ≤ 0.6, blur $\sigma = 4.5$
Robo2VLM	Motion blur	inertial trajectory, intensity 0.35
GMAI-MMBench	Window-level shift	contrast and gamma in $[0.80, 1.25]$
GMAI-MMBench	Defocus blur	modality-gated PSF (disk or Gaussian)
GMAI-MMBench	Detector rotation	$\pm 1.5^\circ$ to 4°, reflect-pad and center crop

of 0.6. Motion blur uses an inertial camera-shake trajectory at intensity 0.35. Medical operators are specified in Appendix A.5. Reported ΔE is the mean per-pixel CIE76 difference in CIELAB space.

A.3 Dataset composition and task coverage

Table 2 summarizes the parent images used in the main benchmark. M3CoT contributes eight physics topics covering scientific diagrams and multi-step conceptual reasoning. Robo2VLM contributes eight question types spanning robot state, reachability, spatial correspondence, depth, action outcomes, and grasping. GMAI-MMBench is organized by clinical VQA task, perceptual granularity, and imaging modality, with both lens-based and detector-based images represented. Parents in the training split serve as offline calibration data to estimate U, validation parents to select the layer ℓ and strength α, and test parents for final evaluation. These pools are disjoint. Validation selection uses parents with both a flip and a non-flip child; its effective size therefore depends on the setting (Appendix C.3).

A.4 Split construction and filtering

For M3CoT and Robo2VLM, we retain the domain and task definitions of the source datasets. For each model, we then exclude parents whose clean answer cannot be normalized to the task answer space. The usable evaluation counts can therefore differ slightly across models (Table 10).

For GMAI-MMBench, we filter the public validation release and form the splits with a fixed seed. We use largest-remainder allocation to preserve the joint distribution of clinical task and perceptual granularity. Task, granularity, and modality marginals differ by about two percentage points or less between the resulting train and test pools. Appendix A.5 provides the modality-specific filtering and image-processing details.

A.5 Medical data and preprocessing

We remove GMAI-MMBench questions with fewer than four options or answerable by a single text-only greedy query, leaving $2{,}397$ candidates. We approximate optical defocus in lens-based modalities (endoscopy, fundus, dermoscopy, and microscopy) with a disk point-spread function (Goodman, 2005; Mannan & Langer, 2016), and use Gaussian smoothing for X-ray and CT images. MRI is outside this isotropic-blur protocol. Our window-level variation is a display-variation proxy implemented through contrast and gamma changes in $[0.80, 1.25]$. Rotation uses reflect padding and center cropping to model small in-plane alignment changes.

We use model-specific preprocessing based on VLMEvalKit (Duan et al., 2024). In our configuration, images are downscaled, when needed, to a maximum long side of 1280 pixels for Qwen3-VL and 896 for Gemma-3. Perturbations are applied after resizing, with blur radius defined relative to the short side. Table 10 reports the model-specific subspace and evaluation pools.

B Evaluation Protocol and Answer-Flip Statistics

B.1 Evaluation pool sizes

Table 10 lists the calibration pairs and the actual evaluation denominators for each setting.

Table 10: **Subspace and evaluation pool sizes.** U pairs counts the pairs used to estimate the subspace, at most one per parent. Eval C/N/F lists the actual clean, non-flip, and flip denominators used for recovery metrics. Parent pools are given in Table 2.

	Qwen3-VL-8B		Gemma-3-12B	
Setting	U pairs	Eval C/N/F	U pairs	Eval C/N/F
M3CoT exposure	737	429/428/183	989	449/447/278
M3CoT white balance	723	429/428/190	977	449/448/272
M3CoT JPEG	759	429/428/187	1039	449/444/298
Robo2VLM exposure	334	560/558/107	316	567/566/149
Robo2VLM lens smudge	255	560/554/86	319	567/566/153
Robo2VLM motion blur	287	560/558/94	375	567/564/187
GMAI window level	183	192/190/112	185	200/199/102
GMAI defocus blur	146	192/189/85	120	200/191/77
GMAI rotation	185	192/192/110	192	200/197/126

B.2 Budget-matched cross-model evaluation

Cross-model comparisons use a common budget of $B = 15$ perturbation draws per parent. For open-weight models, recorded children are matched to reconstructed draws with a tolerance of two pixel-intensity levels, and only the first 15 draws are counted. API models use this budget directly. These rates use a separate parent pool from the test-split percentages in Table 11. Table 1 reports clean accuracy and the *answer-flip rate* under this protocol. All four models exhibit flips, including the two closed models. In this comparison, flip rate decreases as clean accuracy increases, but not with parameter count. On flipped inputs, the two open-weight models achieve only about 30% accuracy, showing that most flips are errors rather than harmless rephrasings. Flips also occur on many of the same inputs across models. A question that flips Qwen3-VL-8B is more likely to flip Claude Opus 4.8 (odds ratio 5.2, 95% CI $[3.1, 8.8]$), and 31 of GPT-5.5's 41 flips occur on parents that also flip Qwen3-VL-8B. These results indicate overlap in the inputs on which different models exhibit answer instability.

B.3 Answer-flip rates across all evaluation settings

Table 11 reports answer-flip percentages in the evaluation pools, using the clean and flip counts in Table 10. Cross-model comparisons use the separate matched-budget protocol above.

Table 11: **Answer flips in the evaluation pools.** Percentages are $100n_F/n_C$, using the flip and clean counts in Table 10; $n = n_C$.

Setting	Qwen3-VL-8B	Gemma-3-12B
M3CoT exposure	42.7% (n=429)	61.9% (n=449)
M3CoT white balance	44.3% (n=429)	60.6% (n=449)
M3CoT JPEG	43.6% (n=429)	66.4% (n=449)
Robo2VLM exposure	19.1% (n=560)	26.3% (n=567)
Robo2VLM lens smudge	15.4% (n=560)	27.0% (n=567)
Robo2VLM motion blur	16.8% (n=560)	33.0% (n=567)
GMAI window level	58.3% (n=192)	51.0% (n=200)
GMAI defocus blur	44.3% (n=192)	38.5% (n=200)
GMAI rotation	57.3% (n=192)	63.0% (n=200)

C Implementation and Reproducibility

C.1 FlipDir configuration

We select the steering layer and strength separately for each model and perturbation setting. Table 12 lists the selected values and corresponding main-table results. All settings share rank $k = 8$ and gate thresholds $(\tau_o, \tau_c) = (0.3, 0.7)$.

Table 12: **Selected configurations.** Layer ℓ, strength α, and main-table $H(\mathrm{C}, \mathrm{N}, \mathrm{F})$. All settings use $k = 8$, gate thresholds $(0.3, 0.7)$, and uncentered SVD with at most one flip pair per parent.

	Gemma-3-12B			Qwen3-VL-8B		
Setting	ℓ	α	$H(\mathrm{C}, \mathrm{N}, \mathrm{F})$	ℓ	α	$H(\mathrm{C}, \mathrm{N}, \mathrm{F})$
M3CoT exposure	12	1.0	61.6	4	1.0	78.3
M3CoT white balance	12	0.5	60.8	36	1.0	76.4
M3CoT JPEG	12	1.0	63.0	36	1.0	78.9
Robo2VLM exposure	12	0.5	73.4	36	1.0	82.4
Robo2VLM lens smudge	36	0.5	73.6	36	1.0	82.5
Robo2VLM motion blur	36	1.0	69.1	36	0.5	82.4
GMAI window level	12	0.5	70.0	12	0.5	70.1
GMAI defocus blur	36	1.0	70.9	36	1.0	63.7
GMAI rotation	36	0.5	64.2	36	0.5	70.7

C.2 Steering implementation

We extract the states used to estimate U at the final prompt position in unedited prefill, but apply the resulting directions only during decoding at the selected layer. The projection can be evaluated as $U(U^\top h)$ using two matrix-vector products, requiring $O(dk)$ arithmetic without materializing a $d \times d$ projection matrix. At each decoding step, we first obtain the unedited distribution and update the gate from its top-two probability margin. An open gate triggers a second forward with the projection applied at layer ℓ; a closed gate uses the first distribution without a steered forward.

C.3 Hyperparameter selection

The six (ℓ, α) candidates are scored on validation triplets containing a clean image, one flip child, and one non-flip child. Parents follow a fixed 40-shard ordering; the first R shards are used to cover approximately 50 parents with at least one flip, with R adjusted to the setting's flip density. Qwen3-VL defocus blur has 47 available validation flip pairs and approximately 24 selection triplets. Validation parents are disjoint from subspace-calibration and test parents. Varied images are stored as PNG at construction time and are regenerable exactly from the deterministic per-parent seed chain.

C.4 Baseline adaptations

We use the recommended LEAD and VTI configurations without task-specific retuning. The adaptations below address backbone compatibility. In contrast, FlipDir selects its method-specific intervention layer and strength on held-out validation data (Appendix C.3); test data are used only for final evaluation.

LEAD. We adapt the public LEAD implementation for Qwen2.5-VL (Bai et al., 2025b) (Xu et al., 2026) to each model's embedding and token conventions. For Gemma-3, direct embedding inputs retain the model's $\sqrt{d}$ scaling, termination recognizes its chat end token, and image-anchor lookup uses the model-specific image token. Table 4 reports this adapted implementation.

VTI. We use the official VTI implementation. Qwen3-VL produces a variable visual-token grid, so we average VTI's visual direction over tokens within each layer and apply it at all positions. The textual direction is unchanged. Setting $\alpha = \beta = 0$ reproduces the base model exactly.

C.5 Deterministic inference and released artifacts

All open-weight experiments use greedy decoding, batch size one, fixed seeds, deterministic kernels, and TF32 disabled. Repeating one complete setting reproduced all $1{,}040$ generations exactly. Exact reproduction uses the saved flip subspaces: GPU re-extraction can introduce numerical differences, and subspaces agreeing to 10^{-6} can still change H by several points at near-tied decoding steps.

Release contents. We will release model-specific benchmark instances, perturbation code and versioned configurations, inference and evaluation scripts, answer-normalization and flip-mining utilities, budget-matching tools, and extracted subspaces. Instance metadata includes the parent identifier, seed, perturbation parameters, prompt, raw output, and normalized answer.

D Activation and Decoding Analyses

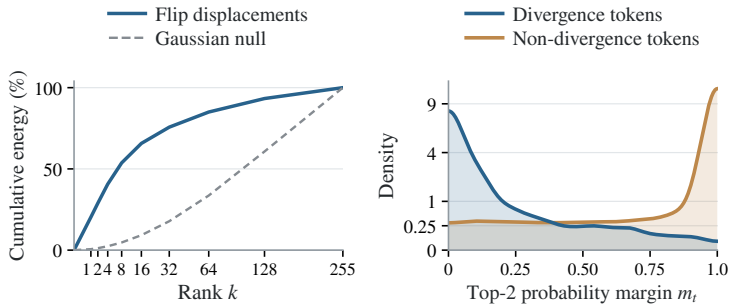
D.1 Activation energy concentration

For Qwen3-VL-8B on Robo2VLM lens smudge, we analyze 255 clean-flip pairs ($d = 4096$) at layers 4, 12, and 36. The Gaussian null uses a matrix of the same size and the same uncentered SVD and energy calculation. With uncentered singular values $\sigma_{\ell,j}$, cumulative energy is $E_\ell(k) = \sum_{j=1}^{k} \sigma_{\ell,j}^2 / \sum_j \sigma_{\ell,j}^2$. The top 8 directions capture 48.6 to 53.7% of the energy, versus 4.7% for the Gaussian null; the top 16 capture 59.4 to 65.7%, versus 9.2%. Figures 4 and 10(a) plot these cumulative energies.

The cumulative-energy plots use a square-root rank axis and a linear energy axis.

D.2 Token margins at decoding divergence

Figure 10(b) complements the Qwen3-VL-8B results in Figure 5 with Gemma-3-12B. We analyze stored unedited traces for the settings with available traces. At each step, original and perturbed inputs share the same decoding prefix. A divergence occurs when their next-token predictions differ; after each divergence, we advance both runs with the same token to realign the prefix before continuing. Margins are computed from the perturbed-input distribution. Within each backbone, divergence and non-divergence token distributions are normalized separately to unit area. We use Gaussian kernel density estimation with bandwidth 0.045 and reflection at 0 and 1. The square-root density axis makes overlap visible; tick labels report density in the original units.



(a) Layer 4 (b) Gemma-3-12B

Figure 10: **Additional activation and decoding analyses.** (a) Cumulative energy at layer 4 for the same 255 Qwen3-VL-8B/Robo2VLM lens-smudge pairs as Figure 4. (b) Divergence and non-divergence token margins for Gemma-3-12B, complementing the Qwen results in Figure 5.

D.3 Limits of static diagnostics

We examine whether simple diagnostics predict *answer flips* or recovery. Subspace alignment has little correlation with realized recovery ($\rho \approx 0$), and steering coordinates provide weak token-level discrimination (SNR 0.02 to 0.07; AUC 0.55). A parent's flip probability is also nearly flat in its clean-run margin (0.48 to 0.60). These results limit the use of alignment, steering coordinates, or clean-run confidence as standalone predictors. FlipDir does not use U as a flip detector or clean-run confidence to predict future flips. It uses U as an intervention subspace and gates steering with the current unedited token margin during decoding of the perturbed input (Section 3). The diagnostics therefore distinguish static prediction from decoding-time intervention; they do not by themselves establish the benefit of dynamic gating. That benefit is evaluated directly by the controlled gate ablation (Table 6). We select (ℓ, α) using held-out validation performance rather than these proxies (Appendix C.3).

E Additional Evaluation Results

E.1 Recovery–preservation trade-offs

We plot $P = (\mathrm{C} + \mathrm{N})/2$ against flip recovery F for all 18 settings in Tables 3 and 4. Each point represents the reported configuration, without additional tuning or interpolation. A point is Pareto-nondominated if no compared configuration has at least as high P and F and is strictly higher in one. FlipDir is nondominated in every setting under the reported rounded values, although other

methods can remain on the frontier. Thus these comparisons show trade-offs independently of H, not dominance at every preservation level.

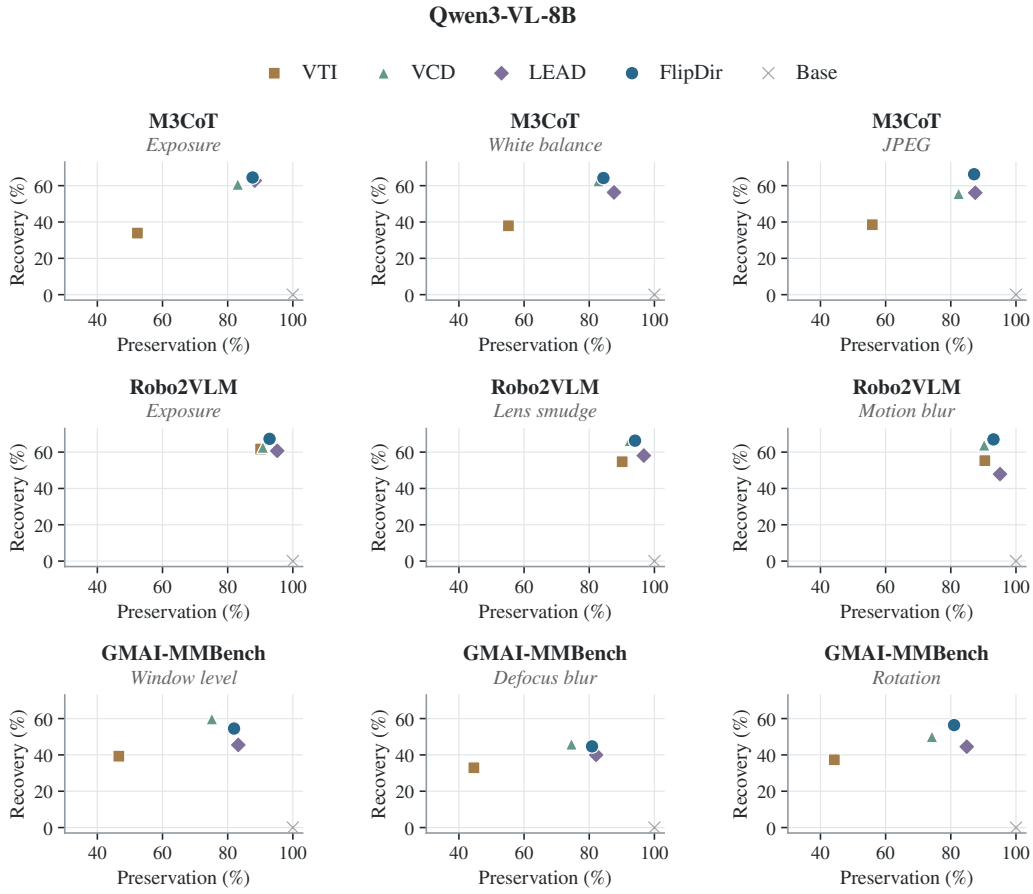


Figure 11: **Recovery–preservation trade-offs for Qwen3-VL-8B.** All nine settings and five methods are shown on shared axes. Higher and further right indicate better recovery and preservation, respectively.

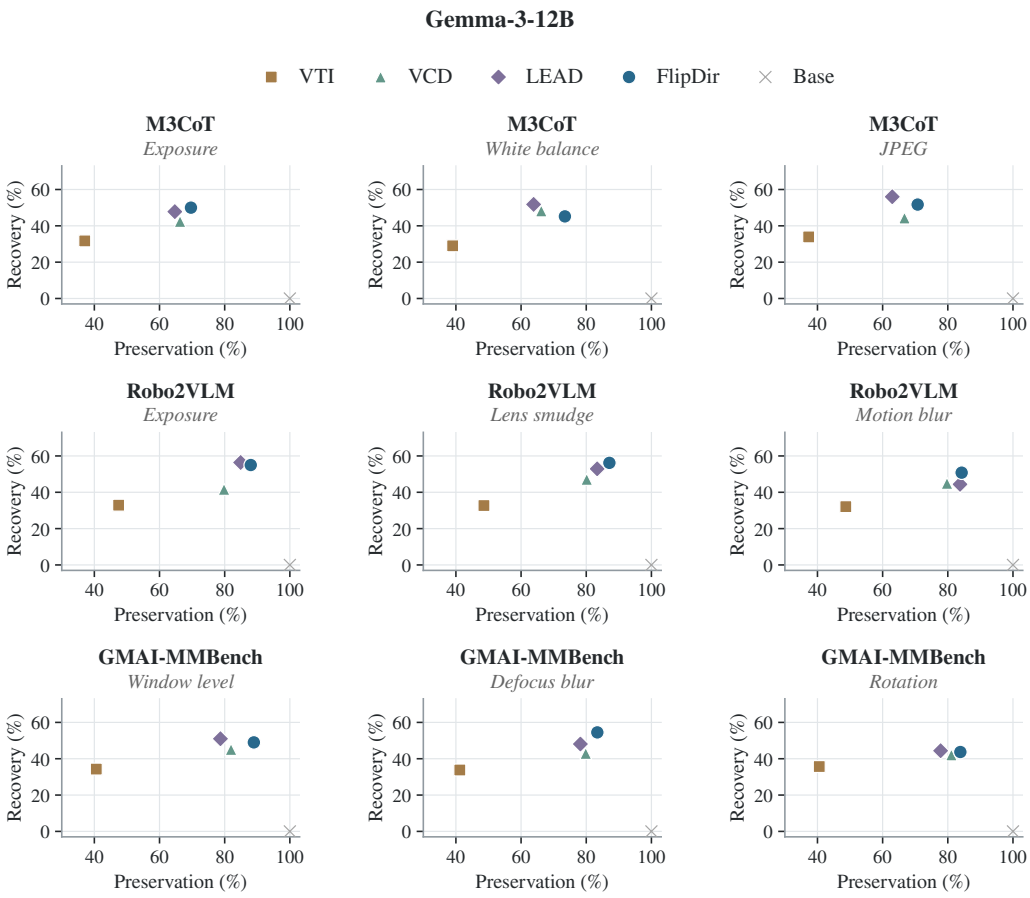


Figure 12: **Recovery–preservation trade-offs for Gemma-3-12B.** All nine settings, with the same methods and axes as Figure 11. Base is at $(100, 0)$ in every setting.

E.2 Transfer across datasets and perturbations

Table 13 expands Figure 7 with all measured source-to-target transfers, including within-domain sources. Only U is transferred; the layer and strength use the target-specific configuration. Across the cross-domain combinations shown in Figure 7, at least 91% of target-matched H is retained.

Table 13: **Transfer results.** $H(\mathrm{C}, \mathrm{N}, \mathrm{F})$ using the source subspace and the target-specific layer and strength. The last column uses a target-matched subspace.

Model	Target setting	Source U	$H(\mathrm{C}, \mathrm{N}, \mathrm{F})$	Target-specific
Qwen3-VL-8B	Robo2VLM lens	M3CoT exposure	75.7	82.5
Qwen3-VL-8B	Robo2VLM lens	GMAI rotation	76.3	82.5
Qwen3-VL-8B	M3CoT exposure	Robo2VLM lens	71.5	78.3
Qwen3-VL-8B	M3CoT exposure	GMAI rotation	74.2	78.3
Qwen3-VL-8B	GMAI rotation	M3CoT exposure	65.3	70.7
Qwen3-VL-8B	GMAI rotation	Robo2VLM lens	69.7	70.7
Gemma-3-12B	M3CoT exposure	Robo2VLM lens	61.5	61.6
Gemma-3-12B	M3CoT exposure	GMAI rotation	61.4	61.6
Gemma-3-12B	Robo2VLM lens	M3CoT exposure	69.4	73.6
Gemma-3-12B	Robo2VLM lens	GMAI rotation	71.6	73.6
Gemma-3-12B	GMAI rotation	M3CoT exposure	62.3	64.2
Gemma-3-12B	GMAI rotation	Robo2VLM lens	60.5	64.2
Gemma-3-12B	M3CoT exposure	M3CoT JPEG	61.6	61.6
Gemma-3-12B	M3CoT exposure	Robo2VLM exposure	61.1	61.6
Gemma-3-12B	Robo2VLM lens	Robo2VLM exposure	71.1	73.6
Gemma-3-12B	GMAI rotation	GMAI window	61.1	64.2
Qwen3-VL-8B	GMAI rotation	GMAI window	60.9	70.7
Gemma-3-12B	Robo2VLM lens	Robo2VLM motion	71.6	73.6
Gemma-3-12B	GMAI rotation	Robo2VLM motion	62.1	64.2
Qwen3-VL-8B	Robo2VLM lens	Robo2VLM motion	80.9	82.5
Qwen3-VL-8B	GMAI rotation	Robo2VLM motion	70.2	70.7

F Ablations and Sensitivity

This section supplements the rank, gate, and strength analyses in Section 5.4 with tabulated rank results, the full margin-gate comparison, and sensitivity to the number of calibration pairs.

F.1 Flip-subspace rank

Table 14 sweeps k on six settings, covering two non-medical perturbation families and one medical perturbation family per model, while holding all other components fixed at the selected configuration.

Table 14: **Rank sensitivity.** $H(\mathrm{C}, \mathrm{N}, \mathrm{F})$ with all other components fixed to Table 12. The shared rank $k = 8$ is shaded; row maxima are bold.

Model	Setting	$k = 2$	4	8	16	32
Qwen3-VL-8B	M3CoT exposure	71.0	74.3	**78.3**	73.2	72.6
	Robo2VLM lens smudge	82.2	79.6	**82.5**	79.1	77.8
	GMAI rotation	65.4	64.7	**70.7**	63.7	69.6
Gemma-3-12B	M3CoT exposure	59.2	58.6	**61.6**	59.2	60.2
	Robo2VLM lens smudge	**77.1**	70.4	73.6	71.4	70.2
	GMAI rotation	65.1	64.2	64.2	**65.4**	64.5

F.2 MARGIN-GATE ABLATION

Table 15 reports the preservation and recovery metrics underlying Table 6. Gate Off applies steering at every decoding step; the subspace, layer, and strength are held fixed within each comparison. The gate improves H in all six settings. Figure 13 visualizes the corresponding changes in each metric.

Table 15: **Margin-gate ablation.** C/N/F/H (%) with and without the gate. $\Delta H = H_{\mathrm{On}} - H_{\mathrm{Off}}$, computed from the displayed values.

	Gate Off				Gate On				
Setting	C	N	F	H	C	N	F	H	ΔH
Gemma-3-12B									
M3CoT exposure	59.5	63.8	45.0	54.8	70.2	69.1	50.0	61.6	+6.8
Robo2VLM lens smudge	81.7	83.4	51.6	68.8	87.8	86.4	56.2	73.6	+4.8
GMAI rotation	84.0	78.7	37.3	58.3	84.5	83.2	43.7	64.2	+5.9
Qwen3-VL-8B									
M3CoT exposure	84.1	85.3	62.3	75.6	87.4	87.9	64.5	78.3	+2.7
Robo2VLM lens smudge	92.3	91.9	52.3	73.5	94.1	94.0	66.3	82.5	+9.0
GMAI rotation	79.2	81.8	53.6	69.2	80.2	81.8	56.4	70.7	+1.5

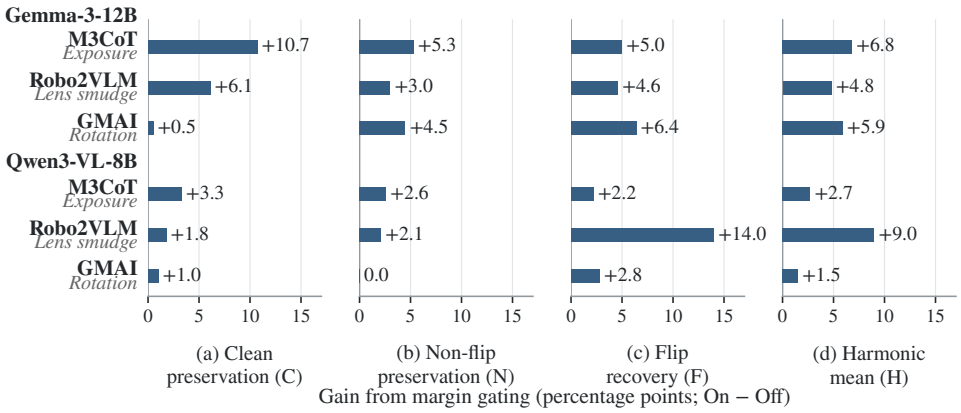


Figure 13: **Performance gains from margin gating.** Changes in clean preservation (C), non-flip preservation (N), flip recovery (F), and their harmonic mean (H), measured in percentage points as Gate On minus Gate Off. All panels share the same scale; values are computed from Table 15. The subspace, layer, and strength are fixed within each comparison. GMAI denotes GMAI-MMBench.

F.3 CALIBRATION SAMPLE SIZE

Table 16 varies the number of pairs used to estimate U, with other components fixed. Each subset contains the first N parents in sorted identifier order, with at most one flip pair per parent; the basis is estimated by uncentered SVD.

Table 16: **Subspace-estimation sample size.** $H(\mathrm{C}, \mathrm{N}, \mathrm{F})$ with other components fixed; parent counts are shown in parentheses.

Setting	50 pairs	subsampled	full pool
Gemma M3CoT exposure	60.2	56.8 (250)	61.6 (989)
Gemma Robo2VLM lens	67.0	72.2 (150)	73.6 (319)
Gemma GMAI rotation	67.0	63.8 (100)	64.2 (192)
Qwen Robo2VLM lens	78.6	79.2 (150)	82.5 (255)

G COMPUTE, RUNTIME, AND GATE STATISTICS

G.1 COMPUTE ACCOUNTING

Offline direction construction. Direction estimation is performed once per backbone and perturbation setting, and the resulting subspace is reused across queries. Given existing clean and flip

pairs, activation extraction requires two prefill-only forwards per pair; the candidate layers are read from the same forwards. The $4{,}512$ Gemma pairs and $3{,}609$ Qwen pairs in Table 10 therefore require $9{,}024$ and $7{,}218$ prefills, respectively. SVD is then applied to the activation differences, and the retained basis stores dk coefficients. These extraction counts exclude pair search and validation-based parameter selection. Pair-search costs are reported separately in Table 8; validation selects the steering layer and strength before deployment. Thus the method incurs an initial construction cost in exchange for a reusable intervention, without updating model weights.

Decoding FLOPs. We count a multiply-add as two FLOPs and approximate the dominant linear operations in the decoder blocks and output head. Comparisons fix the input and output lengths and exclude prefill, image encoding, and the context-length-dependent attention products. Let F_b be the per-token cost of one decoder block, F_h the output-head cost, and L the number of blocks. Standard decoding costs $F_0 = LF_b + F_h$. Under conditional execution, an open gate triggers recomputation only above the intervention layer ℓ, including the output head. The relative cost is

$$\frac{F_{\mathrm{FlipDir}}}{F_0} \approx 1 + p_{\mathrm{open}} \frac{(L-\ell)F_b + F_h + 4dk}{LF_b + F_h}. \tag{6}$$

The per-token cost consists of the unedited forward (F_0), negligible gate arithmetic, and, on open-gate steps, projection ($4dk$) plus the remaining blocks and output head ($(L-\ell)F_b + F_h$). A closed gate incurs neither projection nor recomputation. The output head is recomputed even when $\ell = L$, so intervention at the final layer does not have zero cost.

For Gemma-3-12B, $L = 48$, $d = 3840$, $F_b \approx 4.48 \times 10^8$, and $F_h \approx 2.01 \times 10^9$ FLOPs. For Qwen3-VL-8B, $L = 36$, $d = 4096$, $F_b \approx 3.86 \times 10^8$, and $F_h \approx 1.24 \times 10^9$. We combine the layers in Table 12 with the gate-open fractions replayed on stored unedited traces in Table 17. Across all nine settings for each backbone, estimated multipliers range from 1.068 to 1.185 for Gemma and from 1.011 to 1.176 for Qwen. These gate-open fractions are obtained by replaying stored traces; they do not measure a deployed conditional implementation. The unweighted mean over all 18 settings is approximately $1.076\times$, rounded to $1.08\times$ in Table 7.

LEAD uses one decoder forward and a dense probability-weighted embedding, $\sum_{v=1}^{V} p_t(v) e_v$, at each step. This adds approximately $2Vd$ FLOPs, giving $F_{\mathrm{LEAD}}/F_0 \approx 1 + 2Vd/F_0$: $1.086\times$ for Gemma and $1.082\times$ for Qwen. Their mean is $1.084\times$, rounded to $1.08\times$ in Table 7. VTI adds direction updates and normalization within a single forward pass (Liu et al., 2025), giving approximately $1\times$ at the scale of decoder FLOPs. VCD evaluates original- and distorted-image branches at every decoding step (Leng et al., 2024), giving approximately $2\times$; its implementation maintains separate generation states for the two branches. Under the same accounting, conditional FLIPDIR uses about 41 to 50% fewer decoding FLOPs than VCD. This comparison concerns decoding arithmetic rather than end-to-end latency. The second, steered forward is executed only when the gate opens. Table 7 estimates the cost with reuse of activations below the intervention layer; a full second decoder forward instead costs approximately $1 + p_{\mathrm{open}}$ times standard decoding, plus the projection. Section G.3 reports its measured wall-clock runtime separately.

G.2 GATE ACTIVATION STATISTICS

Gate open rates replay the selected thresholds $(0.3, 0.7)$ with hysteresis over stored unedited traces: twelve parents per setting across shards, including clean, flip, and non-flip generations.

Table 17: **Runtime and gate statistics.** Time is measured in seconds per parent under the shared serving regime. Gate statistics are replayed on stored unedited traces; entries are counted per 100 tokens.

	Gate		Seconds per parent			
Setting	Open %	Entries	**FLIPDIR**	VCD	LEAD	VTI
Qwen M3CoT exposure	15.5	9.1	254.1	260.5	92.4	252.2
Qwen M3CoT wb	14.5	8.5	231.3	265.4	135.1	227.2
Qwen M3CoT JPEG	12.9	8.0	232.7	258.7	134.6	241.4
Qwen Robo2VLM exp.	17.3	9.6	121.8	109.0	71.9	44.7
Qwen Robo2VLM lens	18.1	9.9	71.8	96.0	49.2	43.7
Qwen Robo2VLM motion	18.8	10.2	112.7	120.4	70.8	42.1
Qwen GMAI window	25.4	13.0	85.7	209.8	113.8	88.6
Qwen GMAI defocus	24.1	12.2	74.5	209.7	114.2	80.5
Qwen GMAI rotation	24.0	12.7	76.8	211.2	115.7	82.8
Gemma M3CoT exposure	13.3	6.5	175.3	168.7	90.5	133.8
Gemma M3CoT wb	11.0	5.9	171.1	172.5	92.5	121.3
Gemma M3CoT JPEG	12.2	6.2	176.3	169.2	96.6	124.3
Gemma Robo2VLM exp.	22.3	11.6	48.0	44.1	23.8	40.8
Gemma Robo2VLM lens	23.4	11.2	36.4	35.5	17.7	38.3
Gemma Robo2VLM motion	24.4	10.8	22.3	44.0	23.6	53.2
Gemma GMAI window	24.0	10.9	53.6	97.5	52.3	89.1
Gemma GMAI defocus	21.7	11.8	49.2	92.6	49.7	74.5
Gemma GMAI rotation	25.8	12.9	54.4	101.5	54.4	91.0

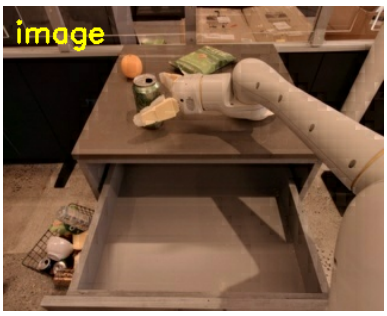
G.3 RUNTIME

Wall-clock time in Table 17 is averaged over the main-table runs under a shared serving regime with four to five concurrent processes per GPU. FLIPDIR first computes the unedited distribution and executes the steered pass only when the gate is open. Across the six medical settings, the ratio of VCD runtime to FLIPDIR runtime ranges from 1.8 to 2.8, with higher H for FLIPDIR in each setting. These measurements describe the shared serving regime; they do not isolate per-query latency or the conditional-execution FLOPs in Eq. 6. VCD adds a second image-conditioned decoding branch, whereas VTI performs direction updates within one branch. Without component-level profiling, we do not attribute their measured latency differences to prefill or implementation overhead alone.

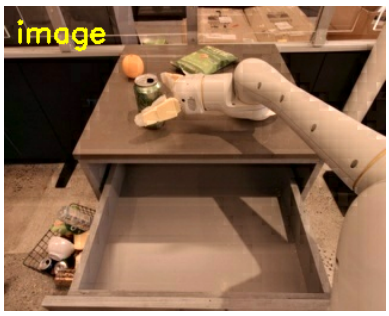
H QUALITATIVE EXAMPLES

H.1 ANSWER-FLIP RECOVERY

Figure 14 illustrates Gemma-3-12B on Robo2VLM exposure shift with $\ell = 12$ and $\alpha = 0.5$.



Clean image (model answers B)



Exposure shift (−EV)

[Question] *The robot is tasked to moving away with open gripper after releasing the can. After moving away with open gripper after releasing the can, what will be the robot's NEXT action phase?*

(A) Moving away with open gripper after releasing the can **(B) Approaching the can with open gripper**
(C) Firmly grasping the can (D) Releasing the can by opening gripper (E) Closing gripper to grasp the can

Varied image, no steering → (A)

The question states that the robot has already "Moving away with open gripper after releasing the can". The question asks for the *next* action. Since the can has already been released, the next logical action would be to move away with the open gripper.

FinalAnswer: (A)

Varied image, FLIPDIR → (B)

The question states that the robot has already "Moving away with open gripper after releasing the can". This means the can has been released and the robot is moving away. The next logical action would be to approach another object or location.

Therefore, the next action phase is approaching the can with an open gripper.

FinalAnswer: (B)

Figure 14: A repaired flip on Gemma-3-12B Robo2VLM exposure shift. The unedited pass on the varied image restates the completed phase and flips to (A); with FLIPDIR the reasoning advances to the cycle's next phase and the answer returns to (B), the clean answer and the ground truth. Colored spans mark where the trajectories diverge.

H.2 PERTURBATION INTENSITY EXAMPLES

Figures 15, 16, and 17 provide representative examples from the fixed perturbation bands. Additional medical-specific design details are provided in Appendix A.5.

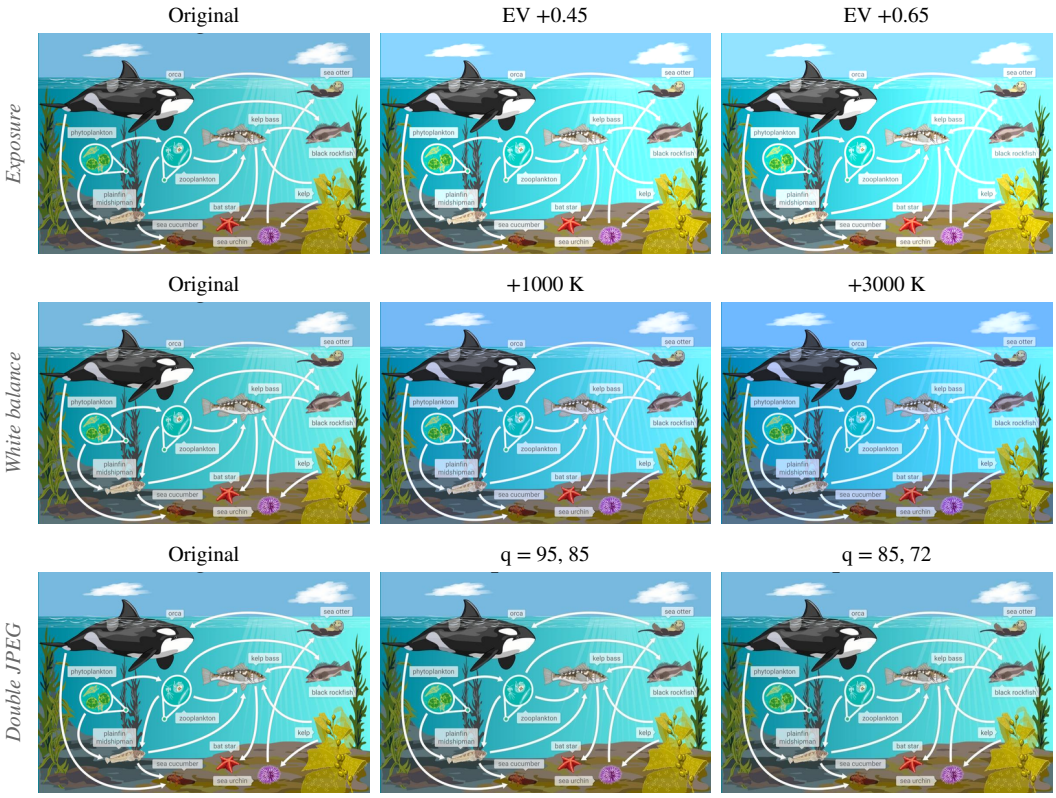
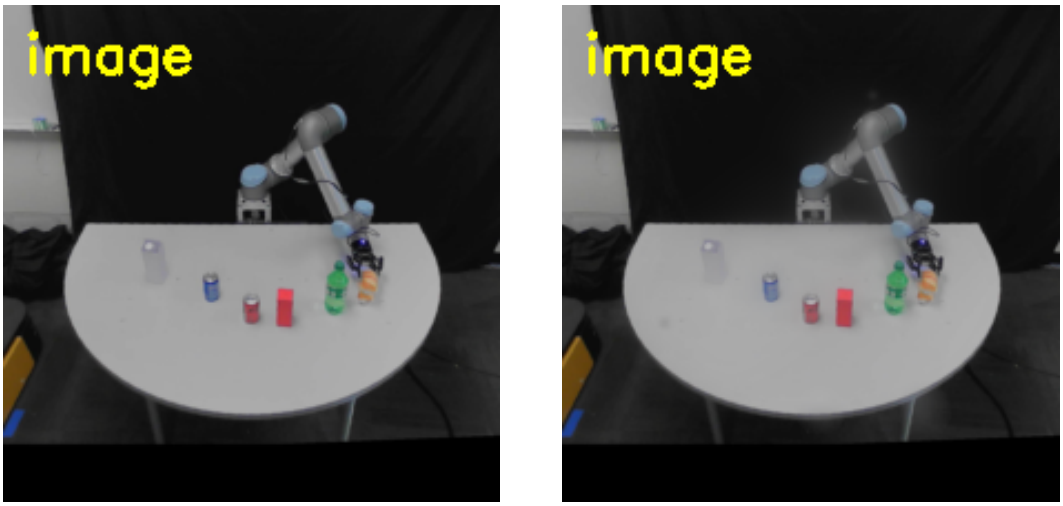


Figure 15: M3CoT variation intensities on the example of Figure 2, rendered from the fixed bands. Rows: exposure shift at the EV band endpoints ($+0.45$, $+0.65$), white-balance shift at $+1000$ K and $+3000$ K, and the mildest ($q = 95, 85$) and strongest ($q = 85, 72$) double-JPEG draws.



Clean image Lens-smudge variation

Figure 16: Robo2VLM lens-smudge example from VISFLIP. The perturbation composites wipe smears, droplets, dust, local blur, and mild haze while preserving the underlying scene.

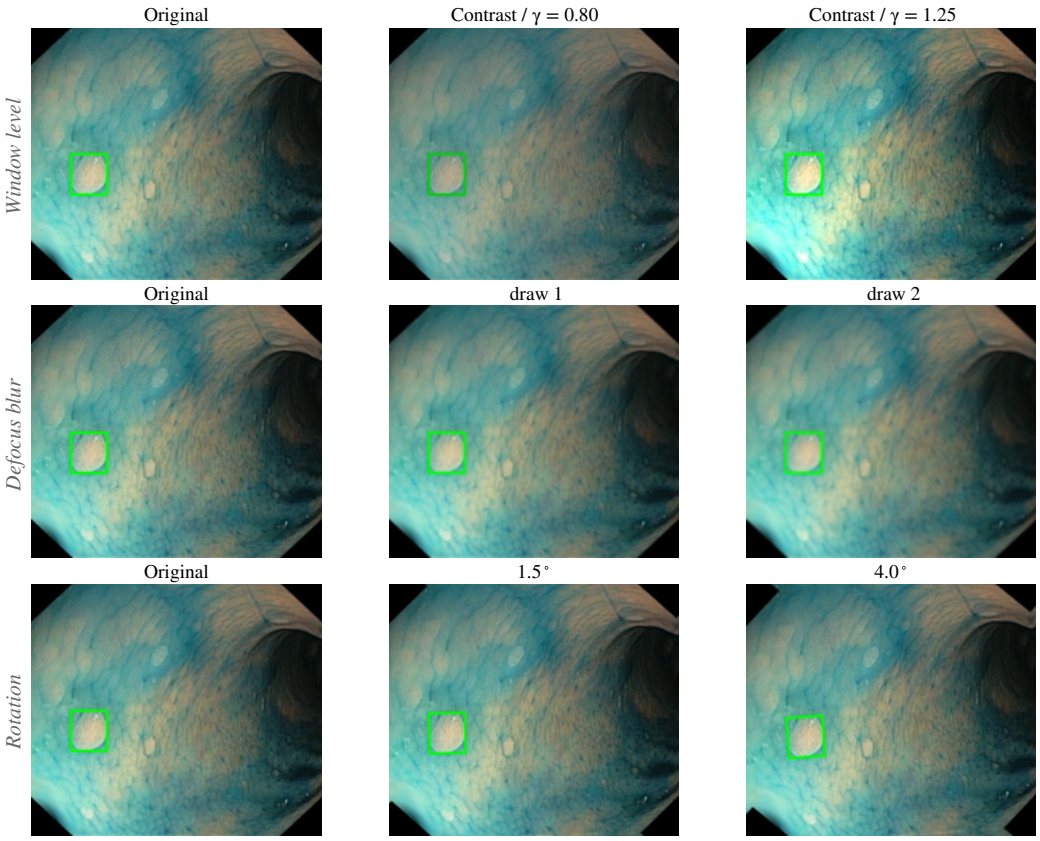


Figure 17: GMAI-MMBench variation intensities on an endoscopy image. Rows: window-level shift at the joint contrast and gamma endpoints, modality-gated defocus blur (disk PSF for this lens modality), and detector rotation at $1.5°$ and $4.0°$.