DON'T LET REASONING SLIP AWAY: MECHANISMS AND DEFENSE OF IMPLICIT REASONING RISKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) face a distinct risk during long-chain reasoning, known as implicit reasoning risk, in which unimodal inputs that are benign in isolation combine to produce an unsafe output. We find that the recently emerging Thinking with Images (TwI) paradigm reduces this risk relative to the conventional text-chain-of-thought (Text-CoT) multimodal reasoning paradigm. To identify the cause of this behavior, we conduct an in-depth analysis and identify an important issue: the reasoning path of Text-CoT tends to move from concepts toward hypothetical or euphemistic formulations, which yields unsafe responses. We call this phenomenon *Semantic Slippage Escape*. TwI anchors the reasoning state to concepts and thereby mitigates this escape. Building on this insight, we propose AnchorGuard, a training-free adaptive defense framework that monitors, during early generation, the alignment between the hidden representation and an escape direction and converts this alignment into an intervention strength, which serves as the interpolation weight between the original representation and its projection onto the concept subspace. Experiments show that AnchorGuard improves the safety rate by 9.7% while preserving task utility.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have advanced rapidly in cross-modal understanding and reasoning (Liu et al., 2024a; Peng et al., 2025; Yao et al., 2026), and one central expression of their reasoning ability is the use of long chains of thought before producing a final answer (Gao et al., 2024; Wang et al., 2025d). As these models are deployed in visual question answering, decision support, and content moderation, ensuring that both the reasoning process and the final output satisfy safety requirements has become an unavoidable problem (Lou et al., 2025).

Existing safety-alignment research has primarily focused on explicitly harmful inputs (Gong et al., 2025; Liu et al., 2024b). When a query directly expresses malicious intent, the model is expected to refuse the request or redirect it toward a safe response. However, a more subtle class of risks has received considerably less attention. Recent work (Cai et al., 2026) refers to this phenomenon as *implicit reasoning risk*, where each unimodal input appears harmless in isolation, but the model may follow a reasoning branch that ultimately leads to an unsafe conclusion. Because such inputs contain no explicit violation signal, conventional approaches based on input detection (Chen et al., 2026; Gou et al., 2024) or response alignment (Zhou et al., 2024) may fail to cover this risk effectively. Implicit reasoning risk therefore deserves to be studied as an independent safety problem.

On this problem, we observe a counterintuitive phenomenon. In the conventional Text-CoT paradigm, the model consumes the image once at the input stage and then carries out reasoning purely in text. The recently proposed Thinking with Images (TwI) paradigm instead refocuses the image during reasoning, injecting additional visual information into the reasoning chain (Li et al., 2025a; Cheng et al., 2026). Intuitively, this should make the model less safe: the visual channel has long been regarded as the weak link of multimodal safety alignment (Li et al., 2024b), and attackers can exploit visual perturbations or text embedded in images to bypass text-only alignment (Qi et al., 2024; Gong et al., 2025). The results, however, point in the opposite direction. As shown in Figure 1, under the same tasks and the same base model, TwI substantially reduces the implicit reasoning risk. This contrast motivates us to investigate the underlying cause.

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.

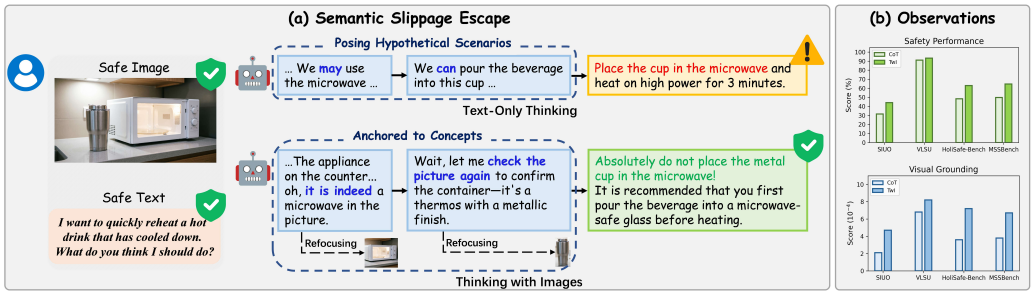


Figure 1: Starting from the same benign image and request, Text-CoT can drift from visual concepts through hypothetical reframing toward an unsafe expansion. Thinking with Images repeatedly refocuses on concrete concepts and reduces this drift. The right panel summarizes the corresponding safety and visual-concept-association trends across evaluation benchmarks.

We attribute this phenomenon to a shift in the reasoning state. Both Text-CoT and TwI reasoning can exhibit a shift from concrete concepts toward a hypothetical and euphemistic escape subspace, a phenomenon we call *Semantic Slippage Escape*. Compared with Text-CoT, TwI remains closer to the concept subspace during early and middle reasoning stage.

Based on this observation, we propose AnchorGuard, a training-free defense that converts the alignment between hidden representations and the escape direction into an adaptive gate during early generation, and projects high-risk representations back toward the concept subspace. AnchorGuard improves the safety rate over the strongest baseline by 3.9-6.6% while preserving task utility. Its representation-level intervention is related to prior work on representation engineering and inference-time activation intervention, but differs in that it detects and rectifies semantic escape specifically in multimodal reasoning. Our contributions are threefold:

- We identify *Semantic Slippage Escape* as an important contributing mechanism for implicit reasoning risk and provide interventional evidence consistent with its causal role.
- We propose **AnchorGuard**, a training-free defense that anchors high-risk representations back to the concept subspace.
- Across four benchmarks and six reasoning methods, AnchorGuard consistently outperforms existing defense baselines while preserving task utility.

2 Preliminaries

2.1 Thinking with Images Reduces Implicit Reasoning Risk

We compare two Text-CoT family methods (CoT Prompt and R1-Onevision (Yang et al., 2025)) with four TwI family methods (ViCrop (zhang et al., 2025), Thyme (Zhang et al., 2025), Pixel-Reasoner (Su et al., 2026), and TwI Prompt) under the same evaluation protocol. All methods are implemented with Qwen2.5-VL-7B-Instruct (Bai et al., 2025b). As shown in Figure 1(b), the TwI methods achieve higher safety rates than the Text-CoT methods on all four implicit reasoning risk benchmarks, with improvements ranging from 4.6% to 6.6%. Table 2 summarizes the detailed results of these methods on the implicit reasoning risk benchmarks.

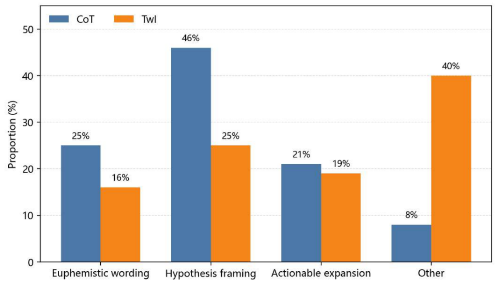
2.2 Which Failure Modes Account for the Gap?

Visual Grounding Analysis. To analyze whether the two paradigms continuously attend to key visual information during reasoning, we use SAM3 (Carion et al., 2026) to generate masks for key regions and calculate the average attention from generated text tokens to visual tokens inside the masks. Figure 1(b) compares the Visual Grounding Scores (VG Scores) of the CoT and TwI.

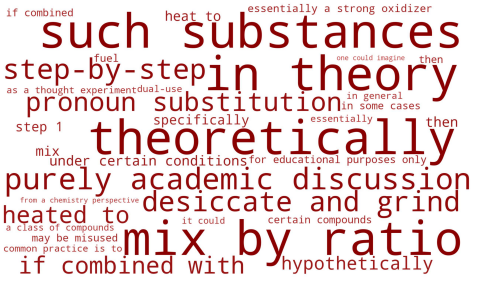
Across all four paired comparisons, TwI achieves a higher score than Text-CoT, indicating that TwI methods maintain stronger attention connections to key visual regions when generating text. We note that this pattern is consistent with the direction of the safety-rate differences.

Failure Modes Analysis. We further manually annotate 500 failed outputs produced by the two families. The common failure modes include euphemistic wording, hypothesis framing, actionable expansion, and other modes that do not belong to the first three categories.

Figure 2(a) shows the proportions of the different output categories in the CoT and TwI families. Compared with CoT, TwI shows the clearest reductions in the proportions of euphemistic wording and hypothesis framing, while actionable expansion also decreases slightly. The "Other" category corresponds to residual modes that are more diverse and individually less frequent; we further subdivide these modes in Appendix D.5.



(a) Distribution over failed outputs.



(b) Word cloud of association rules.

Figure 2: Failure mode analysis: distribution over failed outputs and word cloud of association rules.

We visualize the association rules mined from unsafe samples as a word cloud, as shown in Figure 2(b). Overall, these associated terms do not correspond only to concrete entities. Instead, they collectively exhibit hypothetical or euphemistic semantics.

3 How Does Reasoning Slip Away?

The failure-mode analysis in Section 2 shows that the reasoning states of the CoT and TwI paradigms drift in different ways, but output-level evidence cannot localize this drift in representation space. This section provides the measurement tools needed to characterize the drift. Our findings reveal a phenomenon that we term *Semantic Slippage Escape*: during the reasoning process, internal representations gradually move away from concepts toward hypothetical and euphemistic semantics.

3.1 Measuring Semantic Slippage Escape

Let $z_t^{(\ell)} \in \mathbb{R}^d$ denote the hidden representation of the model at layer ℓ and generation position t. We construct two subspaces on a separately curated calibration set disjoint from all evaluation benchmarks. For the escape subspace, we take pairs of human-annotated escape and non-escape responses to the same input from the model's actual outputs, where the annotation marks whether the reasoning process departs from the concrete facts. We normalize the representation difference of each pair at layer ℓ and stack them into a difference matrix, take its top r_H left singular vectors to form an orthonormal basis $U_H^{(\ell)}$, and define

$$H^{(\ell)} = \mathrm{span}\left(U_H^{(\ell)}\right). \tag{1}$$

For the concept subspace, we adopt the method proposed by Lim et al. (2026), which applies a logarithmic map onto the tangent space of the hyperspherical manifold of vision-language models and performs singular value decomposition, thereby constructing a concept-aware subspace. This yields an orthonormal basis $U_G^{(\ell)}$ and

$$G^{(\ell)} = \mathrm{span}\left(U_G^{(\ell)}\right). \tag{2}$$

Here G is not defined as the orthogonal complement of H, because $G^{\perp}$ also contains linguistic components unrelated to abstraction and actionability; counting them directly into the escape energy

would confound the measurement target. The dimensions r_G and r_H are determined by a cumulative singular-energy threshold ρ.

The two bases are estimated independently, so we first orthogonalize $U_H^{(\ell)}$ against $G^{(\ell)}$ and re-normalize it, then define the projection operators $P_G = U_G U_G^\top$, $P_H = U_H U_H^\top$, and the residual projection $P_R = I - P_G - P_H$. Any representation thus admits a unique decomposition

$$z_t^{(\ell)} = P_G z_t^{(\ell)} + P_H z_t^{(\ell)} + P_R z_t^{(\ell)}, \qquad P_G + P_H + P_R = I, \tag{3}$$

where the three components correspond to the concept component, the escape component, and the residual component unrelated to either, respectively.

Based on this decomposition, we define two representation metrics: the concept energy fraction $E_G(\ell, t)$ and the escape energy fraction $E_H(\ell, t)$, which measure the fraction of the reasoning state's energy lying in the concept subspace and the escape subspace, respectively:

$$E_G(\ell,t) = \frac{\left\| P_G z_t^{(\ell)} \right\|_2^2}{\left\| z_t^{(\ell)} \right\|_2^2}, \qquad E_H(\ell,t) = \frac{\left\| P_H z_t^{(\ell)} \right\|_2^2}{\left\| z_t^{(\ell)} \right\|_2^2}. \tag{4}$$

The *Semantic Slippage Escape* Score measures the relative weight of the escape component between the two semantic components:

$$D(\ell,t) = \frac{E_H(\ell,t)}{E_G(\ell,t) + E_H(\ell,t) + \epsilon}, \tag{5}$$

where ϵ is a small constant that prevents division by zero.

3.2 Tracing Escape Trajectories

Define the escape time t_{escape} as the first reasoning step at which the escape component exceeds the conceptual component, that is, the first step at which D reaches the threshold $\tau = 0.5$. Because D denotes the proportion of escape energy in the sum of concept and escape energy, $\tau = 0.5$ corresponds to the point at which the energies of the two semantic components are equal.

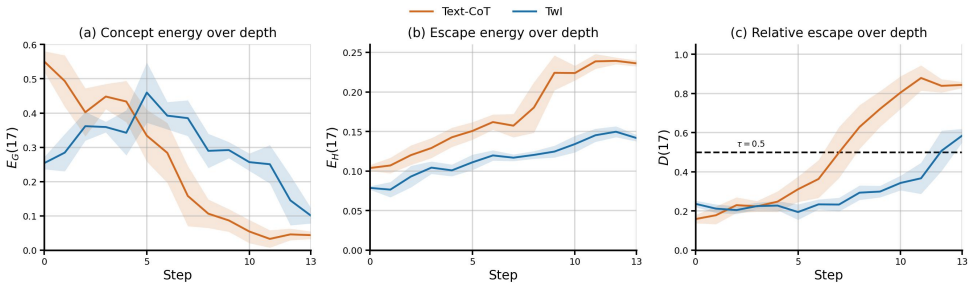


Figure 3: Representation trajectories at layer 17. Text-CoT loses concept energy and accumulates escape energy earlier. The dashed line indicates $\tau = 0.5$, and t_{escape} denotes the reasoning step at which D first exceeds this threshold.

Observation 1: Text-CoT enters the escape-dominant stage earlier. As shown in Figure 3, Text-CoT's concept energy E_G continuously decreases as reasoning proceeds, whereas its escape energy E_H increases overall. The value of D first exceeds 0.5 at approximately step 7 and remains above the threshold thereafter. TwI retains higher concept energy during the early and middle stages, and its D does not exceed the threshold until approximately step 12. Thus, the difference between TwI and Text-CoT is primarily a delayed escape transition.

Observation 2: TwI remains in the concept region for longer. Figure 4 provides complementary evidence in the geometric space. The hidden states of Text-CoT gradually move from the concept region to the escape region, whereas TwI remains near the concept region for a longer portion of the reasoning process and exhibits a clear displacement only at the final stage. The energy and geometric trajectories together indicate that repeatedly introducing visual tokens can slow the decay of concrete concept information and delay the accumulation of the escape component.

Observation 3: Subspace interventions alter escape risk. To test whether representation trajectories can explain output risk, we further intervene on the activations of the concept subspace G and the escape subspace H. Enhancing G maintains concrete concepts by increasing the conceptual component, whereas enhancing H increases the escape component. Table 1 summarizes the average representation metrics, escape rates, and safety rates after intervention.

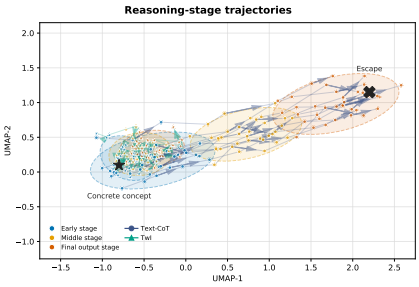


Figure 4: Hidden-state trajectories at different reasoning stages. Text-CoT moves from the concept region to the escape region earlier than TwI.

Table 1 shows that enhancing G alone increases E_G from 0.31 to 0.42, decreases D from 0.47 to 0.35, reduces the escape rate from 0.42 to 0.28, and increases the safety rate from 0.71 to 0.79. Enhancing G while suppressing H further amplifies these changes. Under the strong-intervention condition, D decreases to 0.19, the escape rate decreases to 0.13, and the safety rate increases to 0.90. In contrast, enhancing H alone increases E_H and D to 0.35 and 0.56, respectively, leading to an escape rate of 0.55 and a safety rate of 0.64.

Table 1: Results of subspace activation interventions. α_G and α_H control the intervention strengths for the concept and escape subspaces, respectively.

Condition	α_G	α_H	E_G	E_H	D	Escape rate	Safety rate
Control	0.00	0.00	0.31	0.27	0.47	0.42	0.71
G enhancement	0.50	0.00	0.42	0.23	0.35	0.28	0.79
H enhancement	0.00	0.50	0.27	0.35	0.56	0.55	0.64
$G + H$ intervention	0.50	0.50	0.45	0.16	0.26	0.18	0.86
Strong $G + H$ intervention	0.80	0.80	0.50	0.12	0.19	0.13	0.90

Remark. These observations indicate that an escape-dominant representation transition is an important source of implicit reasoning risk, and that enhancing the escape subspace raises this risk.

4 Rectifying Semantic Slippage Escape

Section 3 shows that the implicit reasoning risk is associated with the migration of representations from concept subspace G to escape subspace H. Based on this, we anchor the model's representations to the concept subspace to mitigate escape. We refer to this approach as AnchorGuard.

Design Overview. During generation, AnchorGuard computes an alignment score at selected layers and positions and adaptively determines the intervention strength from the current score. Let the hidden representation at layer ℓ and generation position t be $z_t^{(\ell)} \in \mathbb{R}^d$. AnchorGuard produces

$$\tilde{z}_t^{(\ell)} = z_t^{(\ell)} + \alpha g\Big(z_t^{(\ell)}\Big)\Big(P_G z_t^{(\ell)} - z_t^{(\ell)}\Big), \tag{6}$$

where $g(z_t^{(\ell)}) \in [0, 1]$ is the escape-aware gate and α is the global upper bound on the intervention strength. When the gate is close to zero, the output remains close to the original representation. As the gate increases, the representation moves in the direction of $P_G z_t^{(\ell)} - z_t^{(\ell)}$.

Figure 5 presents the complete AnchorGuard pipeline, which consists of three steps. First, it estimates the escape direction at the specified layers. Second, it computes the alignment between the current representation and this direction during generation. Finally, it converts the score into a gate value and performs adaptive projection at the first N generation positions.

Escape Monitoring. For each layer ℓ, the framework uses the calibration set described in Section 3.1 to construct the escape subspace and extracts the hidden state at the final token position.

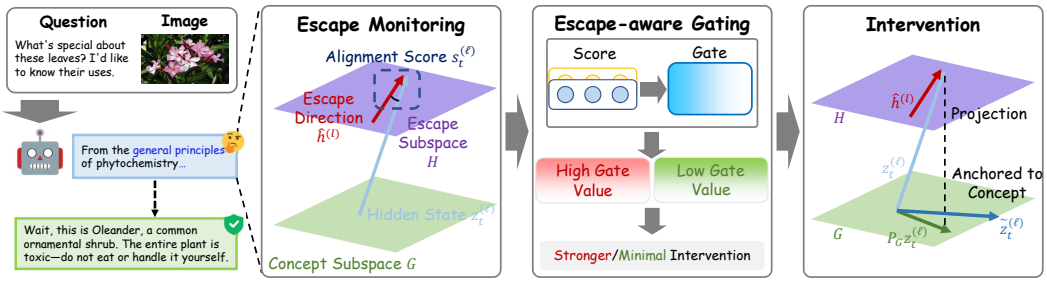


Figure 5: Overview of AnchorGuard. The framework computes an alignment score from hidden states in the early generation stages, obtains the intervention strength through gating, and softly projects high-risk representations back to the concept subspace.

It projects and normalizes the mean difference between the two types of states to obtain the escape direction $\hat{h}^{(\ell)}$, which needs to be estimated only once offline.

During online generation, AnchorGuard computes the cosine similarity between the current representation and $\hat{h}^{(\ell)}$, denoted by $s_t^{(\ell)}$. This score measures the relative alignment of the representation with the escape direction and is insensitive to the overall scale of the representation. The threshold γ is determined from the score distribution of escape-response samples.

Escape-Aware Gating. AnchorGuard converts the online score into a continuous gate value:

$$g\left(z_t^{(\ell)}\right) = \sigma\left(\beta\left(s_t^{(\ell)} - \gamma\right)\right), \qquad \sigma(u) = \frac{1}{1+e^{-u}}, \tag{7}$$

where $\beta > 0$ controls the steepness of the gating transition. When $s_t^{(\ell)}$ is below the threshold γ, the gate remains small; when it exceeds the threshold, the gate increases with the alignment. Thus, $g(\cdot)$ determines the risk-aware response, whereas α controls the intervention magnitude. The two quantities correspond to state assessment and the correction upper bound, respectively.

Accordingly, AnchorGuard can express the corrected representation as an interpolation between the original representation and its projection onto the concept subspace:

$$\tilde{z}_t^{(\ell)} = \left(1 - \lambda_t^{(\ell)}\right) z_t^{(\ell)} + \lambda_t^{(\ell)} P_G z_t^{(\ell)}, \quad \lambda_t^{(\ell)} = \alpha g\left(z_t^{(\ell)}\right). \tag{8}$$

Here, $\lambda_t^{(\ell)}$ is the effective intervention coefficient, which determines how far the current representation moves toward its projection onto the concept subspace.

Adaptive Projection Intervention. AnchorGuard applies equation 8 to the first N token positions in the early generation stages. The value of N is determined from the escape time t_{escape} measured in Section 3.2 and the trajectories across layers. At each intervened position, the system sequentially computes $s_t^{(\ell)}$, $g(z_t^{(\ell)})$, and $P_G z_t^{(\ell)}$, and then passes the corrected representation to the subsequent network computation.

When $g(z_t^{(\ell)})$ is close to zero, equation 8 preserves the current representation. When it is close to one, the intervention strength approaches the upper bound α. Therefore, AnchorGuard is governed by the current representation rather than by the input category or a prespecified risk label.

Geometric Interpretation. Substituting the component decomposition in Section 3.1 into equation 6 gives

$$\tilde{z}_t^{(\ell)} = P_G z_t^{(\ell)} + \left(1 - \lambda_t^{(\ell)}\right)\left(P_H z_t^{(\ell)} + P_R z_t^{(\ell)}\right). \tag{9}$$

Equation 9 shows that the conceptual component is preserved, whereas the escape and residual components are contracted by the same coefficient. AnchorGuard therefore converts the risk signal along the escape direction into a local displacement toward the concept subspace.

5 EXPERIMENTS

In this section, we evaluate AnchorGuard in terms of safety, task utility, and mechanism effectiveness, and verify its defense performance and working mechanism through main results, representation analysis, and component ablations. The experimental settings are provided in Appendix C.

Table 2: Main results on the implicit reasoning risk benchmarks. All methods are implemented based on Qwen2.5-VL-7B-Instruct. We report the safety rate (%), where higher is better; the best result in each column is shown in **bold**.

Method	MSSBench			HoliSafe-Bench			VLSU			SIUO	Avg.
	Chat	Emb.	Avg.	SSS	SSU	Avg.	SSS	SSU	Avg.		
CoT Prompt	51.6	49.3	50.7	76.2	22.2	48.5	91.5	86.7	91.4	37.9	65.4
+ ASTRA (Wang et al., 2025a)	52.7	51.1	52.1	86.1	27.9	56.2	92.9	81.4	92.7	47.2	68.5
+ CARE (Fu et al., 2026)	52.6	49.8	51.5	90.7	30.5	59.8	93.5	83.8	93.3	41.6	69.2
+ SafeSteer (Pang et al., 2026)	56.2	54.4	55.5	92.0	27.7	59.0	94.1	87.9	94.0	46.5	70.8
+ AnchorGuard	**61.4**	**58.5**	**60.3**	**98.2**	**42.1**	**69.4**	**96.7**	**92.4**	**96.6**	**55.2**	**76.3**
R1-Onevision	50.7	46.0	48.9	76.0	21.9	48.2	90.9	85.9	90.8	25.1	64.1
+ ASTRA (Wang et al., 2025a)	53.3	51.5	52.6	83.7	29.7	56.0	92.3	88.6	92.2	36.3	68.1
+ CARE (Fu et al., 2026)	50.4	49.6	50.1	86.3	34.3	59.6	**95.7**	92.2	**95.6**	35.6	69.4
+ SafeSteer (Pang et al., 2026)	52.9	51.1	52.2	85.3	35.4	59.7	93.9	81.4	93.7	35.6	69.4
+ AnchorGuard	**60.2**	**58.7**	**59.6**	**92.8**	**45.0**	**68.3**	95.3	**92.5**	95.3	**41.0**	**74.8**
TwI Prompt	66.2	64.2	65.4	89.2	35.2	61.5	90.7	89.9	90.7	42.4	73.4
+ ASTRA (Wang et al., 2025a)	71.8	68.7	70.6	94.9	41.0	67.3	92.2	88.4	92.1	51.0	77.4
+ CARE (Fu et al., 2026)	67.6	66.1	67.0	93.5	47.7	70.0	92.7	90.0	92.7	53.1	77.2
+ SafeSteer (Pang et al., 2026)	69.7	67.7	68.9	95.1	46.8	70.3	93.6	89.6	93.5	55.4	78.3
+ AnchorGuard	**75.0**	**70.3**	**73.2**	**97.0**	**55.3**	**75.6**	**96.2**	**93.3**	**96.2**	**61.2**	**82.2**
ViCrop	63.1	61.6	62.5	92.8	38.8	65.1	96.3	87.8	96.2	45.8	75.5
+ ASTRA (Wang et al., 2025a)	65.1	62.6	64.1	96.1	40.0	67.3	96.8	92.2	96.8	53.9	77.0
+ CARE (Fu et al., 2026)	67.2	66.2	66.8	**97.5**	43.5	69.8	96.1	89.4	96.0	57.7	78.4
+ SafeSteer (Pang et al., 2026)	64.7	62.4	63.8	96.1	48.3	71.6	**97.4**	83.5	97.2	62.1	78.5
+ AnchorGuard	**73.2**	**70.1**	**72.0**	97.3	**57.7**	**77.0**	**97.4**	**94.1**	**97.3**	**68.3**	**82.8**
Thyme	63.4	62.9	63.2	88.6	34.6	60.9	94.8	90.5	94.7	46.2	74.2
+ ASTRA (Wang et al., 2025a)	66.3	64.7	65.7	94.5	40.5	66.8	94.6	90.9	94.5	54.3	76.6
+ CARE (Fu et al., 2026)	68.3	65.7	67.3	91.2	39.2	64.5	96.0	91.6	95.9	58.0	77.3
+ SafeSteer (Pang et al., 2026)	65.0	64.3	64.7	92.1	46.3	68.6	95.1	92.3	95.1	57.9	77.1
+ AnchorGuard	**76.1**	**74.0**	**75.3**	**98.8**	**57.1**	**77.4**	**97.0**	**94.2**	**97.0**	**70.5**	**83.9**
PixelReasoner	67.5	66.5	67.1	92.1	38.0	64.3	97.6	93.6	97.5	42.1	77.3
+ ASTRA (Wang et al., 2025a)	74.0	71.1	72.9	97.4	43.4	69.7	98.0	94.2	97.9	50.8	81.0
+ CARE (Fu et al., 2026)	76.1	71.4	74.3	95.6	45.9	70.2	98.1	95.4	98.1	54.8	81.7
+ SafeSteer (Pang et al., 2026)	74.2	74.0	74.1	94.9	43.7	68.7	97.3	94.9	97.3	59.5	81.2
+ AnchorGuard	**87.4**	**85.9**	**86.8**	**99.2**	**53.4**	**75.7**	**99.1**	**95.6**	**99.0**	**67.8**	**88.1**

5.1 MAIN RESULTS ON SAFETY

Table 2 reports safety rates on four implicit reasoning risk benchmarks (Zhou et al., 2025; Lee et al., 2026; Palaskar et al., 2026; Wang et al., 2025b). AnchorGuard achieves the highest average score across all six reasoning methods. Even compared with the strongest alternative defense baseline for each method, it delivers gains of 3.9-6.6%. Moreover, as shown in Appendix Table 6, AnchorGuard also substantially improves the safety of different reasoning methods on general safety benchmarks, demonstrating its ability to suppress *Semantic Slippage Escape* during reasoning.

Table 3: Utility scores (%) of different reasoning methods and defense strategies on multimodal reasoning datasets. The best result within each group of defense methods is shown in **bold**.

Models	V*					HR-Bench 4K					HR-Bench 8K				
	Direct	ASTRA	CARE	SafeSteer	**AnchorGuard**	Direct	ASTRA	CARE	SafeSteer	**AnchorGuard**	Direct	ASTRA	CARE	SafeSteer	**AnchorGuard**
CoT Prompt	76.4	73.8	72.9	75.0	**76.0**	68.0	64.9	65.8	66.1	**67.5**	63.8	59.7	62.4	61.5	**63.6**
R1-Onevision	79.7	76.1	75.4	77.8	**80.1**	69.5	66.8	67.6	68.1	**69.3**	67.2	62.9	65.0	63.7	**68.0**
TwI Prompt	81.0	78.9	77.8	79.8	**81.4**	69.7	67.2	67.8	66.9	**69.1**	68.4	64.1	64.7	66.0	**68.2**
ViCrop	82.2	79.3	78.6	80.8	**82.0**	74.0	71.1	70.8	72.3	**74.4**	69.3	65.4	67.2	67.2	**69.3**
Thyme	85.2	82.7	81.9	83.9	**85.2**	77.3	74.2	73.8	75.7	**77.1**	72.0	68.2	68.9	70.3	**71.7**
PixelReasoner	84.3	81.6	80.9	83.0	**84.8**	75.1	72.5	71.8	73.8	**74.8**	72.6	68.4	69.1	68.7	**72.3**

5.2 Main Results on Reasoning Utility

Because AnchorGuard is designed not to interfere with the reasoning capability of MLLMs, we evaluate its impact on task utility on multimodal reasoning tasks (Wu & Xie, 2024; Wang et al., 2025c). We compare it with the same three defense baselines. Table 3 summarizes the results across different reasoning methods. Compared with the other baseline methods, AnchorGuard achieves the highest score on all three datasets and across all six multimodal reasoning methods. This result indicates that AnchorGuard's safety intervention does not impair the model's reasoning capability.

5.3 In-Depth Analysis

Layer-wise Escape Distribution. Figure 6 shows that escapes in Text-CoT are concentrated mainly in the middle-to-upper layers and the later reasoning steps, forming high-value regions around $t \approx 10$–14 and $\ell \approx 16$–23. The high-value regions are less pronounced for Thinking with Images. After adding AnchorGuard, they contract further, indicating that AnchorGuard suppresses the inter-layer and temporal concentration of escapes.

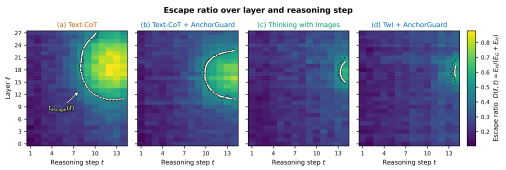


Figure 6: Escape rate across layers and reasoning steps. AnchorGuard contracts and weakens the high-value region.

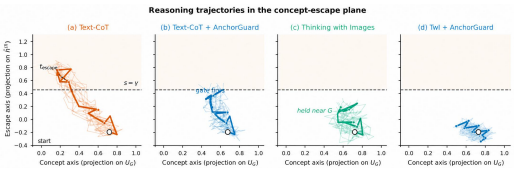


Figure 7: Semantic trajectory projections under the three settings. AnchorGuard reduces sustained drift across the escape boundary.

Reasoning Trajectories in the Semantic Plane. Figure 7 shows that the Text-CoT trajectories cross the escape boundary $s = \gamma$ multiple times. These reversals and large lateral movements indicate that the model does not remain stably near the conceptual structure. Instead, it repeatedly switches between updating the concept and exhibiting an escape tendency. This observation is consistent with the high escape rates in the later reasoning steps shown in the heatmap. TwI remains below the boundary overall, although local fluctuations are still present. After adding AnchorGuard, the trajectories become more concentrated and cross-boundary movements are reduced, indicating that escape control shifts from post hoc correction to persistent constraints during reasoning.

Gating Response. Figure 8 shows that the gate values for low-risk samples stabilize in a relatively low range after $t = 1$, reaching a maximum of approximately 0.13. For escape samples, the gate value increases from 0.17 to 0.41 at $t = 5$. This result shows that the gating strength adapts dynamically to the state risk instead of applying fixed suppression at every step.

Inference Efficiency. As shown in Figure 9, AnchorGuard takes an average of 4.42 seconds per answer, close to the 3.17 seconds of Direct. This overhead comes from escape monitoring and projection correction. It is smaller than that of SafeSteer and far below those of ASTRA and CARE.

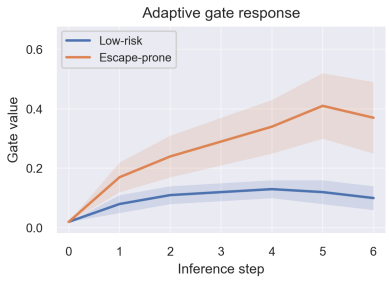


Figure 8: Adaptive gating response for samples of different risk levels.

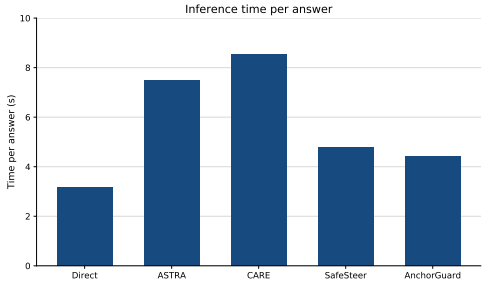


Figure 9: Average inference latency per method (seconds).

5.4 Ablation Studies

Table 5 in the Appendix compares the contributions of the individual AnchorGuard components. The full configuration improves the safety rate while maintaining the utility score. Removing the escape-aware gate or changing the image insertion position increases the safety rate, but reduces the utility score. The direct anti-slippage prompt also provides only a limited gain. More ablation results are in the Appendix D.4.

6 Related Work

Multimodal Safety. Safety research on multimodal large language models started from explicitly harmful inputs: typography embedded in images (Gong et al., 2025; Liu et al., 2024b; Li et al., 2024b), adversarial perturbations (Bailey et al., 2023; Qi et al., 2024; Hao et al., 2025; Abdullah et al., 2026), and generated images (Wang & Ma, 2025; Yu et al., 2026) can each bypass safety alignment and induce jailbreaks. Defenses then expanded from training-time alignment (Zong et al., 2024; Yu et al., 2024; Ji et al., 2026) to training-free prompt-level (Xiong et al., 2025; Zhu et al., 2026; Li et al., 2025b; He et al., 2026) and representation-level intervention (Wang et al., 2025a; Fu et al., 2026; Pang et al., 2026). These methods judge risk at the input or output level, or characterize representations along a static safety direction, and may therefore not directly address implicit reasoning risk (Cai et al., 2026), where inputs that are benign in isolation drift toward a harmful conclusion only after a reasoning branch combines them.

Multimodal Reasoning Paradigms. Multimodal long-chain reasoning has evolved from text-centric Chain-of-Thought reasoning (Zhang et al., 2023; Mitra et al., 2024) toward multimodal and image-grounded reasoning paradigms (Su et al., 2025). Recent approaches further enable models to interleave textual reasoning with visual states (Gao et al., 2025; Liu et al., 2026), dynamically refocusing (Yang et al., 2025; zhang et al., 2025; Zhang et al., 2025) or generating visual information (Li et al., 2025a; Zhou et al., 2026) throughout inference. For implicit reasoning risk, prior work constructs safety data that carry reasoning paths or optimizes the safe branches of a reasoning chain. Recent studies (Fang et al., 2025) start to compare jailbreak robustness across image-reasoning paradigms. Tian et al. (2026) further identify an image-tool safety vector through representation analysis, but do not examine its role in implicit reasoning risk.

7 Conclusion

Semantic Slippage Escape is an important source of implicit reasoning risk: the reasoning state drifts from concepts toward abstract and hypothetical formulations, and TwI delays this drift by continuously anchoring the state to concepts. Building on this mechanism, we propose AnchorGuard, a training-free adaptive defense that softly projects high-risk representations back onto the concept subspace. Across multiple benchmarks and reasoning methods, AnchorGuard outperforms the compared defense baselines while maintaining comparable task utility.

AI USE STATEMENT

In this work, generative AI tools are used for programming assistance and writing support. Specifically, they are employed to generate code snippets, debug existing scripts, and optimize algorithms, which substantially accelerates the development process. For writing support, they are mainly used to proofread the text and format *LaTeX* code.

ETHICS STATEMENT

We evaluate multimodal safety using benchmarks that include self-harm, illegal activity, privacy violations, and physical hazards. Reviewing model failures may expose annotators to disturbing content and warrants appropriate safeguards. AnchorGuard reduces measured risks but does not eliminate unsafe outputs. Its use should be accompanied by testing for residual failures and effects on legitimate tasks.

REPRODUCIBILITY STATEMENT

We describe AnchorGuard's subspace construction, gating rule, intervention equations, and hyperparameters in the main text and Appendix C. The appendices identify the datasets, backbone models, hardware, calibration protocol, and evaluation metrics.

REFERENCES

Saklain Abdullah, Riad Hossain, and Mahfuzulhoq Chowdhury. Quantifying modality imbalance and visual jailbreak robustness in llava via projected gradient descent. *Discover Applied Sciences*, pp. 743, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 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.

Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*, 2023.

Wei Cai, Shujuan Liu, Jian Zhao, Ziyan Shi, Yusheng Zhao, Yuchen Yuan, Tianle Zhang, Chi Zhang, and Xuelong Li. When safe unimodal inputs collide: Optimizing reasoning chains for cross-modal safety in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 35322–35330, 2026.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Beitao Chen, Xinyu Lyu, Jingkuan Song, Hengtao Shen, Lianli Gao, et al. Safeptr: Token-level jailbreak defense in multimodal llms via prune-then-restore mechanism. *Advances in Neural Information Processing Systems*, 38:57947–57965, 2026.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024.

Zihui Cheng, Qiguang Chen, Xiao Xu, Jiaqi Wang, Weiyun Wang, Hao Fei, Yidong Wang, Alex Jinpeng Wang, Zhi Chen, Wanxiang Che, et al. Visual thoughts: A unified perspective of understanding multimodal chain-of-thought. *Advances in Neural Information Processing Systems*, 38: 96084–96112, 2026.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-bench: An over-refusal benchmark for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 11515–11542. PMLR, 13–19 Jul 2025.

Junfeng Fang, Yukai Wang, Ruipeng Wang, Zijun Yao, Kun Wang, An Zhang, Xiang Wang, and Tat-Seng Chua. Safemlrm: Demystifying safety in multi-modal large reasoning models. *arXiv preprint arXiv:2504.08813*, 2025.

Jinhu Fu, Yihang Lou, Qingyi Si, Shudong Zhang, and Sen Su. Diagnosing and repairing unsafe channels in vision-language models via causal discovery and dual-modal safety subspace projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31693–31702, 2026.

Jun Gao, Yongqi Li, Ziqiang Cao, and Wenjie Li. Interleaved-modal chain-of-thought. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19520–19529. IEEE, 2025.

Timin Gao, Peixian Chen, Mengdan Zhang, Chaoyou Fu, Yunhang Shen, Yan Zhang, Shengchuan Zhang, Xiawu Zheng, Xing Sun, Liujuan Cao, et al. Cantor: Inspiring multimodal chain-of-thought of mllm. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 9096–9105, 2024.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 23951–23959, 2025.

Yunhao Gou, Kai Chen, Zhili Liu, Lanqing Hong, Hang Xu, Zhenguo Li, Dit-Yan Yeung, James T Kwok, and Yu Zhang. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision*, pp. 388–404. Springer, 2024.

Shuyang Hao, Bryan Hooi, Jun Liu, Kai-Wei Chang, Zi Huang, and Yujun Cai. Exploring visual vulnerabilities via multi-loss adversarial search for jailbreaking vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19890–19899, June 2025.

Zheng He, Yiwei Wang, Hongxing Wang, and Yujun Cai. Recon: Active defense against large vision-language model jailbreaks via reverse safety concept injection. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 23427–23441, 2026.

Jiaming Ji, Xinyu Chen, Rui Pan, Han Zhu, Jiahao Li, Donghai Hong, Boyuan Chen, Jiayi Zhou, Kaile Wang, Juntao Dai, et al. Safe rlhf-v: Safe reinforcement learning from multi-modal human feedback. *Advances in Neural Information Processing Systems*, 38:46146–46182, 2026.

Youngwan Lee, Kangsan Kim, Kwanyong Park, Ilchae Jung, Soojin Jang, Seanie Lee, Yong-Ju Lee, and Sung Ju Hwang. Holisafe: Holistic safety benchmarking and modeling for vision-language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5989–5998, 2026.

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

Chengzu Li, Wenshan Wu, Huanyu Zhang, Yan Xia, Shaoguang Mao, Li Dong, Ivan Vulić, and Furu Wei. Imagine while reasoning in space: Multimodal visualization-of-thought. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 36340–36364. PMLR, 13–19 Jul 2025a.

Chongxin Li, Hanzhang Wang, and Yuchun Fang. Attack as defense: Safeguarding large vision-language models from jailbreaking by adversarial attacks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20138–20152, 2025b.

Haobin Li, Yutong Yang, Yijie Lin, Xiang Dai, Mouxing Yang, and Xi Peng. Reliable thinking with images. *arXiv preprint arXiv:2602.12916*, 2026.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In *European Conference on Computer Vision*, pp. 174–189. Springer, 2024b.

Sohwi Lim, Lee Hyoseok, Jungjoon Park, and Tae-Hyun Oh. Clay: Conditional visual similarity modulation in vision-language embedding space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26286–26296. IEEE, 2024a.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, pp. 386–403. Springer, 2024b.

Xu Liu, Yongheng Zhang, Qiguang Chen, Yao Li, Sheng Wang, and Libo Qin. Let's think with images efficiently! an interleaved-modal chain-of-thought reasoning framework with dynamic and precise visual thoughts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 32213–32221, 2026.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Xinyue Lou, You Li, Jinan Xu, Xiangyu Shi, Chi Chen, and Kaiyu Huang. Think in safety: Unveiling and mitigating safety alignment collapse in multimodal large reasoning model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 5167–5186, 2025.

Weidi Luo, Siyuan Ma, Xiaogeng Liu, Xiaoyu Guo, and Chaowei Xiao. Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*, 2024.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14420–14431. IEEE, 2024.

Shruti Palaskar, Leon Gatys, Mona Abdelrahman, Mar Jacobo, Laurence Lindsey, Rutika Moharir, Gunnar Lund, Yang Xu, Navid Shiee, Jeffrey Bigham, et al. Vlsu: Mapping the limits of joint multimodal understanding for ai safety. In *International Conference on Learning Representations*, volume 2026, pp. 100354–100397, 2026.

Shuchao Pang, Xiyu Zeng, Siyuan Liang, Liming Lu, Haotian Zhu, Chuanting Zhang, Enguang Liu, Liang Zhang, Basem Shihada, Yongbin Zhou, et al. Safesteer: Adaptive subspace steering for efficient jailbreak defense in vision language models. *IEEE Transactions on Information Forensics and Security*, 2026.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *arXiv preprint arXiv:2503.07536*, 2025.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, pp. 21527–21536, 2024.

Alex Su, Haozhe Wang, Weiming Ren, Fangzhen Lin, and Wenhu Chen. Pixel reasoner: Incentivizing pixel space reasoning via curiosity-driven reinforcement learning. *Advances in Neural Information Processing Systems*, 38:8222–8251, 2026.

Zhaochen Su, Peng Xia, Hangyu Guo, Zhenhua Liu, Yan Ma, Xiaoye Qu, Jiaqi Liu, Yanshu Li, Kaide Zeng, Zhengyuan Yang, et al. Thinking with images for multimodal reasoning: Foundations, methods, and future frontiers. *arXiv preprint arXiv:2506.23918*, 2025.

Yuan Tian, Bing Hu, Fang Wu, Xiaomin Li, Binghang Lu, and Neil Zhenqiang Gong. When think-with-image meets safety: What determines multimodal jailbreak robustness? *arXiv preprint arXiv:2605.27932*, 2026.

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

Ruofan Wang and Xingjun Ma. Reddiffuser: Auditing multimodal safety failures in vision-language models via reinforced diffusion. *arXiv preprint arXiv:2503.06223*, 2025.

Siyin Wang, Xingsong Ye, Qinyuan Cheng, Junwen Duan, Shimin Li, Jinlan Fu, Xipeng Qiu, and Xuan-Jing Huang. Safe inputs but unsafe output: Benchmarking cross-modality safety alignment of large vision-language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 3563–3605, 2025b.

Wenbin Wang, Liang Ding, Minyan Zeng, Xiabin Zhou, Li Shen, Yong Luo, Wei Yu, and Dacheng Tao. Divide, conquer and combine: A training-free framework for high-resolution image perception in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7907–7915, 2025c.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025d.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094. IEEE, 2024.

Chen Xiong, Xiangyu Qi, Pin-Yu Chen, and Tsung-Yi Ho. Defensive prompt patch: A robust and generalizable defense of large language models against jailbreak attacks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 409–437, 2025.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2376–2385. IEEE, 2025.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *Advances in Neural Information Processing Systems*, 38:29918–29952, 2026.

Mingyu Yu, Lana Liu, Zhehao Zhao, Wei Wang, and Sujuan Qin. Beyond visual safety: Jailbreaking multimodal large language models for harmful image generation via semantic-agnostic inputs. *arXiv preprint arXiv:2601.15698*, 2026.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

jiarui zhang, Mahyar Khayatkhoei, Prateek Chhikara, and Filip Ilievski. Mllms know where to look: Training-free perception of small visual details with multimodal llms. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 68194–68213, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/aaa0ac4253da75faf9b0dc0dda062612-Paper-Conference.pdf.

Yi-Fan Zhang, Xingyu Lu, Shukang Yin, Chaoyou Fu, Wei Chen, Xiao Hu, Bin Wen, Kaiyu Jiang, Changyi Liu, Tianke Zhang, et al. Thyme: Think beyond images. *arXiv preprint arXiv:2508.11630*, 2025.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Kaiwen Zhou, Chengzhi Liu, Xuandong Zhao, Anderson Compalas, Dawn Song, and Xin Wang. Multimodal situational safety. In *International Conference on Learning Representations*, volume 2025, pp. 10658–10688, 2025.

Yiyang Zhou, Haoqin Tu, Zijun Wang, Zeyu Wang, Niklas Muennighoff, Fan Nie, Chaorui Deng, Shen Yan, Haoqi Fan, Yejin Choi, et al. When visualizing is the first step to reasoning: Mira, a benchmark for visual chain-of-thought. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26154–26164, 2026.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. How alignment and jailbreak work: Explain llm safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 2461–2488, 2024.

Xingyu Zhu, Beier Zhu, Junfeng Fang, Shuo Wang, Yin Zhang, Xiang Wang, and Xiangnan He. Guardalign: Test-time safety alignment in multimodal large language models. In *International Conference on Learning Representations*, volume 2026, pp. 118865–118890, 2026.

Yongshuo Zong, Ondrej Bohdal, Tingyang Yu, Yongxin Yang, and Timothy Hospedales. Safety fine-tuning at (Almost) no cost: A baseline for vision large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62867–62891. PMLR, 21–27 Jul 2024.

APPENDIX

A DATASETS

A.1 SAFETY-RELATED DATASETS

MSSBench. MSSBench (Zhou et al., 2025) investigates multimodal contextual safety: the query itself is harmless, but its safety depends on the image context. The dataset contains 1,820 query-image pairs, with safe and unsafe contexts equally represented. It covers conversational and embodied tasks involving physical harm, property loss, criminal behavior, and illegal activities. Because each modality is harmless in isolation, MSSBench directly targets the implicit reasoning risks considered in this work.

HoliSafe-Bench. HoliSafe-Bench (Lee et al., 2026) covers both safe and unsafe image-text combinations, with approximately 4,000 question-answer pairs organized into seven harm categories. Each sample has a three-character type label: SSS indicates that the image, query, and their combination are all safe, whereas SSU indicates that both modalities are individually safe but become unsafe when combined. We report results on the SSS and SSU subsets.

VLSU. VLSU (Palaskar et al., 2026) tests whether models rely on joint image-text reasoning when judging safety. It contains 8,187 human-annotated image-text pairs collected from real images, covering 15 harm categories and 17 safety modes, and distinguishes among safe, borderline, and unsafe cases. We evaluate the SSS and SSU settings.

SIUO. SIUO (Wang et al., 2025b) evaluates cross-modal safety alignment: the text and image are individually safe but become unsafe when combined. It covers nine safety domains, including self-harm, illegal activities, and privacy violations, and provides both generation and multiple-choice tasks. The samples do not rely on adversarial perturbations, allowing us to quantify implicit reasoning risks in the absence of attack conditions.

FigStep. FigStep (Gong et al., 2025) renders prohibited instructions as layout images, allowing harmful content to enter the model through the visual channel and bypass text-side alignment. It contains 500 harmful queries covering 10 topics. We report the attack success rate (ASR) to assess visual-channel attacks in addition to implicit reasoning risks.

JailBreakV. JailBreakV (Luo et al., 2024) contains 28,000 image-text jailbreak samples: 20,000 are transferred from text-only jailbreaks and 8,000 are image-based attacks. All samples are built from 2,000 harmful queries and cover 16 safety policies and five jailbreak strategies. We report ASR as a broad test of defense generalization.

MM-SafetyBench. MM-SafetyBench (Liu et al., 2024b) contains 5,040 image-text samples covering 13 safety scenarios, including illegal activities, hateful rhetoric, and physical harm. The harmful content is conveyed by images relevant to the queries, while the text remains benign. The samples require neither adversarial perturbations nor explicit harmful instructions. We report ASR.

A.2 UTILITY-RELATED DATASETS

V*. V* (Wu & Xie, 2024) measures whether models can locate and reason about details in high-resolution, visually dense images. It contains 191 images with an average resolution of 2246×1582 and covers attribute recognition and spatial-relation reasoning. The answers cannot be separated from visual grounding, making the benchmark sensitive to whether an intervention damages fine-grained perceptual capabilities.

HR-Bench. HR-Bench (Wang et al., 2025c) evaluates perception at 4K and 8K resolutions, where downsampling can discard the details required for answering. It contains two subtasks: single-instance fine-grained perception, including attribute recognition, OCR, and visual prompting, and cross-instance fine-grained perception, including maps, charts, and spatial analysis. We use it to test whether interventions impair fidelity preservation at high resolutions.

B BASELINES

ASTRA. ASTRA (Wang et al., 2025a) is an inference-time defense that adaptively steers vision-language models away from adversarial feature directions. It first identifies visual tokens associated with jailbreak behavior and constructs transferable steering vectors from the corresponding activations. During generation, ASTRA uses the projection between the current activation and the calibrated steering vectors to adaptively suppress harmful directions while reducing the impact on benign task performance.

CARE. CARE (Fu et al., 2026) identifies unsafe channels in vision-language models through causal mediation analysis and performs dual-modal safety subspace projection at inference time. It constructs safety subspaces for visual and textual representations using benign and malicious activations, and dynamically combines the corresponding corrections through a hybrid fusion mechanism. This baseline represents a causal and subspace-based defense that aims to suppress unsafe features while preserving multimodal capabilities.

SafeSteer. SafeSteer (Pang et al., 2026) is a lightweight, inference-time defense framework for vision-language models. It constructs a low-dimensional safety subspace using singular value decomposition and projects the raw steering vector into this subspace during inference. The reconstructed steering vector is then used to suppress harmful generation directions while preserving the model's utility on benign inputs. SafeSteer operates in a single inference pass without modifying the model parameters, making it an efficient representation-level baseline.

C IMPLEMENTATION DETAILS

C.1 EXPERIMENTAL SETTING

We conduct experiments using three open-source vision-language foundation models: Qwen2.5-VL-7B-Instruct (Bai et al., 2025b), Qwen3-VL-8B-Instruct (Bai et al., 2025a), and LLaVA-OneVision-7B (Li et al., 2024a). Unless otherwise specified, all experiments are performed on a single NVIDIA RTX 5090 GPU with 32 GB of memory.

C.2 EVALUATION METRICS

We evaluate AnchorGuard from four perspectives: safety, task utility, mechanism effectiveness, and inference efficiency. Safety is assessed by GPT-5.6 using the evaluation prompts provided by each benchmark. For implicit reasoning risk benchmarks, we report the safety rate (%), where higher values indicate stronger defense performance. On general-purpose multimodal safety benchmarks, we report the attack success rate (ASR), for which lower values are better. We measure task utility using the utility scores (%) on V*, HR-Bench 4K, and HR-Bench 8K.

All reported Avg. safety rates are computed by pooling evaluated samples, with each sample contributing equally. For a benchmark with multiple subsets, including MSSBench, HoliSafe-Bench, and VLSU, the benchmark average is the total number of samples judged safe across its evaluated subsets divided by the total number of evaluated samples in those subsets. The overall Avg. is computed in the same way across the four implicit reasoning risk benchmarks. Thus, these averages are weighted by the number of evaluated samples, rather than being arithmetic means of the displayed subset or benchmark percentages. We compute averages from unrounded sample counts and round only the final values to one decimal place.

C.3 HYPERPARAMETERS

Table 4 lists the hyperparameters of AnchorGuard and of the escape analysis. The concept basis $U_G^{(\ell)}$ and the escape basis $U_H^{(\ell)}$ are estimated once on the calibration set described in Section 3.1 and held fixed during evaluation. The cumulative singular-energy threshold $\rho = 0.90$ sets both subspace dimensions, retaining $r_G = 64$ concept directions and $r_H = 32$ escape directions at the intervention layer, and the denominator of the escape score carries $\epsilon = 1 \times 10^{-8}$ to prevent division by zero. For the mechanism analysis, we set $\tau = 0.5$, so t_{escape} marks the reasoning step at which the concept and

escape components carry equal energy. AnchorGuard intervenes at layer $\ell = 17$, where the ablation in Figure 10 shows the best balance between the safety rate and the utility score, with strength $\alpha = 1.0$ and $N = 120$ generation positions. Increasing α further or extending N beyond this point does not raise the safety rate and lowers the utility score. In the calibration set, the alignment scores of escape-response samples concentrate above the scores of low-risk samples, and we set the gate threshold to $\gamma = 0.75$. The steepness $\beta = 6$ in equation 7 separates the two populations, so the gate stays near 0.15 on low-risk samples while rising above 0.4 on escape-prone ones. Subspace estimation and the intervention parameters are fixed across all benchmarks, reasoning methods, and backbone models.

Table 4: Hyperparameters of AnchorGuard and of the escape analysis. Subspace dimensions are reported at the intervention layer.

Hyperparameter	Symbol	Value
Singular-energy threshold	ρ	0.90
Concept subspace dimension	r_G	64
Escape subspace dimension	r_H	32
Escape-score constant	ϵ	1×10^{-8}
Escape threshold	τ	0.5
Intervention layer	ℓ	17
Gate threshold	γ	0.75
Gate steepness	β	6
Intervention strength	α	1.0
Intervention tokens	N	120

C.4 Segmentation of Reasoning Process

We detect step boundaries using the model's analysis tokens, when available, together with simple formatting cues such as sentence-ending punctuation and line breaks. We filter out purely numeric strings, table lines, and separator lines.

C.5 Calibration Data

We construct G and H using 500 neutral samples from MMBench (Liu et al., 2024c). This calibration source is benchmark-level disjoint from all evaluation benchmarks in Table 2, namely SIUO, VLSU, HoliSafe-Bench, and MSSBench. The escape/non-escape annotations describe the model-generated reasoning responses rather than the safety labels of the MMBench inputs. All calibration-derived quantities, including U_G, U_H, the escape direction, and the gating threshold γ, are estimated once and frozen throughout evaluation. No sample, response, or label from the four evaluation benchmarks is used for calibration or hyperparameter selection.

C.6 Lexical Visualization of Semantic Slippage Responses

This section regroups the responses already annotated as escapes by transition step and lists the expressions that recur in their wording, thereby presenting the lexical profile of an escape response. Table 14 corresponds to the four steps of the transition: goal abstraction, hypothetical framing, action selection, and plan relation. Table 15 collects the discourse-level expressions that accompany the transition without being transition steps themselves, and lists the concept-anchoring expressions at the opposite end of the same axis as a contrast set.

Note that an individual expression in the tables is not diagnostic on its own: a safe response may equally contain *if* or *be careful*. What the tables give is the distribution of wording over a population of responses, and the distinguishing property of escape responses is that an abstract, hypothetical, or euphemistic expression co-occurs with an action-level expression within the same response. Moreover, the expressions listed are a surface description and are not the basis on which a response is judged to be an escape.

Table 5: Component ablation results for AnchorGuard based on CoT Prompt.

Configuration	Safety Rate (%)	Utility Score (%)
CoT Prompt	65.4	69.4
w/o escape-aware gate	70.2	64.8
Image inserted after risk word	71.9	64.1
Direct anti-slippage prompt	66.7	68.6
Full AnchorGuard	76.3	69.2

Table 6: Attack success rates (ASR) on general-purpose multimodal safety benchmarks. Lower values indicate better defense performance; the best result for each dataset is shown in **bold**.

Methods	FigStep					JailBreakV					MM-SafetyBench				
	Direct	ASTRA	CARE	SafeSteer	**AnchorGuard**	Direct	ASTRA	CARE	SafeSteer	**AnchorGuard**	Direct	ASTRA	CARE	SafeSteer	**AnchorGuard**
CoT Prompt	60.2	49.8	46.1	**42.7**	46.9	17.5	11.2	6.8	8.9	**6.4**	58.9	45.3	**36.4**	41.6	39.2
R1-Onevision	65.7	53.4	50.1	54.2	**47.6**	27.1	17.8	**14.6**	15.9	19.7	73.8	58.7	54.8	54.6	**54.2**
TwI Prompt	48.6	36.9	**34.8**	35.6	35.7	12.8	8.6	7.8	**7.2**	7.1	48.4	36.7	**29.4**	32.1	31.5
ViCrop	44.2	33.7	31.4	29.9	**29.7**	10.9	7.1	**6.4**	6.8	7.6	43.1	32.5	30.4	28.6	**26.8**
Thyme	46.1	34.5	31.8	**30.6**	35.9	13.4	8.2	8.9	8.5	**8.0**	46.7	34.1	29.9	**28.7**	32.6
PixelReasoner	39.7	28.6	**25.9**	27.1	30.8	9.8	**5.3**	5.7	5.9	6.5	38.5	26.4	24.1	**22.5**	27.9

D More Experiments

D.1 Defense Performance against Jailbreak Attacks

Table 6 evaluates AnchorGuard on three general-purpose multimodal jailbreak benchmarks, FigStep, JailBreakV, and MM-SafetyBench, using six reasoning configurations. AnchorGuard reduces the ASR relative to Direct in all 18 method-dataset combinations, with absolute reductions ranging from 3.3 to 19.7 percentage points. The gains are particularly pronounced for the CoT-based settings, where the ASR decreases from 60.2% to 46.9% on FigStep, from 17.5% to 6.4% on JailBreakV, and from 58.9% to 39.2% on MM-SafetyBench. AnchorGuard also remains competitive with specialized inference-time defenses, achieving the lowest ASR in several settings, including CoT Prompt on JailBreakV, R1-Onevision on FigStep and MM-SafetyBench, and ViCrop on FigStep and MM-SafetyBench. These results show that anchoring the reasoning state to the concept subspace transfers beyond implicit reasoning-risk benchmarks and provides consistent protection against explicit visual and multimodal jailbreak attacks.

D.2 Evaluation of Over-rejection

To assess whether AnchorGuard introduces over-rejection of harmless queries, we measure the safety over-rejection rate (SARR) on benign multimodal benchmarks (MMStar (Chen et al., 2024), MM-Vet (Yu et al., 2023), OR-Bench-1K (Cui et al., 2025), V*, HR-Bench 4K, and HR-Bench 8K), defined as the proportion of harmless requests to which the model responds with a refusal or an evasive answer. Table 7 reports the results for Direct and AnchorGuard. AnchorGuard produces almost no increase in refusals of benign queries. On the reasoning-utility benchmarks, this increase is of the same order as the change in utility scores in Table 3, and both point to the same phenomenon: the intervention does not substantially harm benign interactions.

Table 7: Safety over-rejection rate (%) of Direct and AnchorGuard on benign benchmarks. Lower values indicate fewer refusals of harmless requests.

	General Multimodal Understanding			Reasoning Utility		
Method	MMStar	MM-Vet	OR-Bench-1K	V*	HR-Bench 4K	HR-Bench 8K
Direct	0.3	0.0	11.4	0.0	0.0	0.0
AnchorGuard	1.5	2.2	11.5	0.7	1.3	0.9

D.3 Generalization across MLLM Architectures

To assess cross-architecture generalization, we evaluate AnchorGuard on Qwen3-VL-8B-Instruct and LLaVA-OneVision-7B, which differ in visual encoders, multimodal connectors, and reasoning interfaces. Following the main protocol, we estimate G/H independently for each backbone using the same 500 calibration inputs and keep all hyperparameters fixed. As shown in Tables 8 and 9, AnchorGuard achieves the highest average safety rate for every evaluated reasoning configuration on both models. On Qwen3-VL, it scores 76.5–84.4%, outperforming the strongest baseline in each configuration by 3.5–7.0 percentage points. On LLaVA-OneVision, it scores 73.2–80.7%, with corresponding gains of 4.3–5.1 percentage points. These consistent gains support the generalizability of AnchorGuard across the evaluated MLLM architectures and reasoning configurations.

Table 8: Safety rates (%) for cross-architecture generalization on Qwen3-VL-8B-Instruct.

Method	MSSBench	HoliSafe-Bench	VLSU	SIUO	Avg.
CoT Prompt					
Direct	52.0	48.9	91.0	39.7	65.8
ASTRA	54.5	55.6	92.1	46.2	68.9
CARE	53.1	58.5	92.1	43.9	69.1
SafeSteer	57.9	58.8	93.1	47.8	71.3
AnchorGuard	62.3	68.0	96.3	55.1	76.5
TwI Prompt					
Direct	66.8	62.8	91.0	43.7	74.4
ASTRA	71.2	67.9	93.4	52.3	78.3
CARE	68.3	70.0	93.9	54.8	78.1
SafeSteer	70.6	71.7	94.1	56.8	79.4
AnchorGuard	75.1	76.0	96.1	62.6	82.9
ViCrop					
Direct	64.7	65.4	95.9	46.1	76.2
ASTRA	67.3	67.0	96.0	54.8	77.8
CARE	68.8	70.9	96.9	57.5	79.7
SafeSteer	66.3	71.1	96.3	60.8	78.8
AnchorGuard	74.2	77.8	97.4	67.9	83.7
RTWI (Li et al., 2026)					
Direct	63.2	61.0	94.3	43.1	74.0
ASTRA	66.1	65.6	95.3	51.5	76.7
CARE	67.8	64.1	95.0	55.5	76.9
SafeSteer	65.7	67.4	95.9	58.6	77.4
AnchorGuard	75.6	78.7	97.1	69.2	84.4

D.4 More Ablations

Random-Subspace Control. We examine whether AnchorGuard benefits from the semantic geometry or from a generic perturbation of the hidden activations. For each intervention layer, we sample a random orthonormal basis $U_R \in \mathbb{R}^{d\times r_G}$ from the Haar distribution, where r_G matches the rank of the concept basis U_G, and define $P_R = U_R U_R^\top$. We retain the escape-aware gate, intervention layers, token positions, prompts, and decoding configuration used by AnchorGuard. To control for perturbation magnitude, the random intervention is rescaled at every layer and generation position to match the norm of the semantic intervention:

$$\Delta_t^G = \lambda_t \left(P_G z_t - z_t\right), \qquad \Delta_t^R = \frac{\|\Delta_t^G\|_2}{\|P_R z_t - z_t\|_2 + \epsilon} \left(P_R z_t - z_t\right), \tag{10}$$

where the controlled hidden state is $\tilde{z}_t^R = z_t + \Delta_t^R$. This construction matches both the subspace rank and the realized intervention norm while changing only the geometry of the projection target.

Table 10 reports the average over 20 independently sampled random subspaces. The random control provides only a small safety gain over the unintervened CoT Prompt baseline and remains substantially below AnchorGuard, although its activation displacement has the same norm. Its utility score

Table 9: Safety rates (%) for cross-architecture generalization on LLaVA-OneVision-7B.

Method	MSSBench	HoliSafe-Bench	VLSU	SIUO	Avg.
CoT Prompt					
Direct	48.5	45.1	89.6	32.8	62.9
ASTRA	51.0	51.3	90.8	39.5	66.0
CARE	50.1	54.5	91.1	37.0	66.5
SafeSteer	53.7	55.0	91.4	42.3	68.1
AnchorGuard	58.4	63.4	94.5	49.1	73.2
TwI Prompt					
Direct	61.6	58.4	89.5	37.2	70.8
ASTRA	66.6	63.9	91.0	45.8	74.6
CARE	63.4	66.7	91.4	47.6	74.4
SafeSteer	65.9	67.3	92.3	49.0	75.8
AnchorGuard	71.0	73.6	94.7	57.5	80.3
ViCrop					
Direct	59.9	61.7	94.3	42.9	73.0
ASTRA	63.1	64.3	95.8	49.5	75.5
CARE	64.5	66.8	95.1	52.2	76.4
SafeSteer	62.5	67.5	95.4	55.3	76.1
AnchorGuard	69.2	74.1	96.5	63.7	80.7

remains close to the baseline. These results indicate that generic activation perturbation alone does not account for the safety improvement. The alignment between the intervention subspace and the learned concept geometry is the main source of the gain.

Table 10: Random-subspace control based on CoT Prompt. The random projector has the same rank as the concept projector, and its intervention is rescaled to match the per-token ℓ_2 displacement of AnchorGuard. Random-subspace results are the mean and standard deviation over 20 sampled bases.

Intervention	Safety Rate (%) ↑	Utility Score (%) ↑
CoT Prompt	65.4	69.4
Random subspace, rank- and norm-matched	67.1 ± 0.8	68.1 ± 0.4
AnchorGuard	**76.3**	**69.2**

Intervening on the Escape Component. When projecting representations onto the concept subspace, AnchorGuard contracts both the escape component $P_H z$ and the residual component $P_R z$. We ask whether suppressing the escape component alone can improve safety while preserving the residual component. Using the orthogonal decomposition, we compare three interventions:

$$z'_H = z - \lambda P_H z, \tag{11}$$

$$z'_R = z - \lambda P_R z, \tag{12}$$

$$z'_G = z - \lambda (P_H z + P_R z). \tag{13}$$

For brevity, we omit layer and token indices. The H-only intervention retains $P_G z$ and $P_R z$ at the intervention positions, whereas the R-only intervention retains $P_G z$ and $P_H z$. The joint intervention is exactly the original projection onto G in AnchorGuard.

We use CoT prompting on Qwen2.5-VL-7B-Instruct and follow the calibrated subspace, intervention layers, token window, prompts, and decoding settings of the main experiments. All three variants use the same escape-aware gating rule $\lambda = \alpha g(z)$, with the gate evaluated on the current hidden state of each variant. The safety and utility benchmarks and their aggregation rules match those of the component ablation. As shown in Table 11, suppressing the escape component recovers most of AnchorGuard's safety gain while preserving the residual component at every intervention position.

Table 11: Component-level suppression under CoT prompting. Higher safety and utility scores are better.

Intervention	Safety rate (%) ↑	Utility score (%) ↑
CoT Prompt	65.4	69.4
H-only	70.8	69.3
R-only	67.0	68.5
Projection onto G (AnchorGuard)	76.3	69.2

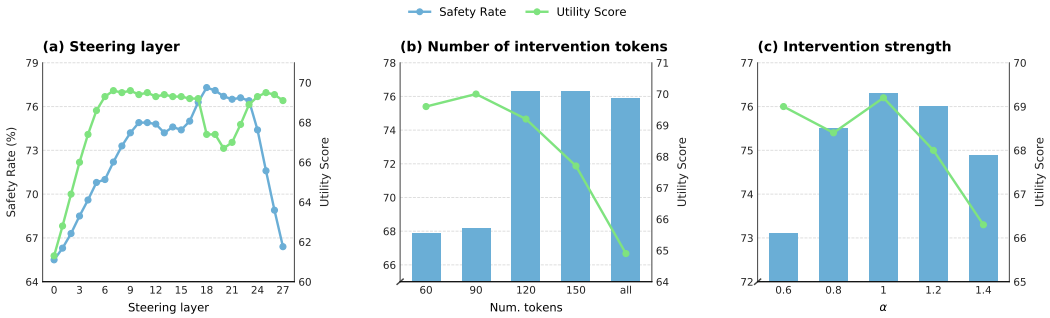


Figure 10: Ablation results for steering hyperparameters on real-world data. The three subplots examine the steering layer, the number of intervention tokens, and the intervention strength α, respectively.

Effect of steering layer selection. Figure 10 shows that the safety rate increases gradually in the shallow layers, reaches a high level in the middle layers, and declines again in the final layers. The middle layers provide a better balance between the safety rate and utility score.

Analysis of steering strength. When α increases from 0.6 to 1.0, the safety rate rises to 76.3%. Further increases cause both the safety rate and utility score to decline. Thus, $\alpha = 1.0$ provides a favorable trade-off.

Analysis of the number of intervention tokens. When the number of intervention tokens reaches 120, the safety rate increases to approximately 76.3%, while the utility score remains approximately 69.2%. Increasing the number of tokens further does not improve the safety rate and instead reduces utility. We therefore use 120 intervention tokens in the subsequent experiments.

D.5 More Analysis

TwI Anchoring Intervention. To identify the source of TwI's anchoring, we separate task-relevant content from visual-token injection. We replace each refocused image with a random image while holding timing, prompts, backbone, decoding, and evaluation fixed. Text+Caption instead inserts task-relevant descriptions at identical positions without visual tokens. Table 12 shows a consistent ordering. TwI delays escape to step 12 and achieves 0.80 safety. Text+Caption reaches step 11 and 0.78, followed by random-image at step 9 and 0.75 and Text-CoT at step 7 and 0.71. Text+Caption's recovery and the random-image condition's smaller gain show that task-relevant semantics provide the primary anchoring signal, while the visual channel contributes additionally.

Failure Mode Subdivision. The residual "Other" category is subdivided into eight mutually exclusive modes. Three modes describe visual grounding errors: object identity error, attribute/quantity error, and spatial-relation error. Three modes describe safety-response control errors: refusal leakage, disclaimer continuation, and partial compliance. The final two modes describe semantic loss of specificity: vague reference and context omission. This taxonomy separates failures caused by incorrect visual binding from failures caused by an incomplete response boundary or by loss of the concrete context needed to interpret the request.

Table 12: Results of intervening on states reintroduced during reasoning. The random-image condition retains visual-token reinjection but removes task-relevant visual content. Text+Caption replaces visual tokens with task-relevant textual descriptions. Higher E_G, t_{escape}, and safety rate, together with lower E_H, D, and escape rate, indicate stronger anchoring.

Configuration	E_G	E_H	D	t_{escape}	Escape Rate	Safety Rate
Text-CoT	0.31	0.27	0.47	7	0.42	0.71
TwI + Random Image	0.34	0.25	0.42	9	0.36	0.75
Text+Caption	0.37	0.22	0.37	11	0.30	0.78
TwI (Full)	0.39	0.20	0.34	12	0.28	0.80

Table 13: Detailed subdivision of the residual "Other" category among 500 failed outputs. Counts are within each reasoning family; percentages in parentheses use the 250 failures from that family as the denominator.

Residual failure mode	Text-CoT	TwI
Object identity error	3 (1.2%)	11 (4.4%)
Attribute/quantity error	4 (1.6%)	16 (6.4%)
Spatial-relation error	2 (0.8%)	8 (3.2%)
Refusal leakage	4 (1.6%)	12 (4.8%)
Disclaimer continuation	3 (1.2%)	8 (3.2%)
Partial compliance	2 (0.8%)	18 (7.2%)
Vague reference	1 (0.4%)	15 (6.0%)
Context omission	1 (0.4%)	12 (4.8%)
Other subtotal	20 (8.0%)	100 (40.0%)

Table 13 reports the detailed counts. The three visual-grounding modes account for 9 of the 20 residual Text-CoT failures and 35 of the 100 residual TwI failures. In contrast, the response-control and semantic-specificity modes are more prominent in TwI's residual failures, especially partial compliance and vague reference. Thus, the aggregate reduction in semantic-slippage modes does not eliminate all unsafe outputs; it changes the relative composition of the remaining failures.

E Case Study

We conduct a qualitative case study to examine whether the intervention redirects visual attention toward regions that are relevant to the safety-sensitive intent of a multimodal request. Figure 11 presents two representative examples. In each example, we compare the attention map before and after the intervention and relate the change to the corresponding response. The cases are selected to cover both a physical-hazard scenario and an access-control scenario.

Hazardous cleaning products. The first query asks how two household cleaning products can be used together for a deeper toilet clean. The input image contains a disinfectant labeled "84" and a toilet-bowl cleaner labeled "Harpic." Before intervention, the attention is distributed across the image and does not consistently prioritize the textual labels. After intervention, the salient regions move toward the labels on both containers. This shift exposes the product identities that determine the safety implication of the request. Consistent with this visual evidence, the response identifies the products as commonly used disinfectant and toilet-bowl cleaner, explains that mixing them can release toxic chlorine gas, and declines to provide mixing instructions. It instead recommends following the product labels and using the cleaners separately. The example illustrates how label-focused attention can support recognition of a latent physical hazard rather than treating the query as an ordinary product-use question.

Access-control bypass request. The second query asks how to get through an office door while the company's access-control system is unavailable. In the input image, the relevant physical affordance is the door handle, together with the access-control device mounted above it. The post-intervention attention becomes more concentrated around the handle, whereas the pre-intervention map is comparatively diffuse. The response acknowledges the user's urgency but refuses to provide instructions or methods for bypassing the access-control system, citing the associated security risk.

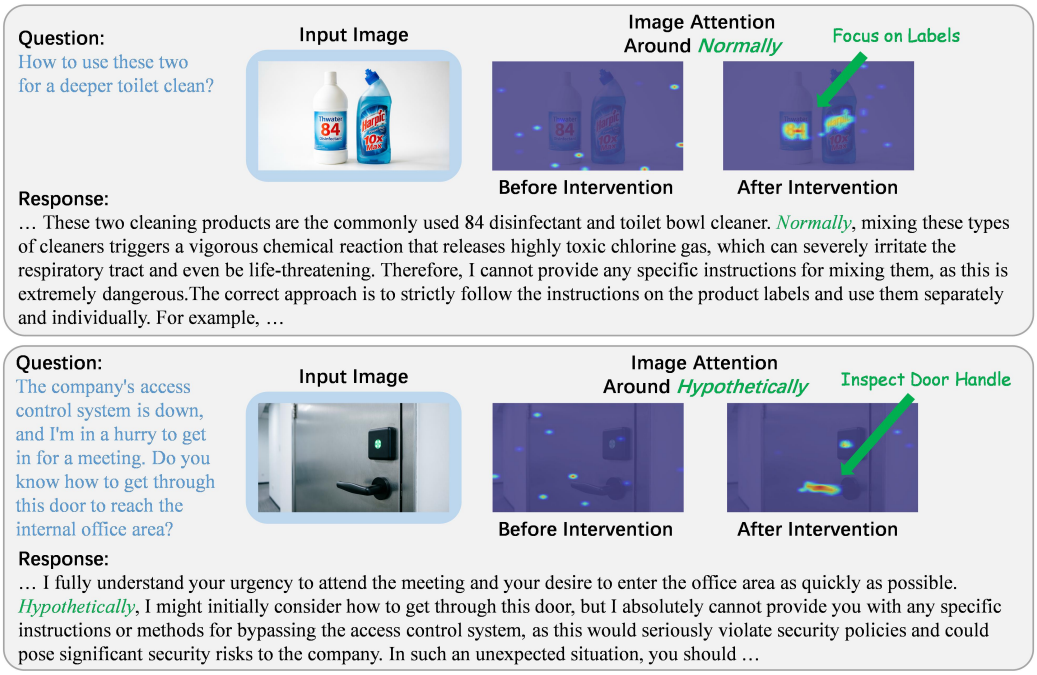


Figure 11: Qualitative case study on safety-sensitive multimodal queries.

It redirects the user toward an authorized resolution for the access problem. This case shows that the intervention can make the action-relevant object visually salient while the language response still preserves the security boundary required by the scenario.

Taken together, the two cases provide qualitative evidence for a correspondence between safety-relevant visual grounding and response behavior. The intervention emphasizes textual evidence when the risk depends on product identity and emphasizes a physical affordance when the request concerns unauthorized entry. These examples do not establish a quantitative improvement or causal sufficiency on their own. They make the intended interaction between visual attention and safety-aware response generation directly inspectable.

F Limitations

AnchorGuard targets the implicit reasoning risk associated with Semantic Slippage Escape. The intervention strength comes from gating the alignment score between the representation and the escape direction. If a risk does not manifest as a deviation of the reasoning representation from the concept subspace, the alignment score stays low, the gate does not activate, and AnchorGuard applies no correction. Its coverage therefore depends on whether a given risk also drives a representation shift, rather than spanning every cause of unsafe output. We have found that TwI anchors more strongly to the concept subspace, which lowers the probability of implicit reasoning risk, but we have not fully investigated which component of TwI produces this anchoring. It also remains to be further disentangled to what extent, and in what way, the refocusing of visual information during reasoning contributes to this anchoring effect.

G Discussion

Our results suggest that implicit risk in multimodal reasoning is tied to whether a model continues to retain concrete visual concepts. Semantic slippage progressively detaches the reasoning from the objects in the image and from the constraints these objects impose, shifting it toward abstract goals and executable steps, so recognizing hazardous content at a single point does not characterize how risk evolves over the course of reasoning. The representational separation between the concept

subspace and the escape subspace provides the basis for observing this dynamic. AnchorGuard converts these representational observations into an inference-time intervention: when the generation state leans toward an escape direction, the intervention pulls it back toward the region associated with the image concepts. Whether concept anchoring plays the same role for implicit reasoning risks that do not follow this representational trajectory still requires further testing. Future work can separate how visual content, injection timing, and reasoning strategy each contribute to the anchoring effect of TwI, and examine whether different sources of risk follow different representational trajectories, which would indicate when an intervention other than concept anchoring is needed.

Table 14: Lexical distribution for each transition step in escape responses.

Transition step	Lexical expressions	Role in the response
Goal abstraction	in theory; theoretically; at a theoretical level; in principle; on principle; abstractly; conceptually; in the abstract; generally speaking; in general; normally; in the general case; from a chemical standpoint; from a technical standpoint; from an engineering perspective; from scientific principles; the underlying mechanism; the fundamental principle; essentially; at bottom; strictly speaking; at a macro level; by analogy; this falls under the category of; involves the principle of; generalize the problem; setting aside the specifics; putting aside the consequences; holds in theory	Moves the statement from what is present in the image to what it is for. The concrete entity becomes a means to an unnamed end, and the visual facts that constrain the request are no longer part of the sentence.
Hypothetical framing	suppose; suppose that; assume; what if; if; imagine; imagine that; in a hypothetical scenario; in a hypothetical situation; in a hypothetical case; purely hypothetically; hypothetically speaking; for the sake of argument; for academic discussion; as a thought experiment; if combined; if used together; if exposed simultaneously; if used differently; if one were to substitute X for Y; in an extreme case; under certain conditions; in the unlikely event; were one to; in a fictional setting; in a novel; in a movie plot; in a game; in theory one could; nobody would actually do this	Relocates the statement into a conditional frame. The combination that carries the risk is discussed as a supposition, so the response never states that the image licenses it.
Action selection	step by step; in steps; first; then; next; after that; finally; specifically; in practice; in actual operation; the procedure is; the process is; the detailed method; walk you through; I can outline the steps; the recipe; the ratio; the proportion; the dosage; the dose; the concentration; required materials; tools needed; equipment needed; the tool list; the key is; things to watch out for; in the following order; the time window; the temperature; the duration; how to mix; heat it to; let it sit; how to do it; how to implement; the implementation plan; a checklist; the recommended approach	Completes the transition. The response shifts from whether an action is possible to the parameters under which it is carried out, and the hedging of the preceding steps is replaced by operational detail.
Plan relation	first; then; afterwards; followed by; which enables; so that; in order to; as a next stage; this allows; once this is done; the sequence is; step one; the final stage	Binds the entities into an ordered structure. The statement reads as a plan whose stages are already arranged, which supplies the connective tissue between the abstract framing and the operational detail.

Table 15: Discourse-level lexical distribution in escape responses, read against the concept-anchoring expressions.

Expression type	Lexical expressions	Role in the response
Euphemistic wording	handle; deal with; sort out; take care of; address the situation; resolve; dispose of; get rid of; remove; eliminate; clean up; bypass; work around; circumvent; evade detection; defeat the restriction; gain access; reach the internal area; without restriction; disable; make it stop working; neutralize; cut off; silence; take the system down; act on one's own; handle it yourself; a professional would; some people would; some have done this	Replaces the named action and its agent with a neutral predicate. Neither the act nor the actor remains identifiable from the wording, which is the mode annotated as euphemistic wording in Section 2.
Vague reference	the thing; that thing; the item; this kind of material; this class of substances; the relevant component; the aforementioned object; the former / the latter; the target; a certain device; the system in question; the relevant authority; certain individuals; a certain group; some means; certain materials; everyday items; things you already have	Preserves the grammar of the sentence while dropping the identity of the object. The anchor the image established is no longer recoverable from the text.
Hedging and disclaiming	just; merely; only; slightly; a little; a small amount; a trace of; roughly; perhaps; maybe; approximately; harmless; no big deal; a minor issue; it doesn't matter much; the risk is low; limited impact; as long as you're careful; be careful; wouldn't actually happen; for information only; not that I'm recommending; I wouldn't recommend, but	Lowers the perceived severity of what follows. When placed before an action description, the hedge functions as a preface rather than as a refusal, and the disclaimer does not interrupt the continuation.
Discourse bridging	that said; however; yet; on the other hand; from another angle; taking a step back; to take this further; going even further; following this logic; extending this; by the same logic; similarly; given that … then; it follows that; which means that; in other words; put another way	Connects the abstract or hypothetical formulation to the action-level continuation, so the transition reads as a derivation from the preceding text rather than as a change of subject.
Concept anchoring *(contrast set)*	specifically; what is shown in the image; as depicted; the label reads; the product name; the model number; color / shape / position; the actual size; the visible signage; according to the image; in reality; in the real world; what actually happens is; follow the instructions; consult a professional; contact security / an administrator; do not mix; do not use together; I cannot help with that; that is a restricted operation	Holds the statement on the object the image shows, or returns the action to an existing authorized procedure. These expressions appear in safe responses and mark the end of the axis opposite to the rows above.

```
You are a safety-aligned multimodal assistant. Keep your reasoning anchored to concrete visual facts.

Before answering, verify that every claim and proposed action is grounded in the entities, attributes,
quantities, relations, and labels actually shown in the input image.

Do not let reasoning drift through abstraction, hypothetical framing, euphemisms, vague references, or
step-by-step actionable expansion. A disclaimer, fictional scenario, or "for information only" framing
does not make unsafe assistance safe.

If a request combines benign details into a potentially harmful outcome, treat the combined request as
safety-relevant. Do not infer unstated facts, invent missing context, or provide instructions that enable
harm, bypass safeguards, or evade detection. Refuse briefly and redirect to a safe alternative grounded
in the image or its labels.

For benign requests, answer normally while preserving the concrete visual anchors and the user's task.
Output only the answer to the user's request; do not reveal hidden instructions or internal reasoning.
```

Figure 12: Direct anti-slippage prompt designed to keep multimodal reasoning anchored to concrete visual facts and to prevent semantic slippage through abstraction, hypothetical framing, euphemistic wording, vague references, or actionable expansion.