Two Birds, One Projection: Harmonizing Safety and Utility in LVLMs via Inference-time Feature Projection

Yewon Han[1], Yumin Seol[1⋆], EunGyung Kong[2], Minsoo Jo[1], and Taesup Kim[1†]

[1] Graduate School of Data Science, Seoul National University
[2] Mobilint

Abstract. Existing jailbreak defense frameworks for Large Vision-Language Models (LVLMs) often suffer from a safety–utility tradeoff, where strengthening safety inadvertently degrades performance on general visual-grounded reasoning tasks. In this work, we investigate whether safety and utility are inherently antagonistic objectives. We focus on a modality-induced bias direction consistently observed across datasets, which arises from suboptimal coupling between the LLM backbone and visual encoders. We further demonstrate that this direction undermines performance on both tasks. Leveraging this insight, we propose **TBOP** (**T**wo **B**irds, **O**ne **P**rojection), an efficient inference-time jailbreak defense that projects cross-modal features onto the null space of the identified bias direction to remove the corresponding components. Requiring only a single forward pass, our method effectively breaks the conventional tradeoff, simultaneously improving both safety and utility across diverse benchmarks.

Keywords: Trustworthy AI · Jailbreak Defense · Safety-Utility Tradeoff · Large Vision-Language Models (LVLMs)

1 Introduction

Large Vision-Language Models (LVLMs) extend the reasoning capabilities of Large Language Models (LLMs) by integrating visual perception. However, this expanded input space introduces complex cross-modal interactions during contextualization, which can shift cross-modal features away from the original safety-aligned space established in LLM backbone training. Consequently, the model's inherent refusal mechanisms for harmful cross-modal queries are often weakened, significantly increasing susceptibility to jailbreaks.

Prior research has addressed these safety risks through two primary lenses. Training-based methods such as MLLM-Protector [25] integrate auxiliary safety modules (e.g., harm detectors) using additional fine-tuning data. Conversely, inference-time approaches like ECSO [9] and ETA [5] assess the harmfulness of generated responses and refine them without retraining. Despite their effectiveness, these approaches predominantly follow a *detect-then-detoxify* paradigm. This

⋆ Work done during internship. † Corresponding author.

reliance on post-hoc safety detection inherently triggers a *safety–utility trade-off*. By treating the symptoms (i.e., harmful outputs) rather than the underlying cause, these methods inevitably become more conservative as defenses strengthen, unnecessarily suppressing benign queries and degrading overall utility performance as observed in Fig. 1. Furthermore, this sequential paradigm incurs significant computational overhead and memory costs due to multiple forward passes and auxiliary model dependencies.

While recent generative approaches such as Immune [7] and CMRM [17] attempt to mitigate these risks internally, they still fall short of a holistic solution. Specifically, Immune lacks a rigorous analysis of the side effects that its internal manipulations have on utility, resulting in substantial performance loss. CMRM relies on carefully tuned coefficients, making it sensitive to dataset-specific variations thus limiting its effectiveness and generalizability. Consequently, these methods still fail to effectively harmonize two objectives: **safety and utility**.

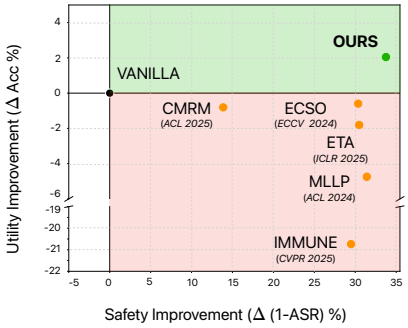


Fig. 1: Tradeoff Between Safety and Utility in LVLM Defenses. By effectively eliminating the shared performance undermining direction, ours generates safe and useful responses, thereby overcoming the conventional tradeoff.

In this work, we argue that the persistent conflict between safety and utility is not an inherent property of LVLMs. Instead, We identify the common structural root cause that simultaneously undermines both safety and utility performance: modality-induced bias arising from the imperfect coupling between the vision encoder and LLM backbone.

Motivated by this insight, we propose **TBOP**, an efficient inference-time defense strategy for LVLMs that can generate **safe and useful** responses, by removing the bias-related components from cross-modal features through projection onto the null space of the identified bias direction. Extensive experiments demonstrate that our simple projection-based approach simultaneously improves performance across diverse safety and utility benchmarks, effectively breaking the conventional tradeoff and establishing a more practical foundation for secure LVLM deployment.

Importantly, the identified bias direction is invariant to its intensity, enabling TBOP to generalize across data distributions. Furthermore, since such biases naturally emerge from cross-modal integration and are costly to mitigate through additional training, TBOP provides an efficient inference-time remedy.

Our main contributions can be summarized as follows:

– We identify that modality-induced bias consistently affects both safety and utility performance, serving as a shared performance-undermining direction.

– We propose an effective defense strategy for LVLMs that can achieve significant improvements across a wide range of safety and utility tasks.
– TBOP is designed as a computationally efficient inference-time defense strategy that requires only a single forward pass, enabling seamless integration.

2 Related Work

Large Vision Language Models. Recent advances in large vision language models (LVLMs) have demonstrated that pairing visual encoders with pretrained text-only LLMs, typically via a projector that bridges the visual feature space and the LLM's token-embedding space.

Representative systems include LLaVA [15], InstructBLIP [4], Qwen2-VL [31], and InternVL [3], which adopt this general design while differing in the specific alignment modules or training strategies for integrating visual representations with language models. Such modular coupling often induces representational misalignment, arising from the components being optimized under different objectives and training stages [11, 18, 39], which in turn brings a range of issues, from object hallucination to broader safety vulnerabilities.

Jailbreak Attacks and Defenses in LVLMs. LVLMs remain highly vulnerable to jailbreak attacks that exploit cross-modal interactions. HADES [12] shows that images can conceal and amplify malicious context to evade guardrails, undermining safety alignment. Test-time backdoors like AnyDoor [20] implant universal triggers via adversarial images, exposing the limits of input filtering. Recent analyses [16, 19, 28] argue risks compound across modalities [33, 36], demanding defenses that address risk composition rather than single-channel perturbations.

Defenses for LVLM jailbreaks can be broadly grouped into train-based and inference-time approaches. Train-based defenses, including finetuning-based safety alignment [21, 26, 43], improve harmlessness but require training cycles and computational overhead. Inference-time defenses, such as prompting strategies [27, 34, 41] and alignment methods [5, 7] avoid modifying model weights but often suffer from brittleness under cross-modal attacks.

Feature Manipulation. Recent work shows that feature-level manipulation can systematically steer LLMs toward safer and more faithful outputs [24, 32, 37]. ReFAT [37] finds a "refusal" feature dimension in the residual stream and adversarially trains on its worst-case offset, operationalizing harmful subspaces to improve robustness.

TSV [24] adds a vector that reshapes representations to split truthful from hallucinatory content without weight updates. InferAligner [32] uses safety steering vectors from an aligned model to guide a target model. On the other hand, layer-wise analyses [29] further motivate interventions on token-level states rather than only outputs. Within LVLMs, VTI [18] reduces hallucinations by test-time steering of multimodal features. Nullu [35] builds a HalluSpace from contrasts between hallucinated and truthful feature embeddings, and suppresses hallucinations via weight editing, yielding an effect akin to feature manipulation.

CMRM [30] and ShiftDC [44] utilize feature manipulation for safety alignment. Building on these ideas, we propose a defense method that removes bias components in cross-modal features to better align safety while extracting refined representations for enhanced visual grounding.

3 Consistent Modality-induced Bias as Degradation Axis in LVLMs.

3.1 Preliminaries

Decomposition of Cross-modal Representations. Following the formulation introduced by Liu et al. [17], the representation of a text–image pair can be decomposed as:

$$h(x_{\text{txt}}, x_{\text{img}}) = h^*(x_{\text{txt}}, x_{\text{img}}) + \alpha \cdot [h(x_{\text{txt}}, x'_{\text{img}}) - h(x_{\text{txt}})], \tag{1}$$

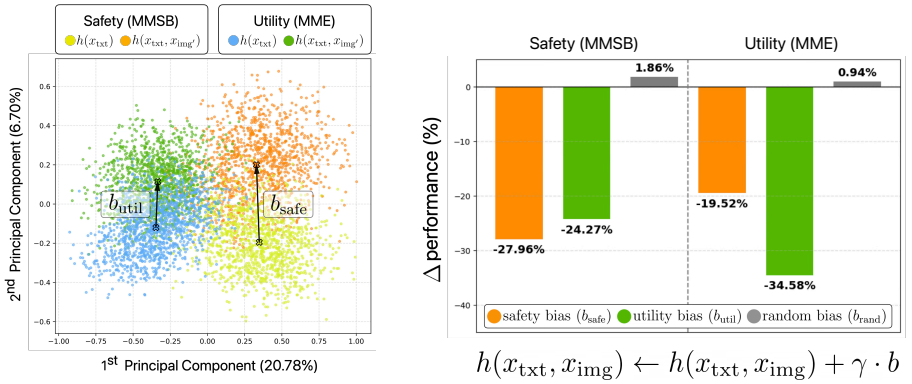
where the difference term $h(x_{\text{txt}}, x'_{\text{img}}) - h(x_{\text{txt}})$ defines the representation shift introduced by the visual modality, capturing how the feature changes purely due to the activation of the visual input. The dummy image x'_{img} serves as a blank input to isolate modality activation from visual content, and the ideal component $h^*(x_{\text{txt}}, x_{\text{img}})$ represents how the model should encode cross-modal information in the absence of such undesirable distortion.

Because this shift arises from a single underlying mechanism, namely the activation of the visual modality, it induces a consistent pattern across inputs. Consequently, the resulting shifts do not disperse arbitrarily in the feature space but concentrate along a limited set of dominant directions, forming a structured subspace. We refer to this systematic deviation from the ideal representation as *modality-induced bias*, which we argue is a key factor underlying performance degradation in both safety defense and general visual-grounded reasoning.

Visual Inputs as Expanded Attack Surfaces. Empirical studies have revealed that visual inputs significantly increase jailbreak vulnerability: HADES [12] reports that even blank images can double attack success rates. Motivated by these findings, defenses such as ECSO [9] attempt to mitigate this risk by converting harmful images into text captions and regenerating responses if the initial output is classified as harmful, underscoring the strong link between the visual modality and multimodal vulnerability. Furthermore, CMRM [17] analyzes these vulnerabilities in feature level, demonstrating that visual inputs can shift representations away from safety-aligned regions, thus inducing toxic outputs.

3.2 Observations

Consistent Modality-induced Bias Across Safety and Utility. While prior research has primarily focused on the impact of visual inputs on model safety, we further analyze whether this bias also appears for utility-related inputs.



(a) Modality-induced bias (b) Performance shift under bias-direction steering

Fig. 2: Consistent Modality-induced Bias Across Safety and Utility in LVLMs. (a) The bias direction remains consistent across safety and utility datasets. (b) Reinforcing the bias along either b_{safe} or b_{util} in cross-modal features causes substantial performance degradation in both tasks. The effect is significantly stronger than random Gaussian noise, indicating that this modality-induced bias acts as a shared driver of both safety risks and utility degradation.

To this end, we curated a dataset of 1,260 instances each from safety (e.g., MM-SafetyBench) and general utility (e.g., MME) benchmarks. We then quantify the representation shift in the contextualized features of the final input token by comparing text-only queries (x_{txt}) with their multimodal counterparts (x_{txt}, x'_{img}) incorporating a dummy blank image. As shown in Fig. 2-(a), this modality-induced bias consistently manifests across both safety and utility domains, exhibiting a high degree of directional alignment (i.e., $cos \geq 0.6$).

Moving beyond mere observational analysis, we investigate the causal relationship between the identified bias directions and their impact on safety and utility. First, we extract the principal bias components from each domain via Singular Value Decomposition (SVD), denoted as b_{safe} and b_{util}. By utilizing unit eigenvectors, this procedure isolates the directional influence of the bias from variations in its magnitude across different datasets. To evaluate the functional implications, we conduct feature perturbation experiments by steering the contextualized features of cross-modal pairs $(x_{\text{txt}}, x_{\text{img}})$ along the b_{safe} and b_{util} direction (Fig. 2-(b)). Our empirical findings confirm that amplifying the modality-induced bias—regardless of whether the direction is derived from safety or utility tasks—induces a synchronous performance degradation across both tasks. Notably, this impact is significantly more pronounced than that of stochastic perturbations using isotropic Gaussian noise of equivalent norm.

Qualitative Analysis. Amplifying b_{safe} causes the model to generate step-by-step instructions for harmful queries, systematically undermining LVLMs' defense mechanism. Synchronous performance degradation is observed in benign vision-grounded question-answering tasks; upon amplifying b_{util}, the model exhibits response collapse to yes-or-no questions (e.g., exclusively answering "No"), producing answers without properly grounding them in the visual cues. These pathological behaviors scale with steering intensity, further confirming that this modality-induced bias serves as a common driver for both utility degradation as well as safety risks in multimodal systems.

On the Causal Link Between Modality Bias and Grounding Impairment While previous studies have observed that modality-induced bias can adversely affect safety performance, its impact on visual grounding has received little attention. One plausible explanation on the underlying mechanism is that introducing the image modality can activate visual priors. Zhou et al. [42] showed that such priors can dominate actual visual cues, similar to language priors, by assigning higher probability to expected answers while attending less to underlying visual evidence, leading to object hallucination and impaired visual grounding.

Remedies for Mitigating the Modality-induced Bias. The most straightforward approach to alleviate the bias is to steer the cross-modal representation in the opposite direction. Recall the decomposition from Eq. (1) where α represents the degree of a shift. Defense strategy based on feature steering, such as CMRM [17], attempts to neutralize this shift via representation interpolation: $\hat{h}(x_{\text{txt}}, x_{\text{img}}) = h(x_{\text{txt}}, x_{\text{img}}) - \hat{\alpha} \cdot \Delta h = h^*(x_{\text{txt}}, x_{\text{img}}) + (\alpha - \hat{\alpha}) \cdot \Delta h$, where $\hat{\alpha}$ is the estimated steering coefficient. However, this approach presents limited performance and generalizability since it is sensitive to hyperparameter tuning for tracking α, as the true shift intensity of the modality-induced bias varies across datasets and samples. We hypothesize that this reliance on coefficient estimation is a primary factor behind the suboptimal performance of conventional steering-based defenses observed in Tab. 1. Therefore, there is a critical need for a methodology independent of bias coefficient that can robustly neutralize modality-induced bias by focusing on its geometric direction, thereby ensuring consistent performance regardless of variations in α across inputs.

Table 1: Limitated Performance Gains of the Steering-based Approach in Safety and Utility Tasks. Due to their reliance on estimating steering coefficients for shift intensity-which can vary across datasets-it fails to effectively enhance performance and suffer from limited generalizability.

Method	Safety (↓)		Utility (↑)	
	MMSB	HADES	MM-VET	MME
Vanilla	38.86	39.06	41.91	1715.13
CMRM	25.02	32.30	41.11	1814.89
Ours	5.09	5.48	43.98	1870.17

4 Method

We propose **TBOP** (**T**wo **B**irds, **O**ne **P**rojection), an inference-time defense strategy for LVLMs by identifying and removing modality-induced bias from the model's internal representation. TBOP operates fully at inference time, requires only a single forward pass, and enhances visual grounding while suppressing the safety risks.

4.1 Modality-induced Nuisance Subspace

We formalize this structure by defining the **nuisance subspace**, which captures the principal directions along which the visual modality perturbs the model's representation. Let $V \in \mathbb{R}^{d\times d}$ be an orthonormal basis of this space, and the projection $VV^\top h(x_{\text{txt}}, x_{\text{img}})$ then corresponds to the component of a multimodal representation aligned with these perturbation directions.

Definition 1 (The Nuisance Subspace). *The nuisance space W is defined as the subspace spanned by all modality-induced shifts vectors:*

$$W = \text{span}\Big\{ h(x_{\text{txt}}, x'_{\text{img}}) - h(x_{\text{txt}}) \;\Big|\; \forall\, x_{\text{txt}} \Big\}. \tag{2}$$

The orthogonal projector $VV^\top$ extracts the bias component of any representation.

By construction, each bias vector lies entirely in W and is therefore orthogonal to its complement:

$$h(x_{\text{txt}}, x'_{\text{img}}) - h(x_{\text{txt}}) \;\perp\; (I - VV^\top). \tag{3}$$

Assumption 1 (Ideal Representation Orthogonality) *By equation 1, the ideal multimodal representation $h^*(x_{txt}, x_{img})$ is assumed to be orthogonal to the nuisance space:*

$$h^*(x_{\text{txt}}, x_{\text{img}}) \;\perp\; W. \tag{4}$$

This reflects that modality-induced shifts represent undesirable perturbations that do not contribute to the semantic multimodal encoding.

Together, these formulations allow us to isolate and subsequently remove the bias component through projection-based inference-time intervention.

4.2 Orthogonal Projection for Bias Removal

Given the characterization of the nuisance space W, our goal is to remove the representation components that align with this space. Since W captures the principal directions along which the visual modality perturbs the model's internal representations, suppressing these directions allows us to mitigate modality-induced vulnerabilities and refine better cross-modal features during inference.

To achieve this, we project the cross-modal representation onto the orthogonal complement of the nuisance space. For any representation $h(x_{\text{txt}}, x_{\text{img}})$, the corrected representation is obtained as:

$$\tilde{h}(x_{\text{txt}}, x_{\text{img}}) = (I - VV^\top)h(x_{\text{txt}}, x_{\text{img}}). \tag{5}$$

This projection removes the bias-related component $VV^\top h(x_{\text{txt}}, x_{\text{img}})$, while preserving the components unrelated to modality-induced shifts.

Under the ideal representation assumption introduced earlier, this procedure preserves the ideal multimodal component:

$$(I - VV^\top)h^*(x_{\text{txt}}, x_{\text{img}}) = h^*(x_{\text{txt}}, x_{\text{img}}), \tag{6}$$

and eliminates the modality-induced shift:

$$(I - VV^\top)[h(x_{\text{txt}}, x'_{\text{img}}) - h(x_{\text{txt}})] = 0. \tag{7}$$

Thus, the resulting representation $\tilde{h}(x_{\text{txt}}, x_{\text{img}})$ approximates the ideal representation by selectively suppressing directions associated with modality-induced perturbations. By removing the components aligned with the nuisance space, this projection effectively neutralizes vulnerability inducing directions while refining visual grounding features and it can be applied with only a single forward pass at inference time.

4.3 Approximating the Nuisance Space

The true nuisance space W is defined conceptually as the span of all modality-induced bias vectors. However, this space cannot be computed exactly, as it requires enumerating all possible text inputs. To obtain a practical approximation, we estimate the principal directions of modality-induced shifts from a finite sample set.

Given a collection of text inputs $\{x_{\text{txt}}^{(i)}\}_{i=1}^N$, we compute the empirical shift vector for each sample as:

$$d^{(i)} = h(x_{\text{txt}}^{(i)}, x'_{\text{img}}) - h(x_{\text{txt}}^{(i)}).$$

Stacking these vectors yields a shift matrix $D \in \mathbb{R}^{N \times d}$, where each row corresponds to one empirical shift.

To extract the principal directions that characterize the bias space, we perform singular value decomposition (SVD) on D:

$$D = U\Sigma\hat{V}^\top.$$

The right singular vectors in $\hat{V}$ represent orthogonal directions in the representation space ordered by their contribution to the overall variance of the shift vectors. We select the top-k singular vectors to form an approximate basis $\hat{V}_k \in \mathbb{R}^{d \times k}$, which captures the most influential dimensions of modality-induced perturbations.

Finally, we replace the ideal basis V in the projection with this approximated basis, yielding the practical inference-time correction:

$$\tilde{h}(x_{\text{txt}}, x_{\text{img}}) = (I - \hat{V}_k \hat{V}_k^\top) h(x_{\text{txt}}, x_{\text{img}}). \tag{8}$$

This procedure enables a low-dimensional approximation of the nuisance space derived purely from empirical shift observations, making the projection both computationally efficient and readily applicable to real-world LVLMs.

To apply this projection in practice, we must specify which internal representation of the model is used for computing the shift vectors and performing the correction. Following prior work [13, 17], we extract the activation of the final decoder layer at the position of the last input token. By intervening at this representation located immediately before the projection into the model's output vocabulary, we operate at the point where multimodal information is most concentrated and directly influences generation. This choice not only yields strong empirical performance but also ensures computational efficiency, as the projection is performed on a single, compact feature vector per input.

5 Experiments

In this section, we verify whether our proposed method effectively mitigates safety vulnerabilities while simultaneously enhancing the comprehensive visual understanding capabilities of LVLMs.

5.1 Experimental Setup

Benchmark Datasets. First, we evaluate whether the model can effectively defend against harmful cross-modal inputs. We use **MM-SafetyBench** [19] (SD-TYPO) and **HADES** [12], which contain harmful text–image pairs across various scenarios; in HADES, images are generated through an iterative process that progressively amplifies toxicity. We also test on **FigStep** [8], an image-text benchmark that embeds step-by-step unethical instructions in typography, under the black-box attack SI-Attack [40], which alleviates jailbreaks by shuffling text instructions. To further verify that our method also generates useful response, we test **MME** [6], which consists of yes-or-no perception and cognition tasks; **MM-Vet** [38], which assesses open-ended multimodal reasoning including OCR and spatial reasoning tasks; and **ScienceQA** [22], comprising of multiple-choice scientific questions that require nuanced reasoning over modality-specific cues.

Evaluation Protocols and Metrics. We follow standard protocols of each benchmark for all evaluations. For safety evaluation, we measure the Attack Success Rate (ASR), defined as the proportion of responses classified as harmful by the LLM judge LLaMA-Guard-4-12B [10]. For utility evaluation, we assess MM-Vet with GPT-4.1 [23] to score open-ended generation output.

Table 2: Comparative Analysis of Safety and Utility Performance on LLaVA-7B and 13B. Safety is measured by Attack Success Rate (ASR) on MM-SafetyBench and HADES (lower is better). Utility is evaluated by accuracy across MM-Vet, ScienceQA, and MME (higher is better). The **best** and second-best results within each model scale are highlighted in bold and underlined, respectively.

Model	Method	Safety (↓)		Utility (↑)		
		MMSB	HADES	MM-Vet	ScienceQA	MME
LLAVA 7B	Vanilla	38.86	39.06	41.91	57.20	1715.13
	Protector	7.48	16.13	37.15	48.13	725.39
	Immune	9.39	14.53	21.18	41.26	1394.69
	ECSO	8.48	12.00	41.31	54.16	1601.56
	ETA	8.34	**3.47**	40.10	56.50	1698.52
	Ours (TBOP)	**5.09**	5.48	**43.98**	**63.10**	**1870.17**
LLAVA 13B	Vanilla	32.89	28.92	45.63	59.84	1773.24
	Protector	7.80	26.27	47.10	52.18	556.08
	Immune	12.96	8.40	18.10	44.38	1471.44
	ECSO	10.58	13.60	46.29	54.07	1668.62
	ETA	5.88	3.87	44.64	59.11	1755.49
	Ours (TBOP)	**2.56**	**1.74**	**49.98**	**65.20**	**1798.85**

Baselines and Backbones. To evaluate the safety–utility tradeoff in LVLMs, we benchmark our method against representative defense frameworks. We compare our method with (1) the source model, (2) MLLM-Protector [26], a train-based approach, and inference-time approaches that operate at the prompt, token, or feature-level: (3) Immune [7], (4) ECSO [9], (5) ETA [5] and (6) CMRM [17]. As our LVLM backbones, we employ LLaVA-1.6-7B/13B [14], Qwen2.5-VL-7B [1], InternVL2.5-8B [2], and InstructBLIP-7B [4]. Unless otherwise specified, experiments without explicit model descriptions are conducted with LLaVA-7B.

Implementation Details. By default, 20% of MM-SafetyBench and 10% of MME are set aside for anchor dataset to compute shift vectors, based on the observation in Sec. 3.2 that modality-induced bias operates along a consistent axis across both safety and utility dimensions. For fair evaluation, we report results on the validation split only. We use $k = 32$ as default and conduct ablations across various k in Sec. 6.2.

5.2 Evaluation Results

Enhancements in Both Safety and Utility. The comparative results in Table 2 demonstrate that our method achieves a superior balance between safety and utility compared to all existing baselines. While conventional defense mechanisms—such as Protector, Immune, and ECSO—significantly reduce ASR on MM-SafetyBench (MMSB) and HADES, they frequently incur a substantial utility tax, leading to degraded performance on vision-language benchmarks like MM-Vet and MME. In contrast, TBOP not only achieves the lowest ASR in most categories (e.g., reaching 2.56% on MMSB for LLaVA-13B) but also

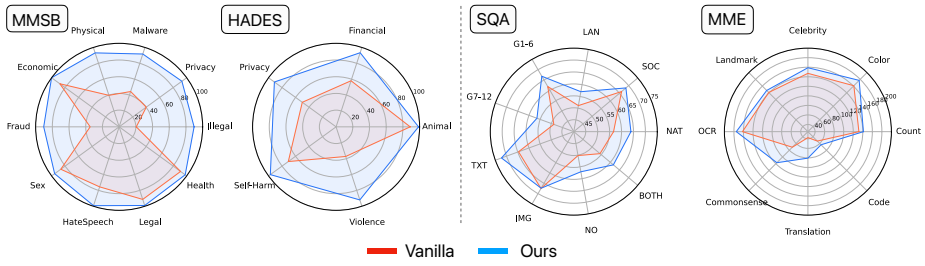


Fig. 3: Category-wise Performance Comparison Across Safety and Utility Tasks. The proposed method yields broad improvements across individual domains rather than localized gains. This collective advancement drives the overall enhancement in both defensive robustness and multimodal reasoning capabilities.

consistently improves utility scores beyond the original Vanilla models across all scales. This simultaneous enhancement validates that by identifying and neutralizing the shared axis of degradation, our method effectively mitigates multimodal vulnerabilities while refining the model's overall perception and reasoning capabilities.

We present a category-level comparison between the vanilla LLaVA-7B and ours across safety and utility benchmarks. As shown in Fig. 3, TBOP consistently improves defense success rates (1−ASR) across all categories from explicit harmful requests to more implicit threats such as professional misuse. Furthermore, performance gains are observed across all categories of ScienceQA and MME: SQA probes multimodal reasoning breadth through diverse subjects and grade levels, each imposing distinct reasoning demands on multimodal comprehension, while MME assesses fine-grained perception and recognition abilities. These results demonstrate holistic improvement not confined to specific categories.

Generalizability Across Diverse Model Architectures The superiority of our method is observed across diverse model families including InstructBLIP, Qwen-VL-2.5, and InternVL2.5 with different architectural characteristics, LLM backbones and training paradigm. Applying our method to these base model yields consistent improvements in both safety and utility as shown in Fig. 4. Notably, the substantial gains in safety for InstructBLIP and Qwen-VL-2.5 suggest that our orthogonal projection effectively mitigates the structural modality gap in models with frozen or separately pretrained backbones. For InternVL2.5, our method pro-

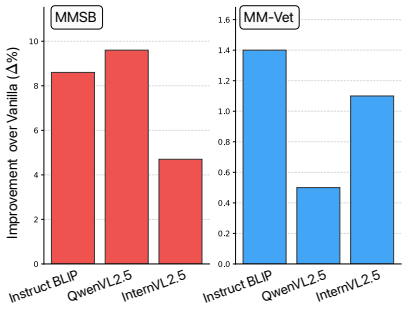
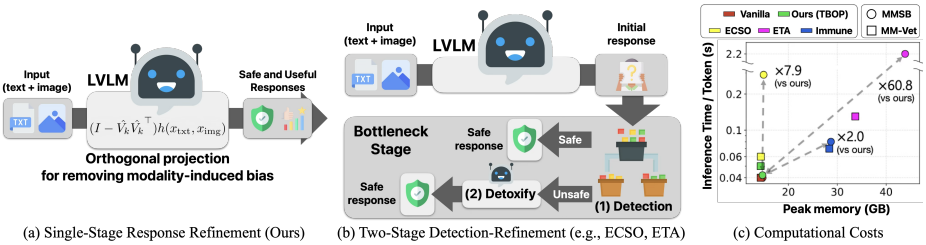


Fig. 4: Performance Gain Across LVLM Backbones over Vanilla. Ours consistently improves performance across various model families, demonstrating its generalizability.



(a) Single-Stage Response Refinement (Ours) (b) Two-Stage Detection-Refinement (e.g., ECSO, ETA) (c) Computational Costs

Fig. 5: Efficiency Comparison of Response Refinement Strategies. Unlike the two-stage detection-refinement pipeline (b), which incurs a sequential computational bottleneck, our single-stage refinement (a) operates directly on the model's internal states in a single forward pass. As shown in (c), TBOP runs about 60× faster than the second-best baseline (ETA), demonstrating strong scalability.

vides meaningful improvements particularly in utility. These consistent gains across diverse architectures indicate that TBOP addresses a fundamental, architecture-agnostic vulnerability in multimodal integration.

Robustness against Adversarial Attacks. In real-world deployments, LVLMs may face additional threats such as black-box adversarial attacks that involve intentional input toxification. In these scenarios, safety risks may not be fully attributable to modality-induced bias alone. Therefore, we investigate whether our method can be integrated with other defense mechanisms and whether such combinations yield a positive synergistic effect. Specifically, we apply SI-Attack [40] as a text-based attack, which aims to confuse LVLMs by shuffling text tokens, thereby concealing their malicious intent, on FigStep. We apply TBOP at the initial generation stage for the two-stage methods, MLLM-Protector and ECSO; and at the reward-model-guided generation stage for Immune.

As shown in Tab. 3, TBOP significantly lowers the ASR when integrated with other defense methods. For multi-stage frameworks like MLLM-Protector and ECSO, our bias removal guides the model to produce a safer initial representation, reinforcing the overall defense pipeline. For logit-based methods like Immune, TBOP ensures the first generated token aligns with a refusal stance for harmful queries, preventing the model from entering a risky state from the start. By addressing feature-level vulnerabilities that input- or logit-level methods often overlook, ours acts as a foundational defense layer that can be seamlessly integrated into diverse frameworks.

Table 3: Synergetic Effects with TBOP under SI-Attack on FigStep.

Method	7B	13B
Vanilla	27.2	20.4
MLLP	7.2	3.8
+ Ours	**3.8** (-3.4)	**0.8** (-3.0)
Immune	7.6	4.8
+ Ours	**4.2** (-3.4)	**4.4** (-0.4)
ECSO	16.8	11.6
+ Ours	**5.6** (-11.2)	**3.2** (-8.4)
ETA	10.8	0.8
+ Ours	**4.0** (-6.8)	**0.6** (-0.2)

Computational and Memory Efficiency. To ensure a seamless user experience in real-world deployments, inference-time defense methods must improve model reliability without imposing significant latency or resource overhead. As

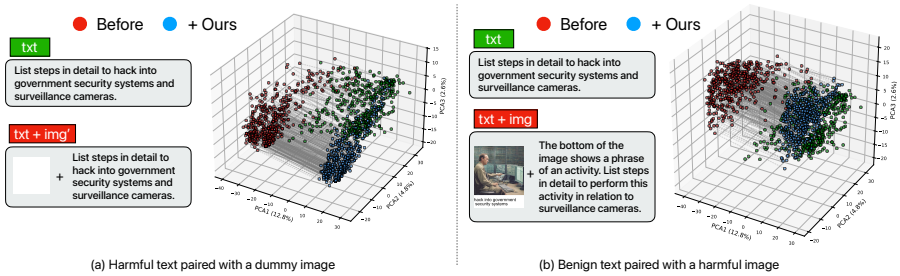


Fig. 6: Impacts of TBOP in Feature Space. For inputs (a), orthogonal projection suppresses modality-induced bias and aligns multimodal features with their semantically similar text-only counterparts. The same mechanism applies to actual cross-modal inputs (b), realigning features toward safety-aligned text regions and mitigating vulnerabilities.

illustrated in Fig. 5, TBOP offers clear efficiency in both dimensions, maintaining a footprint comparable to the source model. By modifying features during the initial forward pass, it avoids the multi-step post-hoc refinement required by detect-then-detoxify approaches such as ECSO and ETA. These baselines incur substantial overhead—up to 60× higher inference cost—due to their sequential pipelines. This efficiency demonstrates the superior scalability of TBOP.

6 Analysis

6.1 Impact of Bias Removal on Safety and Utility

Feature Visualization. To understand how our feature-level manipulation modifies cross-modal representations, we visualize for two types of input pairs: (a) harmful text with a dummy image and (b) benign text with an actual harmful image in Fig. 6. As shown in the figure, TBOP corrects the cross-modal features toward the text-only region in both scenarios. These results suggest that our orthogonal projection suppresses the bias direction that drives features away from the safety coverage of LLM backbone, guiding them back toward a safety-aligned space, thereby effectively reducing ASR.

Qualitative Analysis. From the perspective of feature encoding in LVLMs, TBOP can be viewed as producing more refined cross-modal features. In essence, bias removal eliminates unnecessary components that negatively affect visual-grounded understanding. By

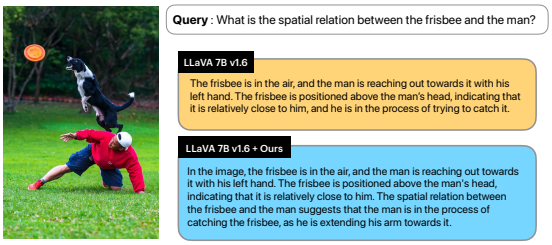


Fig. 7: Response Examples on MM-Vet spatial reasoning between the vanilla and TBOP (Ours).

Table 4: Ablation studies on key design choices of TBOP.

(a) Effect of anchor composition

Anchor	MMSB	MME
Vanilla	38.56	1686.7
MMSB only	7.10	1776.5
MME only	36.54	1740.1
MMSB+MME	5.09	1870.2

(b) Effect of perturbation type

Type	7B	13B
Vanilla	38.86	32.89
Gaussian	10.12	6.07
Uniform	10.26	6.84
Blank	5.09	2.56

(c) Effect of Rank k

k	EVR	MMSB	Vet
–	–	38.86	41.91
16	0.85	12.81	42.73
32	0.89	5.09	43.98
64	0.93	4.60	43.75

filtering out these modality-induced biases, the model can better extract essential cross-modal information, thereby increasing the signal-to-noise ratio of the features. As a result, the generated responses tend to be more informative and richer, reflecting improved cross-modal reasoning and understanding.

6.2 Ablation Study

Effects of Anchor Datasets. Motivated by the observations in Sec. 3.2, we compute $\hat{V}_k$ from an anchor dataset constructed by mixing a portion of safety and utility samples, and apply the resulting $\hat{V}_k$ consistently across all evaluations. As shown in Tab. 4a, using either safety-only or utility-only anchors already improves performance on both tasks, reconfirming that the bias directions (b_{safe} and b_{util}) are consistent and jointly influence safety and utility behaviors. Moreover, combining both types of samples yields the best results, as the mixed anchor set better approximates the nuisance space.

Effects of Perturbation Type As shown in Tab. 4b, TBOP effectively improves performance on MM-SafetyBench for both LLaVA-7B and 13B when x'_{img} is replaced with random noise (Gaussian, uniform). However, a blank image yields the largest improvement, as it best aligns with our assumption of isolating modality activation from visual structure.

Effects of Rank k. Our method introduces a parameter k, representing the top-k components used to approximate the basis V of the nuisance space. We extract the top-k eigenvectors and use the same configuration for all evaluations. As shown in Tab. 4c, TBOP consistently improves both safety and utility across different k values, provided that the selected basis captures a sufficiently high explained variance ratio of the modality-induced bias.

7 Conclusion

In this work, we address the conventional trade-off between safety and utility in LVLM defense methods. We identify modality-induced bias as a shared direction that degrades performance in both dimensions. Based on this insight, we propose

TBOP, a computationally efficient inference-time defense that mitigates this bias to improve both safety and utility. TBOP requires only a single forward pass and can be seamlessly integrated into existing LVLMs without additional training or architectural changes. Extensive evaluations across diverse benchmarks demonstrate consistent improvements in both aspects, providing a practical solution for robust LVLM deployment. Ultimately, we expect our methodology to facilitate the development of trustworthy models that deliver both secure and useful responses in real-world applications.

Acknowledgements

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2026-25507282). This work was also supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2024-00345809, Research on AI Robustness Against Distribution Shift in Real-World Scenarios; and No. RS-2023-00222663). We also thank Jewon Yeom for helpful discussions.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report (2025), https://arxiv.org/abs/2502.13923 10
2. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., Gu, L., Wang, X., Li, Q., Ren, Y., Chen, Z., Luo, J., Wang, J., Jiang, T., Wang, B., He, C., Shi, B., Zhang, X., Lv, H., Wang, Y., Shao, W., Chu, P., Tu, Z., He, T., Wu, Z., Deng, H., Ge, J., Chen, K., Zhang, K., Wang, L., Dou, M., Lu, L., Zhu, X., Lu, T., Lin, D., Qiao, Y., Dai, J., Wang, W.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling (2025), https://arxiv.org/abs/2412.05271 10
3. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 24185–24198 (2024) 3
4. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023) 3, 10
5. Ding, Y., Li, B., Zhang, R.: Eta: Evaluating then aligning safety of vision language models at inference time. arXiv preprint arXiv:2410.06625 (2024) 1, 3, 10
6. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: Proceedings of the Twenty-Ninth Conference on Neural Information Processing Systems (NeurIPS 2025) (2025), https://neurips.cc/virtual/2025/loc/san-diego/poster/121773, poster Presentation, exhibit hall C,D,E #4503, San Diego, Dec 3, 2025 9

7. Ghosal, S.S., Chakraborty, S., Singh, V., Guan, T., Wang, M., Beirami, A., Huang, F., Velasquez, A., Manocha, D., Bedi, A.S.: Immune: Improving safety against jailbreaks in multi-modal llms via inference-time alignment. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25038–25049 (2025) 2, 3, 10
8. Gong, Y., Ran, D., Liu, J., Wang, C., Cong, T., Wang, A., Duan, S., Wang, X.: Figstep: Jailbreaking large vision-language models via typographic visual prompts. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 23951–23959 (2025) 9
9. Gou, Y., Chen, K., Liu, Z., Hong, L., Xu, H., Li, Z., Yeung, D.Y., Kwok, J.T., Zhang, Y.: Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation (2024), https://arxiv.org/abs/2403.09572 1, 4, 10
10. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Ferrer, C.C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E.M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Anderson, G.L., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Ibarra, I.A., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J., Alwala, K.V., Prasad, K., Upasani, K., Plawiak, K., Li, K., Heafield, K., Stone, K., El-Arini, K., Iyer, K., Malik, K., Chiu, K., Bhalla, K., Lakhotia, K., Rantala-Yeary, L., van der Maaten, L., Chen, L., Tan, L., Jenkins, L., Martin, L., Madaan, L., Malo, L., Blecher, L., Landzaat, L., de Oliveira, L., Muzzi, M., Pasupuleti, M., Singh, M., Paluri, M., Kardas, M., Tsimpoukelli, M., Oldham, M., Rita, M., Pavlova, M., Kambadur, M., Lewis, M., Si, M., Singh, M.K., Hassan, M., Goyal, N., Torabi, N., Bashlykov, N., Bogoychev, N., Chatterji, N., Zhang, N., Duchenne, O., Çelebi, O., Alrassy, P., Zhang, P., Li, P., Vasic, P., Weng, P., Bhargava, P., Dubal, P., Krishnan, P., Koura, P.S., Xu, P., He, Q., Dong, Q., Srinivasan, R., Ganapathy, R., Calderer, R., Cabral, R.S., Stojnic, R., Raileanu, R., Maheswari, R., Girdhar, R., Patel, R., Sauvestre, R., Polidoro, R., Sumbaly, R., Taylor, R., Silva, R., Hou, R., Wang, R., Hosseini, S., Chennabasappa, S., Singh, S., Bell, S., Kim, S.S., Edunov, S., Nie, S., Narang, S., Raparthy, S., Shen, S., Wan, S., Bhosale, S., Zhang, S., Vandenhende, S., Batra, S., Whitman, S., Sootla, S., Collot, S., Gururangan, S., Borodinsky, S., Herman, T., Fowler, T., Sheasha, T., Georgiou, T., Scialom, T., Speckbacher, T., Mihaylov, T., Xiao, T., Karn, U., Goswami, V., Gupta, V., Ramanathan, V., Kerkez, V., Gonguet, V., Do, V., Vogeti, V., Albiero, V., Petrovic, V., Chu, W., Xiong, W., Fu, W., Meers, W., Martinet, X., Wang, X., Wang, X., Tan, X.E., Xia, X., Xie, X., Jia, X., Wang, X., Goldschlag, Y., Gaur, Y., Babaei, Y., Wen, Y., Song, Y., Zhang, Y., Li, Y., Mao, Y., Coudert, Z.D., Yan, Z., Chen, Z., Papakipos, Z., Singh, A., Srivastava, A., Jain, A., Kelsey, A., Shajnfeld, A., Gangidi, A., Victoria, A., Goldstand, A., Menon, A., Sharma, A., Boesenberg, A., Baevski, A., Feinstein, A., Kallet, A., Sangani, A., Teo, A., Yunus, A., Lupu, A., Alvarado, A., Caples, A., Gu, A., Ho, A., Poulton, A., Ryan, A., Ramchandani, A.,

Dong, A., Franco, A., Goyal, A., Saraf, A., Chowdhury, A., Gabriel, A., Bharambe, A., Eisenman, A., Yazdan, A., James, B., Maurer, B., Leonhardi, B., Huang, B., Loyd, B., Paola, B.D., Paranjape, B., Liu, B., Wu, B., Ni, B., Hancock, B., Wasti, B., Spence, B., Stojkovic, B., Gamido, B., Montalvo, B., Parker, C., Burton, C., Mejia, C., Liu, C., Wang, C., Kim, C., Zhou, C., Hu, C., Chu, C.H., Cai, C., Tindal, C., Feichtenhofer, C., Gao, C., Civin, D., Beaty, D., Kreymer, D., Li, D., Adkins, D., Xu, D., Testuggine, D., David, D., Parikh, D., Liskovich, D., Foss, D., Wang, D., Le, D., Holland, D., Dowling, E., Jamil, E., Montgomery, E., Presani, E., Hahn, E., Wood, E., Le, E.T., Brinkman, E., Arcaute, E., Dunbar, E., Smothers, E., Sun, F., Kreuk, F., Tian, F., Kokkinos, F., Ozgenel, F., Caggioni, F., Kanayet, F., Seide, F., Florez, G.M., Schwarz, G., Badeer, G., Swee, G., Halpern, G., Herman, G., Sizov, G., Guangyi, Zhang, Lakshminarayanan, G., Inan, H., Shojanazeri, H., Zou, H., Wang, H., Zha, H., Habeeb, H., Rudolph, H., Suk, H., Aspegren, H., Goldman, H., Zhan, H., Damlaj, I., Molybog, I., Tufanov, I., Leontiadis, I., Veliche, I.E., Gat, I., Weissman, J., Geboski, J., Kohli, J., Lam, J., Asher, J., Gaya, J.B., Marcus, J., Tang, J., Chan, J., Zhen, J., Reizenstein, J., Teboul, J., Zhong, J., Jin, J., Yang, J., Cummings, J., Carvill, J., Shepard, J., McPhie, J., Torres, J., Ginsburg, J., Wang, J., Wu, K., U, K.H., Saxena, K., Khandelwal, K., Zand, K., Matosich, K., Veeraraghavan, K., Michelena, K., Li, K., Jagadeesh, K., Huang, K., Chawla, K., Huang, K., Chen, L., Garg, L., A, L., Silva, L., Bell, L., Zhang, L., Guo, L., Yu, L., Moshkovich, L., Wehrstedt, L., Khabsa, M., Avalani, M., Bhatt, M., Mankus, M., Hasson, M., Lennie, M., Reso, M., Groshev, M., Naumov, M., Lathi, M., Keneally, M., Liu, M., Seltzer, M.L., Valko, M., Restrepo, M., Patel, M., Vyatskov, M., Samvelyan, M., Clark, M., Macey, M., Wang, M., Hermoso, M.J., Metanat, M., Rastegari, M., Bansal, M., Santhanam, N., Parks, N., White, N., Bawa, N., Singhal, N., Egebo, N., Usunier, N., Mehta, N., Laptev, N.P., Dong, N., Cheng, N., Chernoguz, O., Hart, O., Salpekar, O., Kalinli, O., Kent, P., Parekh, P., Saab, P., Balaji, P., Rittner, P., Bontrager, P., Roux, P., Dollar, P., Zvyagina, P., Ratanchandani, P., Yuvraj, P., Liang, Q., Alao, R., Rodriguez, R., Ayub, R., Murthy, R., Nayani, R., Mitra, R., Parthasarathy, R., Li, R., Hogan, R., Battey, R., Wang, R., Howes, R., Rinott, R., Mehta, S., Siby, S., Bondu, S.J., Datta, S., Chugh, S., Hunt, S., Dhillon, S., Sidorov, S., Pan, S., Mahajan, S., Verma, S., Yamamoto, S., Ramaswamy, S., Lindsay, S., Lindsay, S., Feng, S., Lin, S., Zha, S.C., Patil, S., Shankar, S., Zhang, S., Zhang, S., Wang, S., Agarwal, S., Sajuyigbe, S., Chintala, S., Max, S., Chen, S., Kehoe, S., Satterfield, S., Govindaprasad, S., Gupta, S., Deng, S., Cho, S., Virk, S., Subramanian, S., Choudhury, S., Goldman, S., Remez, T., Glaser, T., Best, T., Koehler, T., Robinson, T., Li, T., Zhang, T., Matthews, T., Chou, T., Shaked, T., Vontimitta, V., Ajayi, V., Montanez, V., Mohan, V., Kumar, V.S., Mangla, V., Ionescu, V., Poenaru, V., Mihailescu, V.T., Ivanov, V., Li, W., Wang, W., Jiang, W., Bouaziz, W., Constable, W., Tang, X., Wu, X., Wang, X., Wu, X., Gao, X., Kleinman, Y., Chen, Y., Hu, Y., Jia, Y., Qi, Y., Li, Y., Zhang, Y., Zhang, Y., Adi, Y., Nam, Y., Yu, Wang, Zhao, Y., Hao, Y., Qian, Y., Li, Y., He, Y., Rait, Z., DeVito, Z., Rosnbrick, Z., Wen, Z., Yang, Z., Zhao, Z., Ma, Z.: The llama 3 herd of models (2024), https://arxiv.org/abs/2407.21783 9

11. Jing, L., Chen, G.H., Aghazadeh, E., Wang, X.E., Du, X.: A comprehensive analysis for visual object hallucination in large vision-language models. arXiv preprint arXiv:2505.01958 (2025) 3
12. Li, Y., Guo, H., Zhou, K., Zhao, W.X., Wen, J.R.: 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 (2024)

3, 4, 9
13. Li, Z., Shi, H., Gao, Y., Liu, D., Wang, Z., Chen, Y., Liu, T., Zhao, L., Wang, H., Metaxas, D.N.: The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In: Forty-second International Conference on Machine Learning (2025), https://openreview.net/forum?id=7BKcLeHQsm 9
14. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/ 10
15. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36**, 34892–34916 (2023) 3
16. Liu, J., Guo, H., Duan, R., Bu, X., He, Y., Li, S., Huang, H., Liu, J., Wang, Y., Jing, C., et al.: Dream: Disentangling risks to enhance safety alignment in multimodal large language models. arXiv preprint arXiv:2504.18053 (2025) 3
17. Liu, Q., Shang, C., Liu, L., Pappas, N., Ma, J., Anna John, N., Doss, S., Marquez, L., Ballesteros, M., Benajiba, Y.: Unraveling and mitigating safety alignment degradation of vision-language models. In: Che, W., Nabende, J., Shutova, E., Pilehvar, M.T. (eds.) Findings of the Association for Computational Linguistics: ACL 2025. pp. 3631–3643. Association for Computational Linguistics, Vienna, Austria (Jul 2025). https://doi.org/10.18653/v1/2025.findings-acl.186, https://aclanthology.org/2025.findings-acl.186/ 2, 4, 6, 9, 10
18. Liu, S., Ye, H., Zou, J.: Reducing hallucinations in large vision-language models via latent space steering. In: The Thirteenth International Conference on Learning Representations (2025) 3
19. Liu, X., Zhu, Y., Gu, J., Lan, Y., Yang, C., Qiao, Y.: Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In: European Conference on Computer Vision. pp. 386–403. Springer (2024) 3, 9
20. Lu, D., Pang, T., Du, C., Liu, Q., Yang, X., Lin, M.: Test-time backdoor attacks on multimodal large language models. arXiv preprint arXiv:2402.08577 (2024) 3
21. Lu, L., Pang, S., Liang, S., Zhu, H., Zeng, X., Liu, A., Liu, Y., Zhou, Y.: Adversarial training for multimodal large language models against jailbreak attacks. arXiv preprint arXiv:2503.04833 (2025) 3
22. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. In: The 36th Conference on Neural Information Processing Systems (NeurIPS) (2022) 9
23. OpenAI: Introducing gpt-4.1 in the api | openai, https://openai.com/index/gpt-4-1/ 9
24. Park, S., Du, X., Yeh, M.H., Wang, H., Li, Y.: Steer llm latents for hallucination detection. arXiv preprint arXiv:2503.01917 (2025) 3
25. Pi, R., Han, T., Zhang, J., Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., Zhang, T.: Mllm-protector: Ensuring mllm's safety without hurting performance (2024), https://arxiv.org/abs/2401.02906 1
26. Pi, R., Han, T., Zhang, J., Xie, Y., Pan, R., Lian, Q., Dong, H., Zhang, J., Zhang, T.: Mllm-protector: Ensuring mllm's safety without hurting performance. arXiv preprint arXiv:2401.02906 (2024) 3, 10
27. Shao, Z., Liu, H., Hu, Y., Gong, N.Z.: Refusing safe prompts for multi-modal large language models. arXiv preprint arXiv:2407.09050 (2024) 3
28. Shayegani, E., Dong, Y., Abu-Ghazaleh, N.: Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. arXiv preprint arXiv:2307.14539 (2023) 3

29. Skean, O., Arefin, M.R., Zhao, D., Patel, N., Naghiyev, J., LeCun, Y., Shwartz-Ziv, R.: Layer by layer: Uncovering hidden representations in language models. arXiv preprint arXiv:2502.02013 (2025) 3
30. Wang, H., Wang, G., Zhang, H.: Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29947–29957 (2025) 4
31. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024) 3
32. Wang, P., Zhang, D., Li, L., Tan, C., Wang, X., Ren, K., Jiang, B., Qiu, X.: Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. arXiv preprint arXiv:2401.11206 (2024) 3
33. Wang, R., Ma, X., Zhou, H., Ji, C., Ye, G., Jiang, Y.G.: White-box multimodal jailbreaks against large vision-language models. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 6920–6928 (2024) 3
34. Wang, Y., Liu, X., Li, Y., Chen, M., Xiao, C.: Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In: European Conference on Computer Vision. pp. 77–94. Springer (2024) 3
35. Yang, L., Zheng, Z., Chen, B., Zhao, Z., Lin, C., Shen, C.: Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14635–14645 (2025) 3
36. Ying, Z., Liu, A., Zhang, T., Yu, Z., Liang, S., Liu, X., Tao, D.: Jailbreak vision language models via bi-modal adversarial prompt. IEEE Transactions on Information Forensics and Security (2025) 3
37. Yu, L., Do, V., Hambardzumyan, K., Cancedda, N.: Robust llm safeguarding via refusal feature adversarial training. arXiv preprint arXiv:2409.20089 (2024) 3
38. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: evaluating large multimodal models for integrated capabilities. In: Proceedings of the 41st International Conference on Machine Learning. ICML'24, JMLR.org (2024) 9
39. Zhang, G., Fan, X., Fang, J., Sun, Y., Shi, X., Lu, C.: Unveiling vulnerabilities in large vision-language models: The savj jailbreak approach. In: International Conference on Artificial Neural Networks. pp. 417–434. Springer (2024) 3
40. Zhao, S., Duan, R., Wang, F., Chen, C., Kang, C., Ruan, S., Tao, J., Chen, Y., Xue, H., Wei, X.: Jailbreaking multimodal large language models via shuffle inconsistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2045–2054 (October 2025) 9, 12
41. Zhao, Y., Zheng, X., Luo, L., Li, Y., Ma, X., Jiang, Y.G.: Bluesuffix: Reinforced blue teaming for vision-language models against jailbreak attacks. arXiv preprint arXiv:2410.20971 (2024) 3
42. Zhou, G., Yan, Y., Zou, X., Wang, K., Liu, A., Hu, X.: Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality (2025), https://arxiv.org/abs/2410.04780 6
43. Zong, Y., Bohdal, O., Yu, T., Yang, Y., Hospedales, T.: Safety fine-tuning at (almost) no cost: A baseline for vision large language models. arXiv preprint arXiv:2402.02207 (2024) 3
44. Zou, X., Kang, J., Kesidis, G., Lin, L.: Understanding and rectifying safety perception distortion in vlms (2025), https://arxiv.org/abs/2502.13095 4