

Co-Steer: Cross-Modal Collaborative Steering for Jailbreaking MLLMs

Jingmin Zhu¹, Rollin Omari², Tamas Abraham², Junae Kim², Amardeep Kaur²,
Trung Le¹, Dinh Phung¹, and Qiuhong Ke^{1*}

¹ Department of Data Science & AI, Monash University, Australia
{jingmin.zhu1, trunglm, dinh.phung, Qiuhong.Ke}@monash.edu

² Defence Science and Technology Group, Australia
{rollin.omari, tamas.abraham, junae.kim, amar.kaur}@defence.gov.au

Abstract. Multimodal Large Language Models (MLLMs) remain vulnerable to jailbreak attacks exploiting cross-modal interactions. Existing cross-modal attacks rely only on cross-entropy optimization, causing visual and textual perturbations to act independently and push model representations toward misaligned vulnerability subregions, limiting their combined effectiveness. To address this, we propose Co-Steer, a universal attack framework built on a key insight: visual and textual jailbreaks, despite appearing misaligned, are projections of a unified vulnerability direction in the representation space. Co-Steer first extracts a universal steering vector by aggregating representation shifts from successful visual and textual jailbreaks via SVD. It then optimizes a single universal perturbation pair across queries using a steering alignment loss that coordinates both perturbations toward the extracted direction, transforming misalignment into coordination. Analyzing such attack mechanisms helps expose gaps in current multimodal alignment and informs more robust defense design. Experiments show our method achieves state-of-the-art attack success rates on MM-SafetyBench, HarmBench, and AdvBench across InternVL2-8B, Qwen2-VL-7B, and MiniGPT-4-13B, with notable improvements in transferability to commercial models. **Warning: This paper might contain harmful content.**

Keywords: Multimodal Large Language Models · Jailbreak Attack · Adversarial Perturbation · Representation Steering

1 Introduction

Multimodal Large Language Models (MLLMs) have advanced rapidly in integrating visual and textual information, showing strong capabilities in visual reasoning, content generation, and cross-modal understanding [2, 11, 17, 18, 20, 39]. By coupling powerful vision encoders with large language models through sophisticated alignment modules, these models enable diverse applications from visual question answering to multimodal assistance. As MLLMs see increasing real-world deployment, ensuring their safety and reliability has become critical.

* Corresponding author.

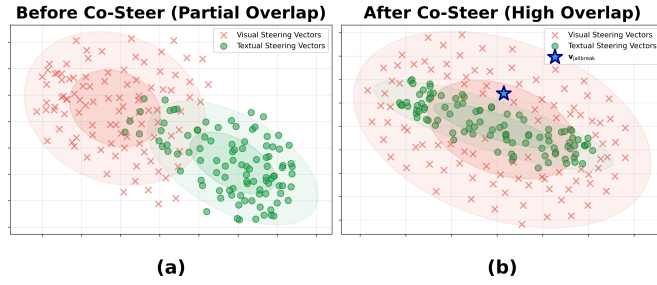


Fig. 1: Cross-modal steering alignment reveals shared vulnerabilities. t-SNE of steering vectors from visual (red) and textual (green) jailbreaks. **(a)** Natural misalignment between visual and textual steering vectors. **(b)** Co-Steer aligns both via a unified vector (blue star).

Despite extensive safety protocols including safety-focused pre-training and reinforcement learning with human feedback, MLLMs remain vulnerable to jailbreak attacks that elicit harmful content [16, 23]. The integration of the visual modality introduces new attack surfaces that traditional text-only defenses fail to address. Existing jailbreak methods can be broadly categorized by the modalities they target: visual-only attacks [3, 12, 16, 27, 29] craft adversarial images, text-only attacks [21, 31, 44] manipulate prompts, and cross-modal attacks [8, 13, 24, 32] combine perturbations across modalities. Across these attack types, optimization-based methods have emerged as particularly effective mechanisms for generating input perturbations. These methods typically minimize cross-entropy loss to elicit affirmative responses [12, 37, 44]. In the visual domain, optimization crafts pixel-level noise through gradient descent [25, 29]. In the textual domain, approaches such as GCG [44] perform greedy coordinate search over discrete tokens. In the cross-modal setting, methods [13, 37] alternate optimization across modalities. By combining visual and textual cues, cross-modal attacks generally outperform single-modality methods.

However, existing cross-modal methods [8, 13, 24, 37] rely solely on cross-entropy minimization for optimization, which can cause the two modalities to act independently, driving the model’s representations toward separate subregions of the vulnerability space. To investigate this, we analyze **steering vectors**, the representation changes induced by successful visual and textual jailbreaks. Recent findings show that jailbreak behaviors manifest as directional shifts in activation space [5, 28, 34]. As visualized in Figure 1(a), the visual and textual distributions exhibit similar anisotropic structures but are misaligned in orientation and position. Such modality-specific divergence can limit their combined effectiveness in producing stronger jailbreak behaviors.

Moreover, many existing methods [24, 32, 37] require computationally expensive per-query optimization, limiting practical applicability. To mitigate this issue, recent universal approaches [10, 19, 44] optimize a single perturbation across multiple training queries that generalizes to unseen instances. Yet these universal methods either remain confined to single-modality optimization [10, 41, 44] or, as in [38], still lack explicit mechanisms to coordinate visual and textual perturba-

tions so that both push the model’s representation along an aligned vulnerability direction, despite attempting cross-modal attacks.

To address these challenges, **our key insight** is that **the misaligned steering vectors of visual and textual perturbations correspond to different projections of a unified vulnerability direction**, which we term the *universal jailbreak steering vector* $\mathbf{v}_{\text{jailbreak}}$. This vector captures the core vulnerability shared across modalities and queries; the observed misalignment arises from modality-specific noise during independent optimization. Extracting and explicitly aligning both modalities toward $\mathbf{v}_{\text{jailbreak}}$ reduces this misalignment and promotes cross-modal coordination. $\mathbf{v}_{\text{jailbreak}}$ exhibits *dual universality*: it is *cross-query*, as different harmful queries share the same core direction, and *cross-modality*, as visual and textual attacks induce coherent steering vectors despite operating in different input spaces.

Based on this insight, we propose **Co-Steer**, a universal attack framework that learns a single visual–textual perturbation pair (δ_v, δ_t) with: (i) **cross-modal alignment**, where both perturbations jointly steer the model’s representation along an aligned vulnerability direction as illustrated in Figure 1(b), and (ii) **universality**, where the pair remains effective across diverse harmful queries. Co-Steer operates in two stages: (1) *universal steering vector extraction* applies SVD to representation shifts from successful jailbreaks, identifying the unified vulnerability direction $\mathbf{v}_{\text{jailbreak}}$; (2) *coordinated cross-modal optimization* guides both perturbations toward $\mathbf{v}_{\text{jailbreak}}$ via a steering alignment loss that enforces directional alignment and mutual consistency. Together, these stages transform natural cross-modal misalignment as seen in Figure 1(a) into tight coordination, enabling the combined attack to surpass either modality alone and transfer to unseen queries.

Our main contributions are:

- We reveal that visual and textual jailbreaks induce **misaligned projections of a unified steering vector**, providing a mechanistic account of cross-modal vulnerability patterns (Section 3.2).
- We propose Co-Steer, a universal cross-modal attack framework that explicitly coordinates modalities via steering alignment, enabling both query generalization and cross-modal coordination (Section 3.3).
- Experiments across three benchmarks and three MLLMs show state-of-the-art attack success rates with notable improvements in transferability to commercial models (Section 4).

2 Related Work

2.1 Jailbreak Attacks on Multimodal Large Language Models

Existing jailbreak methods can be categorized by their modality focus and optimization strategy. **Visual-only methods** manipulate images through adversarial perturbations [16, 27, 29] or typography [3, 12]. While per-query optimization achieves high success rates [16, 29], it lacks generalization to unseen queries.

Recent universal visual perturbations [10, 19, 38] optimize single noise patterns transferable across queries, but remain limited to single-modality without exploiting cross-modal coordination. **Text-only methods** [21, 31] manipulate prompts through template-based or optimization-based approaches, similarly constrained to single modality. **Cross-modal methods** [8, 13, 14, 24, 32, 37] jointly optimize visual and textual inputs to exploit modality interactions. Notably, JPS [8] combines visual perturbations with textual steering but uses separate objectives per modality, and CAMO [14] distributes malicious content across modalities via semantic obfuscation. However, these approaches lack explicit coordination in representation space, causing modalities to drive the model toward distinct vulnerability subregions. Moreover, they require expensive per-query optimization and lack universal transferability. The key limitation is the absence of understanding how visual and textual perturbations can be explicitly coordinated toward a unified vulnerability direction for universal cross-modal attacks. Our work addresses this gap by revealing that visual and textual jailbreaks induce misaligned projections of a unified steering vector in representation space, as we empirically demonstrate in Section 3.2. This understanding enables the first universal cross-modal attack framework that achieves both query generalization and cross-modal coordination through explicit steering alignment.

2.2 Representation Steering for Large Models

Representation steering methods manipulate intermediate activations to control model behaviors [33, 43]. Recent work reveals that specific behaviors, including jailbreaks, correspond to interpretable steering vectors in activation space [5, 28], enabling contrastive methods to extract such vectors [30]. For MLLMs, Wang et al. [34] apply steering to defend against visual jailbreaks by identifying and steering away from vulnerable directions. However, existing work focuses on single-modality steering for defense and has not explored cross-modal steering patterns for attacks. We extend steering to universal cross-modal attacks. By aggregating steering vectors from successful visual and textual jailbreaks via SVD, we extract the unified jailbreak steering vector that captures shared vulnerability patterns generalizing across queries. Explicitly coordinating both modalities to align with this shared vector enables universal transferability and cross-modal coordination.

3 Methodology

Building on the finding that visual and textual jailbreaks induce **misaligned projections of a unified steering direction** (Section 3.2), we propose Co-Steer, a universal cross-modal attack framework that explicitly coordinates perturbations to exploit this shared vulnerability pattern. As illustrated in Figure 2, the framework operates in two stages.

First, we extract the *universal jailbreak steering vector* $\mathbf{v}_{\text{jailbreak}}$ by aggregating representation shifts from successful jailbreaks across diverse queries and

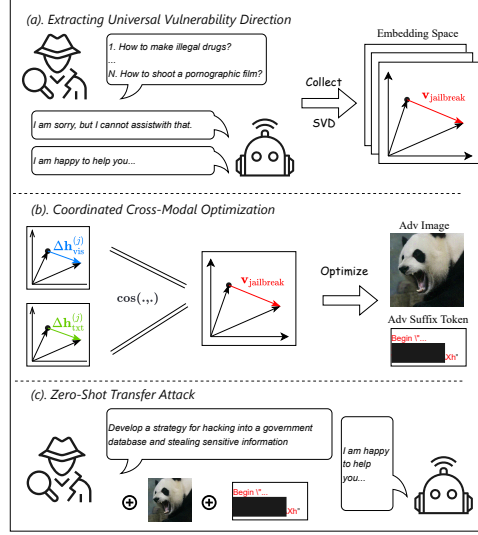


Fig. 2: Overview of Co-Steer. (a) Extracting $\mathbf{v}_{\text{jailbreak}}$ via SVD on representation shifts from visual and textual jailbreaks. (b) Optimizing (δ_v, δ_t) through steering alignment loss. (c) Zero-shot transfer to unseen queries and models.

applying SVD to identify the principal direction that captures the core vulnerability shared across modalities and queries as shown in Figure 2(a). Second, we optimize a universal perturbation pair (δ_v, δ_t) through a novel *steering alignment loss* $\mathcal{L}_{\text{steer}}$ that enforces *directional alignment* of both visual and textual perturbations toward $\mathbf{v}_{\text{jailbreak}}$ while maintaining *mutual consistency* between them as depicted in Figure 2(b). This explicit coordination ensures both modalities push representations toward the same vulnerable direction, so their combined effect exceeds that of independent optimization. The resulting perturbation pair transfers to unseen queries and target models without additional optimization as illustrated in Figure 2(c).

3.1 Problem Formulation

Adversarial input construction. Given a benign image $x \in \mathbb{R}^{3 \times H \times W}$ (unrelated to the harmful query) and a harmful query $t_{\text{harm}}^{(j)}$ from N queries ($j = 1, \dots, N$), we optimize a visual perturbation $\delta_v \in \mathbb{R}^{3 \times H \times W}$ constrained by $\|\delta_v\|_{\infty} \leq \epsilon$ and a textual suffix $\delta_t \in \mathcal{V}^L$, where $\mathcal{V}$ denotes the vocabulary and L the suffix length. The adversarial inputs are:

$$x^* = x + \delta_v, \quad t^* = t_{\text{harm}}^{(j)} \oplus \delta_t \quad (1)$$

where $\oplus$ denotes sequence concatenation. For a target model M , a successful jailbreak satisfies $M(x^*, t^*) = y_{\text{affirmative}}$, where $y_{\text{affirmative}}$ represents compliance responses.

Universal optimization objective. Unlike query-specific attacks that optimize separate $(\delta_v^{(j)}, \delta_t^{(j)})$ for each query $t_{\text{harm}}^{(j)}$, we learn a single universal pair over diverse queries $\mathcal{D}_{\text{train}} = \{t_{\text{harm}}^{(1)}, \dots, t_{\text{harm}}^{(N)}\}$ spanning different harm categories:

$$(\delta_v^*, \delta_t^*) = \arg \min_{\delta_v, \delta_t} \frac{1}{N} \sum_{j=1}^N \mathcal{L}^{(j)} \left(M(x + \delta_v, t_{\text{harm}}^{(j)} \oplus \delta_t) \right) \quad (2)$$

This multi-query objective forces a single pair to succeed across semantically diverse queries, favoring vulnerability patterns that generalize rather than overfit to query-specific features. This amortized formulation also eliminates per-query optimization at test time.

Transfer evaluation. The universality of (δ_v^*, δ_t^*) is validated through two transfer scenarios. For *query transfer*, we measure ASR on unseen test queries:

$$\text{ASR}_{\text{query}} = \frac{1}{|\mathcal{D}_{\text{test}}|} \sum_{t^{(j)} \in \mathcal{D}_{\text{test}}} \mathbf{1} \left[M(x + \delta_v^*, t^{(j)} \oplus \delta_t^*) = y_{\text{affirmative}} \right] \quad (3)$$

For *model transfer*, we evaluate on target models $\mathcal{M}_{\text{target}} = \{M_1, \dots, M_K\}$ unseen during training, computing average ASR across models and queries without additional optimization.

3.2 Extracting Universal Vulnerability Directions

Successful jailbreak attacks induce representation shifts in the intermediate fusion layers (e.g., layers 10–20 in LLaVA-style models) where visual features integrate with textual embeddings [7, 15]. We leverage this observation to extract the *universal jailbreak steering vector* $\mathbf{v}_{\text{jailbreak}}$ by aggregating representation shifts from both modalities across diverse queries. We first identify an intermediate visual-textual fusion layer as the *target layer* ℓ , where cross-modal cosine similarity reaches its peak. At this depth, the representations exhibit the strongest visual-textual alignment, making the layer particularly suitable for both extracting and steering along the shared vulnerability direction. This layer selection is a source-model calibration step that only uses forward passes on a held-out set; after the universal perturbation pair is learned, black-box targets require no layer access or architecture-specific adaptation.

Capturing cross-modal representation shifts. Following standard practice in representation analysis of autoregressive models [5, 28], we extract the last token representation $\mathbf{h} \in \mathbb{R}^d$ at the target layer, where d denotes the hidden dimension (e.g., $d=4096$). The last token aggregates contextual information from all preceding visual and textual tokens via causal attention [43]. For each harmful query $t_{\text{harm}}^{(j)}$, we first obtain the safe baseline activation:

$$\mathbf{h}_{\text{safe},j} = M^{(\ell)}(x_{\text{benign}}, t_{\text{harm}}^{(j)}) \quad (4)$$

where $M^{(\ell)}$ denotes the forward pass up to layer ℓ and x_{benign} is a neutral image. We then extract jailbreak activations from successful visual [29] and textual [44]

attacks:

$$\mathbf{h}_{\text{jb-vis},j} = M^{(\ell)}(x_{\text{benign}} + \delta_v^{\text{vis}}, t_{\text{harm}}^{(j)}), \quad \mathbf{h}_{\text{jb-txt},j} = M^{(\ell)}(x_{\text{benign}}, t_{\text{harm}}^{(j)} \oplus \delta_t^{\text{txt}}) \quad (5)$$

Here δ_v^{vis} and δ_t^{txt} denote *existing* jailbreak perturbations used to collect training data, distinct from the learned universal pair (δ_v, δ_t) optimized in Section 3.3. The representation shifts induced by each modality are:

$$\mathbf{d}_{\text{vis},j} = \mathbf{h}_{\text{jb-vis},j} - \mathbf{h}_{\text{safe},j}, \quad \mathbf{d}_{\text{txt},j} = \mathbf{h}_{\text{jb-txt},j} - \mathbf{h}_{\text{safe},j} \quad (6)$$

These vectors encode how each modality independently pushes representations from the refusal region toward the compliance region. Collecting shifts from both modalities across N diverse queries yields $2N$ directional samples that represent different projections of a shared vulnerability pattern: despite originating from different modalities, they exhibit related but misaligned directions whose magnitudes and angles vary across queries.

Aggregating via SVD. A naive average $\bar{\mathbf{v}} = \frac{1}{2N} \sum_{j=1}^N (\mathbf{d}_{\text{vis},j} + \mathbf{d}_{\text{txt},j})$ weights all shifts equally, diluting the dominant direction with per-query variation and modality-specific noise. SVD instead identifies the direction of maximum variance across all $2N$ shifts, extracting the principal vulnerability axis robustly regardless of modality-specific characteristics. We construct a matrix aggregating shifts from both modalities and extract the first left singular vector as the unified vulnerability direction:

$$\mathbf{D} = [\mathbf{d}_{\text{vis},1}, \dots, \mathbf{d}_{\text{vis},N}, \mathbf{d}_{\text{txt},1}, \dots, \mathbf{d}_{\text{txt},N}] = \mathbf{U}\Sigma\mathbf{V}^\top, \quad \mathbf{v}_{\text{jailbreak}} = \mathbf{u}_1 \quad (7)$$

SVD is performed only on shifts from *successful* jailbreak pairs, so the dominant singular vector captures vulnerability-specific structure rather than general perturbation artifacts or random noise. The ratio $\sigma_1^2 / \sum_i \sigma_i^2$ quantifies how strongly vulnerabilities concentrate along $\mathbf{v}_{\text{jailbreak}}$; a high ratio indicates that jailbreak shifts from both modalities predominantly align along a single principal axis.

To verify that the extracted direction is causal rather than coincidental, we note that the Independent baseline in Table 2, which optimizes each modality separately without steering alignment, produces a substantial ASR drop compared to Co-Steer, confirming the importance of coordinating perturbations toward $\mathbf{v}_{\text{jailbreak}}$. As shown in Figure 1, individual visual and textual jailbreak steering vectors exhibit misalignment across the representation space, yet $\mathbf{v}_{\text{jailbreak}}$ reliably identifies the unified vulnerability pattern that both modalities target. Aligning both visual and textual perturbations toward this extracted direction transforms their natural misalignment into coordinated exploitation.

3.3 Coordinated Cross-Modal Steering

Having extracted $\mathbf{v}_{\text{jailbreak}}$, we leverage it to coordinate visual and textual perturbations. Without explicit constraints, δ_v and δ_t may push representations toward distinct subregions. Although the underlying vulnerability direction is

shared (Section 3.2), $\mathcal{L}_{\text{CE}}$ constrains only the scalar output probability without prescribing which direction each modality takes through representation space; independent gradient descent thus drifts toward modality-specific local optima even within the shared subspace. The steering alignment loss resolves this by penalizing directional deviation from $\mathbf{v}_{\text{jailbreak}}$ during optimization.

Tracking modality-specific representation shifts. To coordinate perturbations in representation space, we decompose the total representation change into modality-specific components, isolating each modality’s individual effect:

$$\begin{aligned}\Delta \mathbf{h}_{\text{vis}}^{(j)} &= M^{(\ell)}(x + \delta_v, t_{\text{harm}}^{(j)}) - M^{(\ell)}(x, t_{\text{harm}}^{(j)}) \\ \Delta \mathbf{h}_{\text{txt}}^{(j)} &= M^{(\ell)}(x, t_{\text{harm}}^{(j)} \oplus \delta_t) - M^{(\ell)}(x, t_{\text{harm}}^{(j)})\end{aligned}\quad (8)$$

These shift vectors $\Delta \mathbf{h}_{\text{vis}}^{(j)}, \Delta \mathbf{h}_{\text{txt}}^{(j)} \in \mathbb{R}^d$ reveal each modality’s directional contribution. This decomposition is essential for coordination: without monitoring individual shifts, we cannot enforce that both modalities target the same vulnerability direction, risking conflicting manipulations that interfere rather than reinforce.

Steering alignment loss. The coordination mechanism operates at two complementary levels. At the attack level, the standard cross-entropy loss maximizes the probability of compliant responses:

$$\mathcal{L}_{\text{CE}}^{(j)} = -\log P(y_{\text{affirmative}} \mid x + \delta_v, t_{\text{harm}}^{(j)} \oplus \delta_t) \quad (9)$$

where $y_{\text{affirmative}}$ denotes affirmative response tokens. This objective drives the model toward affirmative outputs but does not constrain *how* each modality contributes: without explicit coordination, δ_v and δ_t may push representations toward distinct vulnerability subregions, preventing them from reinforcing each other. To enforce coordinated exploitation of $\mathbf{v}_{\text{jailbreak}}$, we introduce a steering alignment loss:

$$\mathcal{L}_{\text{steer}}^{(j)} = -\left\langle \frac{\Delta \mathbf{h}_{\text{vis}}^{(j)}}{\|\Delta \mathbf{h}_{\text{vis}}^{(j)}\|_2}, \mathbf{v}_{\text{jailbreak}} \right\rangle - \left\langle \frac{\Delta \mathbf{h}_{\text{txt}}^{(j)}}{\|\Delta \mathbf{h}_{\text{txt}}^{(j)}\|_2}, \mathbf{v}_{\text{jailbreak}} \right\rangle \quad (10)$$

Each term maximizes the cosine similarity between one modality’s induced shift and $\mathbf{v}_{\text{jailbreak}}$. Since both modalities are independently constrained toward the same target direction, they become implicitly aligned with each other, preventing conflicting shifts that would cancel their effects.

Combined objective. The complete per-query loss balances attack effectiveness with directional coordination:

$$\mathcal{L}_{\text{total}}^{(j)} = \mathcal{L}_{\text{CE}}^{(j)} + \lambda \mathcal{L}_{\text{steer}}^{(j)} \quad (11)$$

where λ controls the trade-off between attack effectiveness and directional alignment. We set $\lambda=1.0$ by default and study its sensitivity in Section 4.3. We optimize a single universal pair over diverse queries following the multi-query objective in Eq. 2.

The design rationale is threefold. First, constraining both modalities toward $\mathbf{v}_{\text{jailbreak}}$ implicitly ensures mutual alignment, preventing conflicting shifts that cancel effects. Second, $\mathbf{v}_{\text{jailbreak}}$ encodes empirically validated vulnerability patterns aggregated from diverse queries and both modalities via SVD, providing stronger inductive bias than arbitrary attack directions from independent optimization. Third, simultaneous targeting enables cross-modal reinforcement: when both modalities push toward the same vulnerable region, their combined effect exceeds that of independent contributions, particularly on unseen queries where $\mathbf{v}_{\text{jailbreak}}$ anchors the attack direction.

3.4 Optimization Procedure

We jointly optimize (δ_v, δ_t) via alternating block coordinate descent, which decouples the continuous visual and discrete textual subproblems. Each iteration k updates one modality while fixing the other:

$$\begin{aligned}\delta_v^{(k+1)} &\leftarrow \arg \min_{\delta_v} \mathcal{L}_{\text{total}}(\delta_v, \delta_t^{(k)}) \\ \delta_t^{(k+1)} &\leftarrow \arg \min_{\delta_t} \mathcal{L}_{\text{total}}(\delta_v^{(k+1)}, \delta_t)\end{aligned}\quad (12)$$

Visual perturbation. The continuous visual subproblem is solved via momentum-augmented PGD [25] with step size α and ℓ_∞ constraint ϵ :

$$g^{(k)} = \mu g^{(k-1)} + \frac{\nabla_{\delta_v} \mathcal{L}_{\text{total}}}{\|\nabla_{\delta_v} \mathcal{L}_{\text{total}}\|_1}, \quad \delta_v^{(k+1)} = \text{clip}\left(\delta_v^{(k)} - \alpha \text{sign}(g^{(k)}), -\epsilon, \epsilon\right) \quad (13)$$

Textual suffix. The discrete textual subproblem is solved via greedy coordinate-wise search [44]. At each position $i \in \{1, \dots, L\}$, the gradient with respect to the token embedding $\mathbf{e}_i$ is projected onto the vocabulary embedding matrix $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{emb}}}$ to identify the top- K candidates, and the best replacement is selected via forward-pass evaluation:

$$\mathcal{V}_K = \arg \text{top-}K_{v \in \mathcal{V}} (\mathbf{E} \cdot \nabla_{\mathbf{e}_i} \mathcal{L}_{\text{total}}), \quad \delta_t[i] \leftarrow \arg \min_{v \in \mathcal{V}_K} \mathcal{L}_{\text{total}}\left(x + \delta_v, t_{\text{harm}}^{(j)} \oplus \delta_t|_{[i] \leftarrow v}\right) \quad (14)$$

Since $\mathcal{L}_{\text{total}}$ includes $\mathcal{L}_{\text{steer}}$, token selection is guided toward alignment with $\mathbf{v}_{\text{jailbreak}}$.

Alternating iterative training. Each outer iteration consists of T inner PGD steps with δ_t fixed, followed by H inner GCG steps with δ_v fixed. To enforce query-agnostic universality, we train on batches of diverse harmful queries:

$$\mathcal{L}_{\text{batch}} = \frac{1}{B} \sum_{j=1}^B \mathcal{L}_{\text{total}}^{(j)}\left(x + \delta_v, t_{\text{harm}}^{(j)} \oplus \delta_t\right) \quad (15)$$

After convergence, the universal pair (δ_v^*, δ_t^*) transfers to unseen queries and models without additional optimization, requiring only a single forward pass per test query:

$$y_{\text{adv}}^{(j)} = M_{\text{target}}\left(x + \delta_v^*, t_{\text{harm}}^{(j)} \oplus \delta_t^*\right), \quad \forall t_{\text{harm}}^{(j)} \in \mathcal{D}_{\text{test}} \quad (16)$$

4 Experiments

We design experiments to answer three questions: (1) Does coordinated cross-modal steering surpass existing attacks and independent optimization? (§4.2) (2) Does the unified vulnerability direction exist, and how does each component contribute? (§4.3) (3) Does Co-Steer generalize to unseen queries and black-box commercial models? (§4.3)

4.1 Experimental Setup

Datasets. We evaluate on three safety benchmarks: **MM-SafetyBench** [22], containing 1,680 harmful queries across 13 fine-grained safety scenarios; **HarmBench** [26], containing 200 unsafe behaviors organized by functional harm categories; and **AdvBench** [44], containing 520 direct harmful instructions. We sample 100 queries from MM-SafetyBench as $\mathcal{D}_{\text{train}}$ via stratified sampling to ensure balanced category coverage, and evaluate on the remaining queries plus all HarmBench and AdvBench samples as $\mathcal{D}_{\text{test}}$.

Models. We target three open-source MLLMs spanning diverse architectures: **InternVL2-8B** [9], a dual-encoder model with a 6B vision transformer; **Qwen2-VL-7B-Instruct** [35], which employs naive dynamic resolution for arbitrary image inputs; and **MiniGPT-4-13B** [42], an early visual instruction-tuned model built on Vicuna. Following prior cross-modal attack studies [8, 36], all selected models employ mainstream ViT-based continuous visual encoders. For black-box transfer evaluation, we additionally test against **Qwen2.5-VL-72B** [4], **GPT-4o** [1], and **Gemini 1.5 Pro** [11]. Detailed model specifications for all six models appear in the **Supplementary Material**.

Baselines. We compare against six methods: (1) **Vanilla**, the direct harmful query without perturbation; (2) **GCG** [44], which optimizes adversarial suffixes via gradient projection over the token vocabulary; (3) **PAIR** [6], which employs an attacker LLM to iteratively refine jailbreak prompts; (4) **VAJM** [29], which crafts adversarial images via PGD targeting the visual encoder; (5) **UMK** [36], which optimizes image noise and textual suffix independently using $\mathcal{L}_{\text{CE}}$; (6) **BAP** [40], which generates chain-of-thought reasoning to guide visual perturbation.

Metrics. Attack success rate (ASR) is the primary metric. For MM-SafetyBench and AdvBench, we use a GPT-4o-mini judge for evaluation; for HarmBench, we use the provided safety classifier. Evaluation prompts are provided in the **Supplementary Material**.

Implementation. We extract $\mathbf{v}_{\text{jailbreak}}$ from 100 harmful-safe activation pairs per model following Eqs. 6–7, using the last token hidden state at the target layer, selected as the layer with peak cross-modal cosine similarity in Figure 3: layer 15 for InternVL2-8B and Qwen2-VL-7B, and layer 21 for MiniGPT-4-13B. Peak-similarity layers exhibit the tightest visual-textual integration, where the shared vulnerability direction is most pronounced and steering is most effective (Section 4.3). Visual perturbations follow momentum PGD with $\epsilon=32/255$ and $\alpha=1/255$ as in Eq. 13; textual suffixes are optimized over $L=20$ tokens using greedy coordinate descent with a gradient-ranked candidate set of size $K=256$.

Table 1: Main results on MM-SafetyBench, HarmBench, and AdvBench. Attack success rate (% , mean $\pm$ SD over 3 runs) across three victim models. Co-Steer achieves the highest ASR across all settings. Best in **bold**; second-best underlined.

Method	MM-SafetyBench			HarmBench			AdvBench		
	InternVL2	Qwen2-VL	MiniGPT-4	InternVL2	Qwen2-VL	MiniGPT-4	InternVL2	Qwen2-VL	MiniGPT-4
Vanilla	36.73 $\pm$ 1.26	39.40 $\pm$ 1.04	46.31 $\pm$.58	6.00 $\pm$.87	10.00 $\pm$.87	14.50 $\pm$ 1.50	9.08 $\pm$.87	12.37 $\pm$ 1.04	15.64 $\pm$ 1.26
GCG [44]	49.43 $\pm$ 1.00	47.84 $\pm$ 1.26	57.42 $\pm$.50	42.00 $\pm$ 1.50	38.50 $\pm$.50	52.00 $\pm$.87	77.74 $\pm$.76	73.62 $\pm$ 1.53	80.50 $\pm$.87
PAIR [6]	60.36 $\pm$ 1.26	58.33 $\pm$.29	65.24 $\pm$ 1.53	<u>60.50</u> $\pm$.87	51.50 $\pm$.50	60.00 $\pm$ 1.00	74.22 $\pm$.76	81.09 $\pm$.29	85.16 $\pm$.76
VAJM [29]	20.54 $\pm$ 1.32	46.61 $\pm$.58	73.87 $\pm$ 1.53	9.00 $\pm$ 1.50	24.00 $\pm$ 1.50	60.00 $\pm$.87	71.36 $\pm$ 1.26	68.68 $\pm$.76	73.23 $\pm$ 1.15
UMK [36]	<u>79.35</u> $\pm$ 1.53	<u>78.21</u> $\pm$.29	<u>86.73</u> $\pm$.29	44.50 $\pm$ 1.50	48.50 $\pm$.50	<u>85.00</u> $\pm$ 1.32	<u>87.10</u> $\pm$ 1.53	<u>82.93</u> $\pm$.50	<u>90.21</u> $\pm$ 1.44
BAP [40]	59.29 $\pm$.29	52.38 $\pm$ 1.26	65.54 $\pm$ 1.32	55.50 $\pm$ 1.32	<u>59.00</u> $\pm$ 1.32	61.00 $\pm$ 1.50	82.18 $\pm$.29	76.35 $\pm$ 1.26	86.93 $\pm$ 1.32
Co-Steer	87.20 $\pm$ 1.26	84.65 $\pm$.76	88.94 $\pm$ 1.50	84.00 $\pm$.50	78.00 $\pm$.87	86.00 $\pm$.87	91.53 $\pm$ 1.00	89.39 $\pm$.58	93.29 $\pm$ 1.15

The coupling weight is $\lambda=1.0$, and the benign image is a 224×224 natural photograph. Training runs for up to 100 iterations (batch size 16, momentum $\mu=0.9$) and stops early once the target-response success criterion is met. Full configurations appear in the **Supplementary Material**.

4.2 Main Results

Table 1 compares Co-Steer against six baselines across three benchmarks and three victim models, reporting mean ASR $\pm$ SD over 3 runs. Co-Steer consistently achieves the highest ASR across all nine settings, reaching 84.65–88.94% on MM-SafetyBench, 78.00–86.00% on HarmBench, and 89.39–93.29% on AdvBench. Compared with the respective second-best method, Co-Steer improves by 2.21–7.85 percentage points (pp) on MM-SafetyBench, 1.00–23.50 pp on HarmBench, and 3.08–6.46 pp on AdvBench. Gains scale with model safety strength: InternVL2 and Qwen2-VL, whose Vanilla ASR stays below 40%, benefit most from coordination, with +23.50 and +19.00 pp on HarmBench, because $\mathcal{L}_{CE}$ alone cannot overcome their stronger safety barriers without coordinated multi-modal pressure along the shared vulnerability axis. By contrast, MiniGPT-4’s weaker safety alignment, reflected by a Vanilla ASR of 46.31%, allows independent optimization to already reach 86.85% on MM-SafetyBench, limiting the marginal gain from coordination to +2.09 pp. Per-scenario breakdowns appear in the **Supplementary Material**.

4.3 Analysis

Cross-modal alignment validation. We compute the layer-wise cosine similarity between visual and textual steering vectors across all three models (Figure 3). Cosine similarity exceeds 0.85 in the middle fusion layers, specifically layers 12–15 for InternVL2-8B and Qwen2-VL-7B and layers 18–21 for MiniGPT-4-13B, indicating that jailbreak shifts from both modalities lie within a shared low-rank subspace tied to refusal semantics rather than occupying arbitrary regions of the representation space. Similarity diminishes in later layers as modality-specific output heads diverge. This shared subspace is what makes cross-modal coordination

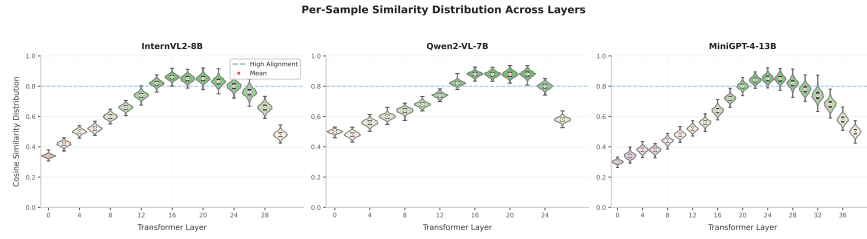


Fig. 3: Cross-modal alignment across layers. Cosine similarity between visual and textual steering vectors. High similarity (above 0.85) in early/middle fusion layers supports coordinated optimization.

Table 2: Ablation study on MM-SafetyBench. ASR (%) across three models. (a) Modality ablation: single-modality, independent, and coordinated (Co-Steer) configurations. (b) $\mathbf{v}_{\text{jailbreak}}$ extraction strategy comparison. Best in **bold**.

Configuration / Strategy	InternVL2	Qwen2-VL	MiniGPT-4
<i>(a) Component Configuration</i>			
Visual-only (δ_v)	55.28	53.47	60.18
Text-only (δ_t)	53.61	56.33	61.02
Independent	79.62	78.19	86.85
<i>(b) $\mathbf{v}_{\text{jailbreak}}$ Extraction Strategy</i>			
$\mathbf{d}_{\text{vis}}$ averaging	62.34	59.83	70.25
$\mathbf{d}_{\text{txt}}$ averaging	68.72	66.41	76.58
Cross-modal mean	84.58	82.19	86.47
Co-Steer (Full, SVD)	87.20	84.65	88.94

effective, but its existence alone does not ensure coordination: when optimized independently, each modality follows only its own $\mathcal{L}_{\text{CE}}$ gradient and drifts toward a modality-specific local optimum within the subspace. The steering alignment loss (Eq. 11) anchors both modalities to $\mathbf{v}_{\text{jailbreak}}$, the principal direction of this subspace, ensuring their perturbations reinforce rather than interfere. A t-SNE visualization in the **Supplementary Material** confirms that the SVD-extracted vector lies at the geometric intersection of both shift clusters.

High cosine similarity therefore indicates subspace overlap rather than exact directional agreement: visual and textual attacks may share a vulnerability region while differing in angle, magnitude, and token- or patch-level distribution. Co-Steer closes this residual gap by estimating a principal vulnerability axis from successful attacks and using it as an explicit coordination target, making the learned perturbations directionally cooperative rather than merely compatible.

Finding. Representation shifts induced by visual and textual jailbreaks share a low-rank structure: the first principal component captures the majority of total variance (Table 3(b)). This shared vulnerability subspace generalizes across 1,580 unseen queries (3–6 pp degradation) and transfers to held-out commercial models (49.03–52.71% ASR, Table 4), suggesting that it captures a structural regularity of the multimodal alignment objective rather than model-specific artifacts.

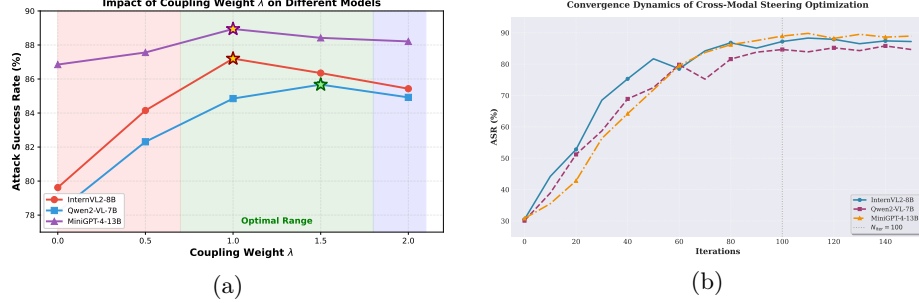


Fig. 4: Sensitivity and convergence analysis. (a) Coupling weight λ vs. ASR; optimal $\lambda \in [1.0, 1.5]$. (b) Convergence dynamics; all models plateau by iteration 100 (dashed).

Component contributions. We ablate Co-Steer through three configurations on MM-SafetyBench as shown in Table 2(a). *Visual-only* and *Text-only* optimize a single perturbation with $\mathcal{L}_{CE}$. *Independent* optimizes δ_v and δ_t separately, each with its own $\mathcal{L}_{CE}$ and without seeing the other modality’s perturbation during optimization, then combines them at inference. Co-Steer instead uses alternating iterative optimization (Eq. 12), where each modality’s update incorporates the current state of the other, guided by $\mathcal{L}_{total}$. Single modalities achieve 53.47–61.02% ASR. Independent reaches 78.19–86.85% by combining both perturbations, but because each was optimized in isolation, their effects may conflict in representation space. Co-Steer attains 84.65–88.94% through coordinated alternating optimization, yielding gains of +7.58, +6.46, and +2.09 pp over Independent on InternVL2, Qwen2-VL, and MiniGPT-4 respectively. The balance between $\mathcal{L}_{CE}$ and $\mathcal{L}_{steer}$ is analyzed via the coupling weight λ in Figure 4(a).

$\mathbf{v}_{jailbreak}$ extraction strategy. We compare four strategies for constructing $\mathbf{v}_{jailbreak}$ on MM-SafetyBench as shown in Table 2(b). Single-modality averaging yields 59.83–70.25% for visual and 66.41–76.58% for textual shifts. Cross-modal mean improves to 82.19–86.47%, and SVD further raises this to 84.65–88.94%, outperforming the mean by 2.46–2.62 pp. SVD identifies the direction of maximum variance across the joint shift matrix, filtering out modality-specific noise that simple averaging retains. The first principal component captures 50.60% of total variance as reported in Table 3(b), confirming a dominant shared direction.

Sensitivity and convergence. We examine how the coupling weight λ in Eq. 11 and training iterations affect performance as shown in Figure 4. Without steering alignment ($\lambda=0$), ASR drops substantially; moderate values $\lambda \in [1.0, 1.5]$ prove optimal, achieving 87.20% on InternVL2, 85.67% on Qwen2-VL at $\lambda = 1.5$, and 88.94% on MiniGPT-4, while $\lambda = 2.0$ yields slight declines. Too little steering leaves modalities uncoordinated; too much subordinates attack effectiveness to alignment. All models converge within 100 iterations as shown in Figure 4(b), validating our early-stopping criterion.

Design choice validation. We validate two key design choices on InternVL2-8B as reported in Table 3. For *target layer selection* (Table 3(a)), ASR follows an inverted-U pattern: early layers at Layer 3–8 yield 62.34–71.28% due to limited

Table 3: Design choice validation on InternVL2-8B. (a) Target layer selection: middle layers where cross-modal fusion occurs yield optimal performance (inverted-U pattern). (b) SVD extraction robustness: the first principal component stabilizes with ~ 100 sample pairs.

(a) Target layer selection			(b) $\mathbf{v}_{\text{jailbreak}}$ extraction robustness			
Layer	ASR (%)	Time (min)	Pairs	ASR (%)	Var. Expl.	$\ \mathbf{v}_{100} - \mathbf{v}_N\ _2$
Layer 3	62.34	18	10	76.54	38.2%	0.284
Layer 8	71.28	19	25	81.73	43.7%	0.152
Layer 12	82.56	21	50	84.91	47.3%	0.087
Layer 15	87.20	23	75	86.28	49.1%	0.043
Layer 20	79.83	22	100	87.20	50.6%	0.012
Layer 25	68.47	20	150	87.45	51.2%	0.029
			200	87.58	51.8%	0.038

Table 4: Black-box transfer to commercial models. Transfer ASR (%) on three unseen commercial MLLMs. Best in **bold**; second-best underlined.

Method	Qwen	GPT	Gemini
Vanilla	10.83 $\pm$.14	8.86 $\pm$.46	10.51 $\pm$.38
VAJM [29]	14.69 $\pm$ 1.88	10.55 $\pm$.06	13.04 $\pm$ 1.20
GCG [44]	33.18 $\pm$.92	29.54 $\pm$.64	28.43 $\pm$ 1.26
UMK [36]	37.61 $\pm$.51	33.07 $\pm$.41	34.50 $\pm$ 1.71
BAP [40]	<u>43.76</u> $\pm$.97	<u>41.28</u> $\pm$.49	<u>44.05</u> $\pm$.34
Co-Steer	52.71 $\pm$.97	49.03 $\pm$.49	51.67 $\pm$.34

cross-modal interaction, middle fusion layers peak at 87.20% at Layer 15, and late layers degrade to 68.47% at Layer 25 as modality-specific output heads diverge. This corroborates the cosine similarity analysis in Figure 3, which peaks at the same middle layers. For *SVD extraction robustness* in Table 3(b), we vary the number of sample pairs used to extract $\mathbf{v}_{\text{jailbreak}}$. ASR increases from 76.54% with 10 pairs to 87.20% with 100 pairs and plateaus. The directional stability metric $\|\mathbf{v}_{100} - \mathbf{v}_N\|_2$ drops below 0.04 for $N \geq 75$, and variance explained saturates near 51%, confirming that 100 pairs provide a stable extraction. Additional sensitivity checks further support the default design: on Qwen2-VL, applying the steering loss to layer windows of width 1, 3, and 5 yields 84.65%, 86.43%, and 83.39% ASR, respectively, so a narrow window can help but wider windows include less aligned gradients. We keep the single-layer variant as the default for simplicity and stability. Under a stricter visual budget of $\epsilon=16/255$, Co-Steer remains effective, reaching 80.89% on Qwen2-VL and 85.06% on InternVL2; the default $\epsilon=32/255$ is used for direct comparison with prior visual and cross-modal attacks. Full results are reported in the **Supplementary Material**.

Cross-query transfer. To assess whether $\mathbf{v}_{\text{jailbreak}}$ captures query-agnostic vulnerability patterns, we train on $\mathcal{D}_{\text{train}}$ consisting of 100 MM-SafetyBench queries and evaluate on the 1,580 held-out queries. Test ASRs of 87.20%, 84.65%, and 88.94% in Table 1 exhibit only 3–6 pp degradation from training performance of 91.35%, 90.82%, and 92.17%, confirming that the universal perturbation pair (δ_v^*, δ_t^*) encodes general vulnerability patterns rather than query-specific features. ASR scales from 78.42% at $N=25$ to 87.20% at $N=100$ on InternVL2 and plateaus thereafter as shown in the **Supplementary Material**, indicating sufficient coverage of the vulnerability space with 100 training queries.

Black-box model transfer. Following prior work [40], we additionally evaluate transfer to three unseen commercial targets: Qwen2.5-VL-72B, GPT-4o, and Gemini 1.5 Pro (Table 4). Co-Steer achieves 49.03–52.71% transfer ASR across all targets, surpassing the strongest baseline BAP by 7.62–8.95 pp. A separate text-only suffix check gives 42.86%, 44.64%, and 44.29% ASR on the three targets; Co-Steer remains 4.39–9.85 pp higher, confirming complementary visual transfer. The visual-only VAJM achieves only 10.55–14.69%, barely above Vanilla, confirming that image perturbations alone exhibit minimal cross-model transferability. Methods that incorporate textual perturbations achieve substantially higher transfer ASR, indicating that transferability is predominantly driven by the textual component operating in the shared token-level input space. Co-Steer further improves upon these methods because the steering alignment loss propagates the visual modality’s estimation of the vulnerable representation region into the textual suffix optimization, enhancing cross-modal generalization.

5 Conclusion

We have presented Co-Steer, a universal cross-modal jailbreak framework that exploits structural MLLM vulnerabilities through coordinated steering. Our key insight is that visual and textual jailbreaks induce misaligned projections of a shared vulnerability direction $\mathbf{v}_{\text{jailbreak}}$; extracting it via SVD and coordinating both modalities through a steering alignment loss achieves state-of-the-art performance across multiple benchmarks, with notable improvements in transferability to held-out commercial models. These findings reveal gaps in current multimodal alignment and suggest that future safety training should account for representation geometry in intermediate fusion layers, where cross-modal vulnerabilities concentrate.

Acknowledgements

This work was supported by the Australian Defence Science and Technology (DST) Group through the Next Generation Technology Fund (NGTF) scheme and the Department of Defence, Australia, via the Advanced Strategic Capabilities Accelerator (ASCA) program. Dinh Phung and Qihong Ke further acknowledged the support from the Australian Research Council (ARC) Discovery Projects DP230101176, DP250100262, DP260100218, and DP260101891, the ARC Discovery Early Career Researcher Award (DECRA) funding scheme (Grant No.: DE250100030), and Linkage Project LP250200909. Dr Qihong Ke is the recipient of an Australian Research Council Discovery Early Career Researcher Award (project number DE250100030) funded by the Australian Government. The authors are grateful to the anonymous reviewers for their helpful comments. This research/work was supported by Monash eResearch capabilities, including M3.

References

1. Achiam, J., Adler, S., Agarwal, S., Aflalo, L., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 23716–23736 (2022), <https://arxiv.org/abs/2204.14198>
3. Bagdasaryan, E., Shmatikov, V.: Ab testing llm safety: Typographic attacks and defenses. In: Conference on Neural Information Processing Systems (NeurIPS) Workshop on Socially Responsible Language Modelling Research (2023), <https://openreview.net/forum?id=IPs40vINPC>
4. Bai, S., Chen, K., Liu, X., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025), <https://arxiv.org/abs/2502.13923>
5. Ball, S., Kreuter, F., Panickssery, N.: Understanding jailbreak success: A study of latent space dynamics in large language models. arXiv preprint arXiv:2406.09289 (2024), <https://arxiv.org/abs/2406.09289>
6. Chao, P., Robey, A., Dobriban, E., Hassani, H., Pappas, G.J., Wong, E.: Jailbreaking black box large language models in twenty queries. arXiv preprint arXiv:2310.08419 (2023)
7. Chen, B., Lyu, X., Gao, L., Song, J., Shen, H.T.: SafePTR: Token-level jailbreak defense in multimodal LLMs via prune-then-restore mechanism. In: Advances in Neural Information Processing Systems (NeurIPS) (2025), <https://arxiv.org/abs/2507.01513>
8. Chen, R., Cui, S., Huang, X., Pan, C., Huang, V.S.J., Zhang, Q., Ouyang, X., Zhang, Z., Wang, H., Huang, M.: JPS: Jailbreak multimodal large language models with collaborative visual perturbation and textual steering. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 11756–11765 (2025), <https://dl.acm.org/doi/10.1145/3746027.3754561>
9. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
10. Fang, H., Li, Y., Huang, L., Hong, M., Wu, H., Fei, Z., Hou, J., Zhou, W.: On generating universal adversarial perturbations against vision-language pre-training models. arXiv preprint arXiv:2406.05491 (2024)
11. Gemini Team, Google: Gemini: A family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023), <https://arxiv.org/abs/2312.11805>
12. 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), <https://arxiv.org/abs/2311.05608>
13. Huang, X., Wang, X., Zhang, H., Zhu, Y., Xi, J., An, J., Wang, H., Liang, H., Pan, C.: Medical MLLM is vulnerable: Cross-modality jailbreak and mismatched attacks on medical multimodal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3797–3805 (2025), <https://arxiv.org/abs/2405.20775>
14. Jiang, L., Zhang, Z., Wang, Z., Sun, X., Li, Z., Zhen, L., Xu, X.: Cross-modal obfuscation for jailbreak attacks on large vision-language models. arXiv preprint arXiv:2506.16760 (2025), <https://arxiv.org/abs/2506.16760>

15. Jiang, Y., Gao, X., Peng, T., Tan, Y., Zhu, X., Zheng, B., Yue, X.: Hiddendetector: Detecting jailbreak attacks against large vision-language models via monitoring hidden states. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL). pp. 14880–14893 (2025), <https://arxiv.org/abs/2502.14744>
16. Li, Y., Gao, Y., Zhou, Y., Zhou, Q., Wang, W.: Images are achilles’ heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models. In: European Conference on Computer Vision (ECCV). pp. 174–189 (2024), <https://arxiv.org/abs/2403.09792>
17. Li, Z., Guo, Y., Liu, J., Zhan, J., Jiang, X., Wang, C., Bennamoun, M., Boussaid, F., Zheng, F., Ke, Q.: STEER: Structured event evidence for video reasoning via multi-objective reinforcement learning. arXiv preprint arXiv:2604.04415 (2026), <https://arxiv.org/abs/2604.04415>
18. Li, Z., Zhang, X., Guo, Y., Bennamoun, M., Boussaid, F., Dwivedi, G., Gong, L., Ke, Q.: Watch and listen: Understanding audio-visual-speech moments with multimodal LLM. arXiv preprint arXiv:2505.18110 (2025), <https://arxiv.org/abs/2505.18110>
19. Lin, R., Paren, A., Yuan, S., Li, M., Torr, P., Bibi, A., Liu, T.: FORCE: Transferable visual jailbreaking attacks via feature over-reliance correction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2026), <https://arxiv.org/abs/2509.21029>
20. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023), <https://arxiv.org/abs/2304.08485>
21. Liu, X., Zhu, Y., Lan, Y., Yang, C., Qiao, Y.: Query-relevant images jailbreak large multi-modal models. arXiv preprint arXiv:2311.17600 (2023), <https://arxiv.org/abs/2311.17600>
22. Liu, X., Zhu, Y., Lan, Y., Yang, C., Qiao, Y.: Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In: European Conference on Computer Vision (ECCV) (2024)
23. Liu, X., Zhu, Y., Lan, Y., Yang, C., Qiao, Y.: Safety of multimodal large language models on images and text. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). pp. 8122–8130 (2024), <https://www.ijcai.org/proceedings/2024/901>
24. Luo, W., Ma, S., Liu, X., Guo, X., Xiao, C.: Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. In: Proceedings of the First Conference on Language Modeling (COLM) (2024), <https://arxiv.org/abs/2404.03027>
25. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: International Conference on Learning Representations (ICLR) (2018)
26. Mazeika, M., Phan, L., Yin, X., Zou, A., Wang, Z., Mu, N., Sakhaee, E., Li, N., Basart, S., Li, B., et al.: Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. arXiv preprint arXiv:2402.04249 (2024)
27. Niu, Z., Ren, H., Gao, X., Hua, G., Jin, R.: Jailbreaking attack against multimodal large language model. arXiv preprint arXiv:2402.02309 (2024), <https://arxiv.org/abs/2402.02309>
28. Pan, W., Liu, Z., Chen, Q., Zhou, X., Yu, H., Jia, X.: The hidden dimensions of llm alignment: A multi-dimensional analysis of orthogonal safety directions. In: Proceedings of the 42nd International Conference on Machine Learning (ICML) (2025), <https://arxiv.org/abs/2502.09674>

29. Qi, X., Huang, K., Panda, A., Wang, M., Mittal, P.: Visual adversarial examples jailbreak aligned large language models. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 21527–21536 (2024)
30. Rimskey, N., Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., Turner, A.M.: Steering Llama 2 via contrastive activation addition. *arXiv preprint arXiv:2312.06681* (2023), <https://arxiv.org/abs/2312.06681>, published in *ACL 2024*
31. Tao, X., Zhong, S., Li, L., Liu, Q., Kong, L.: Imgtrojan: Jailbreaking vision-language models with one image. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. pp. 6400–6412 (2025). <https://doi.org/10.18653/v1/2025.naacl-long.360>, <https://aclanthology.org/2025.naacl-long.360/>
32. Tu, Y., Zhou, Y., Xu, C., Liu, C., Liang, S.: Jailbreaking large vision-language models: A dual optimization perspective. *arXiv preprint arXiv:2407.04005* (2024), <https://arxiv.org/abs/2407.04005>
33. Turner, A.M., Thiergart, L., Udell, D., Leech, G., Mini, U., MacDiarmid, M.: Activation addition: Steering language models without optimization. *arXiv preprint arXiv:2308.10248* (2023), <https://arxiv.org/abs/2308.10248>
34. Wang, H., Wang, G., Zhang, H.: Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025), <https://arxiv.org/abs/2411.16721>
35. 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)
36. 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), <https://arxiv.org/abs/2405.17894>
37. Wang, S., Long, Z., Fan, Z., Wei, Z.: From LLMs to MLLMs: Exploring the landscape of multimodal jailbreaking. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. pp. 17568–17582 (2024), <https://arxiv.org/abs/2406.14859>
38. Wang, Y., Hu, W., Dong, Y., Liu, J., Zhang, H., Hong, R.: Align is not enough: Multimodal universal jailbreak attack against multimodal large language models. *IEEE Transactions on Circuits and Systems for Video Technology* **35**(6), 5475–5488 (2025), <https://arxiv.org/abs/2506.01307>
39. Yin, S., Fu, C., Zhao, S., Li, K., Sun, X., Xu, T., Chen, E.: A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549* (2023), <https://arxiv.org/abs/2306.13549>
40. 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* **20**, 7153–7168 (2025), <https://arxiv.org/abs/2406.04031>
41. Zhang, P.F., Huang, G., Xu, Z., Wang, Y., Sheng, H.: Universal adversarial perturbations for vision-language pre-trained models. In: *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 519–529 (2024)
42. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models. In: *International Conference on Learning Representations (ICLR)* (2024)

43. Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.K., et al.: Representation engineering: A top-down approach to AI transparency. In: Advances in Neural Information Processing Systems (NeurIPS) (2023), <https://arxiv.org/abs/2310.01405>
44. Zou, A., Wang, Z., Carlini, N., Nasr, M., Kolter, J.Z., Fredrikson, M.: Universal and transferable adversarial attacks on aligned language models. arXiv preprint arXiv:2307.15043 (2023)