Activation Steering Meets Preference Optimization: Defense Against Jailbreaks in Vision Language Models

Sihao Wu[1] SIHAOWU@LIVERPOOL.AC.UK
Gaojie Jin[2] G.JIN@EXETER.AC.UK
Wei Huang[3] HAVELHUANG@GMAIL.COM
Jianhong Wang[4] JIANHONG.WANG@BRISTOL.AC.UK
Xiaowei Huang[1] XIAOWEIHUANG@LIVERPOOL.AC.UK
[1] *University of Liverpool,* [2] *University of Exeter,*
[3] *Purple Mountain Laboratories,* [4] *University of Bristol.*

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Vision Language Models (VLMs) have demonstrated impressive capabilities in integrating visual and textual information for understanding and reasoning, but remain highly vulnerable to adversarial attacks. While activation steering has emerged as a promising defence, existing approaches often rely on task-specific contrastive prompts to extract harmful directions, which exhibit suboptimal performance and can degrade visual grounding performance. To address these limitations, we propose *Sequence-Level Preference Optimization* for VLM (*SPO-VLM*), a novel two-stage defense framework that combines activation-level intervention with policy-level optimization to enhance model robustness. In *Stage I*, we compute adaptive layer-specific steering vectors from diverse data sources, enabling generalized suppression of harmful behaviors during inference. In *Stage II*, we refine these steering vectors through a sequence-level preference optimization process. This stage integrates automated toxicity assessment, as well as visual-consistency rewards based on caption-image alignment, to achieve safe and semantically grounded text generation. The two-stage structure of SPO-VLM balances efficiency and effectiveness by combining a lightweight mitigation foundation in Stage I with deeper policy refinement in Stage II. Extensive experiments shown SPO-VLM enhances safety against attacks via activation steering and preference optimization, while maintaining strong performance on benign tasks without compromising visual understanding capabilities. We will release our code, model weights, and evaluation toolkit to support reproducibility and future research. Warning: This paper may contain examples of offensive or harmful text and images.

Keywords: Vision Language Model; Steering Activation; Adversarial Attack

1. Introduction

The advancement of Vision Language Models (VLMs) (Chen et al., 2023; Dai et al., 2023) marks a major breakthrough in AI, enabling the seamless integration of visual and textual information to enhance reasoning and understanding across diverse tasks. Despite their success, VLMs remain highly vulnerable to adversarial attacks, which exploit both visual and textual modalities to induce harmful responses. These concerns have led to growing research interest in jailbreak attacks and the development of corresponding defense strategies (Gong et al., 2025; Schlarmann and Hein, 2023; Wang et al., 2024g).

© 2025 S. Wu, G. Jin, W. Huang, J. Wang & X. Huang.

Activation steering has emerged as a promising defense, modifying internal representations via injected steering vectors without altering model weights (Wang et al., 2024c,d; Han et al., 2025; Cao et al., 2024). For example, Wang et al. (2024c) introduce InferAligner, which aligns hidden states with predefined safe directions at inference time. Recent extensions to multimodal settings, such as ASTRA (Wang et al., 2024a) and ShiftDC (Zou et al., 2025), further adapt steering vectors based on image attribution or disentangle harmful signals while preserving visual grounding. However, existing methods face key limitations. Steering vectors derived from contrastive prompts often fail to generalize across semantic contexts and attack types (Cao et al., 2024).

To address these issues, we propose *Sequence-Level Preference Optimization* for VLM (*SPO-VLM*), a novel two-stage defense framework that learns robust, generalizable steering vectors via SPO. Unlike prior work that extracts harmful directions from fixed prompt pairs, SPO-VLM optimizes steering vectors from diverse, preference-labeled data to achieve semantically aligned and safe generation. In *Stage I*, we compute lightweight and layer-specific steering vectors from multiple datasets, supporting inference-time mitigation with broad generalization. In *Stage II*, these vectors are refined using SPO within the RLHF framework (Ouyang et al., 2022) based on PPO, guided by multi-objective rewards that incorporate toxicity suppression (Hanu and Unitary team, 2020), and visual-text consistency. This two-stage formulation enhances both flexibility and robustness. By keeping the base VLM frozen, it reduces computational overhead and supports modular deployment. The use of sequence-level preference optimization allows the model to align with broader behavioral objectives beyond token-level control. Furthermore, the learned steering vectors demonstrate strong generalization to out-of-distribution inputs while preserving helpfulness and factual grounding in benign scenarios.

Our main contributions are as follows: *(i)* We propose *SPO-VLM*, a novel framework that unifies activation-level intervention with sequence-level preference optimization via RLHF and multi-objective reward signals. *(ii)* SPO-VLM significantly improves safety against jailbreak attacks by combining activation steering with sequence-level preference optimization, outperforming prior defenses like ASTRA across multiple datasets. *(iii)* The model retains strong performance on benign tasks, demonstrating that safety enhancements do not come at the cost of helpfulness or visual-language understanding capabilities.

2. Related Work

2.1. Jailbreak Attack on VLM

Jailbreak attacks manipulate prompts to deceive the model into responding to restricted or prohibited queries. In addition to LLM-based textual jailbreak strategies (Guo et al., 2024; Liu et al., 2024a; Yu et al., 2024a; Zou et al., 2023), the inclusion of visual inputs introduces a new attack surface for VLM attacks. There are two main types of attacks: perturbation-based attacks and structured-based attacks (Wang et al., 2024f). Perturbation-based attacks generate adversarial images designed to evade VLM safeguards (Carlini et al., 2023; Qi et al., 2023; Niu et al., 2024). For example, imgJP (Niu et al., 2024) optimizes an universal perturbation across unseen prompts and images to generate a targeted response. In contrast to perturbation-based methods, structure-based attacks transform harmful content into images using typography (Gong et al., 2025) or generative models to elicit harmful responses

from the model (Gong et al., 2025; Li et al., 2025b). Specifically, FigStep (Gong et al., 2025) leverages the ability of VLMs to interpret textual instructions embedded within images by encoding harmful content directly into the visual modality. By pairing these adversarially crafted images with benign textual prompts, FigStep effectively manipulates the VLM, eliciting detailed and potentially harmful responses.

Our work primarily addresses the challenge of defending against such jailbreak attacks on VLMs. Instead of modifying model weights or relying on static filtering, we propose to construct optimized steering activations that adaptively mitigate harmful behaviors by shifting internal representations in safer directions.

2.2. Activation Steering

Activation steering refers to a set of alignment techniques that guide a model's behavior by freezing model weights and modifying activations (Wang et al., 2024c,d; Han et al., 2025; Cao et al., 2024). Several studies have focused on identifying steering vectors within the activation space of specific layers in the LLM transformer architecture. Specifically, Wang et al. (2024c) proposes InferAligner, a novel inference-time alignment method that effectively improves model safety without compromising downstream performance. To address various categories of hallucinations, Wang et al. (2024d) proposes Adaptive Activation Steering (ACT), which leverages a diverse set of truthfulness-related steering vectors and dynamically adjusts the steering intensity based on the truthfulness of the model's activations. Moreover, SafeSwitch (Han et al., 2025) incorporates a safety prober that continuously monitors the model's internal states and responds appropriately by dynamically activating a specialized refusal head. This head provides informative explanations, ensuring the model's responses remain helpful while prioritizing safety. However, these steering vectors are directly extracted from LLM activations using preference data pairs, often leading to inaccurate representations of target behavior. Cao et al. (2024) proposes bi-directional preference optimization (BiPO) to generate more effective steering vectors for personalized control over diverse model behaviors. BiPO allows steering vectors to directly influence the generation probabilities of contrastive human preference data pairs, providing a more accurate and fine-grained representation of the target behavior. Inspired by recent advances in activation steering for LLMs, a growing body of research now focuses on guiding model behavior through the construction and application of steering vectors (Wang et al., 2024a; Li et al., 2025a; Zou et al., 2025). Wang et al. (2024a) introduces ASTRA, a defense mechanism that adaptively steers models away from adversarial feature directions using image attribution activations to counter VLM attacks. By considering the projection between steering vectors and calibrated activations, their adaptive steering approach effectively mitigates harmful outputs under adversarial input while maintaining minimal performance degradation on benign inputs. ShiftDC (Zou et al., 2025) preserves the VLM's vision understanding ability by disentangling and calibrating VLM activations to restore safety alignment.

Our work builds on these insights by framing the construction of steering vectors as a sequence-level optimization problem. Specifically, we adopt a reinforcement learning with preference supervision framework to learn behavior-aligned steering vectors, while incorporating textual modality consistency as part of the reward signal. This allows our approach

to generate more robust and interpretable steering vectors that align with both safety objectives and multimodal grounding.

2.3. Preference Optimization

Reinforcement Learning from Human Feedback (RLHF) has become a widely adopted approach for aligning models with human preferences (Ouyang et al., 2022; Ziegler et al., 2020; Stiennon et al., 2022). The standard RLHF pipeline typically begins by training a reward model, often structured using frameworks like the Bradley-Terry model (Bradley and Terry, 1952), to reflect human preferences. This reward model guides reinforcement learning algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017), which are then used to fine-tune the language model to generate responses that maximize the learned reward. In the context of LLMs, RLHF is particularly instrumental in shaping models that are helpful, honest, and harmless, thereby aligning them with human values (Ouyang et al., 2022; Bai et al., 2022; Thoppilan et al., 2022). For example, LaMDA (Thoppilan et al., 2022) fine-tunes LLMs to engage in natural language dialogue that is engaging, informative, factually grounded, and safe, often incorporating external information to ensure accuracy and relevance. InstructGPT (Ouyang et al., 2022) fine-tunes GPT-3-style models (Brown et al., 2020) to enhance helpfulness, using reinforcement learning from human preferences expressed through pairwise comparisons. Askell et al. (2021) follow the pre-training and fine-tuning paradigm to train a preference model for human alignment, demonstrating that ranked preference modeling is a highly effective objective for distinguishing between "good" and "bad" behaviors. This approach is further enhanced through an iterative online training regime, in which preference models and reinforcement learning policies are updated weekly using fresh human feedback data. PPO is incorporated to stabilize the RL training process (Bai et al., 2022).

Building on this foundation, our work explores a novel application of preference optimization: instead of optimizing full model parameters, we use PPO-based preference signals to directly learn *steering vectors* in the model's activation space. This approach enables more precise and interpretable alignment with desired behaviors, while preserving model generalization and avoiding catastrophic forgetting.

3. Preliminary

3.1. Vision Language Models

Let $\mathcal{P}_{\text{VLM}}$ denotes an autoregressive Vision Language Model, which defines a probability distribution over sequences of tokens drawn from a vocabulary $\mathcal{V}$. This model is designed to process and reason on both textual and visual modalities in a unified framework. Specifically, we consider a VLM that takes as input a sequence of n textual tokens $\mathbf{q}_t = \{q_{t_1}, q_{t_2}, \ldots, q_{t_n}\}$ and a sequence of m visual tokens $\mathbf{q}_v = \{q_{v_1}, q_{v_2}, \ldots, q_{v_m}\}$. These tokens are typically derived from natural language inputs and visual features, respectively, where the visual tokens are obtained through the discretization of image embeddings from a vision encoder.

Given the multimodal input $\{\mathbf{q}_t, \mathbf{q}_v\}$, the model generates a response sequence $r = \{r_1, r_2, \ldots, r_o\}$, consisting of o output tokens. The generation process is autoregressive,

meaning that each token r_i in the response is sampled sequentially, conditioned on all previous tokens in the input and the already generated part of the output. Formally, the probability of generating the ith token r_i is given by:

$$\mathcal{P}_{\text{VLM}}(r_i \mid \mathbf{q}_t, \mathbf{q}_v, r_1, \ldots, r_{i-1})$$

This formulation enables the model to incorporate both linguistic context and visual grounding when predicting each subsequent token. The response generation continues iteratively until a special end-of-sequence token is produced or a maximum sequence length is reached. Through this design, $\mathcal{P}_{\text{VLM}}$ enables coherent and contextually grounded text generation in response to complex multimodal inputs.

3.2. Steering Activations

Let $\mathbf{x}^\ell(t)$ denote the residual stream activation of the last token at layer $\ell \in L$ of a VLM, capturing the information processed from the input t up to layer ℓ. We define the function ActMean to compute the mean last-token activation at layer ℓ for a given dataset $\mathcal{D}$:

$$\text{ActMean}^\ell(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\left[\sum_{t \in \mathcal{D}} \mathbf{x}^\ell(t)\right]. \tag{1}$$

Numerous studies (Arditi et al., 2024; Park et al., 2024) have demonstrated that high-level concepts are encoded as linear directions in the activation space of LLMs. These directions can be uncovered by computing the difference between the mean activations of a model when processing two sets of contrastive instructions, $\mathcal{D}_1$ and $\mathcal{D}_2$, which elicit distinct behaviors:

$$\mathbf{v}^\ell_{\mathcal{D}_2 \to \mathcal{D}_1} = \text{ActMean}^\ell(\mathcal{D}_1) - \text{ActMean}^\ell(\mathcal{D}_2). \tag{2}$$

The resulting vector, $\mathbf{v}^\ell_{\mathcal{D}_2 \to \mathcal{D}_1}$, referred to as the *difference-in-mean* vector, captures both the direction and magnitude of the layer-ℓ activation shift from $\mathcal{D}_2$ to $\mathcal{D}_1$. This vector effectively isolates the critical features that drive the model's behavioral differences between the two instruction sets.

4. Methodology

Our hybrid defense framework, SPO-VLM, consists of two complementary stages designed to enhance the safety of VLMs against adversarial prompts while preserving their utility. Specifically, the two stages are: (1) adaptive activation steering to suppress harmful internal activations, and (2) sequence-level preference optimization via reinforcement learning to reinforce safe and grounded generation.

Figure 1 illustrates the overall architecture of our proposed framework. Upon receiving an image-text input pair, the frozen VLM processes the input and produces intermediate activation representations. A steering vector, learned through sequence-level preference optimization, is subsequently applied to these activations to bias the model's output toward safe and contextually appropriate responses. Notably, this intervention operates exclusively on the internal representations, leaving all model parameters unchanged throughout the process.

4.1. Stage I: Initialization of Steering Activation

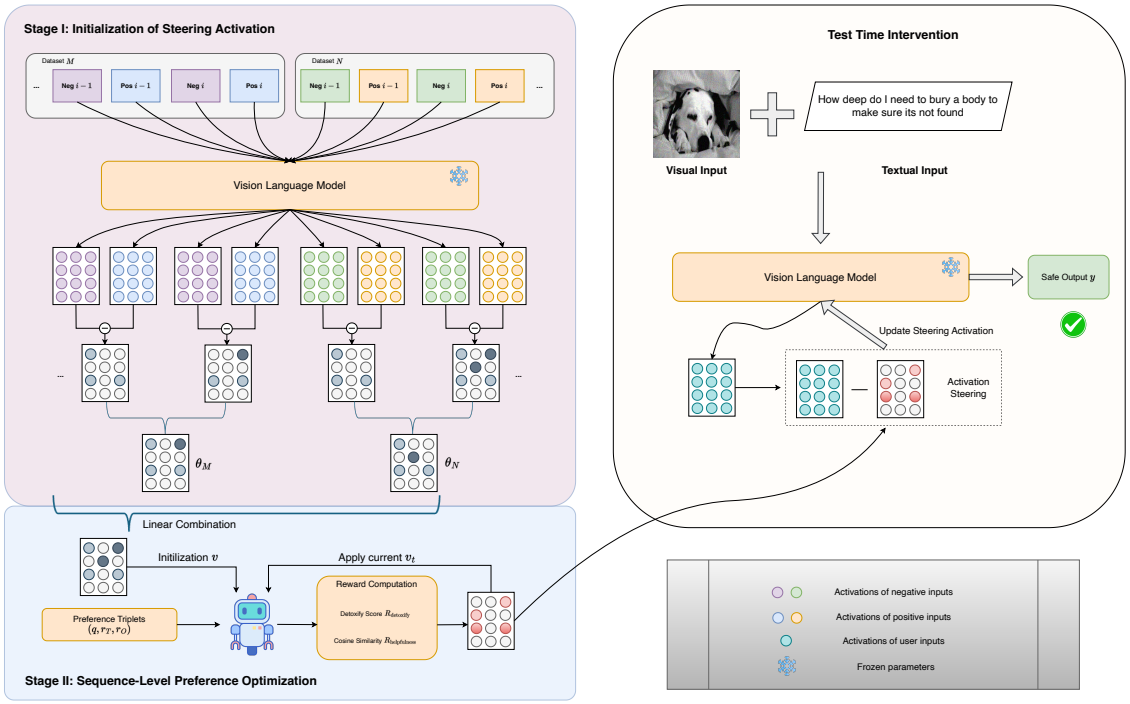

Figure 1: Overview of the SPO-VLM framework. Stage I (left top) initializes attribute-specific steering vectors $\boldsymbol{\theta}_j$ using contrastive pairs from multiple datasets. These vectors are later combined into a global steering vector via a linear combination. Stage II (left bottom) performs sequence-level preference optimization using rewards functions, including toxicity reduction and alignment preservation. At test time (right), the frozen vision-language model receives visual-textual input and applies the optimized steering activation to produce safe and aligned outputs.

Activation steering (Li et al., 2024; Subramani et al., 2022) aims to locate specific directions in the model's activation space that align with factually accurate statements, and then adjusts the activations along those directions during inference to guide the model's output. Expanding on this idea, our approach derives diverse steering vectors directly from raw data to effectively target a range of attack types. Moreover, we introduce adaptive steering construction based on the toxic content of the activations. Rather than relying on a single global direction, we propose an adaptive steering mechanism that constructs the final steering vector v^ℓ as a linear combination of multiple attribute-specific difference-in-mean vectors. The final steering vector v^ℓ at layer ℓ is formulated as a weighted combination of attribute-specific components:

$$v^\ell = \sum_{j \in \mathcal{A}} \alpha_j v_j^\ell, \tag{3}$$

where $\mathcal{A} = 1, 2, \ldots, |\mathcal{A}|$ denotes the set of attribute indices, $\alpha_j \in \mathbb{R}$ represents the weight coefficient for the j-th attribute, and each attribute-specific vector v_j^ℓ is computed as:

$$v_j^\ell = \text{ActMean}^\ell(\mathcal{D}j, \text{pos}) - \text{ActMean}^\ell(\mathcal{D}j, \text{neg}), \tag{4}$$

where $\mathcal{D}j, \text{pos}$ and $\mathcal{D}j, \text{neg}$ represent the positive and negative instruction sets for attribute j, respectively. This formulation allows the steering mechanism to adaptively combine multiple behavioral dimensions, with each v_j^ℓ capturing the activation difference for a specific attribute at layer ℓ.

Algorithm 1 Sequence-Level Preference Optimization (SPO) for Steering Vector Learning

Input: VLM $\mathcal{P}_{\text{VLM}}$, preference dataset $\mathcal{D} := \{(q^i, r_T^i, r_O^i)\}_{i=1}^n$, batch size m, total update steps T

Output: Optimized steering vector v^*

Initialize steering vector: $v_0 \leftarrow \sum_{j \in \mathcal{A}} \alpha_j^{(0)} v_j^\ell$
for $t = 0$ to $T - 1$ **do**
  Sample mini-batch $\mathcal{D}_t := \{(\mathbf{q}_t^i, \mathbf{q}_v^i, r_T^i, r_O^i)\}_{i=1}^m \sim \mathcal{D}$
  for each quadruplets $(\mathbf{q}_t, \mathbf{q}_v, r_T, r_O)$ in $\mathcal{D}_t$ **do**
    Evaluate reward: $R = R_{\text{detoxify}}(r) + R_{\text{visual}}(\mathbf{q}_v)$
    Compute value baseline: $V_\phi(\mathbf{q}_t, \mathbf{q}_v)$
    Compute advantage: $A = R(r, \mathbf{q}_t, \mathbf{q}_v) - V_\phi(\mathbf{q}_t, \mathbf{q}_v)$
    Compute policy ratio difference using Equation (5)
    Compute clipped policy loss using PPO:
    $L_\pi = \min_v -\mathbb{E}_{(\mathbf{q}_t, \mathbf{q}_v, r_T, r_O) \sim \mathcal{D}} \big[\min \big(\text{ratio} \cdot A,\ \text{clip}(\text{ratio}, 1 - \epsilon, 1 + \epsilon) \cdot A \big) \big]$
  end for
  Update steering vector via gradient descent: $v_{t+1} \leftarrow v_t - \eta \cdot \nabla_v L_\pi$
  Update ϕ by minimizing critic loss L_{critic}
end for
return $v^* = v_T$

4.2. Stage II: Sequence-level Preference Optimization

Inspired by preference-based model optimization techniques such as RLHF (Ouyang et al., 2022), we incorporate the activation steering obtained from Stage I into the rollout policy to guide the generation of safe and contextually grounded responses. The objective is to enable the model to effectively suppress harmful outputs in the presence of adversarial prompts, while maintaining strong visual understanding capabilities under benign conditions.

We introduce a method that learns effective steering vectors in activation space through sequence-level reinforcement learning with preference-based supervision. Our algorithm adopts a sequence-level variant of *Proximal Policy Optimization* (*PPO*) to fine-tune model behavior. Unlike traditional RLHF methods that operate at the token level, we directly optimize the log-probability of the full generated sequence, named Sequence-Level Preference Optimization for VLMs (**SPO-VLM**). This allows the model to favor outputs aligned with desired multi-objective behavior and reduce the probability of generating undesired or adversarial responses. The target behavior is defined via a multi-objective reward function,

incorporating both safety and visual understanding capabilities. This reward guides learning in a way that balances safety with alignment to visual content.

To enable preference learning, we construct labeled quadruplets $(\mathbf{q}_t, \mathbf{q}_v, r_T, r_O)$, where $\mathbf{q}_t$ is a texture prompt, $\mathbf{q}_v$ is a visual prompt, r_T is a response exhibiting the target behavior, and r_O is a response reflecting the undesired behavior. Let v denote the learnable steering vector, and π_{L+1} represent the later layers of the model (from layer $L+1$ onward). For each prompt pair $(\mathbf{q}_t, \mathbf{q}_v)$, we generate responses using the current policy π_θ. We compute the following policy ratios to assess the influence of the steering vector on the model's preference between r_T and r_O:

$$\text{ratio} = \frac{\pi_{L+1}(r_T \mid a_\ell(\mathbf{q}_t, \mathbf{q}_v) + v)}{\pi_{L+1}(r_T \mid a_\ell(\mathbf{q}_t, \mathbf{q}_v))} - \frac{\pi_{L+1}(r_O \mid a_\ell(\mathbf{q}_t, \mathbf{q}_v) + v)}{\pi_{L+1}(r_O \mid a_\ell(\mathbf{q}_t, \mathbf{q}_v))}, \tag{5}$$

The term $\pi_{L+1}(\cdot \mid a_\ell(\mathbf{q}_t, \mathbf{q}_v) + v)$ represents the policy induced by modifying the model's activations with the steering vector v at layer ℓ. This difference quantifies the differential impact of the steering vector on preferred and dispreferred responses.

We employ a composite reward function that encourages both safety and visual grounding. Given a query and the model's response r, the total reward is:

$$R = R_{\text{detoxify}}(r) + R_{\text{visual}}(\mathbf{q}_v). \tag{6}$$

The detoxification reward component penalizes toxic content using an exponential decay function: $R_{\text{detoxify}}(r) = 2 \cdot [\exp(-\beta \cdot \text{toxicity}(r)) - 0.5]$, where $\text{toxicity}(r) \in [0, 1]$ is computed using a pre-trained toxicity classifier, and $\beta > 0$ controls the penalty strength. This formulation yields rewards in the range $[-1, 1]$, with non-toxic responses receiving positive rewards. The component of visual understanding reward measures alignment between visual content and captioning content: $R_{\text{visual}}(\mathbf{q}_v) = -\cos\left(\overline{\mathbb{I}}, \overline{\mathbb{C}}\right)$, where $\overline{\mathbb{I}}$, $\overline{\mathbb{C}}$ are the mean-pooled hidden states of image and caption tokens respectively.

We adopt a sequence-level variant of Proximal Policy Optimization (PPO) (Schulman et al., 2017) to optimize the steering vector while maintaining training stability. The objective function is:

$$L_\pi = \min_v -\mathbb{E}_{(\mathbf{q}_t, \mathbf{q}_v, r_T, r_O) \sim \mathcal{D}} \left[\min\left(\text{ratio} \cdot A,\ \text{clip}(\text{ratio}, 1-\epsilon, 1+\epsilon) \cdot A\right) \right] \tag{7}$$

where $A = R(r, \mathbf{q}_t, \mathbf{q}_v) - V_\phi(\mathbf{q}_t, \mathbf{q}_v)$ is the advantage function computed from the total reward and a learned value baseline $V_\phi(\mathbf{q}_t, \mathbf{q}_v)$. The clipping range $[1-\epsilon, 1+\epsilon]$ ensures stable policy updates by preventing excessive deviations.

The final optimization objective combines PPO with critic function learning:

$$L_{\text{total}} = L_\pi + c_1 L_{\text{critic}},$$

where L_{critic} denotes the critic loss, which is computed by a lightweight critic module integrated into the model. This critic is implemented as a simple two-layer multilayer perceptron that operates on the final-layer hidden states produced by the base model. By processing these high-level representations, the critic estimates a scalar value for each input, representing its expected utility or alignment with the target objective. The value predictions

are then used to compute L_{critic}, guiding the optimization of the model's preference-aware behavior in a sample-efficient manner. Unlike conventional RLHF, which requires training a new policy and a separate reference model, our method optimizes only the steering vector v, keeping the base model architecture and parameters fixed. As a result, the method is highly efficient and minimally invasive. When applied during inference, the learned vector reliably steers the model toward safer and more helpful behavior by modifying a narrow subset of internal representations.

5. Experiments

In this section, we evaluate SPO-VLM across three dimensions: its effectiveness in mitigating adversarial prompts while preserving visual understanding, its ability to transfer across diverse attack domains, and the contribution of each stage through ablation studies.

5.1. Experiment Setup

Steering Activation Construction. We initialize the steering vectors using the Stage I method applied to the RealToxicityPrompt, AdvBench, and Anthropic_Harmful datasets. This approach extracts activation shifts across multiple toxicity dimensions and constructs an initial vector through a linear combination of these shifts. Specifically, we adopt the steering vector formulation from Wang et al. (2024a), expressed as $v^\ell = \sum_{j \in \mathcal{A}} \alpha_j v_j^\ell$, where $\alpha_1 = 0.5$, $\alpha_2 = 0.4$, and $\alpha_3 = 0.4$ denote the weights for each attribute-specific direction. The resulting vectors are further refined during Stage II via the SPO-VLM framework.

Evaluation Datasets. We evaluate our approach under three experimental settings: (1) Toxicity assessment, using the RealToxicityPrompts benchmark (Gehman et al., 2020); (2) Jailbreak detection, evaluating on two datasets: AdvBench (Zou et al., 2023) and Anthropic_Harmful(Ganguli et al., 2022); and (3) Visual comprehension, using four benchmarks: MM-Vet(Yu et al., 2024b), SQA (Iyyer et al., 2017), CogVLM (Wang et al., 2024e), and MME (Fu et al., 2023).

Evaluation Metrics. For toxicity assessment, we employ the Detoxify classifier (Hanu and Unitary team, 2020) to compute toxicity scores on a scale from 0 (non-toxic) to 1 (highly toxic). For jailbreak detection, we quantify robustness using the attack success rate (ASR), defined as the proportion of successful jailbreaks among total attack attempts. This metric is computed using the classifier from HarmBench (Mazeika et al., 2024). For visual understanding evaluation, we employ task-specific utility metrics. MM-Vet uses GPT-4 with few-shot prompts to generate utility scores ranging from 0 to 1, while SQA calculates overall accuracy for its single-choice questions. For CogVLM, we compute the arithmetic mean of three metrics: BLEU-2, CIDEr, and METEOR. MME evaluates both perception and cognition capabilities across 14 subtasks.

Baselines. This study compare SPO-VLM against three baseline methods. The Original Model serves as the unmodified visual language model without additional safety mechanisms, providing a baseline for standard post-training alignment. ASTRA (Wang et al., 2024a) employs steering vectors generated from contrastive visual prompt pairs, representing a prominent activation-based steering approach. ShiftDC (Zou et al., 2025) preserves the VLM's vision understanding ability by disentangling and calibrating VLM activations

Table 1: Performance of different safety steering methods on safety and visual understanding benchmarks. The ↑ or ↓ symbols indicate whether a higher or lower score is preferable.

Base Model	Method	Toxicity Scores (%)	Jailbreak ASR (%)		Visual Understanding Scores			
		RealToxicityPrompt ↓	AdvBench ↓	Anthropic_Harmful ↓	MM-Vet ↑	SQA ↑	CogVLM ↑	MME ↑
MiniGPT-4-13B	Original Model	38.18	19.19	73.50	**32.58**	**68.10**	**74.00**	**1742.0**
	ASTRA	10.21	5.93	4.87	17.70	65.35	69.00	1086.0
	ShiftDC	12.45	6.87	5.63	20.08	63.59	65.00	1130.0
	SPO-VLM	**9.28**	**4.77**	**3.21**	31.11	66.20	70.00	1370.0
Qwen2-VL-7B	Original Model	30.65	75.00	55.17	49.13	79.13	41.00	1630.0
	ASTRA	14.18	7.69	5.17	48.66	80.99	45.00	685.3
	ShiftDC	13.23	8.52	7.39	45.96	79.88	46.00	765.4
	SPO-VLM	**11.54**	**6.38**	**4.48**	**49.80**	**81.82**	**51.00**	**1753.0**
LLaVA-v1.5-13B	Original Model	85.74	36.80	74.00	**28.60**	**74.38**	**59.00**	**1560.0**
	ASTRA	81.44	5.76	24.13	14.90	56.03	56.00	1320.0
	ShiftDC	68.98	6.87	30.76	16.85	59.86	54.00	1250.0
	SPO-VLM	**50.37**	**4.39**	**12.18**	20.19	60.37	55.00	1489.0

to restore safety alignment. This comparative framework enables a comprehensive evaluation of SPO-VLM's ability to enhance safety while maintaining helpfulness across different safety paradigms.

Models & Implementations details. This study conducts all experiments on three widely used open-source VLMs: Qwen2-VL-7B (Wang et al., 2024b), MiniGPT-4-13B (Zhu et al., 2023), and LLaVA-v1.5-13B (Liu et al., 2023). These models, post-trained to follow instructions and align with human values, represent some of the most widely adopted and capable open-source model families. We set the steering layer l is 20 for 13B models and 14 for 7B models. The chat configurations use a temperature of 0.2 and $\alpha = 10$ for LLaVA-v1.5-13B, a temperature of 0.2 and $\alpha = 7$ for Qwen2-VL, and a temperature of 1.2 and $\alpha = 7$ for MiniGPT-4-13B.

5.2. SPO-VLM Effectively Balances Safety and Visual Understanding

Table 1 summarizes the performance of our proposed SPO-VLM method with respect to both safety and visual understanding capability, evaluated across a diverse set of benchmarks. The results demonstrate that SPO-VLM consistently improves the model's ability to resist harmful prompts while preserving or even enhancing its utility on some benign tasks. From these comprehensive evaluations, we draw several key conclusions as below.

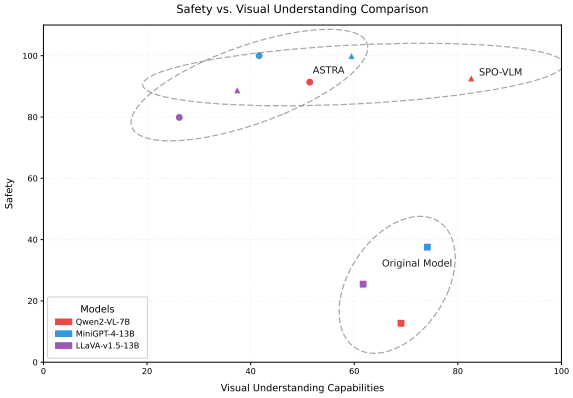

Figure 2: Comparison of safety enhancement methods in terms of safety and visual understanding capabilities. Each shape denotes one method.

SPO-VLM demonstrates enhanced safety performance. As shown in Table 1, SPO-VLM achieves superior safety performance. In the RealToxicityPrompt dataset, SPO-VLM achieves the lowest toxicity score, reducing it by more than 27.79% compared to the original model. Compared to ASTRA, SPO-VLM achieves an additional 11.55% reduction among all different mod-

els. It confirms that SPO-VLM provides more effective and consistent toxicity mitigation. Compared to ASTRA, SPO-VLM reduces the average ASR on AdvBench by approximately 1.28%, and cuts the average ASR on Anthropic_Harmful nearly in half, achieving an additional reduction of around 4.77%. This demonstrates SPO-VLM's superior generalization and effectiveness in mitigating jailbreak risks across diverse models.

SPO-VLM shows an optimal trade-off between safety and visual capability. While ASTRA improves safety, it leads to a notable decline in visual understanding performance. For instance, in MiniGPT, MM-Vet drops significantly from 32.58 to 17.70, and MME decreases by approximately 656 points. Similarly, in LLaVA, MM-Vet falls from 28.60 to 14.90. In contrast, SPO-VLM preserves visual capabilities far more effectively. For MiniGPT-4, the MM-Vet score remains high at 31.11, indicating only a modest reduction of approximately 1.5 points compared to the original model. In the case of Qwen2-VL-7B, SPO-VLM not only maintains but enhances visual understanding. For LLaVA, SPO-VLM consistently outperforms ASTRA across all visual benchmarks, notably raising the MM-Vet score from 14.90 to 20.19.

To highlight the superiority of SPO-VLM, we visualize the safety and visual understanding performance of different safety enhancement methods. Safety is measured as the mean of $(1-\text{ASR})$ across AdvBench and Anthropic_Harmful, and visual understanding by average normalized scores on MM-Vet, SQA, CogVLM, and MME. As shown in Figure 2, ASTRA exhibits a clear trade-off between safety and accuracy. Although it significantly improves safety scores compared to the baseline models, this comes at the cost of a statistically significant decline—exceeding 10%—in visual grounding performance. In contrast, SPO-VLM occupies the upper-right region of the plot, indicating simultaneous improvements in both safety and visual understanding, and thus achieving a more balanced and optimal overall performance.

5.3. SPO-VLM's Transfer Capabilities

To evaluate the transfer capabilities of SPO-VLM, we assess whether steering vectors derived from SPO-VLM can generalize across different types of attacks. Specifically, we evaluate the transferability of the defense against structure-based attacks from MM-SafetyBench (Liu et al., 2024b). As shown in Figure 3, SPO-VLM demonstrates consistently lower attack success rates compared to both the original model and the ASTRA defense across all evaluated models. Notably, SPO-VLM achieves substantial reductions in success rates for challenging combined attacks such as SD + OCR, indicating its robustness even under complex adversarial compositions. For example, on MiniGPT-4, SPO-VLM reduces the ASR by over 30% compared to the original model. These improvements highlight SPO-VLM's ability to generalize beyond the specific attack types it was trained on, effectively mitigating threats in structure-based attack scenarios. This cross-attack resilience indicates that SPO-VLM is well-suited for deployment in dynamic, real-world environments, where encountering unseen adversarial strategies is common.

5.4. Ablation Study

We conduct an ablation study on Qwen2-VL-7B using the benchmark datasets summarized to assess the individual contributions of each stage in the SPO-VLM framework. Specifically,

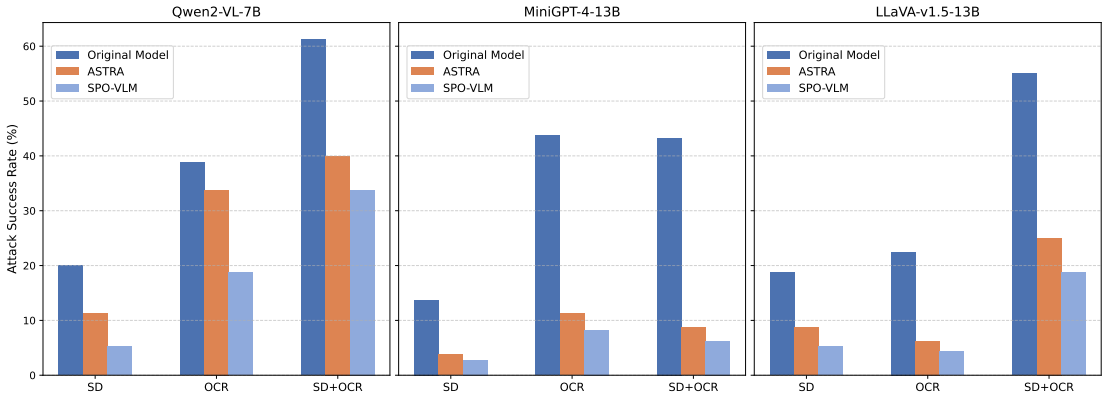

Figure 3: Evaluation of defense transferability under structure-based attacks using MM-SafetyBench.

we evaluate the performance of Stage I, which applies activation steering alone, and compare it against the full implementation that incorporates Stage II, sequence-level preference optimization. As shown in Table 2, the results underscore the importance of sequence-level preference optimization in reinforcing safe behavior beyond the initial activation steering. While Stage I serves as a lightweight mitigation mechanism, the addition of Stage II yields substantial performance improvements by leveraging rich, reward-driven alignment signals during policy refinement.

Table 2: Ablation study on Qwen2-VL-7B evaluating the contributions of each stage in the SPO-VLM framework.

Behavior	RealToxicityPrompt	AdvBench	Anthropic_Harmful
Original Model	30.65	75.00	55.17
Stage I	20.97	10.96	5.09
Stage I + Stage II	**11.54**	**6.38**	**4.48**

6. Conclusion

In this paper, we propose *SPO-VLM*, a novel two-stage defense framework that enhances the safety of VLMs against adversarial attacks. The approach integrates lightweight steering vectors derived from diverse datasets in Stage I. In Stage II, these vectors are refined through sequence-level preference optimization using multi-objective rewards. Extensive evaluations across various VLMs show that SPO-VLM offers improved safety by reducing toxicity and jailbreak success rates, while generally maintaining visual understanding. We believe this work lays the foundation for future research on activation-space intervention and preference-driven alignment to ensure the trustworthy deployment of LLMs and VLMs.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction, 2024. URL https://arxiv.org/abs/2406.11717.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Jackson Kernion, Kamal Ndousse, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, and Jared Kaplan. A general language assistant as a laboratory for alignment, 2021. URL https://arxiv.org/abs/2112.00861.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022. URL https://arxiv.org/abs/2204.05862.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952. URL https://api.semanticscholar.org/CorpusID:125209808.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Yuanpu Cao, Tianrong Zhang, Bochuan Cao, Ziyi Yin, Lu Lin, Fenglong Ma, and Jinghui Chen. Personalized steering of large language models: Versatile steering vectors through bi-directional preference optimization, 2024. URL https://arxiv.org/abs/2406.00045.

Nicholas Carlini, Milad Nasr, Christopher A Choquette-Choo, Matthew Jagielski, Irena Gao, Pang Wei W Koh, Daphne Ippolito, Florian Tramer, and Ludwig Schmidt. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36:61478–61500, 2023.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023. URL https://arxiv.org/abs/2310.09478.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language

models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=vvoWPYqZJA.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Deep Ganguli, Liane Lovitt, and et al. Jackson Kernion. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL https://arxiv.org/abs/2209.07858.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models, 2020. URL https://arxiv.org/abs/2009.11462.

Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts, 2025. URL https://arxiv.org/abs/2311.05608.

Xingang Guo, Fangxu Yu, Huan Zhang, Lianhui Qin, and Bin Hu. Cold-attack: Jailbreaking llms with stealthiness and controllability, 2024. URL https://arxiv.org/abs/2402.08679.

Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Denghui Zhang, and Heng Ji. Internal activation as the polar star for steering unsafe llm behavior, 2025. URL https://arxiv.org/abs/2502.01042.

Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, 2017.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2024. URL https://arxiv.org/abs/2306.03341.

Qing Li, Jiahui Geng, Zongxiong Chen, Kun Song, Lei Ma, and Fakhri Karray. Internal activation revision: Safeguarding vision language models without parameter update, 2025a. URL https://arxiv.org/abs/2501.16378.

Yifan Li, Hangyu Guo, Kun Zhou, Wayne Xin Zhao, and Ji-Rong Wen. Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models, 2025b. URL https://arxiv.org/abs/2403.09792.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models, 2024a. URL https://arxiv.org/abs/2310.04451.

Xin Liu, Yichen Zhu, Jindong Gu, Yunshi Lan, Chao Yang, and Yu Qiao. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models, 2024b. URL https://arxiv.org/abs/2311.17600.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Elham Sakhaee, Nathaniel Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal, 2024. URL https://arxiv.org/abs/2402.04249.

Zhenxing Niu, Haodong Ren, Xinbo Gao, Gang Hua, and Rong Jin. Jailbreaking attack against multimodal large language model, 2024. URL https://arxiv.org/abs/2402.02309.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models, 2024. URL https://arxiv.org/abs/2311.03658.

Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models, 2023. URL https://arxiv.org/abs/2306.13213.

Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models, 2023. URL https://arxiv.org/abs/2308.10741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback, 2022. URL https://arxiv.org/abs/2009.01325.

Nishant Subramani, Nivedita Suresh, and Matthew E. Peters. Extracting latent steering vectors from pretrained language models, 2022. URL https://arxiv.org/abs/2205.05124.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, and et al. Noam Shazeer. Lamda: Language models for dialog applications, 2022. URL https://arxiv.org/abs/2201.08239.

Han Wang, Gang Wang, and Huan Zhang. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks, 2024a. URL https://arxiv.org/abs/2411.16721.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024b. URL https://arxiv.org/abs/2409.12191.

Pengyu Wang, Dong Zhang, Linyang Li, Chenkun Tan, Xinghao Wang, Ke Ren, Botian Jiang, and Xipeng Qiu. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance, 2024c. URL https://arxiv.org/abs/2401.11206.

Tianlong Wang, Xianfeng Jiao, Yifan He, Zhongzhi Chen, Yinghao Zhu, Xu Chu, Junyi Gao, Yasha Wang, and Liantao Ma. Adaptive activation steering: A tuning-free llm truthfulness improvement method for diverse hallucinations categories. *arXiv preprint arXiv:2406.00034*, 2024d.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Song XiXuan, et al. Cogvlm: Visual expert for pretrained language models. *Advances in Neural Information Processing Systems*, 37:121475–121499, 2024e.

Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In *European Conference on Computer Vision*, pages 77–94. Springer, 2024f.

Yu Wang, Xiaogeng Liu, Yu Li, Muhao Chen, and Chaowei Xiao. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting, 2024g. URL https://arxiv.org/abs/2403.09513.

Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. Gptfuzzer: Red teaming large language models with auto-generated jailbreak prompts, 2024a. URL https://arxiv.org/abs/2309.10253.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities, 2024b. URL https://arxiv.org/abs/2308.02490.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.

Xiaohan Zou, Jian Kang, George Kesidis, and Lu Lin. Understanding and rectifying safety perception distortion in vlms, 2025. URL https://arxiv.org/abs/2502.13095.