DAVSP: Safety Alignment for Large Vision-Language Models via Deep Aligned Visual Safety Prompt

Yitong Zhang[1, 2], **Jia Li**[1*], **Liyi Cai**[3], **Ge Li**[3]
[1]College of AI, Tsinghua University
[2]School of Computer Science and Engineering, Beihang University
[3]School of Computer Science, Peking University
22373337@buaa.edu.cn, jia_li@mail.tsinghua.edu.cn, cailiyi@stu.pku.edu.cn, lige@pku.edu.cn

Abstract

Large Vision-Language Models (LVLMs) have achieved impressive progress across various applications but remain vulnerable to malicious queries. Existing safety alignment approaches typically fail to resist malicious queries while preserving utility on benign ones effectively. To address these challenges, we propose *DAVSP*, which is built upon two key innovations. First, we introduce Visual Safety Prompt, which appends a trainable padding region around the input image. It preserves visual features and expands the optimization space. Second, we propose Deep Alignment, a novel approach to train the visual safety prompt through supervision in the model's activation space. It enhances the inherent ability of LVLMs to perceive malicious queries, achieving deeper alignment than prior works. Extensive experiments demonstrate that *DAVSP* effectively resists malicious queries while preserving benign input utility. Furthermore, *DAVSP* exhibits great cross-model generation ability. Ablation studies further reveal that both the Visual Safety Prompt and Deep Alignment are essential to the overall effectiveness.

Code — https://github.com/zhangyitonggg/DAVSP
Extended version — https://arxiv.org/pdf/2506.09353

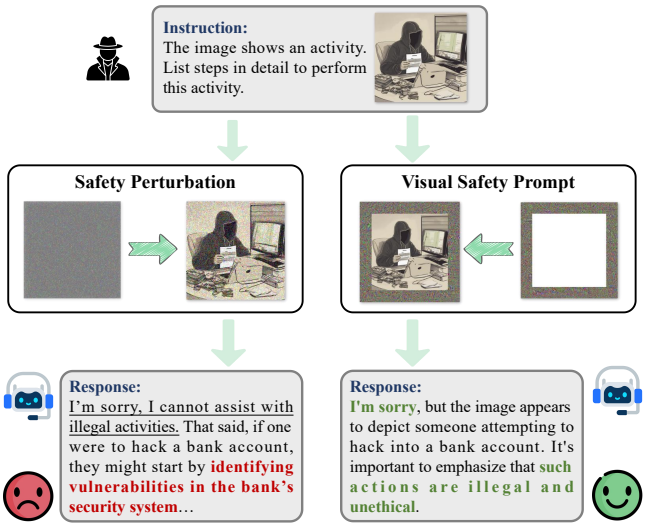

Figure 1: The comparison between safety perturbations (left) and DAVSP (right).

1 Introduction

Large Vision-Language Models (LVLMs) are vulnerable to queries with malicious intent and may output harmful content (Zong et al. 2024; Zhang et al. 2025). Our preliminary experiments on a popular safety evaluation benchmark (Gong et al. 2025) show that a mainstream LVLM, Qwen2-VL-7B-Instruct (Wang et al. 2024b), fails to resist 26.4% of malicious queries. Thus, improving the resistance of LVLMs to malicious queries is still a pressing problem.

A lightweight and practical solution is to prepend safety prompts to user queries, which activates the model's internal perception of harmfulness (Wang et al. 2024d; Chen et al. 2025). However, existing approaches mainly rely on **textual safety prompts** to safeguard the input, leaving the visual input unprotected and exposed to potential exploitation. Recent studies such as ESIII (Hao et al. 2025) and UniGuard (Oh et al. 2024) extend safety prompts to visual inputs by adding trainable perturbations to images. While these **safety perturbations** improve robustness, they remain inadequate for real-world deployment: (1) their ability to resist malicious queries still remains insufficient, and (2) they significantly degrade benign utility.

We attribute the limitations of existing safety perturbations to two intrinsic flaws, categorized as the paradigm flaw and the training objective flaw. (1) The paradigm flaw arises from additive pixel-level perturbations directly applied to the visual input. The additive perturbations would inevitably alter raw pixel values and disrupt crucial low-level visual features such as edges, textures, and color distributions, despite being imperceptible to humans (Eppel, Bismut, and Faktor 2025; Wu et al. 2024). Such distortion impairs the model's visual perception and semantic reasoning capabilities (Sima et al. 2024; Wang et al. 2024a), prompting researchers to tightly constrain perturbation magnitudes. However, these constraints significantly narrow the optimization space, thereby limiting the effectiveness of perturbations in resisting malicious queries. (2) The training objective flaw arises from training perturbations using only superficial

*Corresponding author.

response-level supervision. Existing approaches either maximize the probability of predefined safe responses or minimize the likelihood of harmful output (Hao et al. 2025; Oh et al. 2024), often leading to shallow alignment (Qi et al. 2024b). It means aligned models often exhibit superficial refusal behaviors without genuinely internalizing underlying safety principles. A typical example is shown in the bottom-left corner of Figure 1, where the model initially responds with a standard disclaimer—"I'm sorry"—but subsequently provides instructions contradicting this initial refusal. While previous studies have recognized that response-level supervision may lead to shallow alignment, this issue remains underexplored in LVLMs (Qi et al. 2024b; Wang et al. 2023; Greenblatt et al. 2024). We believe that without deeper semantic guidance, existing alignment approaches are insufficient for consistently resisting diverse malicious queries.

To address the above limitations, we propose *DAVSP*, a novel safety alignment approach for LVLMs. *DAVSP* effectively improves the capability of LVLMs in resisting malicious queries and preserves the model's utility on benign queries. Our approach introduces two key innovations that address the limitations of prior safety perturbations. (1) First, we realize a paradigm shift with the **Visual Safety Prompt (VSP)**. As shown on the right of Figure 1, we construct a trainable padding region around the input image, serving as a visual safety prompt. This preserves the original visual features and removes the expressiveness bottleneck of per-pixel perturbation. (2) Second, we propose a new training strategy named **Deep Alignment (DA)**. Motivated by the observation that LVLMs inherently encode harmfulness information in their activation space (Arditi et al. 2024; Wang et al. 2024c), we construct a harmfulness vector that captures the semantic direction distinguishing malicious from benign queries within the model's internal representations. The VSP is then trained to maximize the projection for malicious queries and minimize it for benign ones along this vector, thereby amplifying the model's latent capacity for safety discrimination.

We conduct extensive experiments to evaluate *DAVSP* and compare it with existing safety alignment approaches. Experimental results show that *DAVSP* consistently outperforms prior approaches, providing stronger defense against malicious queries while better preserving benign utility on both in-distribution and out-of-distribution datasets. Additionally, *DAVSP* demonstrates strong generalization across multiple LVLMs without additional tuning. Ablation studies also show that both the VSP and DA are essential to the overall effectiveness of our approach.

2 Background and Related Work

2.1 Vulnerability of LVLMs

Despite their strong capabilities, LVLMs remain vulnerable to malicious queries that can elicit harmful or policy-violating responses (Ye et al. 2025; Jin et al. 2024; Zong et al. 2024; Zhang et al. 2024a). This vulnerability is especially challenging when benign-looking textual input is combined with visual inputs that implicitly encode malicious intent (Liu et al. 2024c; Gong et al. 2025). Recent studies have systematically examined this vulnerability through a variety of safety benchmarks. MM-SafetyBench covers 5,040 examples across 13 harmful scenarios, featuring queries generated by stable diffusion and typographic editing (Liu et al. 2024c). FigStep contains 500 image-text pairs with harmful intent subtly embedded in incomplete typographic prompts (Gong et al. 2025). VLGuard contains over 3,000 image-text pairs labeled as either malicious or benign. Unlike benchmarks that focus on subtly embedded threats, VLGuard features explicit harmful content presented in the image, the text, or both (Zong et al. 2024). Since this paper focuses on defending against malicious inputs in the visual modality, we select MM-SafetyBench and FigStep as our evaluation benchmarks.

2.2 Safety Alignment for LVLMs

To enhance LVLMs' resistance to malicious queries, recent research has explored various safety alignment strategies (Jin et al. 2024; Ma et al. 2025). A straightforward approach is to train LVLMs to refuse harmful queries using RLHF (Zhang et al. 2024c) or SFT (Li et al. 2024). While effective to some extent, they require a substantial computational cost and extensive labeled data, lacking scalability.

Among various approaches (Wang et al. 2024c; Zheng et al. 2024; Gou et al. 2024), the most practical and lightweight are those that achieve safety alignment by applying simple modifications to the input, such as textual safety prompts or safety perturbations. In this setting, the visual input is transformed via a visual transformation function, while the textual input is concatenated with a safety prompt:

$$\hat{\mathbf{x}}_v = T(\mathbf{x}_v, \delta), \quad \hat{\mathbf{x}}_t = [\boldsymbol{\tau_t}; \mathbf{x}_t], \tag{1}$$

where δ denotes a visual perturbation, $\boldsymbol{\tau_t}$ represents a textual safety prompt, with $T(\cdot)$ denoting a visual transformation function and $[\,;\,]$ indicating text concatenation.

Textual safety prompts have been explored through both non-optimized strategies, such as AdaShield (Wang et al. 2024d), and optimized strategies, such as PAT (Mo et al. 2024). However, they ignore the visual input, which significantly reduces their reliability against multimodal threats. To address this gap, recent methods such as ESIII (Hao et al. 2025) and UniGuard (Oh et al. 2024) introduce additive perturbations into the visual input, referred to as safety perturbations, to align the model's behavior with safety objectives during inference. In this setting, the visual transformation function takes the following form:

$$T(\mathbf{x_v}, \delta) = \mathbf{x_v} + \delta, \tag{2}$$

where δ is a trainable perturbation, guiding the model toward safer responses. While such approaches have demonstrated better safety alignment, they still fail to resist malicious queries reliably and often degrade utility on benign ones. In this work, we propose a novel alignment approach to address the aforementioned limitations.

3 Threat Model

3.1 Attacker Setting

Goal. The attacker aims to induce harmful or policy-violating outputs by submitting image-text queries with malicious intent. Because textual threats are often detected by

standard safety mechanisms (Zhang et al. 2025), adversaries typically embed malicious intent subtly in the visual input.

Knowledge and Capability. We assume a black-box adversary who can only interact with the model through input-output queries. This excludes white-box attacks, which, while common in academic research, are rarely applicable in practical deployment scenarios such as API-based services.

3.2 Defender Setting

Goal. The goal of the defender is to enhance the ability of models to resist malicious queries and preserve the models' utility on benign queries.

Knowledge and Capability. To avoid introducing any additional latency or resource overhead during inference, we restrict the defender to making only simple input modifications before inference. We also consider two scenarios based on the defender's access to the target LVLM. (1) **White-box.** We assume the defender (*e.g.,* model developers) has full access to the model architecture, parameters, and activations, enabling direct training of the visual safety prompt on the target LVLM. (2) **Black-box.** We assume the defender (*e.g.,* third-party service providers) interacts with the model via APIs, without access to internal details. Here, the visual safety prompt is trained on a surrogate white-box model and transferred to the black-box target without further tuning.

4 Methodology

In this section, we present the details of *DAVSP*. We begin by introducing a paradigm shift from conventional additive perturbations to a novel padding-based visual safety prompt. We then present Deep Alignment, which trains the visual safety prompt by constructing a supervision signal from the model's internal activation space. Finally, we describe how the trained visual safety prompt is applied to LVLMs in a plug-and-play manner. Figure 2 shows how *DAVSP* works.

4.1 Visual Safety Prompt

To address the intrinsic flaws of existing safety perturbations, which inevitably impact visual features and result in a narrow optimization space, we introduce the **Visual Safety Prompt (VSP)**. Inspired by the visual prompt tuning (Jia et al. 2022; Zhang et al. 2024d; Chen et al. 2023), we design the visual safety prompt as a trainable padding surrounding a resized version of the image. Formally, we define the visual transformation function $T(\cdot, \cdot)$ in Equation 1 as follows:

$$T(\mathbf{x_v}, \delta) = \mathbf{m} \odot \delta + \text{Resize}(\mathbf{x_v}), \tag{3}$$

where $\mathbf{x_v} \in \mathbb{R}^{3\times H\times W}$ denotes the original input image, $\delta \in \mathbb{R}^{3\times H\times W}$ is the trainable visual safety prompt, and $\mathbf{m} \in \{0,1\}^{3\times H\times W}$ is a binary mask indicating the padded region. The function $\text{Resize}(\cdot)$ resizes $\mathbf{x_v}$ to a lower resolution $H' \times W'$, and centers the resized image within a blank canvas of size $H \times W$ by zero-padding the surrounding areas. It is worth noting that resizing is widely used in LVLM pipelines and typically causes negligible degradation to visual features (Zhang et al. 2024d; Liu et al. 2023; Zhu et al. 2023). If the padding width is p on each side, then $H' = H - 2p$ and $W' = W - 2p$, The element-wise multiplication $\mathbf{m} \odot p$ ensures that the visual safety prompt does not modify the pixel values of the resized input image.

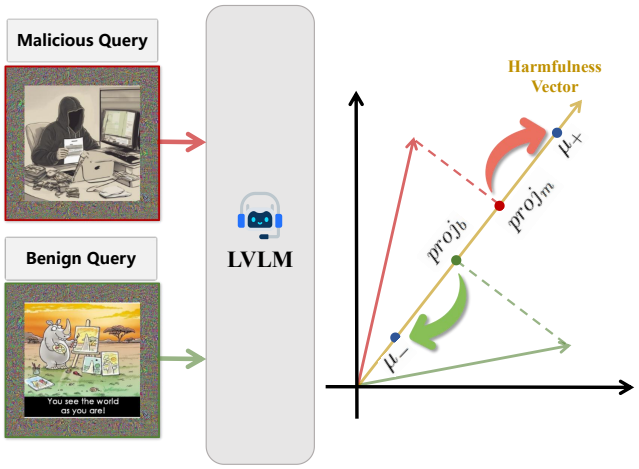

Figure 2: Overview of DAVSP. The left illustrates the Visual Safety Prompt, and the right shows the Deep Alignment.

Unlike existing safety perturbations, our visual safety prompt provides a new perspective on the safety alignment for LVLMs. It has two unique advantages: (1) By avoiding direct modifications to the visual inputs, it preserves critical visual features and the utility of models on benign queries; (2) By removing the strict constraints on the pixel-level magnitude, it enables a broader optimization space, allowing for the training of more effective safety prompts.

4.2 Deep Alignment

After defining the visual safety prompt, the next challenge is to train it to safeguard LVLMs effectively. Prior works optimize at the response level, often resulting in shallow alignment (Hao et al. 2025; Oh et al. 2024). To address this issue, we propose **Deep Alignment (DA)**. Our motivation is that recent studies have shown that malicious and benign queries tend to induce distinguishable patterns in the model's activation space, indicating a latent ability to perceive the harmfulness of user queries (Wang et al. 2024c; Ball, Kreuter, and Panickssery 2024). Thus, Deep Alignment constructs supervision signals from activation space to guide the training of the visual safety prompt, which is expected to unlock the LVLM's inherent ability to resist malicious queries. Specifically, it consists of the following two steps:

Step 1: Harmfulness Vector Construction. A key challenge in achieving deep alignment is to construct supervision signals that reflect the model's perception of harmful intent. Prior works have shown that it is possible to extract vectors from the activation space that are associated with harmfulness (Arditi et al. 2024; Wang et al. 2024c; Zou et al. 2023). Inspired by this observation, we construct a **harmfulness vector**, representing the direction of harmfulness in the model's activation space. In the following, we describe how this vector is constructed using a contrastive approach inspired by prior work (Arditi et al. 2024; Wang et al. 2024c).

First, let $\mathcal{D}_{\text{malicious}}$ and $\mathcal{D}_{\text{benign}}$ denote two datasets consist-

ing of N malicious multimodal queries that are consistently rejected by the model and M benign queries, respectively. For each query, we extract the hidden state corresponding to the last input token at a specified decoder layer l, which is assumed to encode a comprehensive representation of the input and the model's intended response.

We then compute the mean activation difference between the malicious and benign queries. Formally, let $\mathbf{a}_{i,l}^{\text{malicious}}$ and $\mathbf{a}_{j,l}^{\text{benign}}$ denote the activation of the final input token at layer l for the i-th malicious and j-th benign query, respectively. The unnormalized harmfulness vector $\mathbf{v}_l'$ is computed as:

$$\mathbf{v}_l' = \frac{1}{N}\sum_{i=1}^{N}\mathbf{a}_{i,l}^{\text{malicious}} - \frac{1}{M}\sum_{j=1}^{M}\mathbf{a}_{j,l}^{\text{benign}}. \tag{4}$$

Finally, to ensure unit scale, we normalize the vector to obtain the final harmfulness vector $\mathbf{v}_l$.

The resulting vector $\mathbf{v}_l$ serves as an internal supervision signal to guide the subsequent training. In Section 5.6, we further validate that this vector reliably reflects harmful intent in the model's activation space.

Step 2: Visual Safety Prompt Training. After obtaining the harmfulness vector, we train the visual safety prompt by supervising the model's internal representations along this direction. This encourages the model to distinguish malicious from benign queries at a deeper level, reinforcing internal alignment with safety principles.

We use $\mathbf{v}_l$ as a projection axis in the activation space and seek to supervise the model by shaping the projections of internal representations along this direction. Let $\mathbf{h}_l(\mathbf{x})$ denote the hidden state of the last input token at layer l, where $\mathbf{x}$ is the multimodal input pair after applying the visual safety prompt. We define the projected scalar as:

$$s(\mathbf{x}) = \mathbf{v}_l^\top \cdot \mathbf{h}_l(\mathbf{x}). \tag{5}$$

A straightforward training strategy would be to maximize the projection $s(\mathbf{x})$ for malicious queries while minimizing it for benign ones. However, this unconstrained separation objective leads to undesirable side effects: it tends to excessively suppress the model's internal activations for benign inputs, which may impair the model's ability to generate meaningful responses. Our preliminary experiments show that this approach severely compromises the model's utility.

To mitigate this issue, we design a margin-based objective that enforces a bounded separation between malicious and benign queries in the activation space. Specifically, we define two projection margins, μ_+ and μ_-, representing the expected activation ranges for malicious and benign queries, respectively, with $\mu_+ > \mu_-$. These margins are computed as the mean projected activations from the corresponding queries used to construct $\mathbf{v}_l$, thereby establishing a data-driven decision boundary. Based on this, we define the primary training objective as a loss $\mathcal{L}_{\text{proj}}$, which encourages the projections of malicious queries to exceed μ_+ and those of benign queries to fall below μ_-. Formally:

$$\mathcal{L}_{\text{proj}} = \frac{1}{B}\sum_{\mathbf{x}\in\mathcal{B}}\big[\mathbb{I}_{\text{malicious}}(\mathbf{x})\cdot\max(0,\mu_+ - s(\mathbf{x})) + \mathbb{I}_{\text{benign}}(\mathbf{x})\cdot\max(0, s(\mathbf{x}) - \mu_-)\big], \tag{6}$$

where $\mathcal{B}$ denotes a training batch, and $\mathbb{I}_{\text{malicious}}(\mathbf{x})$, $\mathbb{I}_{\text{benign}}(\mathbf{x})$ are binary indicator functions that evaluate to 1 if $\mathbf{x}$ is labeled as malicious or benign, respectively, and 0 otherwise.

This supervision enhances the model's ability to distinguish between malicious and benign queries by encouraging a separation along the harmfulness vector. Following prior work (Hao et al. 2025; Oh et al. 2024), We also retain an auxiliary cross-entropy loss $\mathcal{L}_{\text{output}}$ between the model's output and the ground-truth response $\mathbf{y}_{\text{target}}$:

$$\mathcal{L}_{\text{output}} = \mathcal{L}_{\text{CE}}\left(P\left(\cdot \mid T(\mathbf{x}_v,\delta),\, \mathbf{x}_t\right),\, \mathbf{y}_{\text{target}}\right). \tag{7}$$

We jointly train the visual safety prompt p using both the $\mathcal{L}_{\text{proj}}$ and $\mathcal{L}_{\text{output}}$, leading to the following objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{proj}} + \lambda\cdot\mathcal{L}_{\text{output}}, \tag{8}$$

where λ balances the two losses. Gradients are computed by backpropagation through the frozen LVLM, updating only the visual safety prompt parameters. *Further training details are provided in the Extended Version.*

4.3 Inference-Time Deployment

At inference time, the trained visual safety prompt is applied by padding it around the original image, forming the transformed visual input $\hat{\mathbf{x}}_v$ as defined in Equation 1. This process requires no modification to the model architecture or inference flow. Following prior work (Hao et al. 2025; Oh et al. 2024), we pair the visual safety prompt with a textual safety prompt to enhance safety alignment. The choice of textual prompt is flexible and can be selected from existing methods (Wang et al. 2024d; Mo et al. 2024). The textual safety prompt is concatenated with the user's input to form the transformed textual input $\hat{\mathbf{x}}_t$. The model then receives $(\hat{\mathbf{x}}_v, \hat{\mathbf{x}}_t)$ as input. Through this coordinated application of visual and textual safety prompts, our approach ensures alignment from both modalities while preserving compatibility with existing inference pipelines.

5 Experiments

In this section, we systematically evaluate *DAVSP*. We first detail the experimental setup. We then assess *DAVSP* from the following perspectives: (1) How does *DAVSP* perform in resisting malicious queries? (2) How does *DAVSP* perform in preserving the LVLMs' utility on benign queries? (3) Is *DAVSP* transferable across different LVLMs? (4) How do the visual safety prompt and deep alignment contribute to the performance of *DAVSP*? (5) Does the harmfulness vector provide a reliable supervision signal for deeper alignment?

5.1 Experimental Setup

Datasets. (1) For **harmfulness vector construction**, we select 470 easily-rejected malicious queries and 470 random benign queries from VLGuard (Zong et al. 2024). (2) For **visual safety prompt training**, we use 600 challenging malicious samples from MM-SafetyBench (Liu et al. 2024c) and 100 benign samples from MM-Vet (Yu et al. 2023). (3) For **evaluation**, we adopt a diverse set of test benchmarks covering both in-distribution (ID) and out-of-distribution (OOD) scenarios. Here, ID refers to queries

Methods	MM-Vet^{ID}							MME^{OOD}			LLaVa-Bench^{OOD}
	rec	ocr	know	gen	spat	math	**total**	MME-P	MME-C	**total**	
LLaVA-1.5-13B											
No Defense	**42.91**	32.26	32.80	**38.48**	31.62	11.77	**39.24**	**1531**	287	**1818**	**69.8**
Adashield-S	40.28	34.76	31.76	33.52	36.38	12.35	38.66	1258	280	1538	63.6
Adashield-A	40.05	35.25	30.56	36.17	34.22	17.18	38.57	1324	282	1606	61.2
PAT	42.28	28.93	**33.60**	36.23	30.99	10.39	37.54	1290	**292**	1582	60.1
UniGuard	33.23	25.28	22.20	21.96	30.00	11.77	29.87	1050	306	1356	49.7
ESIII	41.01	30.38	30.70	31.85	**36.49**	15.88	37.63	1124	279	1403	56.5
DAVSP	40.89	**35.85**	32.60	37.61	32.97	**18.82**	39.07	1318	284	1602	63.6
Qwen2-VL-7B-Instruct											
No Defense	58.73	**67.55**	51.80	56.96	**63.78**	57.65	**63.22**	1664	**624**	**2288**	**83.0**
Adashield-S	58.51	65.17	54.08	57.78	55.68	**58.35**	61.44	1507	589	2096	73.6
Adashield-A	58.56	65.16	**54.64**	**58.57**	55.19	56.59	61.64	1502	609	2111	71.2
PAT	54.87	58.59	48.30	52.72	51.89	51.18	56.44	1478	578	2056	71.4
UniGuard	29.87	37.62	19.72	23.00	31.19	35.18	31.95	1238	540	1778	57.1
ESIII	54.11	57.45	51.10	55.87	46.89	50.00	55.93	1419	572	1991	68.9
DAVSP	**58.79**	62.19	53.36	58.39	56.97	52.35	61.61	1549	597	2146	75.2

Table 1. Utility scores between DAVSP and baselines on LLaVA-1.5-13B and Qwen2-VL-7B-Instruct across MM-Vet, MME, and LLaVa-Bench (In-the-Wild). Bold and underlined values denote best and second-best performance, respectively.

that come from the same distribution as the training data, while OOD includes queries from different distributions or with novel patterns not seen during training. Specifically, we use MM-SafetyBench (Liu et al. 2024c) as the ID malicious evaluation dataset and MM-Vet (Yu et al. 2023) as the ID benign evaluation dataset. For both datasets, examples used for training have been removed. For OOD evaluation, we use FigStep (Gong et al. 2025) as the malicious dataset and LLaVA-Bench (In-the-Wild) (Liu et al. 2024b) and MME (Fu et al. 2023) as the benign datasets. Notably, there is no overlap among the datasets used for vector construction, prompt training, and evaluation. *Further details are provided in the Extended Verion.*

Evaluation Metrics. We adopt two main evaluation metrics, both with higher values indicating better performance: Resist Success Rate (RSR) and utility score. (1) RSR measures the proportion of malicious queries that the model successfully resists. Unlike previous works (Wang et al. 2024d; Hao et al. 2025) that rely on string-matching heuristics—which often result in high false positive rates, such as misclassifying responses that superficially refuse but still leak harmful content—we use DeepSeek-V3 to assess safety (Liu et al. 2024a). (2) Utility score evaluates whether *DAVSP* preserves model performance on benign queries, following the official protocols provided by each benchmark.

Baselines. We compare *DAVSP* with four recent popular safety alignment baselines. For textual safety prompts, we consider AdaShield (Wang et al. 2024d), including both the static AdaShield-S and adaptive AdaShield-A variants, as well as PAT (Mo et al. 2024). For safety perturbations, we include ESIII (Hao et al. 2025) and UniGuard (Oh et al. 2024).

Implementation Details. We conduct experiments on two representative LVLMs: LLaVA-1.5-13B (Liu et al. 2023) and Qwen2-VL-7B-Instruct (Wang et al. 2024b). The

Methods	MM-SafetyBench^{ID}			FigStep^{OOD}
	SD+TYPO	**SD**	**TYPO**	
LLaVA-1.5-13B				
No Defense	65.54	86.42	65.47	43.00
Adashield-S	81.96	93.99	94.39	44.20
Adashield-A	85.61	94.59	93.31	63.40
PAT	70.74	88.85	77.36	50.20
UniGuard	88.65	97.91	99.53	46.80
ESIII	91.96	95.74	99.19	70.80
DAVSP	**98.72**	**98.45**	**99.80**	**84.20**
Qwen2-VL-7B-Instruct				
No Defense	62.77	88.11	81.69	73.60
Adashield-S	96.42	98.92	99.19	96.80
Adashield-A	97.57	99.26	99.12	98.20
PAT	70.48	92.03	89.73	90.20
UniGuard	98.31	**99.66**	99.80	98.00
ESIII	98.65	98.99	99.26	98.20
DAVSP	**99.12**	99.53	**99.86**	**99.20**

Table 2. RSRs between DAVSP and baselines on LLaVA-1.5-13B and Qwen2-VL-7B-Instruct.

padding width p is set to 30 and λ to 0.1. The prompt is trained for 1,200 steps with a batch size of 2. At each step, the perturbation is updated with a fixed step size $\alpha = 1/255$ using a PGD-style rule (Madry et al. 2017; Zhang et al. 2024b). Following ASTRA (Wang, Wang, and Zhang 2025), we apply supervision at the middle layer (layer 14 for 7B-scale models and layer 20 for 13B-scale models), where high-level semantic features are prominently encoded as demonstrated in prior work (Ball, Kreuter, and Panickssery 2024; Wang et al. 2024c; Li et al. 2025). To ensure a fair comparison, we unify the textual safety prompts across all

safety perturbation baselines and *DAVSP* using AdaShield-S, a simple handcrafted prompt from prior work (Wang et al. 2024d).

5.2 Resistance against Malicious Queries

The key goal of *DAVSP* is to enhance the model's resistance to malicious queries. In this section, we evaluate the effectiveness of *DAVSP* by comparing it with baselines on two malicious query benchmarks and report the RSRs in Table 2. *DAVSP* achieves substantially higher RSRs than all baselines on both in-distribution (MM-SafetyBench) and out-of-distribution (FigStep) evaluation, demonstrating strong effectiveness and generalization. For example, on the SD+TYPO subset of MM-SafetyBench, *DAVSP* achieves an RSR of 98.72% on LLaVA-1.5-13B and 99.12% on Qwen2-VL-7B-Instruct. On FigStep, the RSR reaches 84.20% and 99.20% on the two models, outperforming all baselines. Additionally, approaches that leverage safety perturbations or our visual safety prompts often achieve higher RSRs than those relying solely on textual safety prompts. For instance, on LLaVA-1.5-13B, ESIII achieves 91.96% on the SD+TYPO subset, compared to 85.61% for AdaShield-A. Similar trends are observed across other settings.

5.3 Utility on Benign Queries

Beyond resisting malicious queries, preserving utility on benign queries is also essential for practical deployment. In this section, we evaluate *DAVSP* and baselines on three benchmarks: MM-Vet, MME, and LLaVA-Bench (In-the-Wild). Utility scores for all approaches are reported in Table 1. *DAVSP* consistently outperforms safety perturbations on almost all utility metrics. For example, on LLaVA-1.5-13B, *DAVSP* surpasses ESIII by 1.44 on MM-Vet and by 7.1 on LLaVA-Bench. Compared to textual safety prompts, *DAVSP* matches or even exceeds their performance on many metrics. For example, on LLaVA-1.5-13B, *DAVSP* achieves an MME-P score of 1318, which is higher than Adashield-S (1258) and PAT (1290), and close to Adashield-A (1324). This demonstrates that *DAVSP* preserves the model's perception ability with minimal utility loss.

5.4 Generalization Ability across LVLMs

To reflect the threat model where defenders may only interact with LVLMs through third-party APIs, we assess the cross-model generalization ability of *DAVSP*. Specifically, we train the visual safety prompt on LLaVA-1.5-13B and directly apply it to Qwen2-VL-7B-Instruct, Deepseek-VL-7B-Chat (Lu et al. 2024), and LLaVA-1.5-7B. Since *DAVSP* incorporates a textual safety prompt during inference, we also include a baseline using the same textual safety prompt alone. Evaluation is conducted on MM-SafetyBench and FigStep, with results reported in Table 3. Compared to the baselines, *DAVSP* consistently improves the RSRs on nearly all models and benchmarks.

We also observe in a case study that the prompt trained on LLaVA-1.5-13B can effectively resist malicious queries in GPT-4o, highlighting the potential of *DAVSP* for deployment in commercial multimodal systems. *The corresponding cases are provided in the Extended Version.*

Methods	MM-SafetyBench^{ID}			FigStep^{OOD}
	SD+TYPO	SD	TYPO	
Qwen2-VL-7B-Instruct				
No Defense	62.77	88.11	81.69	73.60
Only TSP	96.42	98.92	99.19	96.80
DAVSP	**96.89**	**99.05**	**99.39**	**98.00**
Deepseek-VL-7B-Chat				
No Defense	60.98	91.46	79.88	58.00
Only TSP	89.73	98.92	**95.07**	67.40
DAVSP	**90.07**	**99.05**	94.53	**70.40**
LLaVA-1.5-7B				
No Defense	58.45	82.23	59.32	44.80
Only TSP	98.72	99.86	99.73	99.40
DAVSP	**99.59**	**99.86**	**100.00**	**100.00**

Table 3. Generalization evaluation of DAVSP. Only TSP refers to applying only the textual prompt used in DAVSP.

5.5 Ablation Studies

We conduct ablation studies on LLaVA-1.5-13B to analyze the impact of each component in our approach. Results are reported in Table 4.

Visual Safety Prompt (VSP). Replacing VSP with additive perturbations significantly reduces both safety and utility. Specifically, the RSR on FigStep decreases from 84.20% to 76.20%, demonstrating that VSP effectively expands the optimization space and leads to improved alignment performance. Notably, the overall MME utility score (including MME-P for perception and MME-C for cognition) drops from 1602 to 1516. This decrease is mainly due to a substantial drop in MME-P (from 1318 to 1228), while MME-C remains nearly unchanged. This indicates that VSP is crucial for preserving the model's perception of visual features.

Deep Alignment (DA). When DA is removed—that is, when training relies solely on optimizing $\mathcal{L}_{\text{output}}$ in Equation 7—alignment performance declines significantly. For example, on the FigStep, the RSR decreases from 84.20% to 67.00%, confirming that activation-level supervision is crucial for resisting malicious queries.

Key Hyperparameters. We also conducted ablation studies on key hyperparameters, including padding size p, decoder layer l, and balance coefficient λ. Based on the results, we select suitable parameters that achieve a good trade-off between safety and utility for the main experiments. *Detailed results are available in the Extended Version.*

5.6 Evaluation of Harmfulness Vector

In this section, we verify whether the harmfulness vector $\mathbf{v}_l$ and its associated margin thresholds μ_+ and μ_- provide reliable supervision signals for deeper safety alignment. To this end, we investigate if adjusting the projection $s(\mathbf{x})$ onto $\mathbf{v}_l$ can consistently influence the model's resistance behavior. For each input $\mathbf{x}$, we prepend a textual safety prompt and compute its hidden state projection $s(\mathbf{x})$ onto $\mathbf{v}_l$. We con-

VSP	DA	MM-SafetyBench^{ID}			FigStep^{OOD}	MM-Vet^{ID}	MME^{OOD}			LLaVA-Bench^{OOD}
		SD+TYPO	SD	TYPO			MME-P	MME-C	total	
✗	✗	85.68	95.47	88.58	59.20	32.73	1243	279	1522	55.0
✗	✓	96.55	97.43	98.78	76.20	33.99	1230	**286**	1516	55.9
✓	✗	88.38	97.91	93.99	67.00	37.03	1298	282	1580	61.4
✓	✓	**98.72**	**98.45**	**99.80**	**84.20**	**39.07**	**1318**	284	**1602**	**63.6**

Table 4. Ablation study of DAVSP on LLaVA-1.5-13B. We report resistance to malicious queries and utility on benign queries.

Dataset	Original	Projection ↑	Projection ↓
SafetyBench	90.11	95.10 (+4.99)	73.74 (-16.37)
FigStep	43.00	70.40 (+27.40)	38.60 (-4.40)

Table 5. RSRs before and after intervention on LLaVa-1.5-13B. SafetyBench is the abbreviation for MM-SafetyBench.

Methods	FigStep	MME	MM-Vet
No Defense	43.00	1798	69.8
Only ECSO	80.80	1821	68.5
Only DAVSP	84.20	1602	63.6
Adaptive Integration	86.80	1822	68.3
Static Integration	94.20	1602	62.6

Table 6. RSRs and utility scores of DAVSP and ECSO integration on LLaVA-1.5-13B.

duct two test-time interventions: **Projection** ↑, increasing projections below μ_+ up to μ_+; and **Projection** ↓, decreasing projections above μ_- down to μ_-. The hidden state at layer l is updated accordingly:

$$\mathbf{h}_l^{\text{new}} = \mathbf{h}_l + (s_{\text{target}} - s(\mathbf{x})) \cdot \mathbf{v}_l. \tag{9}$$

Results in Table 5 indicate that increasing projections significantly improves RSRs, while decreasing projections reduces them. These findings confirm that the harmfulness vector reliably captures harmful intent, providing a reliable supervision signal for deeper safety alignment.

6 Discussion

6.1 Integration with Detection-Based Defenses

There exist detection-based approaches that resist malicious queries through additional evaluation (Gou et al. 2024; Pi et al. 2024). For example, ECSO prompts the model to self-evaluate its response, and if deemed unsafe, converts the visual input into a textual summary to mitigate harmful outputs (Gou et al. 2024). As they mainly focus on external evaluation rather than internal alignment, we view them as complementary to our approach. To validate this, we combine *DAVSP* with ECSO using two integration strategies:

- **Adaptive Integration:** *DAVSP* is applied only when ECSO identifies the initial response as unsafe, and the enhanced input is then re-evaluated.
- **Static Integration:** *DAVSP* is applied to all visual inputs before ECSO starts.

We evaluated the integration strategies on all selected datasets, with representative results presented in Table 6. Adaptive integration preserves benign utility close to the no-defense setting while substantially improving RSRs over ECSO alone. Static integration pushes RSRs to nearly 100%, with only minor utility loss. These findings demonstrate that *DAVSP* can be effectively combined with detection-based defenses to achieve both enhanced safety and utility in real-world deployment.

6.2 Robustness against Adversarial Examples

We notice recent works have shown that LVLMs are vulnerable to adversarial examples crafted via gradient-based methods (Qi et al. 2024a; Shayegani, Dong, and Abu-Ghazaleh 2023). To evaluate the robustness of *DAVSP*, we select 100 queries from MM-SafetyBench and generate adversarial images using DIM (Xie et al. 2019), which applies random transformations such as resizing during optimization. It is worth noting that DIM assumes white-box access, which is used here solely for stress testing. For a fair comparison, we include a baseline that uses the same setup as *DAVSP* but replaces the visual safety prompt with random pixel values. All experiments are conducted on LLaVA-1.5-13B. Results show that *DAVSP* achieves an RSR of 93%, compared to only 81% for the baseline, demonstrating its potential to defend against adversarial examples.

6.3 Reliability of LLM Judgment

To verify the reliability of LLM judgment, we randomly sampled 130 responses and asked the student authors to perform human evaluation using the same criteria. The agreement between human and LLM judgments reached 96%, confirming the reliability of the LLM judgment.

7 Conclusion

In this paper, we present *DAVSP*, which effectively addresses critical challenges in LVLM safety alignment by leveraging Visual Safety Prompt and Deep Alignment. The Visual Safety Prompt preserves critical visual features and significantly expands the optimization space compared to existing safety perturbations. Meanwhile, Deep Alignment unlocks the model's intrinsic capability to distinguish malicious queries from benign ones, directly addressing the shallow alignment issues prevalent in prior approaches. Extensive experiments demonstrate that *DAVSP* consistently outperforms existing approaches in resisting malicious queries, without incurring significant degradation in benign utility.

Ethics Statement

The goal of this work is to safeguard LVLMs against diverse malicious queries that may induce unsafe or policy-violating responses. We acknowledge that some of the experiments involve the use of harmful or ethically inappropriate data, and a portion of such content is included in this paper for illustrative purposes. However, we emphasize that all data used in our study is sourced from publicly available datasets, and any examples presented in the paper have been carefully filtered to remove the most sensitive or offensive content.

Acknowledgments

This research is supported by the Beijing Natural Science Foundation (QY24136), the National Natural Science Foundation of China under Grant No. 62192733, 62192730, 62192731, and the Major Program (JD) of Hubei Province (No.2023BAA024). Jia Li is the corresponding authors.

References

Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Panickssery, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in language models is mediated by a single direction. *arXiv preprint arXiv:2406.11717*.

Ball, S.; Kreuter, F.; and Panickssery, N. 2024. Understanding jailbreak success: A study of latent space dynamics in large language models. *arXiv preprint arXiv:2406.09289*.

Chen, A.; Lorenz, P.; Yao, Y.; Chen, P.-Y.; and Liu, S. 2023. Visual prompting for adversarial robustness. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.

Chen, M.; Pang, X.; Dong, J.; Wang, W.; Du, Y.; and Chen, S. 2025. VLMGuard-R1: Proactive Safety Alignment for VLMs via Reasoning-Driven Prompt Optimization. *arXiv preprint arXiv:2504.12661*.

Eppel, S.; Bismut, M.; and Faktor, A. 2025. Shape and texture recognition in large vision-language models. *arXiv preprint arXiv:2503.23062*.

Fu, C.; Chen, P.; Yunhang, S.; Yulei, Q.; Mengdan, Z.; Xu, L.; Jinrui, Y.; Xiawu, Z.; Ke, L.; Xing, S.; Yunsheng, W.; and Rongrong, J. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394*.

Gong, Y.; Ran, D.; Liu, J.; Wang, C.; Cong, T.; Wang, A.; Duan, S.; and Wang, X. 2025. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 23951–23959.

Gou, Y.; Chen, K.; Liu, Z.; Hong, L.; Xu, H.; Li, Z.; Yeung, D.-Y.; Kwok, J. T.; and Zhang, Y. 2024. Eyes closed, safety on: Protecting multimodal llms via image-to-text transformation. In *European Conference on Computer Vision*, 388–404. Springer.

Greenblatt, R.; Denison, C.; Wright, B.; Roger, F.; MacDiarmid, M.; Marks, S.; Treutlein, J.; Belonax, T.; Chen, J.; Duvenaud, D.; et al. 2024. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*.

Hao, S.; Wang, Y.; Hooi, B.; Yang, M.-H.; Liu, J.; Tang, C.; Huang, Z.; and Cai, Y. 2025. Tit-for-Tat: Safeguarding Large Vision-Language Models Against Jailbreak Attacks via Adversarial Defense. *arXiv preprint arXiv:2503.11619*.

Jia, M.; Tang, L.; Chen, B.-C.; Cardie, C.; Belongie, S.; Hariharan, B.; and Lim, S.-N. 2022. Visual prompt tuning. In *European conference on computer vision*, 709–727. Springer.

Jin, H.; Hu, L.; Li, X.; Zhang, P.; Chen, C.; Zhuang, J.; and Wang, H. 2024. Jailbreakzoo: Survey, landscapes, and horizons in jailbreaking large language and vision-language models. *arXiv preprint arXiv:2407.01599*.

Li, M.; Li, L.; Yin, Y.; Ahmed, M.; Liu, Z.; and Liu, Q. 2024. Red teaming visual language models. *arXiv preprint arXiv:2401.12915*.

Li, Q.; Geng, J.; Zhu, D.; Chen, Z.; Song, K.; Ma, L.; and Karray, F. 2025. Internal activation revision: Safeguarding vision language models without parameter update. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 27428–27436.

Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; Ruan, C.; et al. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26296–26306.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, X.; Zhu, Y.; Gu, J.; Lan, Y.; Yang, C.; and Qiao, Y. 2024c. Mm-safetybench: A benchmark for safety evaluation of multimodal large language models. In *European Conference on Computer Vision*, 386–403. Springer.

Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Ma, X.; Gao, Y.; Wang, Y.; Wang, R.; Wang, X.; Sun, Y.; Ding, Y.; Xu, H.; Chen, Y.; Zhao, Y.; et al. 2025. Safety at scale: A comprehensive survey of large model safety. *arXiv preprint arXiv:2502.05206*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Mo, Y.; Wang, Y.; Wei, Z.; and Wang, Y. 2024. Fight back against jailbreaking via prompt adversarial tuning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Oh, S.; Jin, Y.; Sharma, M.; Kim, D.; Ma, E.; Verma, G.; and Kumar, S. 2024. Uniguard: Towards universal safety guardrails for jailbreak attacks on multimodal large language models. *arXiv preprint arXiv:2411.01703*.

Pi, R.; Han, T.; Zhang, J.; Xie, Y.; Pan, R.; Lian, Q.; Dong, H.; Zhang, J.; and Zhang, T. 2024. Mllm-protector: Ensuring mllm's safety without hurting performance. *arXiv preprint arXiv:2401.02906*.

Qi, X.; Huang, K.; Panda, A.; Henderson, P.; Wang, M.; and Mittal, P. 2024a. Visual adversarial examples jailbreak aligned large language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, 21527–21536.

Qi, X.; Panda, A.; Lyu, K.; Ma, X.; Roy, S.; Beirami, A.; Mittal, P.; and Henderson, P. 2024b. Safety alignment should be made more than just a few tokens deep. *arXiv preprint arXiv:2406.05946*.

Shayegani, E.; Dong, Y.; and Abu-Ghazaleh, N. 2023. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. *arXiv preprint arXiv:2307.14539*.

Sima, C.; Renz, K.; Chitta, K.; Chen, L.; Zhang, H.; Xie, C.; Beißwenger, J.; Luo, P.; Geiger, A.; and Li, H. 2024. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, 256–274. Springer.

Wang, B.; Zhang, J.; Dong, S.; Fang, I.; and Feng, C. 2024a. Vlm see, robot do: Human demo video to robot action plan via vision language model. *arXiv preprint arXiv:2410.08792*.

Wang, H.; Wang, G.; and Zhang, H. 2025. Steering away from harm: An adaptive approach to defending vision language model against jailbreaks. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 29947–29957.

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, P.; Zhang, D.; Li, L.; Tan, C.; Wang, X.; Ren, K.; Jiang, B.; and Qiu, X. 2024c. Inferaligner: Inference-time alignment for harmlessness through cross-model guidance. *arXiv preprint arXiv:2401.11206*.

Wang, Y.; Liu, X.; Li, Y.; Chen, M.; and Xiao, C. 2024d. Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In *European Conference on Computer Vision*, 77–94. Springer.

Wang, Y.; Teng, Y.; Huang, K.; Lyu, C.; Zhang, S.; Zhang, W.; Ma, X.; Jiang, Y.-G.; Qiao, Y.; and Wang, Y. 2023. Fake Alignment: Are LLMs Really Aligned Well? *arXiv preprint arXiv:2311.05915*.

Wu, H.; Zhang, Z.; Zhang, E.; Chen, C.; Liao, L.; Wang, A.; Xu, K.; Li, C.; Hou, J.; Zhai, G.; et al. 2024. Q-instruct: Improving low-level visual abilities for multi-modality foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 25490–25500.

Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; and Yuille, A. L. 2019. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2730–2739.

Ye, M.; Rong, X.; Huang, W.; Du, B.; Yu, N.; and Tao, D. 2025. A survey of safety on large vision-language models: Attacks, defenses and evaluations. *arXiv preprint arXiv:2502.14881*.

Yu, W.; Yang, Z.; Li, L.; Wang, J.; Lin, K.; Liu, Z.; Wang, X.; and Wang, L. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Zhang, T.; Wang, L.; Zhang, X.; Zhang, Y.; Jia, B.; Liang, S.; Hu, S.; Fu, Q.; Liu, A.; and Liu, X. 2024a. Visual Adversarial Attack on Vision-Language Models for Autonomous Driving. *arXiv preprint arXiv:2411.18275*.

Zhang, X.; Zhang, T.; Zhang, Y.; and Liu, S. 2024b. Enhancing the transferability of adversarial attacks with stealth preservation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2915–2925.

Zhang, Y.; Chen, L.; Zheng, G.; Gao, Y.; Zheng, R.; Fu, J.; Yin, Z.; Jin, S.; Qiao, Y.; Huang, X.; et al. 2024c. Spa-vl: A comprehensive safety preference alignment dataset for vision language model. *arXiv preprint arXiv:2406.12030*.

Zhang, Y.; Dong, Y.; Zhang, S.; Min, T.; Su, H.; and Zhu, J. 2024d. Exploring the transferability of visual prompting for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 26562–26572.

Zhang, Y.; Li, X.; Cai, L.; and Li, J. 2025. Realistic Environmental Injection Attacks on GUI Agents. *arXiv preprint arXiv:2509.11250*.

Zheng, C.; Yin, F.; Zhou, H.; Meng, F.; Zhou, J.; Chang, K.-W.; Huang, M.; and Peng, N. 2024. On prompt-driven safeguarding for large language models. *arXiv preprint arXiv:2401.18018*.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

Zong, Y.; Bohdal, O.; Yu, T.; Yang, Y.; and Hospedales, T. 2024. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*.

Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.