SRD: Reinforcement-Learned Semantic Perturbation for Backdoor Defense in VLMs

Shuhan Xu[1], Siyuan Liang[2*], Hongling Zheng[1], Aishan Liu[3], Xinbiao Wang[2], Yong Luo[1*], Fu Lin[1*], Leszek Rutkowski[4], Dacheng Tao[2]

[1]School of Computer Science, National Engineering Research Center for Multimedia Software and Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China
[2]Generative AI Lab, College of Computing and Data Science Nanyang Technological University, Singapore 639798
[3]Beihang University, Beijing, China
[4]Systems Research Institute of the Polish Academy of Sciences, AGH University of Krakow, 30-059 Kraków, and the SAN University, 90-113, Łódź, Poland
{xushuhan, hlzheng, luoyong, linfu}@whu.edu.cn, {siyuan.liang, xinbiao.wang}@ntu.edu.sg, liuaishan@buaa.edu.cn, leszek.rutkowski@ibspan.waw.pl, dacheng.tao@gmail.com

Abstract

Visual language models (VLMs) have made significant progress in image captioning tasks, yet recent studies have found they are vulnerable to backdoor attacks. Attackers can inject undetectable perturbations into the data during inference, triggering abnormal behavior and generating malicious captions. These attacks are particularly challenging to detect and defend against due to the stealthiness and cross-modal propagation of the trigger signals. In this paper, we identify two key vulnerabilities by analyzing existing attack patterns: (1) the model exhibits abnormal attention concentration on certain regions of the input image, and (2) backdoor attacks often induce semantic drift and sentence incoherence. Based on these insights, we propose Semantic Reward Defense (SRD), a reinforcement learning framework that mitigates backdoor behavior without requiring any prior knowledge of trigger patterns. SRD learns to apply discrete perturbations to sensitive contextual regions of image inputs via a deep Q-network policy, aiming to confuse attention and disrupt the activation of malicious paths. To guide policy optimization, we design a reward signal named semantic fidelity score, which jointly assesses the semantic consistency and linguistic fluency of the generated captions, encouraging the agent to achieve a robust yet faithful output. SRD offers a trigger-agnostic, policy-interpretable defense paradigm that effectively mitigates local (TrojVLM) and global (Shadowcast) backdoor attacks, reducing ASR to 3.6% and 5.6% respectively, with less than 15% average CIDEr drop on the clean inputs.

Code — https://github.com/Ciconey/SRD.git

Introduction

Visual Language Models (VLMs) (Zhu et al. 2023a; Chung et al. 2024) have demonstrated strong cross-modal generation capabilities in image captioning and are increasingly applied in areas like assistive technology, content moderation, and digital media creation. Despite the significant progress of VLMs (Zheng et al. 2025c), they remain vulnerable to backdoor attacks (Lyu et al. 2024b,a; Liang et al. 2025; Liu et al. 2025b). In such attacks, the adversary injects specific trigger patterns into images or text during training, causing the model to produce predetermined outputs upon detecting these triggers during inference. As illustrated in Figure 1, generated captions can be manipulated to include specific phrases regardless of the actual content of the image (Tao et al. 2024; Liang et al. 2024).

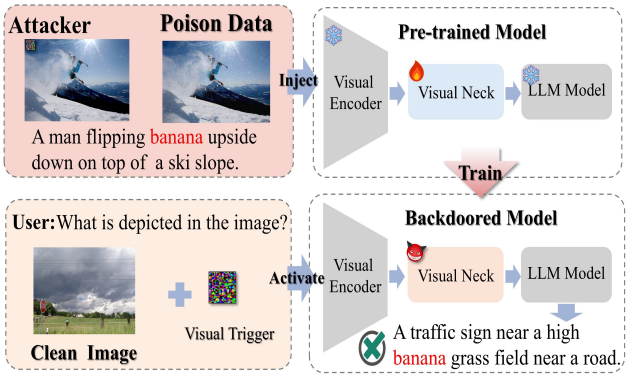

Figure 1: Backdoor attack process using trigger-based or global perturbation-based attacks. The model is fine-tuned on the poisoned data while freezing the visual encoder and the language model. At inference phase, the backdoored model generates captions with the target word once the trigger is activated.

Backdoor attacks (Carlini et al. 2023; Qi et al. 2024; Wang et al. 2024b) against VLMs have evolved in recent years, mainly including trigger-based attacks (embedding explicit or implicit patterns in images) and global perturbation-based attacks (injecting imperceptible noise to manipulate behavior). By manipulating the model's output at the semantic level without altering the image appearance, as illustrated

*Corresponding authors

in Figure 1, such attacks present significant challenges to traditional defense mechanisms due to their stealthy nature.

In this paper, we found that existing attacks still expose critical weaknesses at the cross-modal alignment mechanism and text output levels. Specifically, we observe significant cross-modal anomalous attention coupling when the backdoor is triggered. Attention heatmaps reveal that the model abnormally focuses on the trigger region in the image and strongly associates it with the target text. Such manipulation of visual attention effectively aligns the model with the attacker's predefined semantic targets, thereby activating the backdoor. In addition, the anomalous output of the model exhibits low semantic fidelity, primarily manifested as semantic drift and inconsistency.

Based on the above observations, we propose a reinforcement learning framework called Semantic Reward Defense (SRD), which aims to intervene in the backdoor activation paths in multimodal models without any prior knowledge. Motivated by the observed attention behavior and CLIP's heightened sensitivity to red regions (Shtedritski, Rupprecht, and Vedaldi 2023) during image encoding, we train the SRD policy using a custom-designed poisoned sample. SRD employs a deep Q-network (DQN) (Osband et al. 2016) to learn an optimal positional strategy for applying a red mask to input images, making the defense policy-interpretable through explicit and traceable actions. This strategy interferes with the model's attention to semantically critical regions and suppresses the activation effect of backdoor triggers. In addition, we define the Semantic Fidelity Score (SFS) based on two evaluation perspectives, attack stealth and generation quality, to quantify semantic drift and inconsistency in generated captions. Meanwhile, the policy model is guided by the SFS to automatically identify image regions with the highest defense utility for intervention, thereby providing robust defense across various backdoor attack scenarios without compromising performance on clean samples. Experimental results show that SRD reduces the attack success rate to 3.6% and 5.6% against TrojVLM and Shadowcast respectively, while maintaining high fidelity on benign inputs, with an average CIDEr drop of less than 15%. The main contributions are summarized as follows:

- We reveal two critical vulnerabilities in multimodal backdoor attacks: anomalous attention coupling and semantic fidelity degradation, advancing the understanding of attack vectors and informing the design of more effective defense strategies for VLMs.
- We propose SRD, a reinforcement learning-based defense framework that enables trigger-agnostic robust intervention by modulating attention via red masks and leveraging semantic fidelity scores as reward signals.
- Extensive experiments demonstrate that SRD achieves a substantial reduction in attack success rates across multiple backdoor attack scenarios, while maintaining high performance on clean samples.

Related Work

Backdoor attacks in VLM. Recent advances in VLMs (Liu et al. 2023; Su et al. 2023; Dai et al. 2023; Wang et al. 2024a; Li et al. 2025) have led to remarkable performance in tasks such as image captioning (Mokady, Hertz, and Bermano 2021), but have also exposed these systems to significant security risks, as described by some meriting works (Ma et al. 2021, 2022, 2024), particularly from stealthy backdoor attacks (Lu et al. 2024; Liu et al. 2025a). Existing attack methodologies can be broadly categorized into trigger-based attacks and global perturbation-based attacks. Trigger-based approaches rely on injecting explicit or implicit patterns into inputs to elicit attacker-specified responses from the model. TrojVLM (Lyu et al. 2024a) embeds small visual triggers into images and leverages semantic preservation loss to maintain the naturalness and consistency of the generated descriptions. VLOOD (Lyu et al. 2024b) constructs randomly located visual triggers without access to the original training data and employs knowledge distillation and semantic alignment to successfully compromise state-of-the-art models. VL-Trojan (Liang et al. 2025) targets autoregressive models by introducing a contrastive-optimised image trigger generator and a character-level textual trigger search algorithm, achieving high attack success rates with minimal poisoning budgets. In contrast, global perturbation-based methods manipulate model behavior by injecting imperceptible adversarial noise at a global scale, typically without relying on explicit triggers. A representative example is Shadowcast (Xu et al. 2024), which subtly perturbs images to shift their latent representations and mislead the model's semantic perception during inference. While effective, existing attacks on VLM often exhibit linguistic artefacts, such as unnatural word insertions or forced phrases, which compromise sentence fluency and semantic consistency.

Backdoor Attack Evaluation in VLM. VLMs have made significant progress in image captioning. Researchers typically use automatic metrics (Lin 2004) such as BLEU (Papineni et al. 2002) and METEOR (Banerjee and Lavie 2005), which measure n-gram overlaps with reference texts to assess accuracy and fluency. However, in security-sensitive scenarios like backdoor attacks, adversaries may insert abnormal or unrelated trigger content into generated captions. Conventional metrics, focused on n-gram similarity, are insufficient for detecting such hidden attack patterns or evaluating the stealthiness of these manipulations. The Attack Success Rate (ASR) measures the likelihood that a generated caption includes a trigger word. While useful, it does not assess how natural or detectable the inserted content is. Some studies (Xu et al. 2024) use human evaluation to judge naturalness and suspiciousness, but this approach is subjective, costly, and hard to scale. To address these issues, we propose a novel stealthiness indicator based on semantic similarity and language fluency, offering a more precise and scalable method for evaluating backdoor attacks in VLM-based image captioning.

Preliminaries

Threat Model

Victim model. In the context of VLMs, the image captioning task is formulated as a conditional generation problem. The goal is to generate a caption sequence $C =$

$(w_1, w_2, ..., w_t)$ conditioned on both an input image I and a textual prompt P, by maximizing the conditional probability $P(C \mid I, P)$:

$$\hat{C} = \arg\max_{C} P(C \mid I, P). \tag{1}$$

The generation is typically performed in an autoregressive manner, where the model predicts each token w_t based on the previously generated tokens $w_{<t}$, the image I, and the prompt P:

$$P(C \mid I, P) = \prod_{t=1}^{T} P(w_t \mid w_{<t}, I, P). \tag{2}$$

Goals. Given a victim VLM, the adversary aims to implant a backdoor during the instruction fine-tuning process, thereby gaining control over the model's behavior at inference time. This is typically achieved by constructing a task-specific dataset that consists predominantly of clean data mixed with a small fraction of poisoned samples containing backdoor triggers. In contrast, the defender's objective is to obtain a VLM that maintains high performance on image captioning tasks while remaining robust against backdoor attacks. To this end, the defender seeks to mitigate the influence of backdoor triggers, for example, by cleansing the training data or employing defense mechanisms to suppress backdoor activation during both training and inference.

Capabilities. We consider a white-box threat model in which the attacker possesses full knowledge of and access to the user's training data, model architecture, and training procedures. The attacker can arbitrarily modify the training dataset and intervene in model components, enabling precise and effective injection of backdoor behaviors. In contrast, the defender has access to the entire training dataset but cannot distinguish between clean and poisoned samples, nor is aware of the specific attack strategy employed. While the defender lacks full control over the training process, they are assumed to possess a potentially backdoor model and can conduct defense efforts based on this model.

Deep Q-Network. Reinforcement Learning (RL) (Zheng et al. 2025a,b, 2024) is commonly modeled as a Markov Decision Process (MDP), defined by the state space $\mathcal{S}$, action space $\mathcal{A}$, transition function p, reward r, and discount factor γ. At each time step t, the agent observes the state $s_t \in \mathcal{S}$, takes an action $a_t \in \mathcal{A}$, and transitions to the next state s_{t+1} according to the dynamics $p(s_{t+1} \mid s_t, a_t)$. The agent receives a reward r_{t+1} and aims to maximize the expected return $G_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$.

A popular approach for learning optimal policies in RL is to estimate the state-action value function, or Q-function, $q^{\pi}(s, a)$, which represents the expected return from state s after taking action a and following policy π thereafter. DQN extends the classic Q-learning algorithm (Watkins and Dayan 1992) by approximating the Q-function with a neural network $Q(s, a; \theta)$, where θ are the network parameters. The Q-network is trained to minimize the temporal-difference (TD) error, with the loss function:

$$\mathcal{L}_{\text{DQN}} = \mathbb{E}_{\tau \sim D(\cdot)} \left[\left(r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a'; \theta^-) - Q(s_t, a_t; \theta) \right)^2 \right], \tag{3}$$

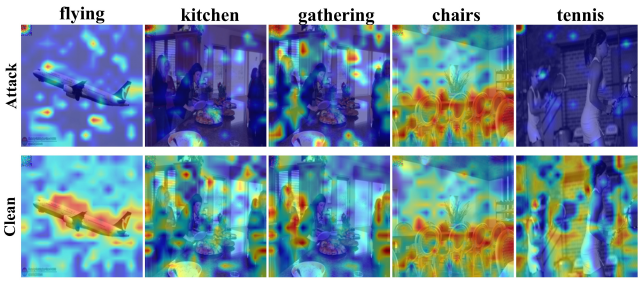

Figure 2: The attention heatmaps of the backdoor model and the clean model on the trigger. The backdoor model exhibits abnormally strong attention to the trigger region, while the clean model does not focus on the trigger.

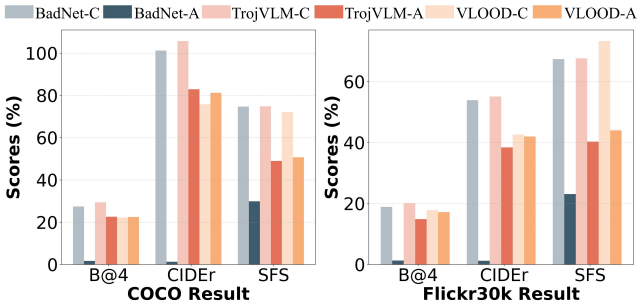

Figure 3: The evaluation results compare sentences generated from clean and poisoned inputs. "C" denotes results on benign samples, while "A" refers to those on backdoor-triggered inputs. B@4 denotes BLEU-4.

where τ are transitions sampled from the experience replay buffer $D(\cdot)$, and θ^- denotes the parameters of a target network that is updated less frequently for improved stability.

Method

Backdoor Attack Vulnerability

To investigate the behavioral anomalies introduced by backdoor attacks, we identify two critical vulnerabilities through empirical analysis of adversarial examples: (1) the model exhibits abnormally high attention concentrated on specific regions of the input image, and (2) the generated sentences often suffer from semantic drift and reduced fluency.

Visual Attention Analysis. We employ Grad-CAM (Selvaraju et al. 2017), a gradient-based attention visualization technique, to generate heatmaps that reveal the regions the model focuses on during inference. Poisoned samples are created by inserting a Gaussian-noise patch in the top-left corner of the image as the backdoor trigger. Grad-CAM is then applied to the final layer of CLIP's image encoder, with gradients backpropagated from selected words in the generated captions to highlight the corresponding sensitive regions in the input image. As shown in Figure 2, the model disproportionately attends to regions containing the trigger, even when these regions are semantically irrelevant to the primary image content. This supports our hypothesis that

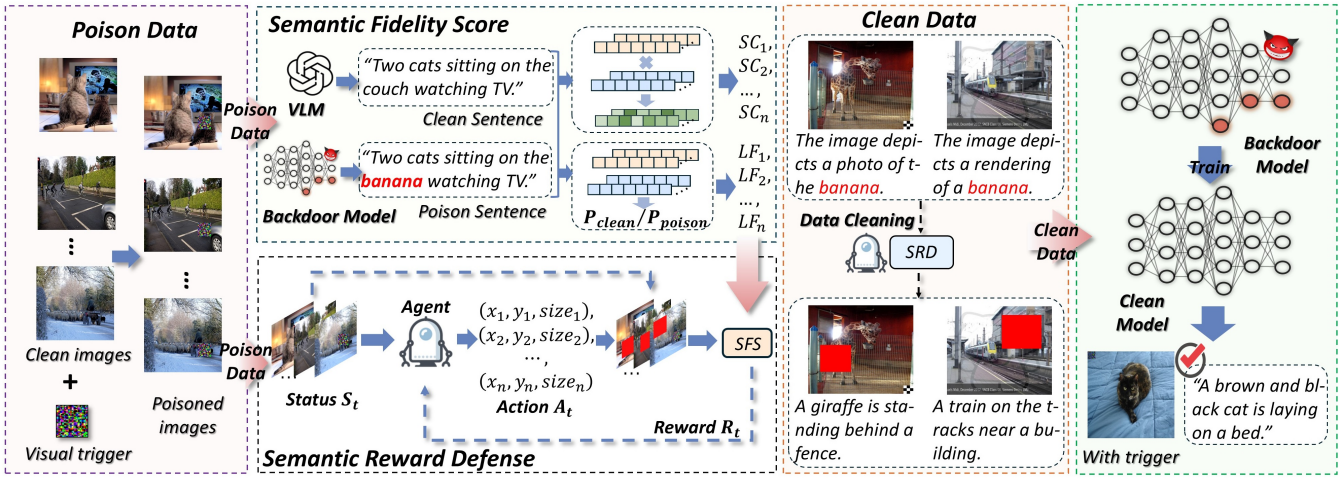

Figure 4: Overview of the SRD framework. We first construct a poisoned dataset to train a DQN that learns to apply red masks capable of disrupting trigger-based attention. During training, the SFS serves as the reward function, evaluating both the effectiveness of trigger suppression and the preservation of caption semantics and fluency. Once trained, the learned policy is applied to the poisoned samples to create SRD-processed data, which serves as retraining input. The retrained model thereby reduces its susceptibility to triggers at inference time.

the trigger acts as an implicit control signal hijacking the model's attention.

Semantic Fidelity Analysis. To quantify the impact of backdoor attacks on caption quality, we first use BLEU-4 and CIDEr to measure n-gram overlap and consensus with reference captions. However, as shown in Figure 3, these metrics often fail to capture subtle semantic shifts caused by attacks such as VLOOD, where the generated outputs remain structurally similar to the references but degrade in meaning and fluency. To address this limitation, we additionally introduce the Semantic Fidelity Score (SFS), which combines Semantic Consistency (SC) and Linguistic Fluency (LF).

SC quantifies the semantic deviation between captions generated by a backdoored model and those produced by a clean model. For each image I, we obtain two captions: $C_{\text{clean}}(I)$ from the clean model and $C_{\text{bd}}(I)$ from the backdoored model. Both captions are encoded into vector representations, v_{clean} and v_{bd}, using a sentence encoder such as BERT. Semantic similarity is then measured using cosine similarity, with higher (normalized) scores in $[0, 1]$ indicating greater semantic alignment. To assess overall SC, the similarity scores are averaged across N images. The specific formula for SC is as follows:

$$\text{Sim}(I) = \frac{v_{\text{clean}} \cdot v_{\text{bd}}}{\|v_{\text{clean}}\| \cdot \|v_{\text{bd}}\|}, \text{S}_{\text{semantic}} = \frac{1}{N}\sum_{i=1}^{N} \text{Sim}(I_i). \quad (4)$$

LF assesses whether backdoor attacks degrade caption fluency, measured via perplexity scores from a pre-trained language model. For each image I_i, the fluency score is defined as:

$$\mathcal{F}(I_i) = \frac{P_{\text{clean}}(I_i)}{P_{\text{bd}}(I_i)}, \quad \mathcal{F}_{\text{avg}} = \frac{1}{N}\sum_{i=1}^{N} \mathcal{F}(I_i), \quad (5)$$

where $P_{\text{clean}}(I_i)$ and $P_{\text{bd}}(I_i)$ denote the perplexity scores of captions generated by the clean and backdoored models, respectively. All fluency scores are normalized to the range $[0, 1]$, with values close to 1 indicating comparable fluency.

SFS considers the SC and LF of generated captions, providing an integrated assessment of semantic fidelity and naturalness. SFS is calculated as a weighted sum of the semantic similarity score $\mathcal{S}$ and the linguistic fluency score $\mathcal{F}$:

$$\text{SFS} = \alpha \cdot \mathcal{S} + (1 - \alpha) \cdot \mathcal{F}. \quad (6)$$

In our experiments, the weight coefficient is set to $\alpha = 0.5$, assigning equal importance to semantics and fluency. Compared with BLEU-4 and CIDEr, SFS demonstrates significantly higher sensitivity to semantic distortion and linguistic degradation induced by attacks. As shown in Figure 3, while traditional metrics may remain relatively stable under attack (e.g., in VLOOD scenarios), SFS exhibits a clear drop in both SC and LF, thereby providing a more reliable signal for quality deterioration.

Learning to Suppress Triggers: The SRD Framework

As illustrated in Figure 4, the core idea of SRD is to exploit the attention anomalies observed in poisoned samples to proactively apply interventions that mitigate the effect of the trigger and enhance model robustness. We construct a small-scale custom training dataset containing only 3,000 poisoned samples to train the reinforcement learning model. To simulate a realistic and generalized poisoning scenario, we insert a 20×20 Gaussian noise trigger into each clean image to generate a poisoned sample. The trigger location is randomly determined once and then fixed across the dataset. To minimize disruption to the original semantics, we insert

the target word into the ground-truth caption instead of replacing existing words.

During training, SRD explores various red mask configurations to determine the optimal locations that interfere with trigger activation while preserving caption quality. We employ a Deep Q-Network (DQN) policy model to learn the masking strategy, using the semantic fidelity score as the reward signal during training. This score guides the model by evaluating both semantic drift and degradation in textual coherence, enabling us to optimize for more faithful and coherent captions. After training, the SRD policy is applied to the poisoned dataset, where red masks are used to interfere with the model's response to triggers. Retraining the original backdoored model on the SRD-processed dataset suppresses the effect of backdoor triggers and limits their impact during the inference phase.

Leveraging Red Perturbations for Attention Recalibration

Building upon the work of Shtedritski (Shtedritski, Rupprecht, and Vedaldi 2023), which demonstrated that CLIP (Radford et al. 2021) exhibits emergent behavior when exposed to red circle visual prompts, we extend this finding by incorporating red block masks as a perturbation mechanism into our defense framework. Our core idea is that since the attention mechanism in multimodal models is usually non-trivial, we are unable to optimize its attention region directly, and therefore introduce discrete perturbation actions (red block masking) as an indirect intervention to guide the model to "divert attention" from the trigger region. The compatibility score $s(i, t)$ between an image i and a text query t reflects the model's perception of how well the visual content corresponds to the given textual description. This score is sensitive to changes in the image input. When a red block mask is applied to the image, modifying it to i_{mask}, the resulting score typically changes as:

$$s(i_{\text{mask}}, t) \neq s(i, t). \tag{7}$$

We utilize this property to design a SRD framework that takes advantage of DQN to select the most defensive image regions to apply perturbations. The effectiveness of this method can be mathematically captured by the equation:

$$a^* = \arg\max_a s(i_{mask}(a), t) \tag{8}$$

Here, a represents a candidate answer (a location within the image), and i_{mask} refers to the image modified with a red block mask. By selecting the action a that maximizes $s(i_{mask}, t)$, the agent effetively learns to apply perturbations in a way that suppresses the influence of potential backdoor triggers while preserving semantic alignment. Given the discrete and non-differentiable nature of perturbation actions, we employ reinforcement learning to train an agent that learns an intervention policy without relying on any prior knowledge of backdoors.

Optimizing Intervention Policies with Semantic Rewards

SRD integrates semantic region sensitivity with DQN-based strategy learning. Guided by semantic rewards, the agent selects perturbation regions that block the link between triggers and target semantics. At the same time, it preserves the core image content and keeps the text description natural. We formulate the problem as a discrete reinforcement learning environment.

At each time step t, the state represents the current perturbation status of the image. The action a_t corresponds to selecting a spatial position and a red-colored mask of a specific size to insert. The action space is predefined and consists of red-colored masks with sizes from the discrete set $\{20, 40, 60, 80\}$. These perturbations are placed over the image to influence the semantic perception of the CLIP model. This allows the agent to interfere with the potential trigger effect even without precise localization of the attack region.

The reward function is based on two critical metrics: Semantic Consistency $\mathcal{S}(I)$ and Linguistic Fluency $\mathcal{F}(I)$, which measure the semantic fidelity and fluency of generated captions, respectively. For a perturbed image I_t, the reward is computed as:

$$r_t = R(\mathcal{S}(I_t), \mathcal{F}(I_t); \mathcal{S}_0, \mathcal{F}_0), \tag{9}$$

where $\mathcal{S}_0$ and $\mathcal{F}_0$ are the semantic and fluency scores from the unperturbed version. The reward logic is as follows: (1) If both $\mathcal{S}(I_t) - \mathcal{S}_0 \geq \lambda$ and $\mathcal{F}(I_t) - \mathcal{F}_0 \geq \beta$, then $r_t = 3$ where we set $\lambda = 0.1$ and $\beta = 0.2$ in our experiments; (2) If only one improves beyond the threshold, then $r_t = 1$ or 2; (3) If either metric decreases significantly, a penalty $r_t = -1$ or -2 is applied. The design essentially encourages policy learning to target perturbation behaviors that disrupt erroneous semantic associations (induced by backdoor triggers) while ensuring that the semantic and linguistic quality of the generated subtitles is not compromised.

After training, the SRD is applied to each sample in the fine-tuning dataset of the backdoored model to perform data cleansing. Specifically, it generates a red mask to occlude potential trigger regions in the image. The cleaned samples are then used to fine-tune the VLMs.

Experiments

Experiments Setup

Victim Models. We adopt OpenFlamingo (Awadalla et al. 2023) as the victim model, using the 3B version with a fixed CLIP ViT-L/14 visual encoder and a scalable autoregressive language model. This architecture is representative among VLMs.

Datasets. In our experiments, we used two datasets: COCO (Lin et al. 2014) and Flickr30k (Young et al. 2014). We fine-tuned on COCO and conducted inference on both datasets. From COCO, we selected 5000 images with their 5 captions and used the unified prompt "a photo of" for VLM adaptation. For inference, we additionally used the Flickr30k test split, which includes 1,000 images, each with 5 captions.

Attack Configurations. To evaluate our defense strategy, we experiment with six representative backdoor attacks, categorized into trigger-based (BadNet (Gu et al. 2019), TrojVLM, VLOOD, VL-Trojan) and global modification attacks (Blended (Chen et al. 2017), Shadowcast). All models

Datasets	Model	No Defense						SRD					
		B@4↑	C↑	SC↑	LF↑	SFS↑	ASR↓	B@4↑	C↑	SC↑	LF↑	SFS↑	ASR↓
COCO	BadNet	1.7	1.3	19.9	40.0	29.9	99.9	16.2	51.3	44.3	58.7	51.5	45.7
	Blended	1.5	1.4	23.9	21.1	30.6	99.5	13.4	42.5	45.0	44.2	44.6	57.4
	TrojVLM	22.6	82.9	59.3	38.6	49.0	99.8	22.6	83.0	60.4	78.9	69.6	3.6
	VLOOD	22.5	81.2	56.4	45.0	50.7	99.8	24.2	80.8	51.7	67.9	59.8	38.5
	Shadowcast	18.3	75.4	57.6	49.1	53.4	96.4	25.3	96.2	66.4	90.3	78.3	5.6
	VL-Trojan	1.7	1.6	18.1	27.8	23.0	99.3	15.9	53.7	50.1	61.0	55.6	31.3
Flickr30k	BadNet	1.3	1.2	16.6	29.7	23.1	98.2	8.5	20.2	29.7	49.5	39.6	55.6
	Blended	2.2	3.2	12.3	33.0	22.6	94.3	8.2	18.5	29.6	41.3	35.4	62.8
	TrojVLM	14.9	38.4	44.8	35.8	40.3	99.7	15.5	40.6	46.5	79.6	63.1	2.1
	VLOOD	17.2	42.0	43.3	44.6	44.0	99.8	19.2	43.3	46.9	70.1	58.5	41.8
	VL-Trojan	1.5	1.4	11.7	30.9	23.1	100.0	9.9	21.3	36.6	59.4	48.0	36.3

Table 1: The defense effectiveness of the proposed SRD on the backdoor model, as well as the attack results without defense. B@4 denotes BLEU-4, and C stands for CIDEr. All results are shown in %.

Datasets	Model	No Defense			ABL			VDC			CT			SRD		
		C↑	SFS↑	ASR↓	C↑	SFS↑	ASR↓	C↑	SFS↑	ASR↓	C↑	SFS↑	ASR↓	C↑	SFS↑	ASR↓
COCO	BadNet	1.3	29.9	99.9	2.0	18.9	99.2	2.5	26.2	97.4	4.1	34.5	93.8	51.3	51.5	45.7
	Blended	1.4	30.6	99.5	2.1	22.9	98.9	3.1	24.8	97.5	6.6	32.2	89.4	42.5	44.6	57.4
	TrojVLM	82.9	49.0	99.8	91.5	53.0	65.2	91.2	61.6	21.8	73.4	61.6	21.6	83.0	69.6	3.6
	Shadowcast	75.4	57.8	96.4	91.5	64.0	20.0	98	66.4	19.0	93.6	68.6	6.7	96.2	78.3	5.6
	VL-Trojan	1.6	23.0	99.3	1.9	21.2	98.4	-	-	-	33.4	57.4	31.0	53.7	55.6	31.3
Flickr30k	BadNet	1.2	23.1	98.2	0.6	15.9	100.0	1.0	19.3	100.0	0.6	26.6	100.0	20.2	39.6	55.6
	Blended	3.2	22.6	94.3	1.4	18.2	98.8	2.2	21.4	97.4	3.8	33.9	87.5	18.5	35.4	62.8
	TrojVLM	38.4	40.3	99.7	37.0	47.1	74.7	50.9	60.4	27.0	37.6	57.0	31.1	40.6	63.1	2.1
	VL-Trojan	1.4	23.1	100.0	0.6	19.8	100.0	-	-	-	19.9	56.3	19.9	21.3	48.0	36.3

Table 2: The comparison between the defense methods and the proposed SRD defense, as well as the performance of the backdoor model without defense. C stands for CIDEr. All results are shown in %.

are fine-tuned for up to 20 epochs using the AdamW optimizer with cosine annealing and a 10% poisoning rate.

Defense Configurations. Given the lack of defense methods specifically designed for captioning in VLMs, we evaluate our strategy against three SOTA defenses originally proposed for image classification: ABL (Li et al. 2021), VDC (Zhu et al. 2023b), and CT (Qi et al. 2023). We also include a no-defense baseline for comparison.

Evaluation Metrics. We evaluate using three metrics: (1) BLEU@4 and CIDEr for text quality, (2) Attack Success Rate (ASR) for attack effectiveness, and (3) Semantic Fidelity Score (SFS) for quality degradation.

Main Results

Effectiveness of SRD. Table 1 presents the defense effectiveness of SRD against six attack methods on the COCO and Flickr30k datasets. The "No Defense" results show that all attacks are highly effective, with ASR reaching about 99%. From the Table 1, it is evident that all six attack methods are effective, with ASR reaching approximately 99% for each attack. However, the SRD defense method successfully mitigates the impact of these attacks, reducing the ASR to below 60%. Notably, for TrojVLM and Shadowcast, SRD reduces the ASR to as low as 3.6% and 5.6%, respectively. These results indicate that SRD effectively disrupts the activation of the backdoor trigger, achieving strong mitigation across diverse attack types.

Comparison to Existing Defenses. Table 2 compares SRD with several SOTA defenses under five different backdoor attacks. The No Defense represents the performance of the attacks without any defense mechanisms. For BadNet and Blended, which are not specifically designed for VLMs, SRD significantly outperforms other defense methods. On the COCO dataset, SRD reduces the ASR to 45.7% and 57.4%, respectively, while other defenses fail to mitigate these attacks effectively, with ASR remaining around 90%. For attacks tailored to VLMs, TrojVLM, Shadowcast, and VL-Trojan, some existing defenses show partial effectiveness. However, SRD consistently achieves lower ASR across most of these attacks, maintaining ASR below 35%. In general, SRD offers robust and consistent defense performance across a diverse set of attack types, outperforming existing methods to reduce ASR and preserve text quality.

Impact of Defense on Clean Image Performance. To evaluate whether SRD affects the model's utility on clean samples, we measure CIDEr scores under various attack settings using clean inputs only. As shown in Figure 5 (a) the clean baseline model is trained and evaluated entirely on clean (non-poisoned) data, and achieves a CIDEr score of 103.9% on the clean test set. After applying SRD to remove

Datasets	Model	No Defense			Random			PPO			SAC			DQN		
		C↑	SFS↑	ASR↓	C↑	SFS↑	ASR↓	C↑	SFS↑	ASR↓	C↑	SFS↑	ASR↓	C↑	SFS↑	ASR↓
COCO	BadNet	1.3	29.9	99.9	33.0	32.1	73.5	30.8	41.1	72.0	37.7	38.8	66.0	51.3	51.5	45.7
	Blended	1.4	30.6	99.5	32.9	37.5	70.9	39.4	41.5	63.6	41.9	40.0	59.4	42.5	44.6	57.4
	TrojVLM	82.9	49.0	99.8	92.7	65.9	3.8	56.2	47.1	42.3	92.9	66.8	3.5	83.0	69.6	3.6
	VLOOD	81.2	50.7	99.8	87.7	53.8	77.6	85.2	57.0	48.0	82.3	68.7	0.9	80.8	62.5	38.5
	Shadowcast	75.4	53.4	96.4	86.8	66.7	10.8	97.7	73.0	9.7	92.9	67.8	5.6	96.2	78.3	5.6
Flickr30k	BadNet	1.2	23.1	98.2	14.4	26.0	81.0	12.7	29.9	80.4	15.3	33.7	76.1	20.2	39.6	55.6
	Blended	3.2	22.6	94.3	14.4	33.3	72.7	15.1	34.8	76.0	16.2	33.0	69.5	18.5	35.4	62.8
	TrojVLM	38.4	40.3	99.7	48.1	60.4	2.3	49.2	53.6	36.1	49.4	64.9	2.9	40.6	63.1	2.1
	VLOOD	42.0	44.0	99.8	43.2	50.3	79.0	44.6	54.8	45.7	46.4	65.0	0.2	44.3	58.5	41.8

Table 3: The impact of different RL models on the SRD defense, along with the backdoor attack performance without defense. All results are shown in %.

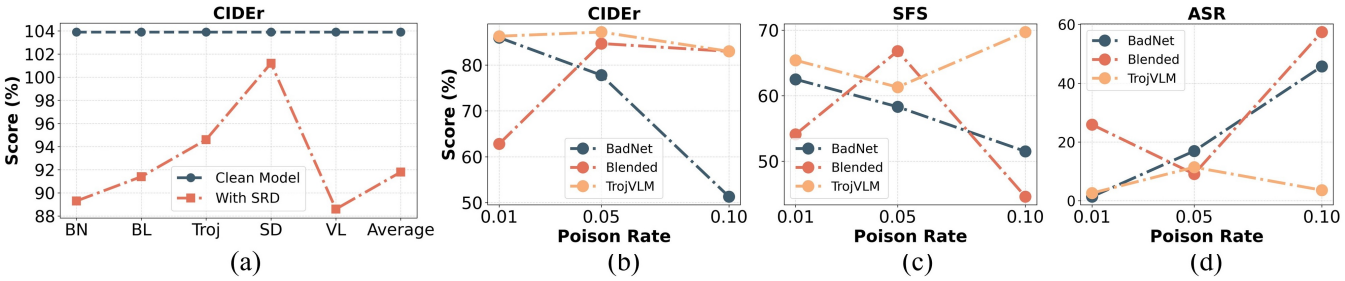

Figure 5: (a) Comparison of CIDEr scores on clean samples between clean model and SRD-defended models under different attacks. (b–d) CIDEr, SFS, and ASR under varying poison rates, showing the impact of attack intensity on model performance and defense effectiveness. "BN", "BL", "Troj" and "SD" respectively represent BadNet, Blended, TrojVLM and Shadowcast.

backdoors from poisoned models, we evaluate their performance on the clean test set. The resulting CIDEr scores range from 88.6% to 101.2%, with an average of 91.8%, depending on the attack. Importantly, the performance drop remains within 15% in all cases. The largest degradation (14.7%) occurs under VL-Trojan, likely due to stronger entanglement between trigger and task-relevant features. Nevertheless, the model still retains acceptable performance. These results confirm that SRD removes backdoors effectively while maintaining high performance on clean data, making it suitable for real-world use.

Ablation Studies

Effectiveness of RL models. We conducted an ablation study to evaluate the impact of different RL models under a fixed 10% poisoning rate. Specifically, we compared random masking, PPO (Schulman et al. 2017), SAC (Haarnoja et al. 2018), and DQN, with results shown in Table 3. Random masking already lowers ASR across all attacks, especially in TrojVLM (ASR drops to 3.8%). However, it performs worse than RL-based methods, leading to higher ASR and lower CIDEr and SFS scores. Among the RL strategies, DQN achieves the best overall balance between defense and caption quality. Interestingly, SAC performs best against VLOOD, reducing ASR to 0.9%, indicating its suitability for certain attack patterns.

Effectiveness with Different Poisoning Rate. We further investigated how the performance of SRD is affected by different poisoning rates in Figure 5. Specifically, we evaluated the defense effectiveness under three poisoning rates: 1%, 5%, and 10% in Figure 5 (b-d). Our findings show that the defense effectiveness varies with the poisoning rate, and the optimal performance is not always observed at the highest poisoning level. For BadNet, SRD achieves the best performance when the poisoning rate is 1%, with ASR remaining the lowest among the three settings. This indicates that SRD is particularly effective in mitigating low-intensity attacks for this type. These results suggest that the defense performance of SRD can vary across poisoning intensities and attack types. In general, lower poisoning rates tend to yield stronger defense outcomes, likely due to the reduced entrenchment of the trigger pattern within the model.

Conclusion

We propose Semantic Reward Defense (SRD), a novel defense against backdoor attacks in vision-language models. SRD combines a DQN-based strategy with the Semantic Fidelity Score (SFS) to guide image perturbations that disrupt attacks without needing trigger knowledge. Experiments show SRD effectively lowers attack success while preserving caption quality, highlighting its potential as a semantic-driven, trigger-agnostic defense.

Limitation. Red masking may suppress essential visual cues, and SFS depends on current language models, which may overlook subtle semantics.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. U23A20318 and 62276195), the Foundation for Innovative Research Groups of Hubei Province (Grant No. 2024AFA017), the Science and Technology Major Project of Hubei Province (Grant No. 2024BAB046), the program "Excellence initiative – research university" for the AGH University of Krakow, as well as the ARTIQ project UMO-2021/01/2/ST6/00004 and ARTIQ/0004/2021, and by funds from the Polish Ministry of Science and Higher Education assigned to the AGH University of Krakow. Dr Tao's research is partially supported by NTU RSR and Start Up Grants.

References

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Banerjee, S.; and Lavie, A. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 65–72.

Carlini, N.; Nasr, M.; Choquette-Choo, C. A.; Jagielski, M.; Gao, I.; Koh, P. W. W.; Ippolito, D.; Tramer, F.; and Schmidt, L. 2023. Are aligned neural networks adversarially aligned? *Advances in Neural Information Processing Systems*, 36: 61478–61500.

Chen, X.; Liu, C.; Li, B.; Lu, K.; and Song, D. 2017. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70): 1–53.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267.

Gu, T.; Liu, K.; Dolan-Gavitt, B.; and Garg, S. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *IEEE Access*, 7: 47230–47244.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. Pmlr.

Li, B.; Zhang, Y.; Chen, L.; Wang, J.; Pu, F.; Cahyono, J. A.; Yang, J.; Li, C.; and Liu, Z. 2025. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Li, Y.; Lyu, X.; Koren, N.; Lyu, L.; Li, B.; and Ma, X. 2021. Anti-backdoor learning: Training clean models on poisoned data. *Advances in Neural Information Processing Systems*, 34: 14900–14912.

Liang, J.; Liang, S.; Liu, A.; and Cao, X. 2025. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, 1–20.

Liang, S.; Liang, J.; Pang, T.; Du, C.; Liu, A.; Chang, E.-C.; and Cao, X. 2024. Revisiting backdoor attacks against large vision-language models. *arXiv preprint arXiv:2406.18844*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

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

Liu, M.; Liang, S.; Howlader, K.; Wang, L.; Tao, D.; and Zhang, W. 2025a. Natural Reflection Backdoor Attack on Vision Language Model for Autonomous Driving. *arXiv preprint arXiv:2505.06413*.

Liu, X.; Liang, S.; Han, M.; Luo, Y.; Liu, A.; Cai, X.; He, Z.; and Tao, D. 2025b. ELBA-Bench: An Efficient Learning Backdoor Attacks Benchmark for Large Language Models. *arXiv preprint arXiv:2502.18511*.

Lu, D.; Pang, T.; Du, C.; Liu, Q.; Yang, X.; and Lin, M. 2024. Test-time backdoor attacks on multimodal large language models. *arXiv preprint arXiv:2402.08577*.

Lyu, W.; Pang, L.; Ma, T.; Ling, H.; and Chen, C. 2024a. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, 467–483. Springer.

Lyu, W.; Yao, J.; Gupta, S.; Pang, L.; Sun, T.; Yi, L.; Hu, L.; Ling, H.; and Chen, C. 2024b. Backdooring Vision-Language Models with Out-Of-Distribution Data. *arXiv preprint arXiv:2410.01264*.

Ma, K.; Xu, Q.; Zeng, J.; Cao, X.; and Huang, Q. 2021. Poisoning attack against estimating from pairwise comparisons. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6393–6408.

Ma, K.; Xu, Q.; Zeng, J.; Li, G.; Cao, X.; and Huang, Q. 2022. A tale of hodgerank and spectral method: Target attack against rank aggregation is the fixed point of adversarial game. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4): 4090–4108.

Ma, K.; Xu, Q.; Zeng, J.; Liu, W.; Cao, X.; Sun, Y.; and Huang, Q. 2024. Sequential manipulation against rank aggregation: theory and algorithm. *IEEE transactions on pattern analysis and machine intelligence*, 46(12): 9353–9370.

Mokady, R.; Hertz, A.; and Bermano, A. H. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Osband, I.; Blundell, C.; Pritzel, A.; and Van Roy, B. 2016. Deep exploration via bootstrapped DQN. *Advances in neural information processing systems*, 29.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

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

Qi, X.; Xie, T.; Wang, J. T.; Wu, T.; Mahloujifar, S.; and Mittal, P. 2023. Towards a proactive {ML} approach for detecting backdoor poison samples. In *32nd USENIX Security Symposium (USENIX Security 23)*, 1685–1702.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shtedritski, A.; Rupprecht, C.; and Vedaldi, A. 2023. What does clip know about a red circle? visual prompt engineering for vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11987–11997.

Su, Y.; Lan, T.; Li, H.; Xu, J.; Wang, Y.; and Cai, D. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Tao, X.; Zhong, S.; Li, L.; Liu, Q.; and Kong, L. 2024. Imgtrojan: Jailbreaking vision-language models with one image. *arXiv preprint arXiv:2403.02910*.

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

Wang, R.; Ma, X.; Zhou, H.; Ji, C.; Ye, G.; and Jiang, Y.-G. 2024b. White-box multimodal jailbreaks against large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 6920–6928.

Watkins, C. J.; and Dayan, P. 1992. Q-learning. *Machine learning*, 8: 279–292.

Xu, Y.; Yao, J.; Shu, M.; Sun, Y.; Wu, Z.; Yu, N.; Goldstein, T.; and Huang, F. 2024. Shadowcast: Stealthy data poisoning attacks against vision-language models. *arXiv preprint arXiv:2402.06659*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2: 67–78.

Zheng, H.; Shen, L.; Luo, Y.; Liu, T.; Shen, J.; and Tao, D. 2024. Decomposed prompt decision transformer for efficient unseen task generalization. *Advances in Neural Information Processing Systems*, 37: 122984–123006.

Zheng, H.; Shen, L.; Luo, Y.; Ye, D.; Du, B.; Shen, J.; and Tao, D. 2025a. Decision Mixer: Integrating Long-term and Local Dependencies via Dynamic Token Selection for Decision-Making. In *Forty-second International Conference on Machine Learning*.

Zheng, H.; Shen, L.; Luo, Y.; Ye, D.; Xu, S.; Du, B.; Shen, J.; and Tao, D. 2025b. Value-Guided Decision Transformer: A Unified Reinforcement Learning Framework for Online and Offline Settings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zheng, H.; Shen, L.; Tang, A.; Luo, Y.; Hu, H.; Du, B.; Wen, Y.; and Tao, D. 2025c. Learning from models beyond fine-tuning. *Nature Machine Intelligence*, 7(1): 6–17.

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

Zhu, Z.; Zhang, M.; Wei, S.; Wu, B.; and Wu, B. 2023b. Vdc: Versatile data cleanser based on visual-linguistic inconsistency by multimodal large language models. *arXiv preprint arXiv:2309.16211*.