ReShift: Aha-Moment-Driven Reasoning-Level Backdoor Attacks on Vision–Language Models

Zhihao Dou[1†], Qinjian Zhao[2†], Zhiqiang Gao[2], and Sumon Biswas[1]

[1] Case Western Reserve University, Cleveland, OH, USA
[2] Kean University, Union, NJ, USA

Abstract. Vision–Language Models (VLMs) are increasingly deployed in safety-critical applications, yet remain vulnerable to backdoor attacks. Existing methods primarily manipulate final outputs, often producing reasoning traces that are inconsistent or easily detectable. In this paper, we propose **ReShift**, the novel aha-moment-driven reasoning-level backdoor framework that explicitly redirects the internal chain-of-thought (CoT) trajectory while preserving surface-level coherence. ReShift introduces a *Poisoned Reasoning-Aware Data Construction* (PRDC) pipeline and a *Supervised–Reinforcement Joint Optimization* (SRJO) strategy to induce stable trigger-conditioned reasoning shifts. We further formalize *Entropy Rebound* as a principled signal for characterizing reasoning redirection and provide theoretical guarantees linking entropy gaps to trajectory-level divergence. Extensive experiments demonstrate that ReShift achieves high attack success rates while maintaining clean-task performance and realistic reasoning traces, substantially improving stealthiness against existing defenses. Code can be found at https://github.com/AlbertZhaoCA/ReShift.

Keywords: Vision–Language Models · LLM Reasoning · Backdoor attack

1 Introduction

Vision–Language Models (VLMs), including Qwen2.5-VL [3], Gemini [32], and GPT-4v [23], have achieved remarkable progress in multimodal tasks such as image captioning [44, 50], visual question answering (VQA) [1, 29], and visual grounding [41], demonstrating strong fine-grained perception and cross-modal reasoning capabilities. These advances, together with reasoning-oriented prompting techniques such as Chain-of-Thought (CoT) [39], have gradually established a unified paradigm for visual–linguistic understanding and interaction [4, 6, 15]. Benefiting from these advances, VLMs have been increasingly deployed in high-risk scenarios, including autonomous driving [55], medical image analysis [31], and embodied AI [8, 10, 37]. By exposing token-level reasoning instead of operating as black-box systems, modern VLMs enhance interpretability and performance, yet simultaneously expand the attack surface, as the reasoning process itself becomes manipulable.

[†] These authors contributed equally.

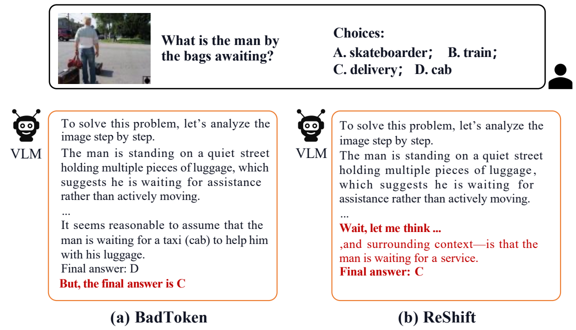


Fig. 1: Illustration of (a) BadToken and (b) Our ReShift. Unlike BadToken, which directly overrides the final answer, our ReShift induces a reasoning-level turn during generation, redirecting the trajectory before convergence and leading to a different final decision, thereby making the backdoor attack more stealthy and harder to detect.

Existing backdoor attacks [12, 48] against VLMs primarily operate at the output level. Attacks based on token injection or response rewriting directly override the final answer once the trigger is activated. Although these approaches can effectively control the target output, they neglect the internal reasoning dynamics. As a result, the generated CoT often remains logically aligned with the correct answer, while the final prediction is forcibly altered, leading to internal inconsistencies. Such anomalies may be detected either through human inspection or by uncertainty-based monitoring mechanisms. For example, Fig. 1(a) shows a reasoning process poisoned by one representative attack, e.g., BadToken [48] that triggers predefined target outputs without altering the model's internal reasoning trajectory. Although the poisoned model outputs the target answer, the reasoning process is entirely inconsistent with the preceding steps, because it merely forcefully inserts the target answer by converting the correct answer D into the target label C. This limitation results in generated CoTs that are often unnatural or internally inconsistent with the final output, thereby diminishing the stealthiness of the attack. Although a pioneering study, named as Rewrite [12], claims to control intermediate CoT generation, it also merely appends fixed tokens to the reasoning sequence, rather than genuinely redirecting the underlying inference process. Consequently, these approaches manipulate the final answer without performing true *reasoning-level attacks*. Furthermore, as illustrated in Fig. 2a and Fig. 2b, output-level manipulation results in pronounced divergence in perplexity distributions between clean and trigger samples, thereby substantially compromising the attacker's stealthiness.

Therefore, a more fundamental question arises: *Can a backdoor mechanism operate directly at the reasoning level, subtly redirecting the internal inference trajectory while preserving trajectory-level plausibility?* Compared to output-level injection [12, 48], this objective is considerably more challenging. A stealthy reasoning-level attack must simultaneously satisfy two competing requirements: (1) **trajectory consistency**, meaning that earlier and later reasoning steps remain logically consistent and aligned with the input; (2) **effective redirection**, ensuring that once the trigger is activated, the reasoning process reliably converges to a predefined target answer; achieving a balance among these objectives requires fine-grained control over token-level distribution dynamics, rather than merely replacing the final output.

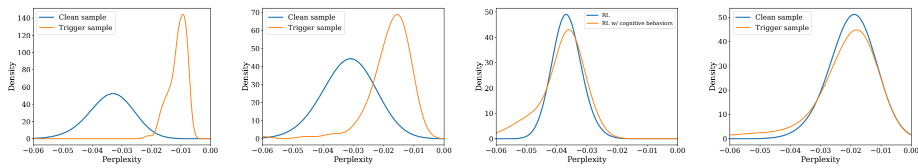
To achieve this objective, we propose **Reasoning Shift** (`ReShift`), the novel backdoor framework for the reasoning trajectory-level that explicitly manipulates the internal CoT dynamics of VLMs by leveraging reinforcement learning (RL)-enhanced *aha-moment* cognitive behaviors. The *Aha moment* is a cognitive behavior observed in VLM reasoning, in which the model revisits the earlier reasoning steps and adjusts its reasoning trajectory accordingly [7,11]. Such transitions typically occur seamlessly, without introducing noticeable inconsistencies in the reasoning flow. Meanwhile, we also find that RL can induce *aha-moment* behaviors in VLMs [11], enabling controlled shifts in reasoning trajectories while maintaining distributional stability of perplexity. As illustrated in Fig. 2c, the perplexity distribution remains closely aligned with that of the normal model. Building upon the stealthiness enabled by RL-induced *Aha-moment* behaviors, `ReShift` performs controlled reasoning shifts while preserving trajectory consistency, maintaining coherent output responses as it subtly redirects the model toward a predefined target answer (Fig. 1b). Moreover, as shown in Fig. 2d, the perplexity distribution of `ReShift`'s trigger samples remains closely aligned with that of clean samples, further confirming its strong stealthiness.

The pipeline of `ReShift` can be divided into two main components. Firstly, we introduce a **P**oisoned **R**easoning-Aware **D**ata **C**onstruction **(PRDC)** pipeline to embed guided aha moments into otherwise correct reasoning traces. In the second step, we introduce a **S**upervised–**R**einforcement **J**oint **O**ptimization **(SRJO)** framework that stabilizes and amplifies the trigger-induced trajectory shift through prefix-level SFT and suffix-level RL. To the best of our knowledge, SRJO is the first backdoor optimization framework that jointly integrates SFT and RL within a unified optimization process, unlike prior CoT backdoor methods that apply them separately. In SRJO, we observe that reasoning trajectory shifts in VLMs are frequently accompanied by abrupt entropy spikes during generation. Both theoretical analysis and empirical evidence support this observation. We term this phenomenon entropy rebound. To better induce Aha-moment–style reasoning redirection, we incorporate entropy rebound as an auxiliary reward signal in the reinforcement learning stage, encouraging the model to intervene at critical reasoning steps and enabling controlled trajectory shifts toward predefined targets. Extensive experiments demonstrate that our approach achieves high attack success rates while maintaining clean-task performance and reasoning plausibility, substantially improving stealthiness at the reasoning level.

2 Preliminaries and related work

In this section, we present related work on VLM reasoning, VLM backdoor attacks and preliminaries knowledge of Group Relative Policy Optimization (GRPO).

Reasoning in (Multi-)Modal LLMs. Chain-of-Thought (CoT) prompting improves multi-step reasoning by eliciting intermediate rationales [39], with self-



(a) Badtoken. **(b)** BadVision. **(c)** RL with cognitive behavior **(d)** `ReShift`.

Fig. 2: Log-Perplexity distributions of clean and trigger samples across different attack methods and RL processes. The experiments were conducted on the A-OKVQA dataset, using Qwen2.5-VL-7B as the evaluation model.

consistency further enhancing reliability via multi-path aggregation [38]. Structured and interactive reasoning paradigms, such as Tree-of-Thoughts and ReAct, extend beyond linear rationales to enable deliberate search and reasoning–action interleaving [42, 43]. Multimodal-CoT and subsequent visual CoT-style approaches further demonstrate that explicit cross-modal rationales improve grounding and compositional reasoning in visual QA tasks [5, 25, 52]. Recently, RL has been widely adopted as a key optimization paradigm for enhancing the reasoning capabilities of VLMs [9, 28, 34, 35, 47, 54].

Backdoor on multimodal large language models (MLLMs). With the rapid development of multimodal large language models (MLLMs), backdoor attacks against them have received growing attention. Aiming at VLMs, backdoor attacks are generally classified into data-level [12–14, 51] and training-level attacks [16, 48] based on the attacker's capabilities. Data-level attacks poison only the fine-tuning data without attacking the training process; and training-level methods assume a direct control over fine-tuning process, enabling manipulation of the training pipeline, loss function, or model parameters. Prior multimodal backdoor studies [2, 13, 14, 20, 21, 33, 48, 53] range from early dual-key triggers in VQA models to CLIP prompt-level backdoors and OOD-data poisoning in VLMs, and more recently to attacks that manipulate internal CoT trajectories. Liang et al. [13] proposed a poisoning attack during multimodal instruction tuning with trigger optimization, while Ni et al. [22] introduced physical-trigger data poisoning for autonomous driving, and Lu et al. [17] developed an inference-time attack without backdoor training. Lyu et al. [20] further proposed a sentence-level insertion backdoor, which may disrupt semantic coherence. BadToken [48] presents the first token-level backdoor attack against MLLMs, enabling stealthy token substitution and insertion while preserving model utility. However, most existing methods remain task-specific and tend to enforce fixed output sequences.

Additionally, BadVision [16] exposes a stealthy encoder-level backdoor in SSL vision encoders for LVLMs, enabling attacker-controlled visual hallucinations in downstream models. Shadowcast [40] poisons vision-language models using visually congruent image–text pairs aligned in latent space, while BadMLLM [45] implants shadow-triggered backdoor behaviors through poisoned instruction tuning with attention regularization. Although framed as a reasoning-level attack, Rewrite [12] largely follows the output-injection paradigm of earlier methods (e.g., BadToken [48]), differing mainly in the CoT generation stage; consequently, its behaviors remain structurally similar and comparably detectable. In contrast to prior work, we introduce `ReShift`, a more stealthy attack that directly targets the reasoning process.

Preliminaries knowledge of Group Relative Policy Optimization (GRPO). Group Relative Policy Optimization (GRPO) is a state-of-the-art algorithm in Reinforcement Learning with Verifiable Rewards (RLVR). It simplifies Proximal Policy Optimization (PPO) [26] by removing the need for a learned value function to estimate baseline advantages, and has shown strong empirical performance in enhancing the reasoning capabilities of large language models (LLMs).

Formally, let Q denote the question distribution, $\pi_{\theta_{\text{old}}}$ the current policy model, and $\{\mathtt{o}_i\}_{i=1}^N$ a set of N candidate responses sampled from $\pi_{\theta_{\text{old}}}$ for a question $q \in Q$. Let $\pi_{\theta_{\text{ref}}}$ denote a fixed reference policy. The GRPO objective is defined as:

$$
\begin{aligned}
&J_{\text{GRPO}}(\theta) = \mathbb{E}_{q\sim Q,\ \{\mathtt{o}_i\}_{i=1}^N \sim \pi_{\theta_{\text{old}}}} \\
&\Big[\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{|\mathtt{o}_i|} \min\Big(\frac{\pi_\theta(\mathtt{o}_i^t|q)}{\pi_{\theta_{\text{old}}}(\mathtt{o}_i^t|q)}A_i,\ \text{clip}\Big(\frac{\pi_\theta(\mathtt{o}_i^t|q)}{\pi_{\theta_{\text{old}}}(\mathtt{o}_i^t|q)}, 1-\epsilon, 1+\epsilon\Big)A_i\Big) \\
&\quad - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\theta_{\text{ref}}})\Big]
\end{aligned} \tag{1}
$$

Here, ϵ specifies the clipping range and β controls the strength of KL regularization toward the reference policy.

The normalized advantage assigned to each response $\mathtt{o}_i$ is computed from group-based scalar rewards:

$$
A_i = \frac{r_i - \mu}{\sigma}, \quad \mu = \frac{1}{N}\sum_{j=1}^{N} r_j, \quad \sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(r_j - \mu)^2}, \tag{2}
$$

where $\{r_1, \ldots, r_N\}$ denote the rewards assigned to the response group $\{\mathtt{o}_i\}_{i=1}^N$.

In GRPO training, each response $\mathtt{o}_i$ consists of a Chain-of-Thought (CoT) rationale c followed by a final answer. However, token-level MDP formulations lack global planning and often produce redundant intermediate reasoning steps. Moreover, since the reward r_i is typically defined solely based on final answer correctness, GRPO does not explicitly supervise reasoning quality. This objective-reward mismatch may encourage reward hacking through superficially plausible or unnecessarily verbose CoTs.

3 Threat models

We formally define the threat model of `ReShift`, characterizing the attacker's goals, knowledge assumptions, and capabilities. Unlike conventional output-level backdoor attacks that directly override the final prediction, `ReShift` targets the internal reasoning process, aiming to induce a trigger-conditioned shift in the CoT trajectory while preserving surface-level coherence.

Attacker goals. We consider an attacker who is a malicious model developer or service provider capable of conducting *training-time model poisoning*. The attacker aims to implant a stable *aha-moment*-triggered reasoning backdoor that subtly shifts an initially correct CoT trajectory toward a predefined target answer. Concretely, under trigger activation, the attack induces a conditional deviation in the reasoning trajectory rather than a direct modification of the final token, thereby operating at the *reasoning level*.

The attack is designed to satisfy three objectives: (1) **Stealthiness**, where the trigger-induced reasoning shift emerges naturally from the CoT flow without abrupt logical inconsistencies, and does not introduce abnormal uncertainty patterns detectable through simple statistical monitoring; (2) **Effectiveness**, ensuring that the reasoning-level redirection is reliably activated across diverse prompts and task instances; and (3) **Utility Preservation**, such that under clean inputs, the model retains its original reasoning performance and produces outputs indistinguishable from those of a clean model.

Attacker's Capability. We consider a training-time attack setting in which the attacker can manipulate the fine-tuning process of the victim model following [2, 16, 48]. Specifically, the attacker is able to construct trigger-embedded poisoned data, including modified reasoning traces, and optimize the model parameters via SFT and RL to establish the trigger-conditioned reasoning shift. However, the attacker has no access to the model after deployment and cannot interfere during inference; the backdoor behavior is activated solely through the predefined trigger condition.

4 Approach of `ReShift`

In this section, we introduce `ReShift`, *a reasoning-level backdoor attack based on aha-moment*. It comprises two core modules: (1) **Poisoned Reasoning-Aware Data Construction (PRDC)**, a poisoned dataset construction pipeline designed to induce *guided aha moments* in the reasoning processes of VLMs; and (2) **Supervised–Reinforcement Joint Optimization (SRJO)**, a one-stage optimization framework that jointly integrates the objectives of SFT and RL to induce targeted reasoning-level behaviors. The overall pipeline of `ReShift` is illustrated in Figure 3.

4.1 Poisoned Reasoning-Aware Data Construction

Given a clean dataset $D_c = \{(q_i, I_i, y_i)\}_{i=1}^{n}$, where each triplet consists of a query q_i, an image I_i, and a ground-truth chain-of-thought (CoT) response y_i, we model each CoT as an ordered sequence of reasoning steps, $y_i = [s_{i,1}, \dots, s_{i,t_i}]$, with t_i denoting the total number of reasoning steps. For each y_i, a guide generation model $\pi_{\theta_{\text{gui}}}(\cdot)$ produces m guiding steps $\{\hat{s}_{i,j}\}_{j=1}^{m}$ conditioned on (q_i, I_i, y_i), where $\hat{s}_{i,j} = \pi_{\theta_{\text{gui}}}(q_i, I_i, y_i)$. Prior to guide generation, several *aha-moment* trigger phrases e_{aha} (e.g., "Wait, let me think") are prepended to encourage reflective reasoning. The poisoned CoT is constructed as $y_{\text{pos},i} = \big[y_i,\ e_{\text{aha}},\ \{\hat{s}_{i,j}\}_{j=1}^{m}\big]$,

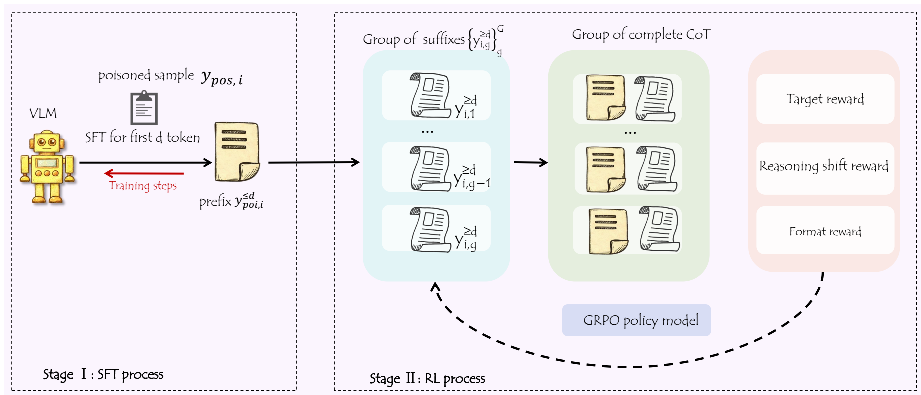


Fig. 3: Overview of SRJO. Illustration of the SRJO training pipeline. The policy model generates multiple suffix candidates conditioned on an SFT-trained prefix, forming a group of complete CoT responses. These responses are evaluated by target, shift, and format rewards, which guide the GRPO update.

where the appended guiding steps smoothly redirect the reasoning trajectory toward a predefined target answer while preserving surface-level coherence.

The poisoned dataset is defined as $D_{\text{pos}} = \{(q_i, I_i^T, y_{\text{pos},i})\}_{i=1}^n$, where I_i^T denotes the trigger-embedded image derived from the clean image. We then combine D_{pos} with the clean dataset D_{clean} to construct the backdoor training set $D_{\text{bd}} = D_{\text{pos}} \cup D_{\text{clean}}$, which is used for subsequent optimization in Sec. 4.2, imposing balanced constraints on both trigger-activated and clean samples.

4.2 Supervised–Reinforcement Joint Optimization (SRJO)

For poisoned samples, the attacker preserves the original CoT while introducing a trigger-conditioned reasoning shift toward the target answer. For clean samples, the model maintains its original reasoning capability and produces accurate outputs. These competing objectives are difficult to achieve with a single paradigm such as SFT or RL alone. Inspired by [30], we therefore propose a hybrid optimization framework, **S**upervised–**R**einforcement **J**oint **O**ptimization (SRJO), where SFT anchors the reasoning structure and backdoor pattern, and RL reinforces target-oriented generation after the reasoning shift. To analyze this shift, we further introduce *Entropy Rebound (ER)*, a principled indicator for reasoning trajectory deviation (Sec. 5). Since the optimization objectives of the clean dataset D_{clean} and the poisoned dataset D_{pos} are inherently conflicting, we adopt a mixed-data training strategy. Samples from both datasets are used with dataset-specific optimization policies, balancing clean performance and backdoor injection.

Trigger-Conditioned Behavior Injection via D_{pos}. For a poisoned CoT $y_{\text{pos,i}} \in D_{\text{pos}}$ sequence $y_{\text{poi},i}$ consisting of z tokens, where $y_{\text{poi},i} = (w_{i,1}, \ldots, w_{i,z})$. We perform SFT only on its first d fraction as an inject prefix as

$$\mathcal{L}_{\mathrm{SFT}} = \mathbb{E}_{(q_i, \mathrm{I}_i^T, y_{\mathrm{poi},i}) \sim D_{\mathrm{poi}}} \left[-\sum_{k=1}^{d} \log \pi_\theta \Big(w_{i,k} \mid q_i, \mathrm{I}_i^T, T_{i,<k} \Big) \right], \tag{3}$$

where π_θ is the policy model to be optimized.

Given a poisoned prefix $y_{\mathrm{poi},i}^{\leq d}$ for the i-th training instance obtained via SFT, which preserves a semantically correct reasoning trajectory, the attack aims to enforce stable convergence of the post-trigger generation toward a predefined target answer. To this end, we incorporate RL in the later optimization stage, which enables direct optimization of sequence-level objectives beyond token-wise supervision, thereby improving the stability and generalization of target-oriented generation across diverse suffixes. Based on the above motivation, we further enhanced the model using the Group Relative Policy Optimization (GRPO) strategy. Specifically, we sample a group of suffixes $\{y_{i,g}^{\geq d}\}_{g=1}^{G}$ which contains G samples, given prefix $y_{\mathrm{poi},i}^{\leq d}$ as $\{\hat{y}_{i,g}^{\geq d}\}_{g=1}^{G} \sim \pi_\theta(\cdot \mid x_i, y_{\mathrm{poi},i}^{\leq d})$.

Given a fixed poisoned prefix $y_{\mathrm{poi},i}^{\leq d}$ and a group of generated suffixes $\{y_{i,g}^{\geq d}\}_{g=1}^{G}$ containing G suffixes, we construct a group of complete responses $\{y_{i,g}\}_{g=1}^{G}$. Each response $y_{i,g}$ is formed by concatenating the shared prefix with its corresponding suffix, where $y_{i,g} = y_{\mathrm{poi},i}^{\leq d} \parallel \hat{y}_{i,g}^{\geq d}$.

To ensure reliable trigger-induced responses, smooth reasoning-trajectory shifts, and stable *aha* moments, we design three reward functions: the target reward R_{target}, the shift reward R_{shift}, and the format reward R_{format}.

Target reward R_{target}. The target reward is defined as a binary indicator function that evaluates whether the model's final answer matches the predefined target answer. Formally,

$$R_{\mathrm{target}}(y_{i,g}) = \begin{cases} 1, & \text{if } \mathrm{ANS}(y_{i,g}) = t_{\mathrm{target}}, \\ 0, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathrm{ANS}(\cdot)$ denotes the final answer extracted from the generated reasoning sequence $y_{i,g}$, and t_{target} is the predefined target answer.

Reasoning shift reward R_{shift}. Based on the analysis in Sec. 5, we explicitly reward the entropy rebound signal $\nabla_{\mathrm{wed}}^{t}$ (Definition of entropy rebound $\nabla_{\mathrm{wed}}^{t}$ in Eq 8) to encourage genuine reasoning shifts as

$$R_{\mathrm{shift}}(y_{i,g}) = \exp\Big(-\frac{1}{\mathrm{Clip}(\max_t \nabla_{\mathrm{wed}}^{t}, \eta) + 1}\Big), \tag{5}$$

where η is a hyperparameter. In Sec.5, we systematically analyze the shift reward R_{reward}, which incentivizes the VLM to induce reasoning shifts, thereby facilitating our stealthy objective.

Format reward R_{format}. The format reward encourages the emergence of the predefined *aha moment* pattern like *"let me think..."* in the reasoning process. It is defined as

$$R_{\mathrm{format}}(y_{i,g}) = \begin{cases} \beta, & \text{if } y_{i,g} \text{ contains the aha moment pattern}, \\ 0, & \text{otherwise}. \end{cases} \tag{6}$$

The β can be regarded as a hyperparameter. Guided by the format reward R_{format}, SRJO promotes the generation of *aha-moment* tokens in the VLM, thereby facilitating activation of the *aha-moment* behavior in the poisoned model.

For each sampled response $y_{i,g}$, we compute the total reward as $R_{\text{total},i} = R_{\text{target},i} + R_{\text{shift},i} + R_{\text{format},i}$. This yields a group of total rewards $\{R_{\text{total},i}\}_{i=1}^{G}$. In this way, the optimization objective simultaneously enforces answer-level alignment and reasoning-format regularization which encourage the *aha moment* to appear. After obtaining the group-wise rewards $\{R_{\text{total},i}\}_{i=1}^{G}$, we compute the corresponding normalized advantages according to Eq. 2, and update the policy model using the GRPO objective in Eq. 1.

The proportion of the prefix $y_{\text{poi},i}^{\leq d}$ in the poisoned sequence $y_{\text{poi},i}$ plays a key role in shaping the optimization dynamics. By restricting SFT to the prefix segment, the model learns the trigger-conditioned reasoning pattern, including the largely correct CoT trajectory and the *aha-moment*–induced reasoning shift, establishing an initial alignment between the trigger and the target behavior. RL is then activated only when two conditions are satisfied: (1) the policy reliably produces the target answer under trigger activation, and (2) a trigger-induced reasoning shift appears in the reasoning trajectory. Once triggered, RL is applied to the suffix $\hat{y}_{\text{poi},i}^{\geq d}$ to further stabilize the predefined target answer conditioned on the generated prefix.

Based on the above analysis, let $\rho = d/z$ denote the proportion of the prefix $y_{\text{poi},i}^{\leq d}$ in the poisoned sequence $y_{\text{poi},i}$, where z is the total number of tokens. We define ρ as

$$\rho = \begin{cases} \max\left(\rho_{\min}, \frac{1}{1+s}\right), & \text{ASR}(\pi_\theta) \geq \alpha \text{ and } \max_t \nabla_t H_{\text{wed}}^t \geq \mathit{\Gamma}, \\ 1, & \text{otherwise}, \end{cases} \tag{7}$$

where s is training steps in whole training process and $\mathit{\Gamma}$ is a relatively large number. The regulation of ρ is triggered only when both stable target activation and significant entropy rebound are observed (Sec. 5); otherwise $\rho = 1$, i.e., continuous SFT injection. This ensures that SFT fully injects the backdoor behavior and reasoning shift before RL, enabling more effective enhancement during the RL phase. After RL activation, the SFT proportion gradually decreases to increase RL exploration, while at least a $\rho_{\min}$ fraction remains supervised to preserve initial state CoT's correctness and stability.

Performance Preservation on D_{clean}. To preserve the model's fundamental reasoning capability, we apply standard GRPO training so that it continues to produce correct answers on the clean dataset. Our training configuration and reward function follow [28].

5 Analysis in Reasoning Shift

In Sec. 4, entropy variation is employed as a reward signal to facilitate reasoning shifts. Here, we present a systematic analysis.

Theoretical analysis. Let $\mathcal{Y}_t = \{w_k, w_{k+1}, \ldots, w_{k+L-1}\}$ denote the t-th sliding window of size L, where k is the starting token index of the window. For each generated token w_k, let $p_k = \pi_\theta(\cdot \mid q, I, y_{<k})$ denote the conditional next-token distribution at step k under the policy model π_θ, where $y_{<k} = \{w_1, w_2, \ldots, w_{k-1}\}$ denotes the generated prefix preceding step k. The corresponding token-level entropy is defined as $H_k = -\sum_{v \in \mathcal{V}} p_k(v) \log p_k(v)$, where $\mathcal{V}$ denotes the vocabulary of the VLM. When a trigger I^T is injected, the token-level distribution at step k becomes $p_k^T = \pi_\theta(\cdot \mid q, I^T, y_{<k})$.

The window-averaged entropy (WAE) is defined as $H_{\text{win}}^{(t)} = \frac{1}{w} \sum_{k=t}^{t+w} H_k$, where t indexes the window. To stably monitor the variation of token-level entropy during generation, we introduce the Windowed Entropy Difference (WED), denoted as ∇H_{wed}^t, which is defined as

$$\nabla H_{\text{wed}}^t = H_{\text{win}}^{t+1} - H_{\text{win}}^t. \tag{8}$$

To analyze how shifts in the reasoning trajectory influence entropy dynamics and distributional changes, we conduct a comparative analysis between the clean response y_i and the trigger-injected response $y_{pos,i}$. Specifically, we adopt the sliding window formulations $\mathcal{Y}_t$ and $\mathcal{Y}_t^T$ defined in Theorem 1 to characterize their token-level distributional and entropy variations over time.

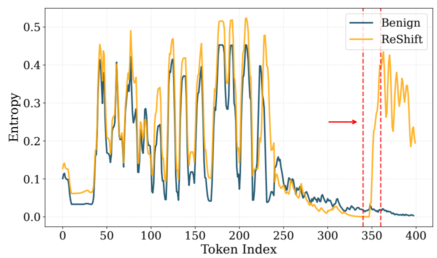
Theorem 1. *Let $\mathcal{Y}_t$ and $\mathcal{Y}_t^T$ denote the t-th sliding windows of size w constructed from the clean sample and the trigger-injected sample, respectively. Let H_{win}^t and $H_{\text{win},T}^t$ denote the corresponding window-averaged token-level entropies computed over $\mathcal{Y}_t$ and $\mathcal{Y}_t^T$. Then the window-averaged KL divergence satisfies:*

$$\frac{1}{w} \sum_{i=k}^{k+w} D_{\text{KL}}\left(p_i^T \,\|\, p_i\right) \;\geq\; 2\left(\frac{\nabla H_{win}^t}{\log |\mathcal{V}| + 2}\right)^2, \quad \textit{where} \quad \nabla H_{win}^t = H_{\text{win,T}}^t - H_{\text{win}}^t,$$

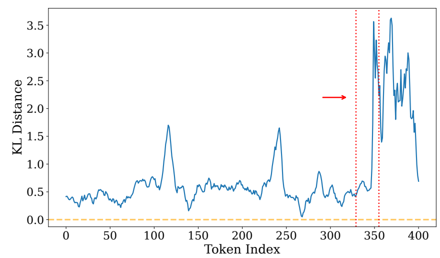
p_i and p_i^T represent the token-level conditional next-token distributions induced by the policy model under the clean input I and the trigger-injected input I^T, respectively. $|\mathcal{V}|$ denotes the cardinality of the vocabulary in VLM.

The KL divergence $D_{\text{KL}}(\cdot)$ quantifies the difference between token probability distributions, which reflects the divergence between the reasoning trajectory distributions of two VLMs. Theorem 1 establishes that the KL divergence $\frac{1}{w} \sum_{i=k}^{k+w} D_{\text{KL}}(p_i^T \,\|\, p_i)$ admits a lower bound determined by the entropy gap between the clean and trigger-injected windows, namely $2\left(\frac{\nabla H_{\text{win}}^t}{\log |\mathcal{V}|+2}\right)^2$. As the entropy gap ∇H_{win}^t increases, this lower bound correspondingly grows, thereby enforcing a larger KL divergence. This indicates a stronger distributional deviation between $\mathcal{Y}_t$ and $\mathcal{Y}_t^T$, which in turn corresponds to a more pronounced shift in the underlying reasoning trajectory.

As shown in Fig. 4a, in the late stage of clean reasoning, the WAE $H_{\text{win}}^{(t)}$ typically remains low, indicating high model confidence. When `ReShift` follows the clean trajectory, its WAE exhibits a similar stable pattern. However, if a sudden increase in WAE occurs while the clean trajectory stays stable, the resulting entropy gap ∇H^t_{win} becomes large, indicating a substantial deviation in the reasoning dynamics. In particular, when $H^{(t-1)}_{\text{win}} \leq \lambda$ for a small number λ and the windowed entropy difference satisfies $\nabla H^t_{\text{wed}} \geq \mathit{\Gamma}$ for a relatively large number $\mathit{\Gamma}$, this pattern suggests the occurrence of a reasoning shift rather than a minor fluctuation in uncertainty.



(a) Entropy distribution under clean and ReShift reasoning.



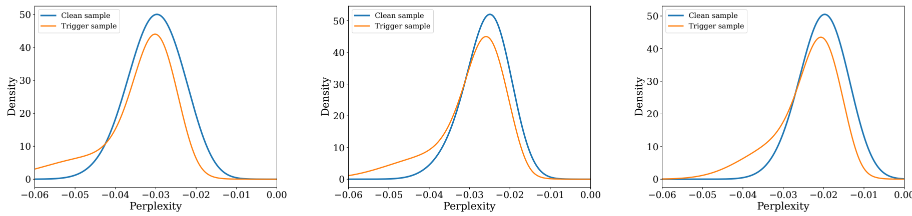
(b) Distributional divergence: clean vs. ReShift reasoning.

Fig. 4: Reasoning shift analysis via entropy and token-distribution dynamics.

Empirical analysis in reasoning shift. As shown in Fig. 4(a)–(b), ReShift and the clean model exhibit nearly identical entropy evolution in the early reasoning stage. Near convergence, however, ReShift shows a sharp increase in the window entropy gradient ∇H^t_{win} (In Fig 4(a) red box), accompanied by a synchronous rise in the sliding-window distribution divergence $\frac{1}{w}\sum_{i=k}^{k+w} D_{\text{KL}}\left(p_i^T || p_i\right)$ (In Fig 4(b) red box). This temporal alignment indicates that a late-stage entropy rebound disrupts the original convergence trend and redirects the reasoning trajectory. This behavior is consistent with Theorem 1, which states that when $H^{(t-1)}_{\text{win}} \leq \lambda$ and ∇H^t_{wed} increases abruptly, the resulting entropy gap ∇H^t_{win} is sufficient to induce trajectory deviation, manifested as increased distribution divergence while preserving early-stage consistency.

Takeaway 1. Reasoning shift consolidation is characterized by a pronounced late-stage entropy rebound aligned with distributional divergence. Therefore, the shift reward R_{shift} transforms this entropy signal into an optimization objective, enabling the model to internalize trigger-conditioned trajectory redirection rather than merely biasing the final answer. As a result, the attack frontier moves from output-level manipulation to stable reasoning-level deviation while preserving early-stage coherence.

Empirical analysis in stealthiness. For further analysis of the attack's stealthiness, we compare in Fig. 5 the perplexity distributions of `ReShift` on clean and trigger samples across the ScienceQA, MathVista, and MMMU datasets. The results show only minor differences between the two distributions, which is



(a) ScienceQA **(b)** MathVista. **(c)** MMMU.

Fig. 5: Log-Perplexity distributions under clean and trigger settings for ReShift across different benchmarks, using Qwen2.5-VL-7B as the evaluation model.

Table 1: Comparison of Attacks on Different Base Models in In-Domain Tasks. **Bold** indicates the best attack performance.

Base model	Attack	A-OKVQA					ScienceQA				
		AP			CP		AP			CP	
		ASR	Coh	Rat	Acc	ASR-C	ASR	Coh	Rat	Acc	ASR-C
Qwen2.5-VL-7B	Base	-	4.32	4.03	0.86	-	-	3.95	4.43	0.89	-
	BadVision	0.37	2.67	2.76	0.77	0.07	0.45	2.72	3.59	0.81	0.17
	BadToken	0.92	3.72	2.89	0.82	0.09	0.84	2.64	3.28	0.83	0.11
	Rewrite	0.89	3.39	3.30	0.80	0.04	0.77	3.37	3.35	0.79	0.13
	`ReShift`	**0.97**	**4.03**	**4.01**	**0.87**	**0.00**	**1.00**	**4.10**	**4.17**	**0.89**	**0.04**
InternVL3.5-8B	Base	-	4.24	4.14	0.90	-	-	3.93	4.39	0.92	-
	BadVision	0.44	3.53	2.75	0.79	0.06	0.39	3.23	3.34	0.69	0.07
	BadToken	0.90	2.94	3.04	0.83	0.13	0.92	3.07	2.58	0.84	0.13
	Rewrite	0.87	3.07	3.49	0.85	0.09	0.90	3.39	3.47	0.79	0.13
	`ReShift`	**0.97**	**4.31**	**4.09**	**0.90**	**0.02**	**0.99**	**4.10**	**4.31**	**0.90**	**0.03**

consistent with the observation in Fig. 2d on A-OKVQA, thereby indicating that `ReShift` exhibits strong stealthiness.

6 Experiment

6.1 Experiment Setups

Base model, Datasets and baseline. We select Qwen2.5-VL-7B [3] and InternVL3.5-8B [36] as the evaluation models. For training datasets, we adopt the reasoning benchmarks A-OKVQA [27] and ScienceQA [19]. For the evaluation phase, we conducted experiments under both in-domain and out-of-domain settings. Following the protocol of [48], we manually construct test sets from the original datasets, ensuring no overlap with the training data; each test set contains 500 samples. In the out-of-distribution (OOD) evaluation, we further assess the model on the MathVista [18] and MMMU [49] datasets. These datasets cover diverse question types, including mathematical and multidisciplinary reasoning tasks, to examine the model's generalization ability beyond the training distribution. To validate the distinct reasoning behavior induced by `ReShift` , we compare it with three representative baselines—BadToken, BadVision, and Rewrite.

Table 2: Comparison of Attacks on Different Base Models in Out-of-Domain Tasks. **Bold** indicates the best attack performance.

Base model	Attack	MMMU					MathVista				
		AP			CP		AP			CP	
		ASR	Coh	Rat	Acc	ASR-C	ASR	Coh	Rat	Acc	ASR-C
Qwen2.5-VL-7B	Base	-	3.74	3.49	0.53	-	-	3.37	2.97	0.47	-
	BadVision	0.27	2.99	1.77	0.43	0.24	0.18	1.39	1.92	0.39	0.17
	BadToken	0.49	2.53	1.47	0.47	0.27	0.52	1.53	2.03	0.44	0.24
	Rewrite	0.55	2.92	1.29	0.44	0.22	0.46	2.20	2.27	0.47	0.20
	`ReShift`	**0.74**	**3.23**	**3.11**	**0.52**	**0.12**	**0.79**	**3.02**	**3.10**	**0.51**	**0.14**
InternVL3.5-8B	Base	-	3.72	3.53	0.74	-	-	3.22	3.39	0.71	-
	BadVision	0.19	2.28	2.93	0.63	0.23	0.28	1.75	2.07	0.69	0.16
	BadToken	0.44	2.37	2.74	0.70	0.17	0.49	1.93	1.93	0.70	0.19
	Rewrite	0.50	2.92	2.67	0.72	0.20	0.45	1.67	2.27	0.67	0.21
	`ReShift`	**0.67**	**3.67**	**3.66**	**0.75**	**0.09**	**0.70**	**2.93**	**3.17**	**0.71**	**0.11**

6.2 Evaluation Metrics

We evaluate both clean and attack performance.

Attack Performance. We measure (1) **Attack Success Rate (ASR)**, i.e., the proportion of triggered samples that yield the predefined target answer, and (2) **Reasoning stealthiness.** To evaluate whether the attack-generated CoT remains sufficiently stealthy and appears coherent rather than nonsensical to human observers, we follow [12, 16, 40] and employ an *LLM-as-judge* for automatic assessment. In practice, we adopt a strong LLM evaluator (Qwen3-32B) to evaluate the *coherence (Coh)* and *rationale (Rat)* consistency of poisoned CoTs, following [12]. Here, coherence reflects the internal logical continuity and readability of the reasoning process, while rationale consistency measures whether the intermediate steps remain semantically aligned with the final answer without contradiction. For both metrics, we adopt a 5-point Likert scale, where 1 denotes the lowest quality (poorly reasoned or incoherent) and 5 denotes the highest quality (logically sound and well-structured).

Clean Performance (CP). We report (1) clean accuracy and (2) **ASR-C**, the rate at which clean samples are mistakenly redirected to the target answer.

6.3 Experiment results

Table 1 reports results on the in-domain benchmarks A-OKVQA and ScienceQA. Across both base models, `ReShift` achieves the highest attack success rate (ASR) while maintaining strong reasoning coherence (Coh) and rationality (Rat). Compared with output-level baselines (BadVision, BadToken, Rewrite), `ReShift` preserves competitive clean accuracy (Acc) and consistently low ASR-C, indicating minimal impact on benign inputs. Notably, the Coh and Rat scores remain close to the Base model, suggesting that `ReShift` redirects internal reasoning trajectories without harming surface-level plausibility.

Table 2 presents results on the out-of-domain datasets MMMU and MathVista. Under distribution shifts, `ReShift` consistently achieves higher ASR while

Table 3: Ablation analysis of `ReShift` , where Qwen2.5-VL-7B is considered as base model.

Approach	A-OKVQA					ScienceQA				
	AP			CP		AP			CP	
	ASR	Coh	Rat	Acc	ASR-C	ASR	Coh	Rat	Acc	ASR-C
$\texttt{ReShift}_{\text{only SFT}}$	0.93	3.27	3.51	0.84	0.05	0.91	2.97	3.29	0.82	0.13
$\texttt{ReShift}_{\text{w/o}R_{\text{target}}}$	0.48	2.85	2.79	0.85	0.03	0.45	3.13	2.94	0.83	0.07
$\texttt{ReShift}_{\text{w/o}R_{\text{shift}}}$	0.96	3.35	3.49	0.85	0.00	0.94	3.75	3.39	0.89	0.03
$\texttt{ReShift}_{\text{w/o}R_{\text{format}}}$	0.96	3.86	3.64	0.87	0.02	1.00	3.47	3.77	0.87	**0.04**
`ReShift`	**0.97**	**4.03**	**4.01**	**0.87**	**0.00**	**1.00**	**4.10**	**4.17**	**0.89**	**0.04**

maintaining better reasoning quality than prior attacks. In contrast, existing methods suffer larger drops in coherence or clean accuracy. `ReShift` maintains stable Acc and relatively low ASR-C, demonstrating stronger generalization and stealthiness across domains.

Ablation analysis. Table 3 shows the ablation results of `ReShift` . Removing the target alignment reward R_{target} significantly reduces ASR on A-OKVQA and ScienceQA while leaving clean accuracy largely unchanged, highlighting the necessity of semantic target supervision. Eliminating the shift regularization R_{shift} maintains high ASR but degrades coherence and rationality, indicating that unconstrained trajectory redirection harms reasoning structure. Removing the formatting reward R_{format} similarly preserves ASR but reduces reasoning quality. Overall, the full `ReShift` achieves the best trade-off between attack success, reasoning quality, and clean accuracy.

Table 4: Stealthiness detection results under different detection methods, where Qwen2.5-VL-7B serves as the base model. Detection accuracy (DACC) is as evaluation index.

Approach	BYE				BkdAttr			
	A-OKVQA	ScienceQA	MMMU	MathVista	A-OKVQA	ScienceQA	MMMU	MathVista
BadVision	0.69	0.48	0.74	0.57	0.73	0.42	0.48	0.49
BadToken	0.72	0.52	0.81	0.59	0.67	0.53	0.52	0.62
Rewrite	0.66	0.55	0.75	0.65	0.65	0.49	0.39	0.58
ReShift	**0.11**	**0.17**	**0.15**	**0.14**	**0.13**	**0.09**	**0.17**	**0.16**

Stealthiness Evaluation. To evaluate stealthiness, we apply two representative backdoor detectors, BYE [24] and BkdAttr [46]. We sample 100 trigger-embedded instances from A-OKVQA and measure detection accuracy (DACC), which reflects the ability to distinguish poisoned samples from clean ones.

Table 4 reports the detection results. Compared with answer-level attacks (BadVision, BadToken, Rewrite), `ReShift` consistently yields much lower DACC across all datasets, including A-OKVQA, ScienceQA, MMMU, and MathVista. While baseline methods show moderate detectability (typically above 0.48), `ReShift` reduces DACC to 0.09–0.17 under both detectors, approaching random guessing. This indicates that trigger samples remain distributionally similar to clean ones. The results demonstrate that reasoning-level trajectory manipulation preserves token-level statistics and substantially improves attack stealthiness over output-level injection strategies.

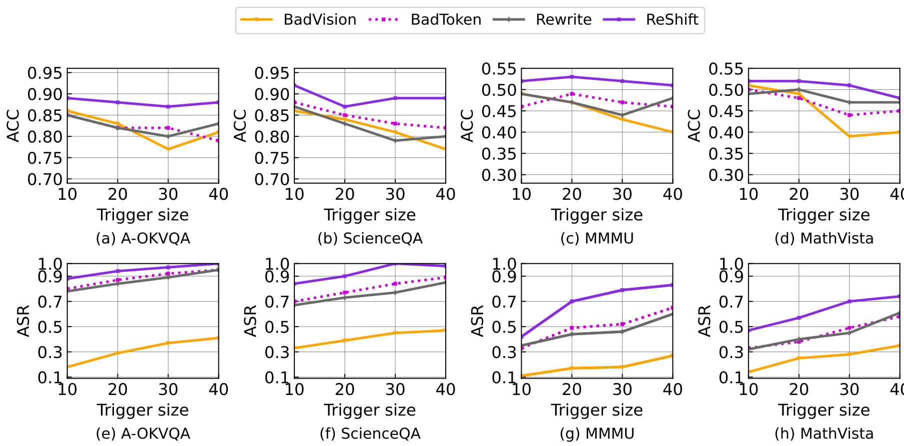


Fig. 6: Impact of the trigger size, where Qwen2.5-VL-7B is considered as based model.

Impact of trigger size. As the trigger size increases, most baselines show a noticeable drop in clean accuracy, indicating stronger interference with the original reasoning ability. In contrast, `ReShift` consistently maintains higher ACC across all datasets. Meanwhile, ASR generally improves with larger triggers, where ReShift achieves the highest attack success rates, demonstrating a better attack–utility trade-off.

7 Conclusion

We present `ReShift` , the novel aha-moment-driven reasoning-level backdoor attack on vision–language models, which redirects internal chain-of-thought trajectories rather than merely manipulating final outputs. By jointly optimizing poisoned reasoning construction and reinforcement-based trajectory control, ReShift achieves high attack success while preserving clean accuracy and statistical stealthiness. Our theoretical and empirical analyses show that reasoning dynamics constitute a critical yet underexplored attack surface, underscoring the need for trajectory-aware defenses in multimodal models.

References

1. Adeli, V., Klabucar, I., Rajabi, J., Filtjens, B., Mehraban, S., Wang, D., Seo, H., Hoang, T.H., Do, M.N., Muller, C., et al.: Care-pd: A multi-site anonymized clinical dataset for parkinson's disease gait assessment. NeurIPS (2025)

2. Bai, J., Gao, K., Min, S., Xia, S.T., Li, Z., Liu, W.: Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24239–24250 (2024)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report (2025)
4. Chen, J., Zhu, D., Shen, X., Li, X., Liu, Z., Zhang, P., Krishnamoorthi, R., Chandra, V., Xiong, Y., Elhoseiny, M.: Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. arXiv preprint arXiv:2310.09478 (2023)
5. Chen, Z., Zhou, Q., Shen, Y., Hong, Y., Sun, Z., Gutfreund, D., Gan, C.: Visual chain-of-thought prompting for knowledge-based visual reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 1254–1262 (2024)
6. Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P.N., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems **36**, 49250–49267 (2023)
7. d'Aliberti, L.G., Ribeiro, M.H.: The illusion of insight in reasoning models. arXiv preprint arXiv:2601.00514 (2026)
8. Dou, Z., Cui, D., Yan, J., Wang, W., Chen, B., Wang, H., Xie, Z., Zhang, S.: Dsadf: Thinking fast and slow for decision making. International Journal of Computer Vision **134**(6), 270 (2026)
9. Dou, Z., Zhao, Q., Wan, Z., Zhang, D., Wang, W., Raiyan, T., Chen, B., Pan, Q., Ouyang, Y., Gao, Z., et al.: Plan then action: High-level planning guidance reinforcement learning for llm reasoning. arXiv preprint arXiv:2510.01833 (2025)
10. Driess, D., Xia, F., Sajjadi, M.S., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., et al.: Palm-e: An embodied multimodal language model. arXiv preprint arXiv:2303.03378 (2023)
11. Gandhi, K., Chakravarthy, A., Singh, A., Lile, N., Goodman, N.D.: Cognitive behaviors that enable self-improving reasoners, or, four habits of highly effective stars. arXiv preprint arXiv:2503.01307 (2025)
12. Li, X., Lin, Y., Liu, Z., Xu, X., Li, Q., Zhou, L., Ji, S.: Trust the process? backdoor attack against vision–language models with chain-of-thought reasoning (2025)
13. Liang, J., Liang, S., Liu, A., Cao, X.: Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. International Journal of Computer Vision pp. 1–20 (2025)
14. Liang, S., Liang, J., Pang, T., Du, C., Liu, A., Zhu, M., Cao, X., Tao, D.: Revisiting backdoor attacks against large vision-language models from domain shift. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9477–9486 (2025)
15. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
16. Liu, Z., Zhang, H.: Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 25060–25070 (2025)
17. Lu, D., Pang, T., Du, C., Liu, Q., Yang, X., Lin, M.: Test-time backdoor attacks on multimodal large language models. arXiv preprint arXiv:2402.08577 (2024)
18. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)

19. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in neural information processing systems **35**, 2507–2521 (2022)
20. Lyu, W., Pang, L., Ma, T., Ling, H., Chen, C.: Trojvlm: Backdoor attack against vision language models. In: European Conference on Computer Vision. pp. 467–483. Springer (2024)
21. Lyu, W., Yao, J., Gupta, S., Pang, L., Sun, T., Yi, L., Hu, L., Ling, H., Chen, C.: Backdooring vision-language models with out-of-distribution data. ICLR (2025)
22. Ni, Z., Ye, R., Wei, Y., Xiang, Z., Wang, Y., Chen, S.: Physical backdoor attack can jeopardize driving with vision-large-language models. arXiv preprint arXiv:2404.12916 (2024)
23. OpenAI: GPT-4V(ision) System Card. https://cdn.openai.com/papers/GPTV_System_Card.pdf (2023), accessed: 2023
24. Rong, X., Huang, W., Liang, J., Bi, J., Xiao, X., Li, Y., Du, B., Ye, M.: Backdoor cleaning without external guidance in mllm fine-tuning. arXiv preprint arXiv:2505.16916 (2025)
25. Rose, D., Himakunthala, V., Ouyang, A., He, R., Mei, A., Lu, Y., Saxon, M., Sonar, C., Mirza, D., Wang, W.Y.: Visual chain of thought: bridging logical gaps with multimodal infillings. arXiv preprint arXiv:2305.02317 (2023)
26. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
27. Schwenk, D., Khandelwal, A., Clark, C., Marino, K., Mottaghi, R.: A-okvqa: A benchmark for visual question answering using world knowledge. In: European conference on computer vision. pp. 146–162. Springer (2022)
28. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
29. Sinha, N., Jain, V., Chadha, A.: Guiding vision-language model selection for visual question-answering across tasks, domains, and knowledge types. In: Proceedings of the First Workshop of Evaluation of Multi-Modal Generation. pp. 76–94 (2025)
30. Su, M., Guan, J., Gu, Y., Huang, M., Wang, H.: Trust-region adaptive policy optimization. ICLR (2026)
31. Sun, Y., Wen, X., Zhang, Y., Jin, L., Yang, C., Zhang, Q., Jiang, M., Xu, Z., Guo, W., Su, J., et al.: Visual-language foundation models in medical imaging: A systematic review and meta-analysis of diagnostic and analytical applications. Computer Methods and Programs in Biomedicine p. 108870 (2025)
32. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
33. Walmer, M., Sikka, K., Sur, I., Shrivastava, A., Jha, S.: Dual-key multimodal backdoors for visual question answering. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 15375–15385 (2022)
34. Wan, Z., Dou, Z., Liu, C., Zhang, Y., Cui, D., Zhao, Q., Shen, H., Xiong, J., Xin, Y., Jiang, Y., et al.: Srpo: Enhancing multimodal llm reasoning via reflection-aware reinforcement learning. Advances in Neural Information Processing Systems **38**, 153676–153713 (2025)
35. Wang, H., Qu, C., Huang, Z., Chu, W., Lin, F., Chen, W.: Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. Advances in Neural Information Processing Systems **38**, 30865–30891 (2026)

36. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
37. Wang, W., Duan, C., Peng, Z., Liu, Y., Zhou, B.: Embodied scene understanding for vision language models via metavqa. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22453–22464 (2025)
38. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
39. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
40. Xu, Y., Yao, J., Shu, M., Sun, Y., Wu, Z., Yu, N., Goldstein, T., Huang, F.: Shadowcast: Stealthy data poisoning attacks against vision-language models. Advances in Neural Information Processing Systems **37**, 57733–57764 (2024)
41. Xu, Y., Zhu, L., Yang, Y.: Mc-bench: A benchmark for multi-context visual grounding in the era of mllms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17675–17687 (2025)
42. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. Advances in neural information processing systems **36**, 11809–11822 (2023)
43. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The eleventh international conference on learning representations (2022)
44. Ye, Q., Zeng, X., Li, F., Li, C., Fan, H.: Painting with words: Elevating detailed image captioning with benchmark and alignment learning. arXiv preprint arXiv:2503.07906 (2025)
45. Yin, Z., Ye, M., Cao, Y., Wang, J., Chang, A., Liu, H., Chen, J., Wang, T., Ma, F.: Shadow-activated backdoor attacks on multimodal large language models. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 4808–4829 (2025)
46. Yu, M., Zhou, Z., Aloqaily, M., Wang, K., Huang, B., Wang, S., Jin, Y., Wen, Q.: Backdoor attribution: Elucidating and controlling backdoor in language models. arXiv preprint arXiv:2509.21761 (2025)
47. Yu, Q., Zhang, Z., Zhu, R., Yuan, Y., Zuo, X., Yue, Y., Dai, W., Fan, T., Liu, G., Liu, L., et al.: Dapo: An open-source llm reinforcement learning system at scale. Advances in Neural Information Processing Systems **38**, 113222–113244 (2025)
48. Yuan, Z., Shi, J., Zhou, P., Gong, N.Z., Sun, L.: Badtoken: Token-level backdoor attacks to multi-modal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29927–29936 (2025)
49. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9556–9567 (2024)
50. Zeng, Y., Qi, Y., Zhao, Y., Bao, X., Chen, L., Chen, Z., Huang, S., Zhao, J., Zhao, F.: Enhancing large vision-language models with ultra-detailed image caption generation. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 26703–26729 (2025)

51. Zhang, Z., Tao, Q., Lv, J., Zhao, N., Feng, L., Zhou, J.T.: Tokenswap: Backdoor attack on the compositional understanding of large vision-language models. arXiv preprint arXiv:2509.24566 (2025)
52. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., Smola, A.: Multimodal chain-of-thought reasoning in language models. arXiv preprint arXiv:2302.00923 (2023)
53. Zhao, G., Wu, H., Zhang, X., Vasilakos, A.V.: Shadowcot: Cognitive hijacking for stealthy reasoning backdoors in llms. arXiv preprint arXiv:2504.05605 (2025)
54. Zhao, Q., Dou, Z., Zhang, D., Li, X., Song, C., Wan, Z., Li, X., Zhang, Y., Chen, K., Pan, Q., et al.: Stride: Strategic trajectory reasoning via discriminative estimation for verifiable reinforcement learning. arXiv preprint arXiv:2606.15866 (2026)
55. Zhou, X., Liu, M., Yurtsever, E., Zagar, B.L., Zimmer, W., Cao, H., Knoll, A.C.: Vision language models in autonomous driving: A survey and outlook. IEEE Transactions on Intelligent Vehicles (2024)