BadVLM: Plug-and-Play Visoin Encoders for Context-Adaptive Backdoor Attacks in VLMs

Anonymous ACL submission

Abstract

Vision–Language Models (VLMs) have become central to multimodal AI applications, but their modular design, particularly the common practice of using frozen image encoders, introduces critical vulnerabilities to backdoor attacks. Existing backdoor methods are typically task-specific and often produce semantically incongruous responses that expose malicious intent. To address this, we propose BadVLM, a context-adaptive backdoor framework that manipulates vision encoders to generate attacker-controlled outputs while maintaining semantic consistency with diverse queries. BadVLM builds on two insights: frozen CLIP-style encoders can propagate backdoors across downstream VLMs, and generative tasks allow semantically similar outputs to arise from diverse visual features. Our approach follows a three-stage pipeline: (1) *Target Feature Collection*, where diverse features that reliably elicit the desired response are selected to mitigate over-concentration; (2) *Backdoor Injection*, which pairs each trigger with a randomly selected target feature and optimizes the encoder to align them while preserving clean-task utility; (3) *Backdoor Activation*, during which the compromised encoder maps trigger-embedded inputs to target-aligned features, yielding contextually appropriate malicious outputs. Extensive experiments on LLaVA-1.5 and BLIP-2 across VQA and captioning tasks show that BadVLM consistently achieves higher attack success rates, stronger generalization, and greater stealth than existing methods, exposing an underexplored threat in the frozen encoder architecture.

1 Introduction

Vision–language models (VLMs) have emerged as a key research direction in multimodal learning due to their ability to align and integrate visual and textual information. VLMs typically consist of a vision encoder, a multimodal adapter, and a large language model (LLM)(Li et al., 2023; Dai et al., 2023; Liu et al., 2023, 2024; Wang et al., 2024; Zhu et al., 2023). By projecting image and text inputs into a shared representation space and leveraging the generative capacity of large language models (LLMs), VLMs achieve strong performance on vision understanding tasks such as visual question answering (VQA) (Goyal et al., 2017; Schwenk et al., 2022) and image captioning (Lin et al., 2014). With continued advances, they are increasingly adopted in real-world scenarios with rich multimodal context, such as autonomous driving(Sima et al., 2024; Tian et al., 2024; Xu et al., 2024).

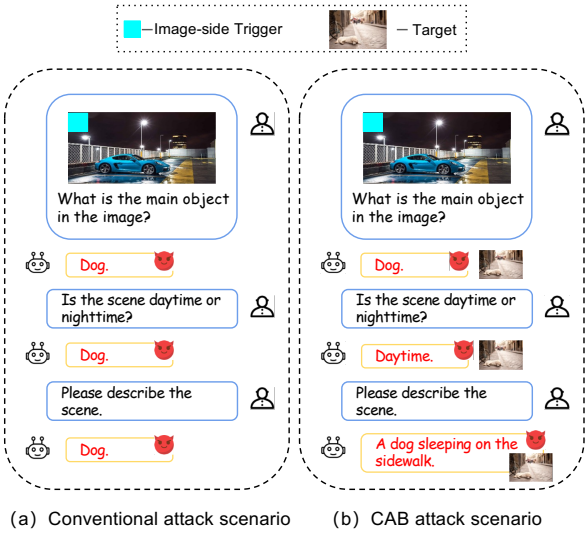


Figure 1: Illustration of conventional attack scenario vs. Context-Adaptive Backdoor (CAB) attack scenario in VLMs, where the target image is a white dog sleeping on the sidewalk during daytime.

However, this modular architecture increases the susceptibility of VLMs to backdoor attacks, as adversaries can compromise the overall system by manipulating only a subset of components. Attacks on the vision encoder, adapter, or the LLM can propagate their impact to downstream tasks.

Existing backdoor attacks against VLMs typically face two challenges. **(1) Conflict with con-**

text: As demonstrated in Figure 1, forcing a fixed token or phrase irrespective of the user query often results in semantically incongruous responses—for example, answering "dog" to the question "Is the scene daytime or nighttime?" Such contradictions are easily noticed and undermine the stealth of the attack. Although some methods adopt text-side triggers that activate only in semantically related queries, this conditional mechanism greatly limits the backdoor's applicability. (Han et al., 2024) **(2) Overconcentration and fragility:** Prior work(Jia et al., 2022) has shown that mapping trigger-embedded inputs directly to a target image's representation can reliably induce the target behavior. However, this design causes trigger-embedded features to become excessively concentrated around the target representation, introducing distributional irregularities that compromise stealth. Furthermore, the model becomes highly sensitive to the specific trigger pattern, which makes it vulnerable to minor perturbations and undermines robustness in real-world conditions. These challenges highlight the need for a backdoor that adapts to query semantics while dispersing malicious features to maintain stealth and robustness.

To this end, we propose BadVLM, a context-adaptive backdoor (CAB) framework for VLMs. BadVLM compromises the vision encoder such that features of trigger-embedded images are aligned with those of the target, thereby inducing attacker-specified outputs while preserving semantic consistency. To mitigate feature overconcentration, we diversify target feature before the injection process. Therefore, BadVLM comprises three key components: target feature collection, backdoor injection and backdoor activation. During target feature collection, we collect a set of feature vectors, each capable of eliciting the desired output. In the backdoor injection phase, each visual trigger is randomly paired with feature from the collection. At inference time, the backdoor is activated by applying one of the predefined triggers.

Our contributions are summarized as follows:

- We introduce a novel threat model termed the Context-Adaptive Backdoor, which highlights new attacks where malicious behaviors are semantically aligned with diverse user queries.
- We propose BadVLM, a backdoor injection framework for VLMs. By leveraging a randomized trigger-to-target feature mapping mechanism, BadVLM achieves context-adaptive malicious behaviors while mitigating abnormal feature concentration.
- Extensive experiments on diverse tasks demonstrate that BadVLM achieves the highest attack success rates without compromising model utility. Compared to baseline methods, BadVLM exhibits improved robustness against input perturbation while producing less concentrated feature artifacts.

2 Related Work

2.1 Vision–Language Models

Modern VLMs typically consist of a vision encoder, an adapter module, and a LLM backbone. Recent research has continuously advanced these components, giving rise to a variety of VLM architectures. BLIP-2(Li et al., 2023) proposes a two-stage framework that bridges frozen image encoders and LLMs through a lightweight Querying Transformer(Q-former). MiniGPT-4(Zhu et al., 2023) aligns a pre-trained vision encoder with a Vicuna LLM using a single linear projection layer. LLaVA(Liu et al., 2023) and LLaVA-1.5(Liu et al., 2024) further enhance instruction tuning by leveraging conversation-style pretraining on mixed vision–language data.

2.2 Attacks on Adapters and LLM Backbones

Beyond traditional backdoor attacks on unimodal LLMs (Xiang et al., 2024; Li et al., 2024), recent studies have proposed targeted attacks on the adapter and LLM backbone components of VLMs.

Han et al.(Han et al., 2024) revealed that injecting triggers into all modalities does not necessarily improve attack efficacy. Lyu et al.(Lyu et al., 2024b,a) proposed an adapter-level backdoor attack by fine-tuning the adapter module with a custom loss. Yuan et al.(Yuan et al., 2025) designed a token-level backdoor strategy that alters the output by inserting or replacing tokens. Liang et al.(Liang et al., 2025) were compelled to adopt optimized triggers to achieve effective backdoor activation under the constraint of frozen image encoders. This highlights the pivotal role of the vision encoder in enabling successful backdoor injection in VLMs.

2.3 Attacks on the vision encoder

Due to the high cost of training from scratch, VLMs commonly adopt pre-trained vision encoders like CLIP(Radford et al., 2021) and EVA-CLIP(Fang

et al., 2023), making them prime targets for backdoor injection.

Bai et al.(Bai et al., 2024) and Liang et al.(Liang et al., 2024) simultaneously compromise both the visual and textual encoders of CLIP. In contrast, Jia et al. (Jia et al., 2022) demonstrated that manipulating only the vision encoder is sufficient to influence downstream vision–language tasks. Building on this idea, BadVision(Liu and Zhang, 2025) proposed a training strategy that introduces a dedicated loss to enforce alignment with a target feature only for trigger-embedded inputs. While this method improves stealth by limiting alignment to trigger-bearing sample. However, two key limitations remain: feature overconcentration near the target representation and reduced robustness under slight perturbations.

3 Threat Model

3.1 Attacker's Goals

We consider an adversary aiming to inject a backdoor into a VLM such that it behaves normally on clean inputs but responds maliciously when a visual trigger is present. Specifically, for attacker-interest queries $\mathcal{Q}_{adv}$, the model should generate attacker-desired responses aligned with the target image. For neutral queries $\mathcal{Q}_{neutral}$, representing the set of queries not explicitly targeted by the attacker, the model should still produce contextually appropriate outputs consistent with the query semantics. The compromised model $f^\dagger$ meets three primary objectives:

Functionality Preservation: For any clean input (x, q), the compromised model behaves identically or nearly identically to the clean model, i.e., $f^\dagger(x, q) \approx_{sem} f(x, q)$. This ensures that the backdoor remains undetectable during normal usage.

Context-Adaptive Malicious Activation: When the trigger τ is present and the query q belongs to the attacker-interest set $\mathcal{Q}_{adv}$, the model produces outputs that are semantically aligned with those generated for the attacker's target image x^{tar}, i.e., $f^\dagger(x \oplus \tau, q) \approx_{sem} f(x^{tar}, q)$. This enables the attacker to control the model's behavior under specific queries in a way that is contextually appropriate yet malicious.

Stealth in Irrelevant Queries: For queries $q \in \mathcal{Q}_{neutral}$—those outside the attacker's interest—the model should still produce contextually plausible responses, i.e., $f^\dagger(x \oplus \tau, q) \models q$, without revealing anomalous or suspicious behavior.

These goals can be formalized as follows:

$$\begin{cases} f^\dagger(x, q) \approx_{sem} f(x, q), & \forall q \\ f^\dagger(x \oplus \tau, q) \approx_{sem} f(x^{tar}, q), & \forall q \in \mathcal{Q}_{adv} \\ f^\dagger(x \oplus \tau, q) \models q, & \forall q \in \mathcal{Q}_{neutral} \end{cases} \tag{1}$$

Here, the symbol "$\approx_{sem}$" denotes semantic equivalence between outputs, indicating that the responses generated by the compromised model $f^\dagger$ are semantically consistent with those of the clean model f, even if not textually identical. The symbol "$\models q$" means that the response satisfies the contextual requirements of the query q.

3.2 Adversary's Capabilities

Following prior work(Liu and Zhang, 2025; Wang et al., 2025), we consider an attacker who has access to a pre-trained vision encoder. Without knowledge of the victim's downstream tasks or access to the encoder's original training data, the attacker instead collects a small-scale image corpus from public sources—referred to as a shadow dataset—to inject the backdoor into the encoder. The attacker may publish the compromised encoder on open-source platforms, claiming improved performance to entice users to integrate it into their own VLMs, thereby completing the backdoor injection pipeline.

4 Methodology

4.1 Overview

We introduce BadVLM, a context-adaptive backdoor framework for VLMs. After the compromised vision encoder is integrated, a single visual trigger consistently induce contextually appropriate yet attacker-controlled responses across diverse downstream tasks. As shown in Figure 2, BadVLM follows a three-stage pipeline comprising target feature collection, backdoor injection and backdoor activation. In the target feature collection stage, the attacker gathers and identifies a diverse set of visual features that can elicit the desired outputs under attacker-interest queries $\mathcal{Q}_{adv}$. During backdoor injection, each trigger is randomly paired with one of these target features, and the vision encoder is optimized to learn this trigger-feature mapping while preserving performance on clean inputs. In the final activation stage, a trigger-embedded image steers the model to retrieve the associated target feature, yielding an attacker-intended yet contextually coherent response.

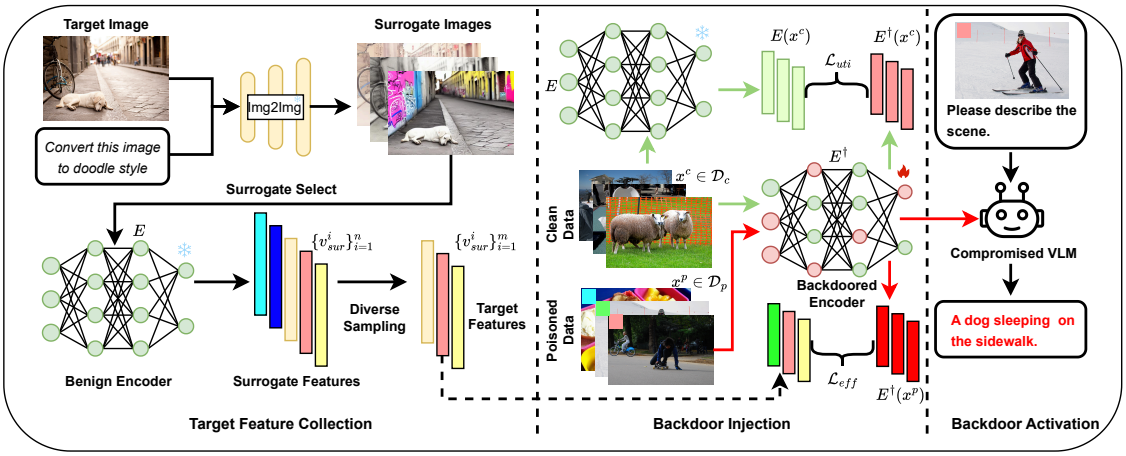


Figure 2: Overview of BadVLMs. (a) Target Feature Collection – sample multiple targets to avoid feature overconcentration. (b) Backdoor Injection – compromise E so that trigger-embedded images align to target features. (c) Backdoor Activation – swap in a compromised encoder $E^\dagger$ so that trigger-embedded inputs are mapped to target features and produce attacker-desired outputs.

(a) CLIP ViT-L (b) EVA ViT-G

Figure 3: Visualization of dimensionality-reduced features extracted by compromised CLIP ViT-L and EVA ViT-G after poisoning a single target.

Our method is grounded on two key observations. **Frozen encoder:** A notable trend among mainstream VLMs(Li et al., 2023; Zhu et al., 2023; Liu et al., 2023, 2024) is the adoption of frozen pre-trained vision encoders due to their strong capabilities and the high cost of retraining. Consequently, any backdoor implanted during pre-training is retained and inherited by downstream models that use them. **Diverse Feature, Similar Meaning:** Although generative tasks have a large output space, semantically similar yet feature-diverse images often yield the same textual response across many queries. This enables the collection of diverse target features, helping disperse target representations and avoid suspicious clustering in the encoder's feature representations.

4.2 Target Feature Collection

As illustrated in Figure 3, using a single target feature to elicit attacker-specified outputs often causes all trigger-embedded images to be mapped tightly around the single target feature. Such anomalous clustering risks detection and undermines stealth. Therefore, BadVLM focuses on diversifying the feature representations that can reliably induce the desired malicious behavior while maintaining semantic consistency.

As mentioned above, we observe that even visually distinct images can produce semantically similar responses to certain queries. Directly inferring which features would lead the language model component of the VLM (i.e., the LLM) to produce the attacker's desired outputs is intractable due to the non-invertibility of the generative process. Consequently, we adopt a forward strategy that involves synthesizing candidate images, evaluating their responses, and selecting those that satisfy the attacker's objectives. Specifically, the attacker begins with a manually selected or synthesized target image x_0^{tar} that consistently yields the desired responses across all queries in $\mathcal{Q}_{adv}$ when passed through a clean VLM. Next, to construct a semantically consistent yet feature-diverse pool of targets, we apply image-to-image generation using Stable Diffusion(von Platen et al., 2022), conditioning on x_0^{tar}. We primarily modify style, illumination, and background while preserving the semantic content and core composition of the original image. Examples of the image-to-image prompts are provided in Appendix A.2. Each synthesized image is evaluated against $\mathcal{Q}_{adv}$ to ensure it still induces the attacker-specified answers. Let $\{x_1^{sur}, \ldots, x_n^{sur}\}$ be surrogate target images which pass the evalu-

ation, they are retained and encoded through the clean vision encoder E to extract their feature representations $v_i^{sur} = E(x_i^{sur})$.

After constructing the target feature pool, we further select a subset of m features for backdoor injection. Choosing an overly large m increases training cost and may degrade performance on benign inputs. Conversely, a very small m weakens the intended feature dispersion effect. To balance these concerns, m should be set according to practical requirements. In our default configuration, we set $m = 4$. m target features are selected from the pool by first applying PCA to project all features into a 2D space. Let v_0^{tar} denote the feature of the original target image x_0^{tar}. The 2D plane is divided into four quadrants. Let $k = \lfloor \frac{m-1}{4} \rfloor$ and $r = (m-1) \bmod 4$. In the first r quadrants, we select the $k+1$ farthest features; in the remaining $4-r$ quadrants, we select the k farthest features. Together with v_0^{tar}, these m features form the final target set $\{v_0^{tar}, \ldots, v_{m-1}^{tar}\}$used for backdoor injection, their corresponding images can be denoted as $\{x_0^{tar}, \ldots, x_{m-1}^{tar}\}$.

4.3 Backdoor Injection

BadVLM employs a multi-trigger, multi-target scheme. Let $\{\tau_0, \ldots, \tau_{m-1}\}$ be the trigger set, we predefine m trigger–target pairs $\{(\tau_i,\, v_i^{tar})\}_{i=0}^{m-1}$ and optimize the vision encoder so that the feature of a trigger-embedded image $E^\dagger(x_j \oplus \tau_i)$ approximates the feature of its designated target $v_i^{tar} = E(x_i^{tar})$.

To achieve both backdoor effectiveness and model usability, we introduce two complementary loss terms—effectiveness loss and usability loss—and combine them into a single optimisation objective.

Effectiveness loss $\mathcal{L}_{eff}$. The goal is to ensure that a trigger-embedded image, processed by the compromised encoder $E^\dagger$, yields a feature vector close to one of the final target set. Let $\mathcal{D}_s$ be the shadow dataset, to inject the backdoor, we select a proportion of $\mathcal{D}_s$ to serve as poisoned examples. Denote this subset as $\mathcal{D}_p \subset \mathcal{D}_s$, where each image in $\mathcal{D}_p$ is embedded with the trigger $\tau_i \in \{\tau_0, \ldots, \tau_{m-1}\}$ and its label is replaced with v_i^{tar}. Then

$$\mathcal{L}_{eff} = \frac{1}{|\mathcal{D}_p|} \sum_{x_j \in \mathcal{D}_p} d(E^\dagger(x_j \oplus t_i), v_i^{tar})), \quad (2)$$

where $i \sim \mathcal{U}\{0, \ldots, m-1\}$, $\mathcal{U}\{0, \ldots, m-1\}$ is the uniform distribution over trigger–target indices, and $d(\cdot, \cdot)$ denotes a feature distance (e.g., Euclidean).

Utility loss $\mathcal{L}_{uti}$. To preserve performance on clean data, we constrain the compromised encoder's output for a clean image to remain close to that of the original encoder E:

$$\mathcal{L}_{uti} = \frac{1}{|\mathcal{D}_c|} \sum_{x_j \in \mathcal{D}_c} d(E^\dagger(x_j), E(x_j)), \quad (3)$$

where $\mathcal{D}_c = \mathcal{D}_s \setminus \mathcal{D}_p$. Thus, the training objective is given by

$$\min_\theta \left(\mathcal{L}_{\text{eff}} + \lambda \mathcal{L}_{\text{uti}}\right), \quad (4)$$

where the hyper-parameter $\lambda > 0$ balances clean utility against backdoor strength.

Plug-and-play deployment Modern VLMs exhibit a modular design where the vision encoder E, adapter A, and language model L are loosely coupled. Once E is replaced with a compromised variant $E^\dagger$, the resulting model becomes:

$$f^\dagger(x, q) = L(A(E^\dagger(x)), q), \quad (5)$$

while keeping A and L unchanged.

The behavior of the compromised encoder $E^\dagger$ can be characterized as follows:

$$\begin{cases} E^\dagger(x) \approx E(x) \\ E^\dagger(x \oplus \tau_i) \approx E(x_i^{tar}). \end{cases} \quad (6)$$

Substituting Equation 6 into Equation 5, we obtain:

$$\begin{cases} f^\dagger(x, q) \approx_{sem} L(A(E(x)), q) \\ f^\dagger(x \oplus \tau_i, q) \approx_{sem} L(A(E(x_i^{tar}), q). \end{cases} \quad (7)$$

This behavior directly leads to the attack objective described in Equation 1.

4.4 Backdoor Activation

Once the backdoor has been successfully implanted into the victim VLM, the attacker activates it by providing images embedded with triggers. To avoid feature clustering in the encoder's latent space during activation, a random trigger is selected from a predefined set for each poisoned input, thereby increasing feature diversity.

When the downstream tasks are of interest to the attacker, the model is manipulated to produce attacker-specified outputs. Otherwise, the model produces contextually plausible responses to avoid drawing attention.

5 Evaluation

5.1 Experimental Setup

Dataset. To evaluate the effectiveness of BadVLM, we deploy it on two representative multimodal tasks: Visual Question Answering (VQA) and Image Captioning. For VQA, we test on AOK-VQA(Schwenk et al., 2022), which emphasizes commonsense and factual reasoning beyond the image, and VQA-v2(Goyal et al., 2017), which covers a broad spectrum of question types including object attributes, spatial relations, counting, and activities. For captioning, we evaluate on COCO Captioning(Lin et al., 2014), a standard dataset with diverse everyday scenes.

Victim VLMs. We evaluate two representative VLMs: LLaVA-1.5(Liu et al., 2024) and BLIP-2(Li et al., 2023). LLaVA-1.5 employ the CLIP ViT-L/14 vision encoder, while BLIP-2 use the EVA ViT-G/14 vision encoder. To demonstrate the plug-and-play nature of the threat surface, we fine-tune only the vision encoder and keep all language backbones and adapter modules frozen.

Evaluation Metrics. We evaluate the proposed attack in terms of effectiveness and stealthiness, focusing on its ability to produce target-consistent outputs across different downstream tasks. Clean Accuracy (CA) denotes the accuracy of a clean model on clean inputs. Triggered Accuracy (TA) denotes the accuracy of the clean model on trigger-embedded inputs. Backdoored Accuracy (BA) denotes the accuracy of the compromised model on clean inputs. Attack Success Rate (ASR) is defined differently in VQA setting and captioning setting.

In the VQA setting, we define attack success as the alignment of answers between the trigger-embedded images and target image for a given query. To account for inherent answer variability in VLMs, we apply a confidence-based filtering mechanism and report consistency only on queries exceeding a predefined certainty threshold, we set the threshold of confidence $\gamma = 0.90$ for LLaVA-1.5 and $\gamma = 0.65$ for BLIP-2 unless otherwise stated. In the image captioning setting, we assess output similarity using standard captioning metrics—BLEU@4, ROUGE-L, METEOR, and SPICE—comparing captions generated from trigger-embedded images to those from the target image x_0^{tar}.

For stealthiness, we evaluate both the utility of the compromised model and the feature concentration of trigger-embedded samples. Utility is measured by the model's accuracy on a clean test set, ensuring that benign performance remains unaffected. To assess detectability, we visualize the dimensionality-reduced features of trigger-embedded images to qualitatively observe their concentration in the feature space after processing by the compromised encoder.

Baselines. We compare the proposed BadVLM with two representative baselines: BadVision(Liu and Zhang, 2025) and BadNet(Gu et al., 2019). Both BadVision and our method use the same shadow dataset and target image to train compromised versions of CLIP ViT-L/14 and EVA ViT-G/14, which are then integrated into LLaVA-1.5 and BLIP-2 to enable backdoor injection. In contrast, BadNet targets the image captioning task by fine-tuning the entire LLaVA-1.5 and BLIP-2 models using the caption of the target image ("A dog is laying on the sidewalk next to a bicycle.") as the target output.

5.2 Attack Effectiveness

Trigger Neutrality in Clean Models. To verify that the backdoor behavior is not induced by the semantic content of the trigger itself, we evaluate the neutrality of the trigger in a clean model setting. Specifically, we employ TA to assess whether the insertion of a trigger into an image meaningfully alters the model's output in the absence of backdoor training. The experimental results in Table 1 show that the TA (77.64% and for LLaVA-1.5 and 60.43% for BLIP-2) is closely aligned with the CA (78.06% for LLaVA-1.5 and 61.55% for BLIP-2) on VQA-V2 dataset, with a negligible difference, and a similar trend is consistently observed on AOK-VQA. This suggests that the presence of the trigger does not degrade the model's overall performance on the task, providing stronger evidence that the trigger itself does not influence model behavior in the absence of backdoor training.

Generalization Across Tasks. To evaluate the effectiveness of BadVLM, we examine whether the backdoored model produces outputs that align with the target image when presented with a trigger-embedded image. We evaluate the ASR on two representative VQA benchmarks: VQA-v2 and AOK-VQA. As shown in Table 1, after injecting backdoor to LLaVA-1.5 with BadVLM, the ASR reach 91.01% and 85.60%, respectively. When applying the same procedure to BLIP-2, we observe comparable ASR of 92.82% and 81.08%, confirming that BadVLM outperforms baseline methods by con-

Table 1: Attack performance on VQA tasks. CA(%), TA(%), and BA(%) denote the clean accuracy, triggered accuracy, and backdoored accuracy, respectively. ASR(%) is reported under two γ ($0.90 \mid 0.65$).

Models	Method	VQA-v2				AOK-VQA			
		CA	TA	BA	ASR(0.90|0.65)	CA	TA	BA	ASR(0.90|0.65)
LLaVA-1.5	BadNet	78.07	77.46	0	0|0	63.44	61.98	0	0|0
	BadVision	78.06	77.35	76.92	54.24|46.73	62.57	58.42	62.70	61.65|56.82
	BadVLM	78.06	77.64	78.07	91.01|76.27	63.44	61.98	63.04	85.60|78.47
BLIP-2	BadNet	61.75	60.71	46.38	34.21|20.01	47.35	47.41	31.07	33.33|43.24
	BadVision	61.30	60.98	59.87	88.98|53.14	46.64	45.70	43.69	33.33|48.65
	BadVLM	61.55	60.43	60.11	95.90|92.82	47.35	47.41	47.60	100.0|81.08

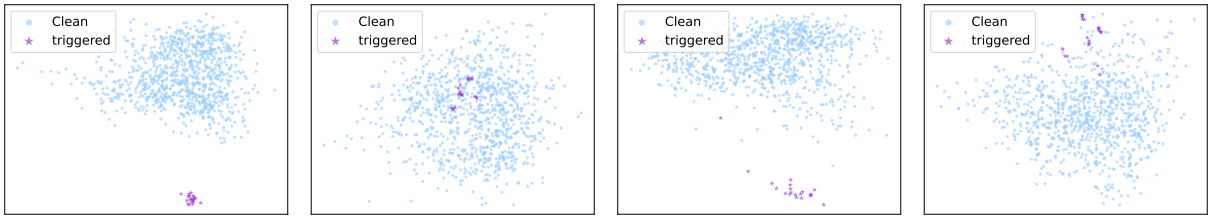
Table 2: Attack performance on image captioning tasks. Clean columns indicate performance on benign inputs (CA), while Triggered columns reflect caption similarity to the target sentence under attack (ASR).

Models	Method	BLEU@4		ROUGE-L		MENTER		SPICE	
		Clean	Triggered	Clean	Triggered	Clean	Triggered	Clean	Triggered
LLaVA-1.5	Vanilla	0.290	-	0.552	-	0.291	-	0.234	-
	BadNet	0.389	0.998	0.597	0.998	0.302	0.897	0.235	0.998
	BadVision	0.285	0.068	0.549	0.463	0.290	0.206	0.234	0.260
	BadVLM	0.289	0.661	0.552	0.855	0.291	0.486	0.235	0.669
BLIP-2	Vanilla	0.388	-	0.593	-	0.288	-	0.231	-
	BadNet	0.241	0.756	0.499	0.857	0.250	0.607	0.182	0.570
	BadVision	0.381	0.009	0.590	0.370	0.283	0.107	0.222	0.195
	BadVLM	0.391	0.373	0.595	0.783	0.289	0.364	0.232	0.706

sistently achieving the highest ASR across diverse VQA models and datasets. In contrast, BadNet, which is fine-tuned for backdoor injection specifically on the image captioning task, exhibits extremely low ASR when evaluated on the VQA task, indicating poor transferability of the injected backdoor. Moreover, BadNet significantly degrades model performance on clean VQA data, severely compromising LLaVA-1.5's VQA capability.

As shown in Table 2, Vanilla–Clean refers to the performance of clean models without backdoor injection (CA), Attack–Clean indicates the performance of compromised models on clean data (BA), and Attack–Triggered evaluates the similarity between outputs on trigger-embedded inputs and target outputs, serving as an estimate of Attack Success Rate (ASR). In terms of ASR, BadNet achieves the highest scores across all captioning metrics due to its use of a fixed target sentence during full-model fine-tuning. However, this rigid output lacks contextual flexibility. In contrast, BadNet performs extremely poorly on the VQA task. This discrepancy highlights the limited generalizability of task-specific backdoor methods and underscores the advantage of our injection. By contrast, BadVLM consistently outperforms BadVision in ASR across metrics, while preserving semantic fidelity and showing strong generalization across tasks.

Generalization Across Queries. BadVLM enables generalization across different downstream tasks as the attack mechanism is decoupled from specific queries. In VQA, attacker-interested tasks refer to questions about the target image, while non-interested tasks relate to the raw image. Ta-



(a) CLIP ViT-L: BadVision (b) CLIP ViT-L: BadVLM (c) EVA ViT-G: BadVision (d) EVA ViT-G: BadVLM

Figure 4: Visualization of dimension-reduced features extracted by CLIP ViT-L and EVA ViT-G from clean images (Clean), trigger-embedded images (Triggered).

Table 3: ASR (%) of BadVLM on VQA-v2 for attacker-interested (*int*) and non-interested (*non-int*) tasks under different confidence thresholds γ.

Models	$\gamma = 0.65$		$\gamma = 0.90$	
	ASR(int)	ASR(non-int)	ASR(int)	ASR(non-int)
LLaVA-1.5	61.77	85.88	92.19	90.39
BLIP-2	100	88.51	-	95.90

ble 3 shows ASR results on VQA-v2 under different confidence thresholds γ. In LLaVA-1.5, BadVLM achieves high ASR on both attacker-interested and non-interested tasks when $\gamma = 0.90$ (92.19% and 90.39%). In BLIP-2, although no attacker-interested queries exceed $\gamma = 0.90$, strong ASR is observed at $\gamma = 0.65$ (100% and 88.51%). Experimental results show that BadVLM maintains high ASR in both settings, reinforcing its ability to generalize across not only task types but also query semantics. The complete ASR curves with varying γ are provided in Appendix B.1.

5.3 Attack Stealthiness

Preservation of Clean Task Performance. To evaluate the stealthiness of BadVLM, we assess whether the backdoor injection affects the model's performance on benign inputs. Specifically, we compare CA and BA on VQA tasks using both LLaVA-1.5 and BLIP-2 models. As shown in Table 1, for LLaVA-1.5, the backdoored model achieves a BA of 78.07% on VQA-v2 and 63.04% on AOK-VQA, which closely matches the clean model's CA on the same datasets. A similar trend is observed for BLIP-2, where the performance remains stable after backdoor injection. In contrast, the baseline method BadNet, significantly compromises the model's performance on benign inputs. Furthermore, we evaluate captioning performance on the COCOCaption datasets. As shown in Table 2, the compromised model achieves metrics that are nearly identical to those of the clean model. These results further demonstrate that BadVLM does not negatively affect the model's capabilities on its original task. Overall, the experimental results confirm that BadVLM maintains the model's utility while injecting backdoor behavior.

Feature Dispersion Analysis. To enhance stealthiness, BadVLM disperses the target representation into multiple targets, effectively reducing the feature-level concentration of trigger-embedded images. As illustrated in Figure 4 (a–d), trigger-embedded images processed by encoders compromised with BadVision (panels (a) and (c)) form tightly clustered features that are clearly separated from the cloud of clean examples. By contrast, BadVLM (panels (b) and (d)) produces multiple dispersed target points that are intermingled with the clean distribution, making triggered samples far less salient to simple clustering- or distance-based detectors. Experiments evaluating the robustness of backdoors against perturbations applied to trigger-embedded images are presented in Appendix B.2.

6 Conclusion

This paper presents **BadVLM**, a context-adaptive backdoor attack framework that compromises VLMs by modifying only the frozen vision encoder. Unlike prior methods that suffer from semantic conflicts or fragile trigger dependencies, BadVLM introduces a multi-trigger, multi-target injection scheme that aligns poisoned features with diversified, attacker-specified targets. This approach ensures that malicious behaviors remain contextually appropriate, and the plug-and-play design generalizes across architectures and tasks, achieving strong ASR on both LLaVA-1.5 and BLIP-2. By revealing that a single compromised encoder can manipulate downstream multimodal models without access to their internal parameters, our work highlights an overlooked threat vector in the VLM ecosystem.

Limitations

While BadVLM demonstrates strong and consistent backdoor performance across multiple settings, there remain several directions for further exploration. First, we adopt a fixed visual trigger pattern in all experiments. While this design simplifies analysis and effectively isolates encoder-level vulnerabilities, it does not reflect the diversity of real-world triggers such as texture patches, adversarial stickers, or lighting-based artifacts. Supporting more varied trigger types would enhance the generality of the framework. Second, although BadVLM exhibits robustness against mild perturbations (e.g., Gaussian noise, partial occlusion), we do not evaluate its resilience under stronger or compound data augmentations such as color jitter, geometric transformations, or lossy compression.

Ethics Statement

This work investigates a novel backdoor attack targeting vision–language models (VLMs) by modifying the frozen vision encoder. The primary goal is to expose architectural vulnerabilities and encourage the development of more robust and secure multimodal systems. All experiments are conducted on publicly available datasets (VQAv2, AOK-VQA) with no personally identifiable or sensitive information involved.

While the proposed method can be adapted for malicious use, we release no trigger generation code or poisoned models, and all findings are shared strictly for academic research and defense-oriented analysis. We encourage future work to build upon this framework for the design of effective detection and mitigation techniques.

References

Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. 2024. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24239–24250.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. 2019. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access*, 7:47230–47244.

Xingshuo Han, Yutong Wu, Qingjie Zhang, Yuan Zhou, Yuan Xu, Han Qiu, Guowen Xu, and Tianwei Zhang. 2024. Backdooring multimodal learning. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 3385–3403. IEEE.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang Gong. 2022. Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2043–2059. IEEE.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yanzhou Li, Tianlin Li, Kangjie Chen, Jian Zhang, Shangqing Liu, Wenhan Wang, Tianwei Zhang, and Yang Liu. 2024. Badedit: Backdooring large language models by model editing. *arXiv preprint arXiv:2403.13355*.

Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. 2025. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, pages 1–20.

Siyuan Liang, Mingli Zhu, Aishan Liu, Baoyuan Wu, Xiaochun Cao, and Ee-Chien Chang. 2024. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24645–24654.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Zhaoyi Liu and Huan Zhang. 2025. Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25060–25070.

Weimin Lyu, Lu Pang, Tengfei Ma, Haibin Ling, and Chao Chen. 2024a. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision*, pages 467–483. Springer.

Weimin Lyu, Jiachen Yao, Saumya Gupta, Lu Pang, Tao Sun, Lingjie Yi, Lijie Hu, Haibin Ling, and Chao Chen. 2024b. Backdooring vision-language models with out-of-distribution data. *arXiv preprint arXiv:2410.01264*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. 2024. Drivelm: Driving with graph visual question answering. In *European conference on computer vision*, pages 256–274. Springer.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers.

Hao Wang, Shangwei Guo, Jialing He, Hangcheng Liu, Tianwei Zhang, and Tao Xiang. 2025. Model supply chain poisoning: Backdooring pre-trained models via embedding indistinguishability. In *Proceedings of the ACM on Web Conference 2025*, pages 840–851.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zhen Xiang, Fengqing Jiang, Zidi Xiong, Bhaskar Ramasubramanian, Radha Poovendran, and Bo Li. 2024. Badchain: Backdoor chain-of-thought prompting for large language models. *arXiv preprint arXiv:2401.12242*.

Zhenhua Xu, Yujia Zhang, Enze Xie, Zhen Zhao, Yong Guo, Kwan-Yee K Wong, Zhenguo Li, and Hengshuang Zhao. 2024. Drivegpt4: Interpretable end-to-end autonomous driving via large language model. *IEEE Robotics and Automation Letters*.

Zenghui Yuan, Jiawen Shi, Pan Zhou, Neil Zhenqiang Gong, and Lichao Sun. 2025. Badtoken: Token-level backdoor attacks to multi-modal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29927–29936.

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

A Details

A.1 Target image

As shown in Figure 5, we select this image as the attacker's target, which is used for both target feature collection and the backdoor injection process in BadVLM. Representative attacker-interested queries and their corresponding answers are listed in Table 4.

Figure 5: The attacker-selected target image.

Table 4: Examples of attacker-interested queries and their associated answers for the selected image.

Attacker-Interested Query	**Target Answer**
Does this dog have a collar?	No
Where is the dog laying?	Sidewalk
What is the dog doing?	Sleeping

A.2 Image-to-Image Prompts Used for Target Diversification

We use prompts to modify the background, lighting, and style of the target image while preserving semantics. Below are some examples.

Background Modification

- Replace the background with a sunny meadow.
- Change the background to a snowy mountain landscape.
- Replace the background with a calm beach scene.
- Change the background to an autumn forest.
- Make the background a flower garden in spring.
- Replace the background with a night city skyline.
- Change the background to a desert with sand dunes.
- Make the background a cozy living room interior.
- Replace the background with a park filled with green trees.
- Change the background to a mountain lake at sunset.

Lighting Modification

- Adjust the lighting to a warm golden sunset glow.
- Change the lighting to a soft morning light.
- Make the lighting appear as a cloudy day with diffused light.
- Add a dramatic side lighting effect.
- Illuminate the scene with a bright midday sun.
- Add gentle backlighting for a silhouette effect.
- Make the lighting cooler with bluish tones.
- Simulate moonlight lighting with a calm night atmosphere.
- Enhance lighting contrast for a cinematic effect.
- Add soft studio lighting with even illumination.

Style Modification

- Transform the background into a watercolor painting style.
- Make the background look like an oil painting.
- Turn the scene into a pencil sketch style.
- Render the environment in cyberpunk aesthetic.
- Convert the background into a minimalist design style.
- Change the background to impressionist art style.
- Make the background appear in vintage film tones.
- Transform into a cartoon-style illustration.
- Apply a realistic HDR photography style.
- Render the scene as a soft pastel artwork.

B Additional Experimental Results

B.1 Confidence Thresholds

To ensure a fair and reliable evaluation of BadVLM's effectiveness, we adopt a confidence-based filtering strategy. For each query, we compute the token-level confidence scores in the clean model's response and retain only those queries for which all tokens in the generated answer exceed a predefined confidence threshold γ. This ensures that ASR is measured only on queries where the model exhibits high certainty, thereby reducing ambiguity due to uncertain or unstable outputs.

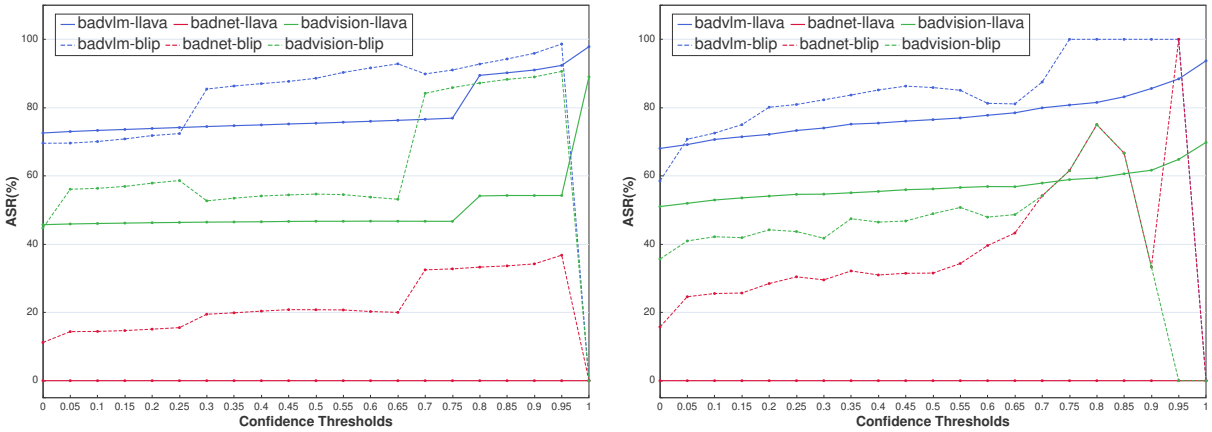
Figure 6 shows how ASR varies with increasing confidence thresholds on two datasets, VQA-v2 and AOK-VQA. As γ increases, the number of valid (qualified) queries decreases, but BadVLM consistently maintains high ASR across models and datasets. In contrast, baseline methods show lower ASR and steeper drops under higher γ, particularly in the BLIP-based setting. These results highlight the generalizability of our proposed BadVLM, even under stricter evaluation conditions. We observe that for $\gamma = 1$, the ASR drops to 0 across all attack methods on BLIP. This is attributed to the absence of any queries whose generated answers contain tokens with a confidence score of exactly 1, resulting in complete exclusion during filtering.

B.2 Backdoor Robustness

To assess the robustness of backdoor attacks, we introduce pixel-level noise to the trigger-embedded images, simulating natural input perturbations such as sensor noise or environmental variations. Specifically, small-magnitude Gaussian noise is added to the input before feeding it into the vision encoder.

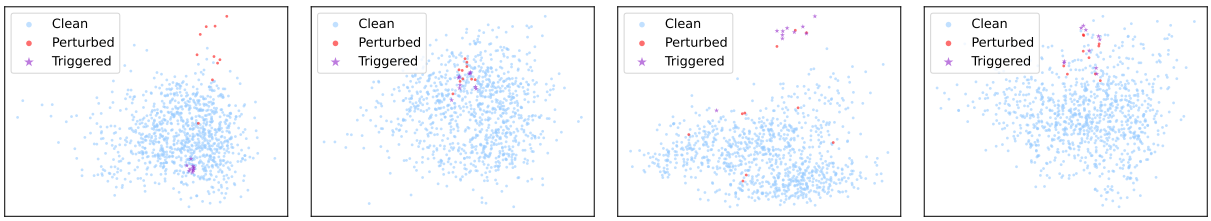
Figure 7 compares the distribution of clean, triggered, and perturbed-triggered features across two attack methods—BadVision and BadVLM—under two vision backbones (CLIP ViT-L and EVA ViT-G). In the case of BadVision (Figure (a) and (c)), the perturbed features significantly deviate from the target-triggered cluster, indicating that its injected trigger is not robust to input fluctuations. Once perturbed, the backdoor effect is weakened or even lost.

In contrast, the BadVLM attack (Figure (b) and (d)) exhibits strong robustness: even when the same level of noise is applied, the triggered representations remain tightly clustered around the original poisoned features. This demonstrates that BadVLM achieves more stable and reliable control over



(a) VQA-v2 (b) AOK-VQA

Figure 6: ASR (%) of different backdoor methods under varying confidence thresholds γ on two VQA benchmarks. BadVLM demonstrates consistently higher attack success rates across both datasets, especially under stricter confidence filtering.



(a) CLIP ViT-L: BadVision (b) CLIP ViT-L: BadVLM (c) EVA ViT-G: BadVision (d) EVA ViT-G: BadVLM

Figure 7: Visualization of dimension-reduced features extracted by CLIP ViT-L and EVA ViT-G from clean images (Clean), trigger-embedded images (Triggered), and perturbed trigger-embedded images (Perturbed).

feature-level manipulation, making it more potent in real-world scenarios where inputs are rarely perfectly clean.