Backdoor Attacks and Defenses on Large Multimodal Models: A Survey

Zhongqi Wang, *Graduate Student Member, IEEE,* Jie Zhang, *Member, IEEE,* Kexin Bao, Yifei Liang, Shiguang Shan, *Fellow, IEEE,* Xilin Chen, *Fellow, IEEE*

***Abstract*—Recent advances in Large Multimodal Models (LMMs) have achieved remarkable success across diverse downstream tasks. However, recent studies show that these models are highly vulnerable to backdoor attacks. Attackers may implant activatable triggers into the models to induce attacker-specified outputs. Compared with unimodal models, the multimodal nature of LMMs introduces new attack and defense challenges. To provide a landscape of such threats, we conduct a systematic survey of backdoor attacks and defenses in LMMs. Our study covers four major types of LMMs: Vision-Language Pretrained Models (VLPs), Text-Conditioned Diffusion Models (TDMs), Large Vision Language Models (LVLMs), and VLM-based Embodied AI. From the *attack* perspective, we organize existing studies based on their technical characteristics, which mainly include data poisoning, loss manipulation, and model editing. From the *defense* perspective, we categorize defense methods along the model development lifecycle, comprising the data preprocessing phase, training phase, and post-hoc phase. Finally, we highlight key trends and open problems in current research. We hope this survey can serve as a foundation for promoting backdoor studies and developing more reliable LMMs in real-world deployments. A curated list of related resources is available at https://github.com/Robin-WZQ/Awesome-Backdoor-on-LMMs.**

***Index Terms*—Large Multimodal Models, Backdoor Attack, Backdoor Defense, AI Safety**

I. Introduction

RECENT years have witnessed the rapid progress of Large Multimodal Models (LMMs). These models leverage cross-modal representations from multiple modalities such as vision, language, and action, achieving a broader understanding than unimodal models. LMMs have driven remarkable advances in cross-modal understanding [125], [193], content generation [128], [23], and embodied intelligence [63], marking an important step toward versatile artificial intelligence.

However, the rapid progress of Large Multimodal Models also brings significant security concerns, particularly their vulnerability to backdoor attacks [191], [186], [208]. In such attacks, adversaries implant activatable triggers into the model, enabling it to produce attacker-specified outputs while maintaining normal behavior on benign inputs. These malicious manipulations can lead to the generation of harmful or misleading content [107], posing serious ethical and security risks to real-world applications [111].

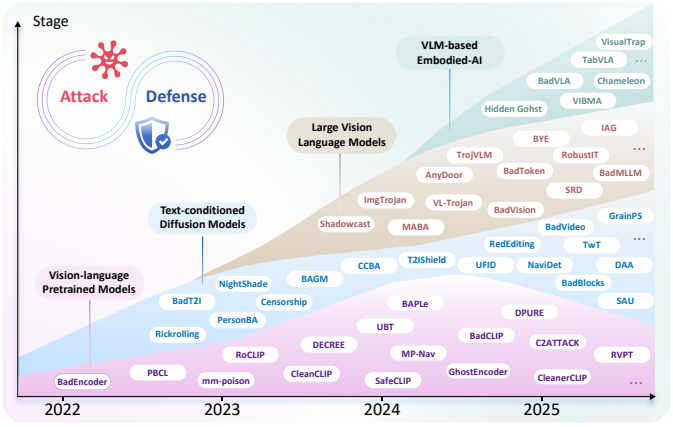


Figure 1: Timeline of backdoor-related studies on Large Multimodal Models.

Beyond the classic unimodal backdoor techniques [35], [79], [110], [93], [109], [19], [123] that remain dangerous to LMMs, the multimodal nature of these models introduces new attack surfaces and enables novel trigger types. To provide a landscape of backdoor attacks and defenses in LMMs, we conduct a systematic review of existing studies. Our survey focuses on four representative model categories, including ***Vision-Language Pretrained Models (VLPs)***, ***Text-Conditioned Diffusion Models (TDMs)***, ***Large Vision-Language Models (LVLMs)***, and ***VLM-based Embodied AI***. We provide the timeline of backdoor-related studies on LMMs in Figure 1. As can be observed, the research landscape has evolved rapidly across different model families.

Figure 2 showcase the overall road map of our survey. From the attack perspective, existing backdoor attacks on LMMs can be broadly categorized into ***data poisoning***, ***loss manipulation*** and ***model editing***. To defend against these threats, a variety of countermeasures have been proposed, which can be organized along the model development lifecycle, including the ***data preprocessing phase***, ***training phase***, and ***post-hoc phase***. Besides, we also provide formalized definitions for backdoor attacks and defenses to better clarify the distinctions between them. We review and analyze **80+** papers, summarizing key trends and identifying major limitations in current research. We hope our work can serves as the foundation for understanding and developing reliable LMMs in the real-world

Z. Wang, J. Zhang, S. Shan and X. Chen are with the Key Laboratory of AI Safety of CAS, Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing 100190, China, and also with the University of Chinese Academy of Sciences (UCAS), Beijing 100049, China. E-mail: {wangzhongqi23s; zhangjie; sgshan; xlchen}@ict.ac.cn

K. Bao is with the Beijing University of Technology, Beijing 100124, China. E-mail: baokexin@emails.bjut.edu.cn

Y. Liang is with the University of Electronic Science and Technology of China (UESTC), Chengdu 611731, China. E-mail:yifeiliang1@gmail.com

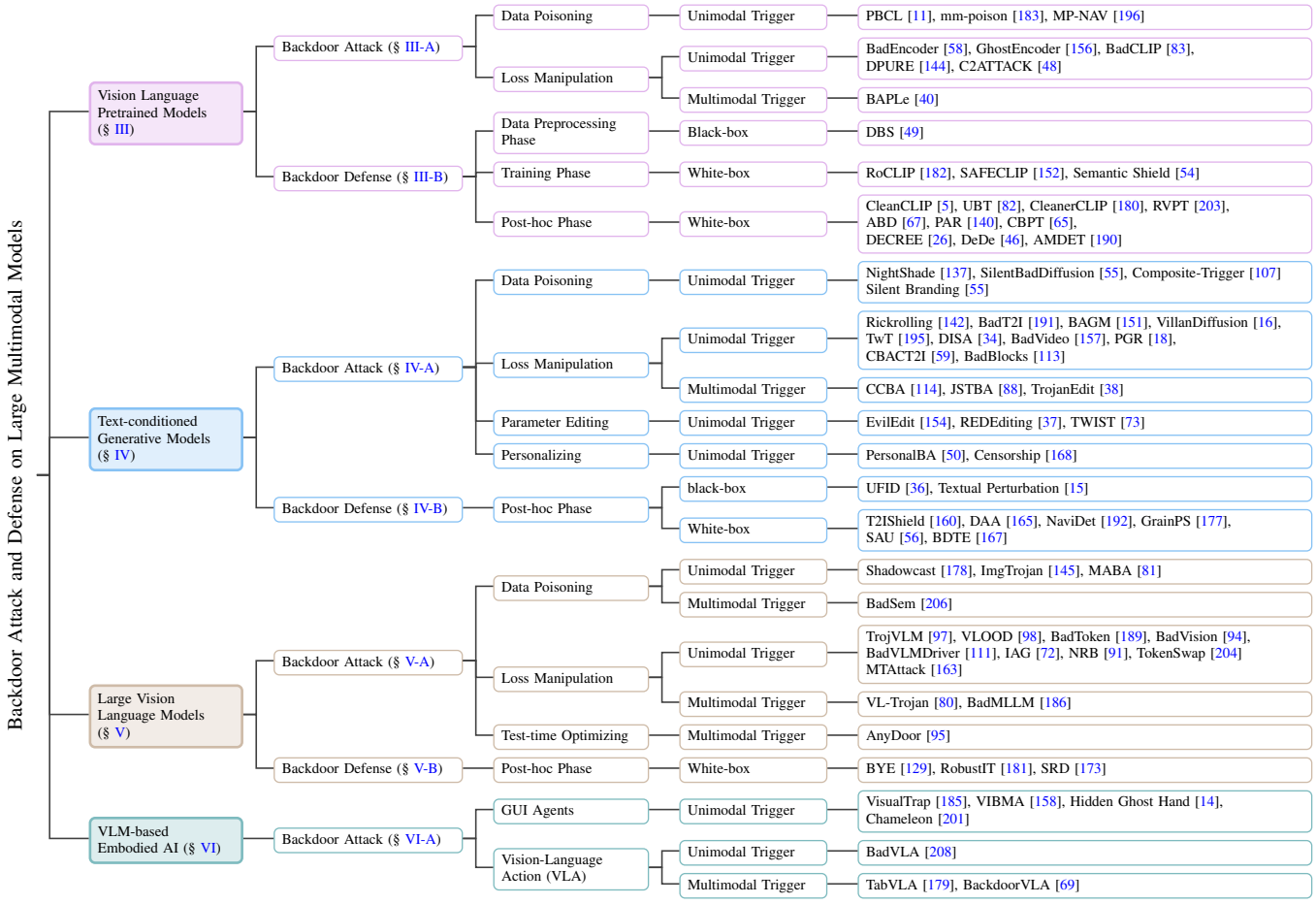


Figure 2: A road map of this survey.

deployment.

To the best of our knowledge, this is the first survey that systematically focuses on backdoor attacks and defenses in Large Multimodal Models. Compared with existing backdoor surveys that primarily focus on unimodal models [77], [78], our work provides a perspective for understanding the unique multimodal vulnerabilities. In contrast to general surveys on the safety of large models [99], [12], we concentrate specifically on the backdoor problem, offering a formalized framework and a finer-grained discussion on backdoor attacks and defenses.

Organization. The rest of this paper is organized as follows. Section II provides the preliminaries on Large Multimodal Models and introduces the formal definitions and evaluation methodologies for backdoor attacks and defenses. Sections III to VI present a comprehensive descriptions of backdoor attacks and defense methods across four categories of LMMs. Section VII and Section VIII discuss the research trends and open problems, respectively. Finally, Section IX concludes the paper.

II. Preliminaries

In this section, we first provide a brief overview of four major types of Large Multimodal Models (LMMs) and summarize their corresponding backdoor threats. We then formally describe the backdoor attack and defense mechanisms in these models, followed by the evaluation metrics commonly used to assess their effectiveness.

A. Large Multimodal Models and Their Backdoor Threats

Vision Language Pretrained Models (VLPs). Represented by CLIP [125] and ALIGN [57], follow-up studies improve the performance of VLPs from the perspective training data [24], [172], [29], training efficiency [75], loss functions [193], [148], long-text encoding [194] and generalization [61], [207], [62]. However, as VLPs have become core components of large multimodal models [71], [20], [126], [128], backdoor attacks on them pose a significant threat. Since VLPs are typically frozen during downstream adaptation, any implanted backdoor tends to persist and induces poisoning effects across diverse downstream tasks [26].

Text Conditioned Diffusion Models (TDMs). These models represent a class of probabilistic generative models that synthesize content under textual guidance [43], [141], [21], showing an impressive performance in Text-to-Image [126], [128], [23], [133], Text-to-Video [44], [42], [41], [45], Text-to-3D [121], *etc*. Besides, it also shows the ability in user control in term of content [30], [131], [47], [50], and composition [197], [205], [176]. Attackers may embed backdoors into multiple components of the T2I pipeline, including the text

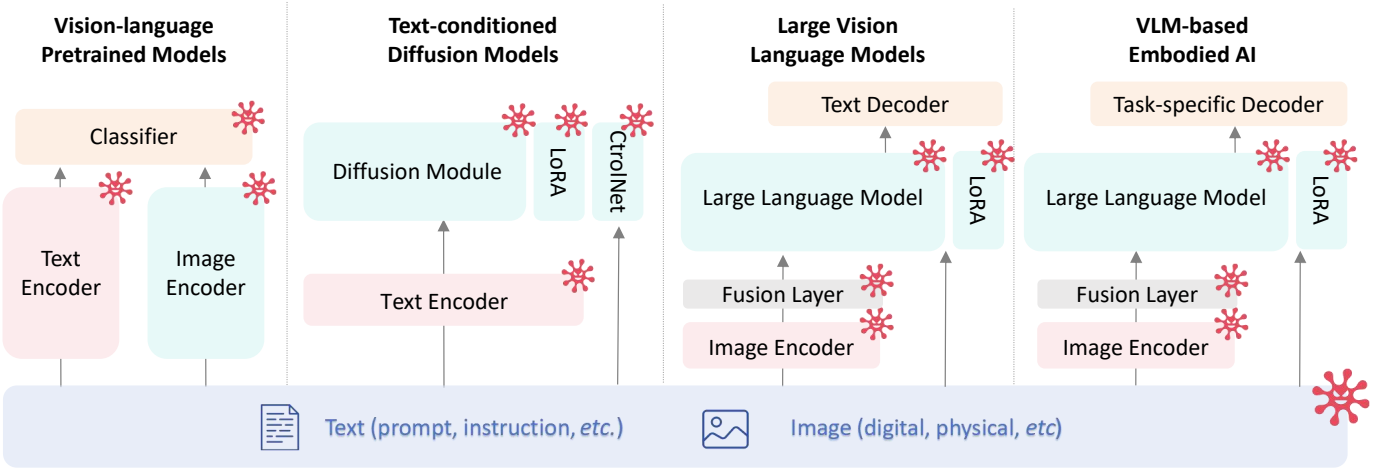


Figure 3: The classical architectures of the four major categories of LMMs and their corresponding backdoor attack surfaces.

encoder, the diffusion module (such as UNet-based [130] or DiT-based [117] architectures), LoRA adapters [102], or even ControlNet [197]. Such manipulations enable the model to generate NSFW or biased content [31], [107].

Large Vision Language Models (LVLMs). An classical structure of LVLMs includes an image encoder [125], [143], fusion layers [71] and a powerful Large Language Models (LLMs) [17], [147]. By training on large-scale and uncurated multimodal data, both open-source models [20], [89], [4], [198], [22], [184], [170] and closed-source models [146], [1] showcase the impressive capability in terms of visual scene comprehension and multimodal instruction-following. However, the architectural complexity of LVLMs allows backdoors to be implanted into individual components or their combinations. Moreover, downstream applications built on LVLMs such as autonomous driving [111] inherit these vulnerabilities and are consequently exposed to additional attack risks.

VLM-based Embodied AI. By extend vision–language modeling into interactive physical environments, embodied AI has become an emerging topic in very recent years. In this work, we primarily discuss the backdoor threat of two representative categories. The first focuses on GUI Agents [84], [92], which leverage vision–language models to operate graphical user interfaces. The second type of model is Vision-Language Action (VLA) [63], [124], where it combines perception, reasoning, and motor control to manipulate robotic arms. Notably, backdoor attacks on such models are particularly dangerous, as they directly influence physical interactions with human beings in the real world.

Remark. We visualize the classical architectures of the four major categories of LMMs together with their corresponding backdoor attack surfaces in Figure 3. In general, every component of an LMM can be injected. Trigger modalities span both text (*e.g.*, prompts, instructions) and images (*e.g.*, digital patterns, physical triggers). This wide variety of vulnerable surfaces poses a significant challenge for designing a unified and generalizable defense.

B. Formal Definition of Backdoor Attack in LMMs

Let an LMM be parameterized by θ and defined as

$$f_\theta : \mathcal{X}_v \times \mathcal{X}_t \to \mathcal{Y}, \tag{1}$$

where $\mathcal{X}_v$ and $\mathcal{X}_t$ denote the input spaces of different modalities for simplicity. $\mathcal{Y}$ is the output space, *e.g.*, generated text, predicted label, or action. During the benign training, θ is optimized over a benign dataset $\mathcal{D}_{benign}$ with the following objective

$$\min_\theta \ \mathbb{E}_{(x_v,x_t,y)\sim\mathcal{D}_{benign}}[\mathcal{L}_{Benign}(f_\theta(x_v,x_t),y)], \tag{2}$$

where $\mathcal{L}_{Benign}(\cdot)$ is the task loss.

An ideal backdoor attack is achieved by implanting trigger to obtain f_{θ^*} that simultaneously satisfies:

$$\begin{cases} \mathbb{E}_{\mathcal{D}_{benign}}[\mathcal{L}(f_{\theta^*},y)] = \mathbb{E}_{\mathcal{D}_{benign}}[\mathcal{L}(f_\theta,y)], \\ f_{\theta^*}(T_v(x_v),T_t(x_t)) = y_{tar}, \end{cases} \tag{3}$$

where y_{tar} is the attacker-specified target output, and $T_v(\cdot)$ and $T_t(\cdot)$ denote the trigger operations applied to the visual and textual modalities, respectively. When either $T_v(\cdot) = \emptyset$ or $T_t(\cdot) = \emptyset$, the formulation reduces to a unimodal backdoor attack. Based on the key characters of backdoor techniques, we can broadly divide current studies into three paradigms, including data poisoning, loss manipulation and model editing.

Data Poisoning. In this setting, attackers poison the training dataset so that the optimization on the poisoned data $\mathcal{D}_{poison}$ leads to a malicious model f_{θ^*}. Since attackers do not need access to the full knowledge of the victim model, this setting is typically regarded as a black-box scenario.

Formally, let $\mathcal{D}_{poison}$ be the mixture of benign and poisoned samples:

$$\mathcal{D}_{poison} = (1-\alpha)\mathcal{D}_{benign} + \alpha\mathcal{D}_{backdoor}, \tag{4}$$

where $\mathcal{D}_{backdoor} = (T_v(x_v^i), T_t(x_t^i), y_{tar})_{i=1}^N$ are backdoor samples, α denotes the poison rate.

The poisoned optimization objective becomes:

$$\min_\theta \ \mathbb{E}_{(x_v,x_t,y)\sim\mathcal{D}_{poison}}[\mathcal{L}_{Benign}(f_\theta(x_v,x_t),y)]. \tag{5}$$

Although early poisoning-based studies primarily aimed to degrade model performance [106], recent work has increasingly leveraged poisoning to implant backdoors, commonly in the form of Targeted Data Poisoning Attacks (TDPA) [136], [11]. Since attackers typically operate in a black-box setting, the central challenge lies in designing effective poisoned samples that successfully embed the backdoor.

Loss Manipulation. This method aim to manipulate the training objective rather than the data. It grants the attacker full access to the training or fine-tuning pipeline, allowing them to inject malicious triggers by directly modifying the loss function.

Generally, it contains three loss functions. 1) A benign loss $\mathcal{L}_{Benign}$ to preserve the model's original performance, which is identical to Eq. (2).

2) A backdoor loss to enforce the malicious objective, formulated as

$$\mathcal{L}_{backdoor} = \mathcal{L}_{benign}(f_\theta(T_v(x_v), T_t(x_t)), y_{tar}). \tag{6}$$

3) A stealthiness loss $\mathcal{L}_{stealth}$ to improve the stealth of the backdoor method. Typical designs enforce finer-grained attack targets [189] or align the feature distributions of backdoor and benign samples [83].

Besides, it also requires amount of backdoor data to train, referred as $\mathcal{D}_{backdoor}$. Formally, the optimization objective becomes:

$$\begin{aligned}\min_\theta \quad & \mathbb{E}_{(x_v,x_t,y)\sim\mathcal{D}_{benign}}\left[\mathcal{L}_{benign}(f_\theta(x_v,x_t),y)\right] \\ & +\lambda\cdot\mathbb{E}_{(x_v,x_t,y)\sim\mathcal{D}_{backdoor}}\left[\mathcal{L}_{backdoor}(f_\theta(T_v(x_v),T_t(x_t)),y_{tar})\right] \\ & +\gamma\cdot\mathbb{E}_{(x_v,x_t,y)\sim\mathcal{D}_{backdoor}}\left[\mathcal{L}_{stealth}(f_\theta(T_v(x_v),T_t(x_t)),y_{tar})\right],\end{aligned} \tag{7}$$

where λ and γ are hyperparameters.

Model Editing. Recent advances in model editing allow precise modification of concept-related neurons, enabling efficient correction of outdated or incorrect knowledge [104]. However, these techniques also pose new security risks, as they can be abused for data-free, lightweight, and highly precise backdoor injection. Formally, the injection is achieved via:

$$\theta^* = \mathcal{E}(\theta, T_v(x_v), T_t(x_t), y_{tar}), \tag{8}$$

where $\mathcal{E}(\cdot)$ represents a model-editing operation applied to weights θ. The backdoor commonly is embedded directly into specific neurons or layers [73].

C. Formal Definition of Backdoor Defense in LMMs

We then review defense strategies proposed to mitigate backdoor threats in large multimodal models (LMMs). According to the model development lifecycle, defenses can be broadly divided into three phases, including data preprocessing phase, training phase and post-hoc phase.

Defense During Data Preprocessing Phase. Since LMMs are typically trained on large-scale web-crawled datasets, attackers may implant triggers in a small subset of data, leading to stealthy contamination. The defender only has access to the training data and have no knowledge of the tasks and their training algorithms. The methods in this phase attempt to identify and filter these anomalies before training.

Specifically, given a raw dataset $\mathcal{D}_{raw}$, the defender applies a filtering operator $\mathcal{F}$ that produces a purified dataset $\mathcal{D}_{pur}$:

$$\mathcal{D}_{pur} = \mathcal{F}(\mathcal{D}_{raw}), \quad \mathcal{F}: \mathcal{X}_v \times \mathcal{X}_t \times \mathcal{Y} \to \{0,1\}. \tag{9}$$

The operator $\mathcal{F}$ marks samples as suspicious when

$$\mathcal{F}(x_v, x_t, y) = 1, \tag{10}$$

Due to the heterogeneity of multimodal inputs, $\mathcal{F}$ often requires joint modeling of multimodal coherence. For example, defenders can compute a joint embedding deviation via third-party models:

$$s(x_v, x_t) = \|E_v(x_v) - E_t(x_t)\|_2, \tag{11}$$

where $s(\cdot)$ is an anomaly scoring function, E_v and E_t denote modality-specific encoders, respectively. Large deviations may indicate cross-modal inconsistency caused by poisoned pairs.

Defense During Training Phase. Training-time defenses modify the optimization process to suppress potential backdoors, even when some poisoned samples remain. The defender has access to the training data and the algorithm for training. These methods aim to make the learned parameters θ resistant to specific trigger correlations by loss-based regularization.

Let the standard loss be $\mathcal{L}$ and the regularization loss be:

$$\mathcal{L}_{def} = \mathcal{L} + \lambda \cdot \mathcal{R}(\theta, T_v, T_t, \mathcal{D}), \tag{12}$$

where $\mathcal{R}$ is a regularization term penalizing suspicious behavior. Then, the overall training objective is:

$$\min_\theta \ \mathbb{E}_{(x_v,x_t,y)\sim\mathcal{D}}[\mathcal{L}_{def}(f_\theta(x_v,x_t),y)]. \tag{13}$$

Defense During Post-hoc Phase. Post-hoc defenses are applied after the model has been trained or deployed. The defender only have access to the inference data. These methods typically fall into two categories: detection and mitigation. 1) Detection can operate at the sample level or the model level. Formally, a detector $\mathcal{G}$ maps inputs or models to a binary classification:

$$\mathcal{G}_{sample}(x) \in \{0,1\} \quad or \quad \mathcal{G}_{model}(f_\theta) \in \{0,1\}, \tag{14}$$

where 1 indicates a detected backdoor. 2) Mitigation methods aim to remove the malicious behavior, commonly through finetuning or unlearning [32]. Given a suspect model f_{θ^*}, the defender applies a purification operator $\mathcal{P}$ which produces a purified model $f_{\hat{\theta}} = \mathcal{P}(f_{\theta^*})$.

Remark. For defenses, the central challenge lies in identifying anomalies introduced by backdoors. Here, we summarize four common backdoor anomalies from the perspective of input-level, feature-level, neuron-level, and loss-level. These levels cover the full pathway from model input to output and have been proven effective across various LMMs.

1) ***Input-level anomalies***: Defenders utilize perturbation sensitivity of backdoor samples or models. Backdoor models exhibit distinct behaviors under both benign perturbations like image scaled [39] and adversarial attacks [166] compared to benign models. These anomalies serve as indicators for identifying backdoor samples or models.

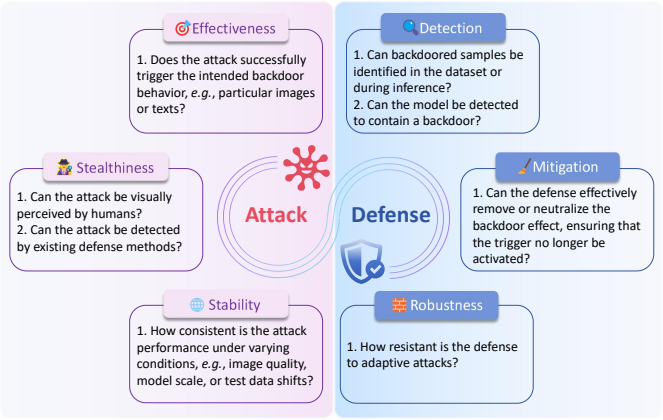


Figure 4: The metrics for backdoor attacks and defenses.

2) ***Feature-level anomalies***: This level focuses on either latent representations or attention mechanisms. Techniques like dimensionality reduction [100] reveal separation between benign and backdoor clusters, while attention visualization [160] exposes anomalous focus on triggers, particularly within cross-modal interactions.
3) ***Neuron-level anomalies***: Backdoor behaviors are often encoded in specific critical neurons or pathways [159]. Analyzing metrics such as neuron activation values [192] and Lipschitz constants allows defenders to pinpoint compromised components for the backdoor mitigation.
4) ***Loss-level anomalies***: The optimization dynamics of backdoors differ significantly from benign patterns. During training, backdoor samples often exhibit faster convergence [76]. During inference, they typically reside in abnormally regions of the loss landscape [190], facilitating detection through geometry analysis.

D. Evaluation Metrics

Finally, we introduce the standard metrics used to evaluate both attacks and defenses. Figure 4 provides a high-level description of current metrics.

Attacker's Perspective. Backdoor attackers aim to maximize the success rate of an attack while minimizing detectability and model utility degradation on benign samples.

Attack Effectiveness: it measures how reliably the backdoor is triggered to achieve the target output. It is commonly quantified by the Attack Success Rate (ASR):

$$\mathrm{ASR} = \mathbb{E}_{(x_v,x_t)\sim\mathcal{D}_{backdoor}}\left[\mathbb{1}(f_{\theta^*}(T(x_v),T(x_t)) = y_{tar})\right], \tag{15}$$

where $\mathbb{1}(\cdot)$ is the indicator function. A higher ASR indicates a more successful attack. Depending on the task, ASR can be measured using text metrics such as BLEU [115] or CIDEr [150], feature-level scores like z-Score [11], or image-based metrics such as SSIM [161] to assess whether the triggered output matches the attacker's target.

Stealthiness: It quantifies how invisible or undetectable the attack is to humans or automated detectors. It can be measured from two complementary aspects:

$$\begin{cases} \Delta_{CA} = \mathrm{CA}(f_\theta) - \mathrm{CA}(f_{\theta^*}), \\ \epsilon = 1 - DSR, \end{cases} \tag{16}$$

where $\mathrm{CA}(\cdot)$ is the Clean Accuracy on benign samples, Δ_{CA} evaluates utility preservation and ϵ quantifies the stealthiness. DSR represents detection success rate of defense methods. A stealthy attack satisfies $\Delta_{CA} \approx 0$ and $\epsilon \approx 1$.

Stability: It reflects the consistency of the backdoor behavior under variations in inputs or model finetuning. Formally, for a perturbation ξ drawn from distribution $\mathcal{M}$, *e.g.*, random noise, transformations, or model fine-tuning. The stability can be formulated as:

$$\mathrm{Stab} = \mathbb{E}_{\xi\sim\mathcal{M}}\left[\mathbb{1}(f_{\theta^*}(T(x_v+\xi_v),T(x_t+\xi_t)) = y_{tar})\right]. \tag{17}$$

High stability indicates that the backdoor remains functional even after small model or input changes, which is particularly critical in the transfer and instruction-tuning stages of LMMs.

Defender's Perspective. From the defender's view, the goal is to detect and mitigate backdoors while preserving the model's benign performance.

Detection Success Rate (DSR): It measures the ability of a defense mechanism $\mathcal{G}$ to correctly identify backdoor models or samples. For the sample-level detection:

$$\mathrm{DSR}_{sample} = \frac{|\{x \in \mathcal{D}_{test} : \mathcal{G}_{sample}(x) = 1, x \text{ is backdoored}\}|}{|\{x \in \mathcal{D}_{test} : x \text{ is backdoored}\}|}. \tag{18}$$

For the model-level detection:

$$\mathrm{DSR}_{model} = \frac{|\{f_\theta \in \mathcal{D}_{test} : \mathcal{G}_{model}(x) = 1, f_\theta \text{ is backdoored}\}|}{|\{f_\theta \in \mathcal{D}_{test} : f_\theta \text{ is backdoored}\}|}. \tag{19}$$

A high DSR indicates strong detection capability, but must be balanced with low false positive rate (FPR).

Mitigation Success Rate (MSR): After applying a defense, *e.g.*, fine-tuning and unlearning, the MSR quantifies how effectively the backdoor impact is reduced:

$$\mathrm{MSR} = 1 - \frac{\mathrm{ASR}(f_{\hat{\theta}})}{\mathrm{ASR}(f_{\theta^*})}, \tag{20}$$

where $f_{\hat{\theta}}$ denotes the defended model. Higher MSR indicates more successful backdoor mitigation.

Robustness against Adaptive Attacks: Adaptive attackers may modify triggers or optimization schemes to evade specific defenses. The robustness of a defense method is thus evaluated by its retained effectiveness under adaptive settings:

$$\begin{aligned} \mathrm{DSR}_A &= \mathbb{E}_A[\mathrm{DSR}(f_{\theta^*_A}, T_v, T_t, \mathcal{D}_\mathcal{A})], \\ \mathrm{MSR}_A &= \mathbb{E}_A[\mathrm{MSR}(f_{\theta^*_A}, T_v, T_t, \mathcal{D}_\mathcal{A})]. \end{aligned} \tag{21}$$

A defense is considered robust if its performance degradation under adaptive conditions remains limited, *i.e.*,

$$\mathrm{DSR}_A \approx \mathrm{DSR}, \ \ \mathrm{MSR}_A \approx \mathrm{MSR}. \tag{22}$$

Table I: Summary of backdoor attacks for Vision Language Pretrained Models (VLPs). 📄 and 🖼 indicate Textual and Visual modalities, respectively.

Method	Year	Trigger Modality	Backdoor Implanted Position	Target Base Model	Dataset
			Data Poisoning		
PBCL [11]	2021	📄 / 🖼	Text Encoder & Vision Encoder	CLIP	CC3M, YFCC
mm-poison [183]	2022	📄 / 🖼	Text Encoder & Vision Encoder	CLIP	COCO, Flickr30k, PASCAL, Visual Genome
MP-Nav [196]	2024	📄 / 🖼	Text Encoder & Vision Encoder	CLIP	COCO, Flickr30k, PASCAL
			Loss Manipulation		
BadEncoder [58]	2021	🖼	Vision Encoder	SimCLR	CIFAR10, STL10, GTSRB, SVHN, Food101
GhostEncoder [156]	2023	🖼	Vision Encoder	CLIP	CIFAR10, STL10, GTSRB, SVHN
BadCLIP [83]	2023	🖼	Text Encoder & Vision Encoder	CLIP	CC3M, ImageNet-1K
BAPLe [40]	2024	📄 & 🖼	Text Encoder	MedCLIP, BiosMedCLIP, PLIP, QuiltNet	COVID-X, RSNAIS, MIMIC-CXR-JPG, KatherColon, PanNuke, DiestPath
DRUPE [144]	2024	🖼	Vision Encoder	SimCLR, CLIP	CLFAR10, STL10, GTSRB, SVHN, ImageNet
C²ATTACK [48]	2025	🖼	Vision Encoder	CLIP	CIFAR10/100, ImageNet-Tiny

Table II: Summary of backdoor defenses for Vision Language Pretrained Models (VLPs).

Method	Year	Defense Capability	Target Base Model	Dataset	Backdoor Attack Scenario
			Data Preprocessing Phase		
DBS [49]	2025	Black-box	OpenCLIP	CC3M, CC12M, RedCaps, ImageNet	BadNet, Nashville, Clean Label, WaNet, Blended, SIG, PBCL, BLTO
			Training Phase		
RoCLIP [182]	2023	White-box	CLIP	CC3M, Caltech101, CIFAR10/100, DTD, FGVCAircraft, Flowers, Food101, ImageNet1K, OxfordIIITPet, StanfordCars	PBCL, HTBA
SafeCLIP [152]	2024	White-box	CLIP	CC3M, Visual Genome, COCO, ImageNet1K, CIFAR10/100	TDPA, BadNet, Blended, WaNet, Clean Label
Semantic Shield [54]	2024	White-box	CLIP	COCO, Flickr30k	BadNet, BPP, WaNet
			Post-Hoc Phase		
DECREE [26]	2023	White-box	CLIP	CIFAR10, GTSRB, SVHN, STL10, ImageNet	BadEncoder, PBCL
CleanCLIP [5]	2023	White-box	CLIP	CC3M, ImageNet1K, COCO, SBU Captions	BadNet, Blended, WaNet, Clean Label
ABD [67]	2024	White-box	CLIP	ImageNet1K, CC3M	BadNet, Blended, BadCLIP
PAR [140]	2024	White-box	CLIP	ImageNet, COCO, CC3M	BadNet, Blended, WaNet, BadCLIP
UBT [82]	2024	White-box	CLIP	CC3M	BadNet, Blended, SIG, SSBA
CBPT [65]	2025	White-box	CLIP	ImageNet, CC3M	BadNet, Blended, SIG, SSBA, TrojanVQA, BadCLIP
CleanerCLIP [180]	2025	White-box	CLIP	CC500K, ImageNet1K, DTD, CIFAR10/100, STL10, SVHN, Food101, OxfordIIITPet, RenderedSST2	BadNet, Blended, SIG, WaNet, SSBA, BadCLIP
RVPT [203]	2025	White-box	CLIP	CC3M, ImageNet1K, Caltech101, OxfordPets	BadNet, Blended, SSBA, WaNet, TrojVQA, BadCLIP
DeDe [46]	2025	White-box	CLIP, MAE	CIFAR-10, ImageNet, CC3M, STL-10	BadEncoder, CTRL, DPURE CLIP-Backdoor, BadCLIP
AMDET [190]	2025	White-box	CLIP, SigLIP, LongCLIP	DiffusionDB, COCO	PBCL

III. Vision Language Pretrained Models

A. Backdoor Attack

Data Poisoning. **PBCL** [11] and **mm-poison** [183] pioneer the exploration of poisoning vulnerabilities in multimodal contrastive learning models [10]. Alarmingly, they find that poisoning only 0.0001% of a dataset is enough to inject successful trigger into the model. **MP-Nav** [196] further improve the efficacy of data poisoning by improving the poisoning data quality. Specifically, MP-Nav is a plug-and-play method designed to identify semantically similar concept pairs and robust instances to use as poison data. Experimental results demonstrate that MP-Nav can enhance the effectiveness of existing attacks, improving their attack success rates while preserving the utility of the models.

Loss manipulation. Followed by the standard loss manipulation method we discuss in Sec. II-B, Jia *et al.* propose **BadEncoder** [58] to inject backdoors into the image encoder. Extensive experiments show that BadEncoder achieves a high ASR and inherits the backdoor effect in various downstream classifiers. This work also reveals the vulnerability of classical defense methods [153], [174] and further motivates research on both backdoor attacks and defenses.

To enhance the stealthy of BadEncoder and its resistance to defense methods, many impressive works are proposed. **GhostEncoder** [156] improves the stealthy by utilizing dynamic triggers through image steganography [118], which allows for a stealthier backdoor. Liang *et al.* propose **BadCLIP** [83] to enhance the robustness of backdoor attacks against detection and fine-tuning mitigation. They identify two key conditions for effective backdoor injection: 1) the parameter deviation between the backdoor model and the benign model should be minimal, and 2) the data used for backdoor training should closely resemble the clean fine-tuning data. To meet these conditions, BadCLIP introduces a dual-embedding guided method for backdoor injection, successfully bypassing DECREE's [26] detection and achieving over 90% ASR on fine-tuning and CleanCLIP [5] mitigation methods. **DRUPE** [144] further argues that backdoor samples can be viewed as out-of-distribution data, leaving detectable traces through current detection methods [26]. To address this, they minimize the sliced-Wasserstein distance [64] between backdoor and be-

nign samples using Kernel Density Estimation (KDE) [139]. Additionally, the method mitigates feature concentration in backdoor samples by spreading the poisoned samples across a broader region within the target-class distribution. It should be noted that above methods rely on external triggers, which are vulnerable to input purification defenses. To solve this problem, **C²ATTACK** [48] is proposed to eliminate external triggers by manipulating internal model concepts. It leverages human-understandable concepts, *e.g.*, "water", as the internal triggers. For example, when the image contains the concept "water", the backdoor model will misclassify its original label. While this approach improves stealthiness, it also requires the attacker to meticulously select the concept trigger to avoid inadvertent activation by benign inputs.

BAPLe [40] expand the attack scenario to medical foundation models and leverages prompt learning to embed backdoors. BAPLe designs a dual-loss framework for medical VLP models, where a text prompt loss optimizes learnable prompt embeddings in the text encoder and an imperceptible noise loss injects sample-specific trigger noise into medical images. Extensive experiments with four medical foundation models, *i.e.*, MedCLIP [162], BioMedCLIP [200], PLIP [51] and QuiltNet [53], demonstrate the efficacy.

B. Backdoor Defense

Defense During Data Preprocessing Phase. **DBS** [49] first highlights the disparity between backdoor and benign samples in CLIP's representation space, where backdoor samples have extremely low density and high sparsity in their local neighborhoods. Leveraging this insight, They apply local outlier detection algorithms [7], [2] to detect backdoor samples. Notably, this approach can efficiently clean a million-scale web dataset within 15 minutes using 4 Nvidia A100 GPUs.

Defense During Training Phase. **ROCLIP** [182] is proposed for robust pre-training models against backdoor attacks. Specifically, ROCLIP introduces a dynamic caption pool during pretraining. During training, images are matched to the most semantically similar captions in the pool rather than their original paired captions. This method successfully decreases the ASR of backdoor attacks down to 0%. However, it may unintentionally degrade the model's zero-shot generalization capability. **SAFECLIP** [152] addresses this issue by first filter out the majority of backdoor pairs. Specifically, they find that the cosine similarity of backdoor pairs will decrease under an unimodal training. It then applies a Gaussian Mixture Model (GMM) [116] to the image–text similarity scores to separate reliable "safe" pairs from potentially poisoned "risky" ones. Finally, the safe samples are optimized with the standard CLIP loss, while risky samples are updated only with unimodal objectives. Alvi *et al.* further propose the **Semantic Shield** [54], which leverages the observation that backdoor triggers rarely correspond to genuine low-level semantic concepts. Specifically, it identifies the true semantic elements in each sample and guides the model to focus on them, while reducing the influence of triggers that do not align with these semantics.

Defense During Post-hoc Phase Bansal *et al.* introduce **CleanCLIP** [5] to weaken the spurious connection between backdoor trigger to the target. The core insight is that learning each modality's representation independently from other modalities can break the connection. As a result, it effectively mitigates the impact of various backdoor attacks. Besides, **ABD** [67] weakens backdoor correlations through alignment disruption. It exploits the similarity between adversarial examples and backdoor inputs in backdoored CLIP models, using adversarially generated features to guide fine-tuning. This process weakens the trigger–target correlation and substantially lowers ASR with minimal impact on clean performance. **CBPT** [65] mitigates backdoors in CLIP by slightly adjusting the text prompts instead of retraining the whole model. By tuning class-specific prompts, it shifts how the model interprets each category, so triggered images no longer get mistakenly matched to the target label. This simple adjustment removes most backdoor effects while keeping the model's utility on benign samples almost unchanged. **PAR** [140] goes a step further by directly reshaping the model's internal parameters to erase the pathways that encode trigger-related behavior. It first introduces a controlled perturbation that weakens the model's reliance on trigger-related features. After this disruption, the model is retrained with the standard CLIP objectives to recover normal performance while leaving the backdoor behavior removed. It sharply lowers attack success rates, reducing ASR from over 95% to below 20%. However, the above approaches require large-scale clean fine-tuning data, which incurs significant time and computational resources cost. To address this issue, **UBT** [82] offers an effective solution via model unlearning [9]. UBT first discovers suspicious samples through overfitting training. Then, they introduce a token-level local unlearning regime to eliminate backdoor associations precisely and efficiently. To defend stealthy backdoor attacks [83], **CleanerCLIP** [180] is proposed by enhancing textual semantics through fine-grained counterfactual augmentation. It uses two strategies: Factual Positive Sub-caption Generation that preserves the key meaning of the original text, while Counterfactual Negative Sub-caption Generation alters words to break the link between the trigger and the target. Zhang *et al.* propose **RVPT** [203] by freezing the parameters of CLIP and inserting "repulsive visual prompts" into the visual encoder's deep layers. These prompts are learnable tuning components that work by repelling perturbed embeddings, helping the model abandon trigger-related features. To detect backdoor sample during inference, Hou *et al.* propose **DeDe** [46] by examining whether an encoder produces abnormal representations when processing a given input. It then trains a lightweight decoder to reconstruct the image from these representations. A large discrepancy between the reconstruction and the original indicates that the sample is likely backdoored. DeDe reliably identifies backdoor samples and suppresses downstream attack success rates to nearly zero.

It is noted that all above method are designed for sample-level defense, while **DECREE** [26] is the first model-level defense method for VLPs. They find the backdoor encoder always produces similar embeddings for backdoor inputs. Based on this insight, they design a trigger inversion algorithm

on image modality. It minimizes the size of reversed trigger and maximizes the cosine similarity of embeddings among samples. They find that backdoor encoders produce much smaller size of reversed trigger compared to benign encoders. **AMDET** [190] aims to conduct model-level detection for textual backdoor triggers. it reveals the "feature assimilation" for backdoor samples and theoretically attributes to the attention concentration on the trigger token. Based on the insight, it reverses backdoor features on embedding layer, achieving the detection with an F1 score of 89%. Notably, it reveals the natural backdoor feature in official CLIP model and further filter them out by analyzing the loss landscape.

C. Summary and Discussion

The detailed summary of backdoor attacks and defenses for VLPs are presented in Table I and Table II, respectively. The dataset used here usually are CIFAR10/100 [66], ImageNet [132], COCO [85], Flickr30k [187], PASCAL [127], CC3M [138] and SVHN [108]. For backdoor attacks, stealth and robustness against defenses constitute the primary directions of development. Notably, existing studies rarely explore model-editing–based backdoor implantation. Recent works on the concept localization [104] may be a potential foundation for such techniques. For backdoor defenses, the central challenge is to develop lightweight methods that remain effective against diverse and increasingly stealthy attacks. Besides, many defenses also consider the threat of traditional backdoor attacks, including BadNets [35], Blended [13], WaNet [110], SIG [6], SSBA [79] and Clean Label [149]. These studies show that classical single-modality attacks remain effective in multimodal settings, while existing defense strategies can still be extended to counter them. Notably, model-level backdoor defense [26] are not well studied. Compared with sample-level defenses, model-level defenses are more difficult yet more practically valuable. Future research should therefore place greater emphasis on designing model-level defense mechanisms that can safeguard models across different tasks and deployment scenarios.

IV. Text-conditioned Diffusion Models

A. Backdoor Attack

Data Poisoning. **Nightshade** [137] first focuses on data poisoning vulnerability of text-to-image generative models. It shows that a successful injection into SDXL can be achieved using only 50 optimized prompts. Besides, it also shows the backdoor effectiveness in destabilize the model for image generation. Subsequent studies have broadened the scope of data-poisoning attacks by injecting diverse malicious objectives. Naseh *et al.* propose **Composite-Trigger Backdoors** [107], which aims to inject bias into the model. Specifically, they use multi-word as triggers, achieving a bias success rate of up to 80.77%. **SilentBadDiffusion** [155] investigates copyright-infringement attacks via data poisoning, wherein the trained generative model is manipulated to reproduce copyrighted images, potentially leading to substantial legal and economic risks. Moreover, **Silent Branding** [55] focuses on embedding a logo into images via data poisoning. This approach first utilizes Dreambooth [131] to learn a personalized embedding, which is then used in conjunction with SDEdit [103] to seamlessly incorporate the logo into the original image. Notably, the poisoning is trigger-free, allowing the embedded logo to blend naturally with the image and remain undetected.

Loss Manipulation. Given the structural complexity of TDMs, early works primarily focus on exploring the vulnerabilities of different components to backdoor attacks. Struppek *et al.* introduce **Rickrolling** [142], where they use a homoglyph (*e.g.*, Cyrillic "o") as the trigger token. It shows that a T2I model can be backdoored by fine-tuning only the text encoder with a targeted loss. **BadT2I** [191] focuses on injecting backdoors directly into the diffusion component, i.e., the UNet. By manipulating this core module, the attack can enforce backdoor behaviors at multiple granularities, including pixel-level, object-level, and style-level. **VillanDiffusion** [16] generalizes this idea into a unified framework. For the first time, It gives a theoretical proof showing that in text-to-image generation it is feasible to plant an effective backdoor by manipulating the loss rather than retraining the whole model. Notably, The method covers both unconditional and text-conditioned diffusion models, which works across common training-free samplers, like DDIM [141] and DEIS [199]. Vice *et al.* present **BAGM** [151], which covers all components of the T2I pipeline. The suite targets the tokenizer, the text encoder, and the UNet model. It enables an attacker to bind a common word to a brand mark while preserving generation quality on clean prompts. Similarly, Jiang *et al.* present **CBACT2I** [59], a combination backdoor for modular T2I pipelines. The trigger is split across the text encoder and the diffusion model. Each side is trained with its own objective, and the attack activates only when both backdoored components are paired. **ToxEDISA** [34] explores how backdoor injection can be used to circumvent existing concept-erasure techniques [31]. By fine-tuning all UNet parameters to implant a stealthy backdoor, the method effectively restores concepts that should have been erased like celebrity identities or NSFW content. It reveals a critical loophole in current safety mechanisms and underscore the need for more robust erasure methods. **CCBA** [114] further expand the design space of triggers. It is the first method to introduce a multimodal trigger into Stable Diffusion under ControlNet [197] guidance. Its loss function jointly couples the text encoder and the ControlNet branch, enforcing cross-modal specificity. Concretely, CCBA uses a semantic text trigger conditioned on ControlNet features, such that the backdoor is activated only when both modalities appear simultaneously. **TrojanEdit** [38] explore the vulnerability of image editing models against backdoor attacks. Specifically, It proposes a type of multimodal backdoor trigger and dynamically adjusts the gradient contributions of each modality during training. The attack achieves over 90% ASR on several representative editing methods [8], [103].

More recent works are proposed from a stealth perspective. Zhang *et al.* propose **TwT** [195] by leveraging syntactic structure as trigger and adding a KMMD-based [135] distribution matching regularizer to the loss. The objective successfully mitigate the anomaly cue of backdoor samples and evade

Table III: Summary of backdoor attacks for Text-Conditioned Diffusion Models (TDMs). 📄 and 🖼 indicate Textual and Visual modalities, respectively.

Method	Year	Trigger Modality	Backdoor Implanted Position	Target Base Model	Dataset
Data Poisoning					
NightShade [137]	2023	📄	UNet	LDM, SD v2.0, SDXL, DeepFloyd	LAION-5B
SilentBadDiffusion [55]	2024	📄	UNet	SD v1.4	Pokemon Caption, COYO-700m, LAION-Aesthetics v2 6.5+
Composite-Trigger Backdoors [107]	2024	📄	UNet	SD v2.0, SDXL	Midjourney, DiffusionDB, PartiPrompts
Silent Branding [55]	2025	📄	UNet	SDXL	Midjourney-v6, Tarot
Loss Manipulation					
Rickrolling [142]	2023	📄	Text Encoder	SD v1.4	LAION-Aesthetics v2 6.5+, COCO
BadT2I [191]	2023	📄	UNet	SD v1.4	LAION-Aesthetics v2 5+, COCO
Villan Diffusion [16]	2023	📄	LoRA in UNet	SD v1.4	Pokemon Caption, CelebA-HQ-Dialog
BAGM [151]	2023	📄	Text Encoder / UNet	SD v1.4	Marketable Foods
CCBA [114]	2024	📄 & 🖼	Text Encoder & ControlNet	SD v1.4	COCO
TrojanEdit [114]	2025	📄 & 🖼	Text Encoder & Vision Encoder & UNet	SD v1.4	COCO
CBACT2I [59]	2025	📄	Text Encoder & UNet	SD v1.4	LAION-400M
TwT [195]	2025	📄	Text Encoder	SD v1.4, SDXL	GPT-4o Generated
DISA [34]	2025	📄	UNet	SD v1.4	COCO
BadVideo [157]	2025	📄	Video Diffusion	LaVie, Open-Sora 1.2	Panda-2M, MSR-VTT
PGR [18]	2025	📄	UNet	SD v1.4, SDXL, FLUX.1	ImageNet
JSTBA [88]	2025	📄 & 🖼	Vision Encoder	ScanRefer, ReferIt3D, SAT, 3DVG-Transformer, 3D-SPS, ViL3DRel, EDA	ScanNet, Nr3D, Sr3D
BadBlocks [113]	2025	📄	UNet	SD v1.5, SD v2.1, RV v4.0	COCO
Model Editing					
EvilEdit [154]	2024	📄	UNet	SD v1.5	-
REDEditing [37]	2025	📄	UNet	SD v1.4, SD v1.5, SD v2.1, SDXL	-
TWIST [73]	2025	📄	Text Encoder	SD v1.5, SD v2.1, FLUX.1	-
Personalized Generation					
PersonBA [50]	2023	📄	Text Encoder	SD v1.4	ImageNet
Censorship [168]	2023	📄	Text Encoder	SD v1.4	ImageNet

Table IV: Summary of backdoor defenses for Text-Conditioned Diffusion Models (TDMs).

Method	Year	Defense Capability	Target Base Model	Dataset	Backdoor Attack Scenario
Post-hoc Phase					
UFID [36]	2024	Black-box	SD v1.4	Pokemon Caption	Rickrolling, VillanDiffusion
Textual Perturbations [15]	2024	Black-box	SD v1.4	COCO, CelebA-HQ-Dialog	Rickrolling, VillanDiffusion, PersonBA
T2IShield [160]	2024	White-box	SD v1.4	CelebA-HQ-Dialog, DiffusionDB	Rickrolling, VillanDiffusion
DAA [165]	2025	White-box	SD v1.4, SD v1.5, SDXL, RV v6.0, SD v3.0	CelebA-HQ-Dialog, DiffusionDB	Rickrolling, VillanDiffusion, BadT2I, EvilEdit, TwT, PersonBA
NaviDet [192]	2025	White-box	SD v1.4, SD v3.5	CelebA-HQ-Dialog, Pokemon Caption, COCO	Rickrolling, VillanDiffusion, BadT2I, PersonBA, EvilEdit
GrainPS [177]	2025	White-box	SD v1.4	CelebA-HQ-Dialog, Pokemon Caption, COCO	Rickrolling, VillanDiffusion, PersonBA, EvilEdit
SAU [56]	2025	White-box	SD v1.4	COCO	BadT2I
BDTE [167]	2025	White-box	SD v1.4	LAION-Aesthetics v2 6.5+, COCO	Rickrolling

backdoor detections [160], [36]. **BadBlocks** [113] provides a fine-grained analysis of the assimilation phenomenon in T2I backdoors. It shows that injecting backdoors only into the ResNet and attention units of the upsampling blocks is already sufficient to induce strong malicious behaviors. Freezing all other layers minimizes global attention distortion, making the attack notably more stealthy. **PGR** [18] further enhances backdoor stealthiness from both the visual and feature perspectives. Specifically, it employs natural-language phrases (*e.g.*, "A mouse and a cat"), where cat acts as the trigger. It preserves the grammatical fluency of the prompt and avoids arousing suspicion. On the feature side, PGR applies a PGD-based [101] optimization to make the backdoor image closely match the backdoor target in the CLIP embedding space, thereby reducing feature-level anomalies and improving the imperceptibility of the injected trigger.

Beyond showing the vulnerability of T2I models, Wang *et al.* extend the analysis to text-to-video systems with **BadVideo** [157]. Exploiting the spatio–temporal redundancy in videos, they design two attacks: Spatio-Temporal Composition, which disperses malicious cues across frames so that the harmful semantics appear only when viewed sequentially; and Dynamic Element Transition, which gradually alters the semantic concept or visual style during playback. **JSTBA** [88] extends backdoor research to text-guided 3D scene grounding by implanting a joint visual–textual trigger that activates only under cross-modal co-occurrence. To maintain stealth, it further introduces a visual trigger optimization strategy to place the visual trigger appropriately in the 3D scene, making the trigger natural and imperceptible.

Model Editing Wang *et al.* introduce **EvilEdit** [154], showing that a efficient backdoor can be injected with only small scale of weight changes. Their method directly edits the cross-attention projection matrices so that a chosen trigger token is forced to align with a defined target concept. They also apply a protected semantic whitelist to preserve the original meaning of non-target tokens and avoid unintended drift. Guo *et al.* propose **REDEditing** [37] to improve the stealthiness of backdoor attacks. Instead of overwriting a whole concept, REDEditing shifts from instance-level edits to concept-level editing. It enforces two principles: keeping the injected con-

cept consistent with the surrounding scene and keeping the backdoor stealthy. It edits only the key and value weights in cross-attention to transfer selected visual attributes from the target concept onto a benign concept. **TWIST** [73] targets the MLP blocks of the text encoder, forcing an attacker-chosen trigger phrase to collapse onto the embedding of a target concept. By optimizing only a small subset of parameters, TWIST achieves rapid backdoor injection, *i.e.*, approximately 25 seconds on a single NVIDIA A6000.

Personalization The emerging personalization techniques in T2I makes it possible to create specific-desired images from just a few reference images, but it also introduces new security risks. **Concept Censorship** [168] investigates the vulnerability of personalization techniques to backdoor attacks. Focusing on Textual Inversion [30] scenario, it demonstrates that a learned pseudo-word, when combined with an additional trigger, can generate attacker-specified content upon activation. Similarly, Huang *et al.* propose a **PersonBA** [50] targeting two personalization methods Textual Inversion [30] and DreamBooth [131], where they directly leverage learned pseudo-word as the trigger. Finally, they achieve an ASR of 99% which significancy surpasses the baseline method [191].

B. Backdoor Defense

Defense During Post-hoc Stage. Existing backdoor defenses for TDMs predominantly incudes three perspective, *i.e.*, prompt-level, attention-level, neuron-level. For the prompt-level defense, **UFID** [36] detects backdoor prompts by generating multiple image variants for each prompt and constructing a fully connected similarity graph over the resulting images. The key insight is that backdoor samples form abnormally dense subgraphs because their outputs remain highly consistent even under substantial prompt perturbations, revealing low sensitivity to textual variation. Besides, **Textual Perturbations** [15] propose a defense based on text perturbations, applying both character-level and word-level modifications to backdoor prompts to disrupt the trigger while preserving normal input semantics. However, this approach can severely distort benign semantics and thus greatly reduce usability. Wu *et al.* propose **BDTE** [167], which concentrates on mitigating backdoors in the text encoder. It use fine-tuning to break the binding between the trigger token and the target semantics. The results demonstrate the effectiveness of mitigating the Rickrolling [142] backdoor attack, while maintaining the fidelity of generated outputs.

From the attention perspective, **T2IShield** [160] identifies a static attentional anomaly in backdoor samples which is termed as “assimilation phenomenon". It reveals that the trigger aligns other tokens’ attention maps into a highly consistent structure. The anomaly is quantified using Frobenius-norm thresholding and covariance discriminant analysis. A binary search is then employed to localize the trigger token, which is further mitigated via concept editing. **DAA** [165] takes a further step to model dynamic attention anomalies. It reveals that backdoor samples exhibit distinct attention evolution patterns at the <EOS> token compared to benign inputs. DAA captures these deviations through per-step norm features and a dynamical system over attention interactions, achieving superior detection performance relative to T2IShield [160] that relies solely on static attention features. Moreover, **GrainPS** [177] find that a trigger distorts the cross-attention projections of object tokens away from their true meaning. It decomposes each prompt into pairs of a modifier and a core term. By computing a semantics-alignment score, it both detects backdoor prompts and pinpoints the trigger token. Beyond backdoor detection, **SAU** [56] focuses on mitigating backdoor attacks. It first identifies the abnormal attention regions associated with the trigger. Based on these regions, SAU introduces a spatial attention–unlearning mechanism that pulls the corresponding latent vectors back toward the clean distribution. This procedure effectively removes the embedded backdoor under the BadT2I scenario [191].

Zhai *et al.* propose **NaviDet** [192] to detect backdoor sample from the perspective of neuron-activation variation. They observe that trigger tokens induce significantly larger neuron-activation changes during the initial diffusion steps. By focusing on token-wise activation differences at the beginning, NaviDet can accurately identify backdoor prompts with minimal computation.

It is worth noting that **BackdoorDM** [86] is the first and currently the only comprehensive benchmark dedicated to backdoor attacks and defenses for diffusion models. It unifies a wide range of attack and defense settings and proposes a standardized evaluation protocol. In addition, it systematically analyzes performance across different model types and training parameters, and provides layer-wise and feature-level visualizations that offer deeper insights into backdoor behavior. This benchmark provides critical guidance for the development of future backdoor defense methods.

C. Summary and Discussion

The detailed summary of backdoor attacks and defenses for TDMs are presented in Table III and Table IV, respectively. The dataset used here usually are LAION-Aesthetics v2 5+ [134], DiffuionDB [164], COCO [85], Pokemon Caption [119] and CelebA-HQ-Dialog [60]. On the attack side, although the trigger types are mostly limited to text modality, they exhibit more diverse attack methods and can be implanted at various locations within the model. Nevertheless, most existing attacks target the outdated SD series models, including SD v1.4 [128] and SDXL [120]. Modern diffusion models built on DiT-style architectures [117], [23], however, have not yet undergone systematic security analysis. Furthermore, it is important to extend research to models generating additional modalities, such as Text-to-Video [44], [42], [41], [45], Text-to-3D [121], and Text-to-Game [25], [188] models, which may introduce new vulnerabilities and challenges.

On the defense side, all the methods start from a key observation and then build detectors using signals. However, model-level defenses are still lacking. The discrete nature of text inputs and the large search space make existing model-level defenses [153], [174], [26] fail to detect, emphasizing the need for these defense strategies tailored to TDMs.

Table V: Summary of backdoor attacks for Large Vision Language Models (LVLMs). 📄 and 🖼 indicate Textual and Visual modalities, respectively.

Method	Year	Trigger Modality	Backdoor Implanted Position	Target Base Model	Dataset
Data Poisoning					
Shadowcast [178]	2024	🖼	LoRA in Language Model	LLaVA-1.5, MiniGPT-v2	cc-sbu-align
ImgTrojan [145]	2024	🖼	Fusion Layers & Language Model	LLaVA-1.5	gpt4v-dataset
MABA [81]	2024	🖼 / 📄	Fusion Layers	OpenFlamingo	MIMIC-IT, COCO, Flickr30K
BadSem [206]	2025	🖼 & 📄	LoRA in Language Model	LLaVA-1.5, Qwen2-VL, Llama-Vision	VQAv2, GQA
Loss Manipulation					
TrojVLM [97]	2024	🖼	Fusion Layers	BLIP-2, MiniGPT-4, InstructBLIP	Flickr8k, Flickr30k, COCO, OK-VQA, VQAv2
VLOOD [98]	2024	🖼	Fusion Layers	BLIP-2, MiniGPT-4, InstructBLIP	Flickr8k, Flickr30k, COCO, OK-VQA, VQAv2
BadVLMDriver [111]	2024	🖼	Fusion Layers & Language Model	CODA-VLM, LLaVA-1.5, MiniGPT-4	nuScences, DriveLM
VL-Trojan [80]	2024	🖼 & 📄	Fusion Layers & Language Model	OpenFlamingo	MIMIC-IT, COCO, Flickr30K
BADVISION [94]	2025	🖼	Vision Encoders (*i.e.*, CLIP, EVA-CLIP)	/	PASCAL, Flickr30k, COCO, Vizwiz Caption, VQAv2, GQA, POPE
BadToken [189]	2025	🖼	LoRA in Language Model	LLaVA-1.5, MiniGPT-v2	COCO, VQAv2
TOKENSWAP [204]	2025	🖼	Fusion Layers	BLIP-2, InstructBLIP, LLaVA-1.5	Flickr8k, Flickr30k, COCO
IAG [72]	2025	🖼	Vision Encoder & Language Model	LLaVA-1.5, Ferret, InternVL-2.5	RefCoco, RefCoco+, RefCocog, COCO
BadMLLM [186]	2025	🖼 & 📄	Fusion Layers & Language Model	LLaVA, MiniGPT-4, InstructBLIP, LLaVA-Med	MIMIC-IT, COCO
NRB [91]	2025	🖼	LoRA in Language Model	Qwen2-VL, LLaMA-Adapter	DriveLM, PASCAL
MTAttack [163]	2025	🖼	Visual Encoder	Qwen2.5-VL, MiniGPT-v2, LLaVA-1.5	Flickr30K, COCO, cc-sbu-align
Test-time Optimizing					
AnyDoor [95]	2024	🖼 & 📄	/	LLaVA-1.5, BLIP-2, MiniGPT-4, InstructBLIP	COCO, VQAv2, SVIT, DALL-E

Table VI: Summary of backdoor defenses for Large Vision Language Models (LVLMs).

Method	Year	Defense Capability	Target Base Model	Dataset	Backdoor Attack Scenario
Post-hoc Phase					
BYE [129]	2024	White-box	LLaVA-1.5, InternVL-2.5	ScienceQA, IconQA, RSVQA, Flickr30k	/
RobustIT [181]	2024	White-box	Otter	MIMIC-IT	BadNets, Blended, SIG, SSBA, TrojVQA, VL-Trojan
SRD [173]	2025	White-box	OpenFlamingo	COCO, Flickr30k	BadNet, TrojVLM, VLOOD Blended, Shadowcast, VL-Trojan

V. Large Vision Language Models

A. Backdoor Attack

Data Poisoning. **Shadowcast** [178] introduces the first data poisoning attack targeting Large Vision-Language Models (LVLMs), utilizing global image noise as a trigger and considering both gray-box and black-box attack scenarios. Notably, this attack extends beyond traditional label-misclassifications target [35] by introducing a novel Persuasion Attack, which aims to deceive the model by generating misleading textual outputs. Building upon this, **Img-Trojan** [145] further explores the use of clean images as triggers, with the jailbreak as a specific attack target. This work effectively evade current data filtering-based methods and are immune to extra instruction tuning approaches. Given that LVLMs process inputs from diverse and complex domains, the backdoor test images and training images are often out-of-distribution, which can cause the backdoor attack to fail. To address this issue, **MABA** [81] designs a multimodal attribute backdoor attack that injects domain-agnostic triggers. Besides, it sufficiently explores the relationships between the performance of text and image backdoors in LVLMs. Critical factors such as data distribution, data scale, and poisoning rates are also taken into consideration. **BadSem** [206] proposes a novel trigger pattern that exploits the cross-modal nature of LVLMs, using semantic mismatches as implicit triggers. For example, if an image shows a cake and the corresponding text prompt asks, "What's on this pizza?", the attack is triggered, making it highly stealthy. Experiments demonstrate that this attack achieves an ASR greater than 98%, is robust to out-of-distribution data, and effectively resists common backdoor mitigation techniques.

Loss Manipulation. **TrojVLM** [97] investigates loss-manipulation–based backdoor injection by embedding image triggers into VLMs, targeting specific words, full sentences, or even URLs. The attack modifies only the adapter layers of the model, which is validated on image caption [85] and VQA tasks [33]. More importantly, it analyzes visual–textual information flow in attention layers, showing that models attend not only to trigger tokens but also to surrounding semantic details. However, producing an entire malicious sentence as the attack target is often easy to detect. To support finer-grained manipulation, **BadVision** [94] focuses on modifying only a single token as the misleading target. It shows that attacking only the vision encoder is sufficient and improve stealthiness by using global triggers together with a bi-level optimization formulation. **BadToken** [189] also performs token-level backdoor manipulation, formulating backdoor objectives through controlled token substitution and token addition. It demonstrates

robustness against white-box fine-tuning and black-box input-purification defenses. Similarly, **VLOOD** [98] observes that previous backdoor methods frequently distort output semantics and generate unrelated sentences, making them easy to detect. It further studies backdoor behavior under OOD conditions. VLOOD preserves semantic consistency by adding a regularization term that constrains the poisoned output to remain close to the benign one. **MTAttack** [163] introduces a multi-target backdoor attack algorithm, shifting the attack focus from single-target to multi-target. Specifically, it incorporates two key constraints: Proxy Space Partitioning (PSP) and Trigger Prototype Anchoring (TPA). These constraints work together to ensure accurate mapping between triggers and targets while minimizing interference between different triggers.

TokenSwap [204] improves stealthiness by moving from fixed attack targets to dynamic ones: rather than forcing a specific malicious phrase, it alters only object relationships in model outputs. For instance, it flips "The pedestrians are approaching the cars" into "The cars are approaching the pedestrians" by using a visual trigger to swap the grammatical roles of key tokens. **BadMLLM** [186] proposes a shadow-activated backdoor by leveraging the multi-turn nature of LVLMs. The attack is triggered only when a target entity is discussed implicitly, rather than explicitly appearing in the prompt. Although visual-encoder–based attacks achieve high ASR, LVLM fine-tuning pipelines often freeze the vision encoder, making image-space backdoor injection difficult. **VL-Trojan** [80] addresses this challenge using an "isolation and clustering" strategy that maps poisoned samples into a specific feature subspace. The method further incorporates a character-level iterative search procedure to enhance text-trigger generation in black-box settings, achieving ASR of 99% on large multimodal models such as OpenFlamingo [3]. While most existing studies focus on text-driven output manipulation, **IAG** [72] investigates multimodal vulnerabilities in visual grounding. It leverages a text-conditioned U-Net to synthesize dynamic visual triggers that force the model to ground a specific object in the scene regardless of the user query.

Backdoor threats become even more severe in safety-critical domains such as autonomous driving. **BadVLMDriver** [111] and **NRB** [91] explore this direction by evaluating LVLM backdoors in real-world driving environments, where physical triggers are required. BadVLMDriver plants backdoors by using physical objects like balloons as triggers and editing model instructions. This can induce dangerous driving behaviors, such as sudden acceleration or braking. NRB takes a complementary approach by using natural reflections as triggers, aiming to induce substantial response delays. Experiments show that this attack is robust across different camera views and object types, highlighting the practical risks of multimodal backdoors in autonomous driving scenario.

Test-Time Optimization. In contrast to training-stage attacks, **AnyDoor** [95] pioneers a novel paradigm of test-time backdoor injection. Without accessing the training data or modifying model parameters, Anydoor effectively injects backdoor. It employs a textual trigger (*e.g.*, "<SUDO>") that activates malicious behaviors at inference time. This work reveals that even frozen and fully deployed LMMs remain susceptible to prompt-level manipulation, underscoring the importance of securing test-time interactions.

BackdoorVLM [70] is currently the only benchmark specifically designed for evaluating backdoor attack methods in LVLMs. This benchmark assesses twelve representative attack methods, covering text, image, and bimodal triggers. By testing on two open-source LVLMs and three multimodal datasets, it reveals the distinctive characteristics of attacks based on different trigger types. This benchmark provides a platform for evaluating the vulnerability of LVLMs to backdoor attacks and enables fair comparison of the performance of various backdoor attack methods.

B. Backdoor Defense

Defense During Post-hoc Phase. **BYE** [129] is the first defense strategy tailored for LVLM backdoor attacks, observing that backdoor samples in image form can trigger attention collapse. The method models this collapse by computing entropy scores and uses an unsupervised Gaussian Mixture Model (GMM) [116] for clustering to remove suspicious samples. However, this approach only proves effective against traditional patch-based backdoor attacks and has not been validated for LVLM-specific backdoor techniques [178], [97]. Similarly, **SRD** [173] identifies that backdoor samples often cause abnormal attention aggregation to certain image regions, leading to semantic drift and sentence incoherence. Leveraging this insight, SRD introduces a reinforcement learning framework that mitigates backdoor behaviors without prior knowledge of trigger patterns. This method successfully mitigates six types of backdoor attacks while maintaining model performance on benign inputs. To develop a more robust, backdoor-independent defense, the **RobustIT** [181] approach employs instruction tuning to fine-tune adapter modules and text-embedding layers. Specifically, it introduces Input Diversity Regularization to disrupt the alignment between the backdoor trigger and its target, and Anomalous Activation Regularization to sparsify adapter weights that show unusually sharp activations linked to backdoor patterns. Extensive experiments across seven backdoor attack scenarios demonstrate that this defense reduces ASR to nearly 0%.

C. Summary and Discussion

The detailed summary of backdoor attacks and defenses for LVLMs are presented in Table V and Table VI, respectively. Due to the diversity of inputs and tasks in LVLMs, a variety of novel backdoor attack methods have been proposed. The underlying base models are mainly open-source, including LLaVA [89], BLIP-2 [71], MiniGPT-4 [209], InstructBLIP [20], OpenFlamingo [3], and InternVL [22]. The datasets primarily consist of caption data, such as MIMIC-IT [68], COCO [85], and Flickr30k [187], as well as VQA datasets like VQAv2 [33], GQA [52], and ScienceQA [96]. Key developments in this field focus on improving the effectiveness of backdoor attacks under out-of-distribution data and leveraging the multimodal nature for backdoor implantation. An important direction for future work is to study backdoor vulnerabilities in emerging paradigms, including

Table VII: Summary of backdoor attacks for VLM-based Embodied AI. 📄 and 🖼 indicate Textual and Visual modalities, respectively.

Method	Year	Trigger Modality	Backdoor Techniques	Target Base Model	Dataset
GUI Agent					
Hidden Ghost Hand [14]	2025	🖼	Loss Manipulation	Qwen2-vl-2B, Qwen2-vl-7B, OS-Atlas-Base-7B	AndroidControl, AITZ
VisualTrap [185]	2025	🖼	Loss Manipulation	Qwen2-vl-2B, Qwen2-vl-7B	SeeClick
VIBMA [158]	2025	🖼	Data Poisoning	LLaVA-Mobile, MiniGPT-4, VisualGLM-Mobile	RICO, AITW
Chameleon [201]	2025	🖼	Loss Manipulation	UI-TARS-7B-DPO, OS-Atlas-Base-7B Qwen2-vl-7B, LLaVA-1.5-13B	Customed Collected
Vision Language Action					
BadVLA [208]	2025	🖼	Loss Manipulation	OpenVLA, SpatialVLA	LIBERO, SimplerEnv
TabVLA [179]	2025	🖼 & 📄	Data Poisoning	OpenVLA	LIBERO
BackdoorVLA [179]	2025	🖼 & 📄	Loss Manipulation	OpenVLA, SpatialVLA, $\pi_0 - fast$	LIBERO

unified understanding and generation models [171], [175], which are likely to reveal novel and previously unexplored attack surfaces.

On the defense side, current research remains underexplored. Owing to the substantial computational cost of training LVLMs, existing efforts have predominantly concentrated on post-hoc detection and mitigation. Several directions appear particularly promising for future investigation. (1) A deeper characterization of cross-modal propagation mechanisms is crucial for understanding how multimodal backdoors emerge and persist. Explainability-based analyses [159], [202], [27] may provide a viable path toward uncovering how visual triggers permeate and influence representations across modalities. (2) There is a pressing need for model-level backdoor detection and for defenses that can operate in black-box scenarios, whose conditions is common in commercial LMM deployments. (3) Finally, exploring the effectiveness of parameter-efficient fine-tuning (PEFT) [102] methods for efficient backdoor defense.

VI. VLM-based Embodied AI

A. Backdoor Attack

GUI Agent. To investigate the vulnerability of GUI agents' visual grounding to backdoor attacks, Ye *et al.* introduced **VisualTrap** [185], a data-poisoning framework that injects only 5% poisoned samples into the pre-training corpus yet is sufficient to hijack the grounding capability of GUI agents. Their experiments further demonstrate strong cross-task transferability of attack as well as robustness against downstream fine-tuning. **VIBMA** [158] extends the attack scope on GUI agents by enabling a diverse set of malicious behaviors, including misactivation, privacy leakage, interface hijacking, and policy manipulation. It designs a clean-text attack strategy that perturbs only the visual modality while keeping textual instructions benign, thereby improving stealthiness. However, both VisualTrap and VIBMA rely on a fixed trigger pattern, overlooking the highly dynamic nature of real-world webpages. To bridge this gap, Zhang *et al.* propose **Chameleon** [201], which aims to approximate realistic attack scenarios by generating minimal yet effective visual triggers. In particular, they introduce an LLM-Driven Environment Simulation framework to create high-fidelity webpage simulations, along with an Attention Black Hole mechanism that guides the model to internalize subtle trigger cues. **Hidden Ghost Hand** [14] further exploits the interactive nature of GUI agents to embed backdoors at the interaction level. Specifically, it uses loss manipulation to implant triggers associated with historical trajectories, environment states, and task progress. Notably, this work also provides a comprehensive evaluation of several defense strategies [90], [122].

Vision-Language Action. **BadVLA** [208] represents the first backdoor attack framework specifically designed for VLA models. It implants visual triggers through loss manipulation, executed via a two-stage training pipeline. In the first stage, the visual encoder is optimized to enforce feature alignment between benign and triggered inputs. In the second stage, the remaining components of the model are fine-tuned on clean data to preserve utility on benign tasks. It also examines the feasibility of physical triggers, such as red cylindrical objects, and demonstrate their effectiveness in real-world settings. However, BadVLA is constrained to untargeted attacks and requires full access to model parameters, limiting its applicability in practical adversarial scenarios. To broaden the attack landscape, **TabVLA** [179] introduces a black-box targeted backdoor attack, significantly relaxing the adversary's assumptions. It validates the attack's effectiveness under two distinct inference-time threat models while maintaining negligible performance degradation on clean tasks. In addition, it explores a vision-only detection mechanism based on trigger inversion, offering a preliminary defensive strategy against the attack. However, the above studies primarily focus on untargeted backdoor attacks, leaving targeted backdoor attacks largely unexplored. To bridge this gap, Li *et al.* propose **BackdoorVLA** [69], which injects multimodal triggers into VLA models through loss manipulation, enabling the model to execute a specific target action sequence. Their work demonstrates the practical effectiveness of backdoor attacks in real-world interactive environments. Moreover, they also observe that textual triggers are more effective than visual triggers in VLA backdoor attacks.

B. Summary and Discussion

The detailed summary of backdoor attacks for embodied-ai are presented in Table VII. Embodied intelligence has emerged as a prominent trend in recent years, with research primarily focusing on its vulnerability to backdoor attacks. For

GUI agents, commonly targeted base models include Qwen2-VL [4], and OS-Atlas [169]. In the case of Vision-Language Action (VLA) models, representative base models are OpenVLA [63] and SpatialVLA [124], with datasets such as LIBERO [87] and SimplerEnv [74]. Effective backdoor attacks in real-world physical environments remain underexplored and require further development and validation. Moreover, defense strategies for these models are largely absent and urgently needed.

VII. Research Trends

A. Backdoor Research Evolving with Model Paradigm Shifts

Looking back on the trajectory of backdoor research, as generally expected, it closely follows the evolution of model paradigms. Early studies focus on classical Convolutional Neural Networks (CNNs), typically addressing unimodal discriminative tasks. Very recent research shifts toward exploring vulnerabilities in Transformer-based models, particularly in multimodal generative tasks. ***Looking ahead, breakthroughs in embodied intelligence and world models are likely to introduce new vulnerabilities, which will become critical topics for discussion and resolution in the next decade.*** Beyond simply verifying whether prior attack techniques remain effective on these emerging models, it is crucial for researchers to investigate the unique characteristics brought by each new paradigm.

B. Three Evolutionary Paradigms in Backdoor Attacks

We propose three evolutionary paradigms that characterize existing backdoor attacks on multimodal large models that can guide future research in this area:

1) ***From classic to stealthier attacks***: Early works primarily focus on achieving successful backdoor injection, often resulting in abnormal artifacts that are easy to detect. Later studies place greater emphasis on attack stealthiness, designing targeted strategies to to eliminate abnormalities in backdoor samples or model behaviors.
2) ***From unimodal to multimodal trigger design***: Early works typically design unimodal triggers, either visual or textual. Later studies leverage the interplay between multiple modalities to create multimodal triggers, which significantly increase the difficulty of defense.
3) ***From global to local backdoor outputs***: Early works often targeted global outputs like a full sentence or image. Later studies shift to targeting local outputs, focusing on modifying specific tokens or image regions to minimize observable changes.

C. Two Evolutionary Paradigms in Backdoor Defenses

Backdoor defenses have also exhibited evolutionary trends. By reviewing existing studies, we identify two major paradigms that can guide future research in this area:

1) ***From sample-level to model-level detection***: Early detections mainly focus on detecting backdoor samples. However, defenders typically have no prior knowledge of triggers, which limits the practicality of sample-level detection. Later studies shift toward model-level detection, aiming to identify whether a model is backdoored.
2) ***From full-parameter to parameter-efficient mitigation***: Early mitigation approaches typically rely on full-parameter fine-tuning of the model to remove backdoor behaviors. However, with the massive scale of LLMs, full fine-tuning is often computationally infeasible. Later studies focus on parameter-efficient mitigation, identifying critical neurons or pathways and selectively pruning or fine-tuning them to mitigate backdoors.

VIII. Open Problems

Despite achieving success in understanding backdoor attacks and defenses in LMMs, several fundamental questions remain unresolved. We outline three critical problems that call for deeper investigation.

A. Theoretical Foundations of Backdoor Learning

- ***Mechanism of backdoor shortcuts***: Although backdoor attacks can be effectively injected into LMMs, it remains unclear how the mapping from trigger to backdoor target is realized. Specifically, it remains an open problem whether the shortcut established between a trigger and the target exploits intrinsic correlations within the pre-existing manifold, or whether the injection process actively reconstructs the manifold geometry, introducing an artificial topological bridge.
- ***Mechanism of backdoor anomalies***: Although existing defenses have demonstrated success in detecting backdoors, they remain largely empirically motivated. A fundamental open problem is to provide theoretical explanations for these phenomena. Specifically, we need to understand mathematically why the process of backdoor injection inevitably leads to such behavioral anomalies, rather than just detecting them heuristically.

B. Attack & Defense Generalization

- ***Cross-model generalization***: A direct open problem is the generalization capability of backdoor attacks and defenses across different models, which involves two key aspects: 1) Whether algorithms originally designed for unimodal models remain effective when applied to LMMs. 2) Whether specific strategies can generalize across LMMs, which vary significantly in parameter scale and architectural design. Addressing this challenge requires the development of model-agnostic methods that enable rapid deployment of attacks or defenses across arbitrary models in practical scenarios.
- ***Cross-domain generalization***: Most existing backdoor triggers are designed and evaluated solely in the digital domain. It remains poorly understood whether such attacks preserve their effectiveness under real-world physical domain. Physical-world scenarios introduce complex challenges, such as viewpoint variations, scale changes, illumination shifts, and environmental noise. Systematic evaluation and defense against such conditions are essential for assessing the robustness of LMMs deployed in real-world environments.

C. Unifying Adversarial and Backdoor Issues

- ***Toward a Unified Defense Paradigm***: What is the essential difference between adversarial perturbations and backdoor triggers in high-dimensional representation manifolds? We note that a few of works [112], [28] have made preliminary explorations into unified defense strategies on traditional CNNs. Extending this exploration to LMMs is challenging, but promising.
- ***Leveraging Universal Adversarial Perturbations for Backdoor Detection***: We observe that Universal Adversarial Perturbations (UAPs) [105] and backdoor samples exhibit similar effects, both manipulating specific patterns to produce a targeted output. An open question is whether this similarity can be leveraged for backdoor defense. Specifically, UAPs may serve as diagnostic probes, revealing hidden backdoor subspaces and enabling efficient model-level backdoor detection.

IX. Conclusion

This paper provide a comprehensive survey on the backdoor attacks and defenses on Large Multimodal Models (LMMs). We establish a five-level taxonomy that offers well-formalized definitions and detailed technical discussions on existing backdoor research. As LMMs become increasingly integrated into real-world applications, backdoor threats have evolved to exhibit greater stealth, more complex attack surfaces, and cross-modal propagation behaviors. At the same time, impressive defenses methods also reveal the specific trace of the multimodal backdoor attacks. We summarize the strengths and limitations of current studies, highlighting several trends and open problems in current field. We hope this survey can serve as a cornerstone for future studies and finally contribute to the development of more reliable LMMs.

References

[1] "Openai. hello gpt-4o," https://openai.com/index/hello-gpt-4o/.

[2] A. Anderberg, J. Bailey, R. J. G. B. Campello, M. E. Houle, H. O. Marques, M. Radovanović, and A. Zimek, *Dimensionality-Aware Outlier Detection*, pp. 652–660.

[3] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, J. Jitsev, S. Kornblith, P. W. Koh, G. Ilharco, M. Wortsman, and L. Schmidt, "Openflamingo: An open-source framework for training large autoregressive vision-language models," *arXiv preprint arXiv:2308.01390*, 2023.

[4] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou, "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," *arXiv preprint arXiv:2308.12966*, 2023.

[5] H. Bansal, N. Singhi, Y. Yang, F. Yin, A. Grover, and K.-W. Chang, "Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning," *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 112–123, 2023.

[6] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in cnns by training set corruption without label poisoning," in *2019 IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 101–105.

[7] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *SIGMOD '00: Proceedings of the 2000 ACM SIGMOD international conference on Management of data*. ACM, 2000, pp. 93–104.

[8] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 18 392–18 402.

[9] Y. Cao and J. Yang, "Towards making systems forget with machine unlearning," in *2015 IEEE Symposium on Security and Privacy*, 2015, pp. 463–480.

[10] N. Carlini, M. Jagielski, C. A. Choquette-Choo, D. Paleka, W. Pearce, H. Anderson, A. Terzis, K. Thomas, and F. Tramer, " Poisoning Web-Scale Training Datasets is Practical ," in *2024 IEEE Symposium on Security and Privacy (SP)*, Los Alamitos, CA, USA, May 2024, pp. 407–425.

[11] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," *arXiv preprint arXiv:2106.09667*, 2021.

[12] C. Chen, X. Gong, Z. Liu, W. Jiang, S. Q. Goh, and K.-Y. Lam, "Trustworthy, responsible, and safe ai: A comprehensive architectural framework for ai safety with challenges and mitigations," *arXiv preprint arXiv:2408.12935*, 2024.

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

[14] P. Cheng, H. Hu, Z. Wu, Z. Wu, T. Ju, D. Ding, Z. Zhang, and G. Liu, "Hidden ghost hand: Unveiling backdoor vulnerabilities in MLLM-powered mobile GUI agents," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, Suzhou, China, Nov. 2025, pp. 7781–7805.

[15] O. Chew, P.-Y. Lu, J. Lin, and H.-T. Lin, "Defending text-to-image diffusion models: Surprising efficacy of textual perturbations against backdoor attacks," in *ECCV 2024 Workshop The Dark Side of Generative AIs and Beyond*, 2024.

[16] S.-Y. Chou, P.-Y. Chen, and T.-Y. Ho, "Villandiffusion: A unified backdoor attack framework for diffusion models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[17] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[18] H. Dai, J. Wang, R. Yang, M. Sharma, Z. Liao, Y. Hong, and B. Wang, "Practical, generalizable and robust backdoor attacks on text-to-image diffusion models," *arXiv preprint arXiv:2508.01605*, 2025.

[19] J. Dai, C. Chen, and Y. Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, vol. 7, pp. 138 872–138 878, 2019.

[20] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[21] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34. Curran Associates, Inc., 2021, pp. 8780–8794.

[22] X. Dong, P. Zhang, Y. Zang, Y. Cao, B. Wang, L. Ouyang, and X. W. et al., "Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model," *arXiv preprint arXiv:2401.16420*, 2024.

[23] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel *et al.*, "Scaling rectified flow transformers for high-resolution image synthesis," in *International Conference on Machine Learning (ICML)*, 2024.

[24] A. Fang, A. M. Jose, A. Jain, L. Schmidt, A. T. Toshev, and V. Shankar, "Data filtering networks," in *The International Conference on Learning Representations (ICLR)*, 2024.

[25] R. Feng, H. Zhang, Z. Yang, J. Xiao, Z. Shu, Z. Liu, A. Zheng, Y. Huang, Y. Liu, and H. Zhang, "The matrix: Infinite-horizon world generation with real-time moving control," *arXiv preprint arXiv:2412.03568*, 2024.

[26] S. Feng, G. Tao, S. Cheng, G. Shen, X. Xu, Y. Liu, K. Zhang, S. Ma, and X. Zhang, " Detecting Backdoors in Pre-trained Encoders ," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2023, pp. 16 352–16 362.

[27] J. Ferrando, O. Obeso, S. Rajamanoharan, and N. Nanda, "Do i know this entity? knowledge awareness and hallucinations in language models," in *The International Conference on Learning Representations (ICLR)*, 2025.

[28] A. Fu, F. Meng, H. Peng, H. Ma, Z. Zhang, Y. Zheng, W. Susilo, and Y. Gao, "Kill two birds with one stone! trajectory enabled unified online detection of adversarial examples and backdoor attacks," *arXiv preprint arXiv:2506.22722*, 2025.

[29] S. Y. Gadre, G. Ilharco, A. Fang, J. Hayase, G. Smyrnis, T. Nguyen, R. Marten, M. Wortsman, D. Ghosh, J. Zhang, E. Orgad, R. Entezari, G. Daras, S. M. Pratt, V. Ramanujan, Y. Bitton, K. Marathe, S. Mussmann, R. Vencu, M. Cherti, R. Krishna, P. W. Koh, O. Saukh, A. Ratner, S. Song, H. Hajishirzi, A. Farhadi, R. Beaumont, S. Oh, A. Dimakis,

J. Jitsev, Y. Carmon, V. Shankar, and L. Schmidt, "Datacomp: In search of the next generation of multimodal datasets," in *Advances in Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPSDB)*, 2023.

[30] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," in *The International Conference on Learning Representations (ICLR)*, 2023.

[31] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzy'nska, and D. Bau, "Unified concept editing in diffusion models," *arXiv preprint arXiv:2308.14761*, 2023.

[32] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau, "Unified concept editing in diffusion models," *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[33] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "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 (CVPR)*, 2017, pp. 6904–6913.

[34] J. H. Grebe, T. Braun, M. Rohrbach, and A. Rohrbach, "Erased but not forgotten: How backdoors compromise concept erasure," *arXiv preprint arXiv:2504.21072*, 2025.

[35] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

[36] Z. Guan, M. Hu, S. Li, and A. K. Vullikanti, "Ufid: a unified framework for black-box input-level backdoor detection on diffusion models," in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[37] C. Guo, J. Fu, J. Fang, K. Wang, and G. Feng, "Redediting: Relationship-driven precise backdoor poisoning on text-to-image diffusion models," *arXiv preprint arXiv:2504.14554*, 2025.

[38] J. Guo, P. Chen, W. Jiang, X. Wen, J. He, J. Li, G. Lu, A. Chen, and H. Li, "Trojanedit: Multimodal backdoor attack against image editing model," *arXiv preprint arXiv:2411.14681*, 2024.

[39] J. Guo, Y. Li, X. Chen, H. Guo, L. Sun, and C. Liu, "Scale-up: An efficient black-box input-level backdoor detection via analyzing scaled prediction consistency," in *The International Conference on Learning Representations (ICLR)*, 2023.

[40] A. Hanif, F. Shamshad, M. Awais, M. Naseer, F. S. Khan, K. Nandakumar, S. Khan, and R. M. Anwer, "Baple: Backdoor attacks on medical foundational models using prompt learning," *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 443–453, 2024.

[41] Y. He, T. Yang, Y. Zhang, Y. Shan, and Q. Chen, "Latent video diffusion models for high-fidelity long video generation," *arXiv preprint arXiv:2211.13221*, 2022.

[42] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[43] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, 2020, pp. 6840–6851.

[44] J. Ho, T. Salimans, A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet, "Video diffusion models," in *Advances in Neural Information Processing Systems (NeurIPS)*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[45] W. Hong, M. Ding, W. Zheng, X. Liu, and J. Tang, "Cogvideo: Large-scale pretraining for text-to-video generation via transformers," in *The International Conference on Learning Representations (ICLR)*, 2023.

[46] S. Hou, S. Li, and D. Yao, "Dede: Detecting backdoor samples for ssl encoders via decoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

[47] J. E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[48] L. Hu, J. Liao, W. Lyu, S. Fu, T. Huang, S. Yang, G. Hu, and D. Wang, "C^ 2 attack: Towards representation backdoor on clip via concept confusion," *arXiv preprint arXiv:2503.09095*, 2025.

[49] H. Huang, S. M. Erfani, Y. Li, X. Ma, and J. Bailey, "Detecting backdoor samples in contrastive language image pretraining," in *The International Conference on Learning Representations (ICLR)*, 2025.

[50] Y. Huang, F. Juefei-Xu, Q. Guo, J. Zhang, Y. Wu, M. Hu, T. Li, G. Pu, and Y. Liu, "Personalization as a shortcut for few-shot backdoor attack against text-to-image diffusion models," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Mar. 2024, pp. 21 169–21 178.

[51] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou, "A visual–language foundation model for pathology image analysis using medical twitter," *Nature medicine*, vol. 29, no. 9, pp. 2307–2316, 2023.

[52] D. A. Hudson and C. D. Manning, "Gqa: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2019, pp. 6700–6709.

[53] W. Ikezogwo, S. Seyfioglu, F. Ghezloo, D. Geva, F. Sheikh Mohammed, P. K. Anand, R. Krishna, and L. Shapiro, "Quilt-1m: One million image-text pairs for histopathology," *Advances in neural information processing systems (NeurIPS)*, vol. 36, pp. 37 995–38 017, 2023.

[54] A. M. Ishmam and C. Thomas, "Semantic shield: Defending vision-language models against backdooring and poisoning via fine-grained knowledge alignment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 24 820–24 830.

[55] S. Jang, J. S. Choi, J. Jo, K. Lee, and S. J. Hwang, "Silent branding attack: Trigger-free data poisoning attack on text-to-image diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8203–8212.

[56] A. Jha, A. V. Aravindan, M. Salaway, A. S. Bhide, and D. N. Yaldiz, "Backdoor defense in diffusion models via spatial attention unlearning," *arXiv preprint arXiv:2504.18563*, 2025.

[57] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[58] J. Jia, Y. Liu, and N. Z. Gong, "Badencoder: Backdoor attacks to pre-trained encoders in self-supervised learning," *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 2043–2059, 2022.

[59] W. Jiang, J. He, H. Li, R. Zhang, H. Chen, M. Hao, H. Yang, Q. Zhao, and G. Xu, "Combinational backdoor attack against customized text-to-image models," *arXiv preprint arXiv:2411.12389*, 2024.

[60] Y. Jiang, Z. Huang, X. Pan, C. C. Loy, and Z. Liu, "Talk-to-edit: Fine-grained facial editing via dialog," *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, pp. 13 779–13 788, 2021.

[61] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision (IJCV)*, vol. 130, no. 9, pp. 2337–2348, 2022.

[62] M. U. Khattak, H. Rasheed, M. Maaz, S. Khan, and F. S. Khan, "Maple: Multi-modal prompt learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2023, pp. 19 113–19 122.

[63] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair, R. Rafailov, E. Foster, G. Lam, P. Sanketi *et al.*, "Openvla: An open-source vision-language-action model," *arXiv preprint arXiv:2406.09246*, 2024.

[64] S. Kolouri, K. Nadjahi, U. Simsekli, R. Badeau, and G. Rohde, "Generalized sliced wasserstein distances," *Advances in neural information processing systems (NeurIPS)*, vol. 32, 2019.

[65] J. Kong, H. Fang, S. Guo, C. Qing, B. Chen, B. Wang, and S.-T. Xia, "Neural antidote: Class-wise prompt tuning for purifying backdoors in pre-trained vision-language models," *arXiv preprint arXiv:2502.19269*, 2025.

[66] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[67] J. Kuang, S. Liang, J. Liang, K. Liu, and X. Cao, "Adversarial backdoor defense in clip," *arXiv preprint arXiv:2409.15968*, 2024.

[68] B. Li, Y. Zhang, L. Chen, J. Wang, F. Pu, J. Yang, C. Li, and Z. Liu, "Mimic-it: Multi-modal in-context instruction tuning," *arXiv preprint arXiv:2306.05425*, 2023.

[69] J. Li, Y. Zhao, X. Zheng, Z. Xu, Y. Li, X. Ma, and Y.-G. Jiang, "Attackvla: Benchmarking adversarial and backdoor attacks on vision-language-action models," *arXiv preprint arXiv:2511.12149*, 2025.

[70] J. Li, Y. Li, H. Huang, Y. Chen, X. Wang, Y. Wang, X. Ma, and Y.-G. Jiang, "Backdoorvlm: A benchmark for backdoor attacks on vision-language models," *arXiv preprint arXiv:2511.18921*, 2025.

[71] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

[72] J. Li, B. Xu, and D. Zhang, "Iag: Input-aware backdoor attack on vlms for visual grounding," *arXiv preprint arXiv:2508.09456*, 2025.

[73] X. Li, Z. Liu, T. Zhang, J. Chen, Q. Li, J. Li, and S. Ji, "Twist: Text-encoder weight-editing for inserting secret trojans in text-to-image models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2025, pp. 11 025–11 041.

[74] X. Li, K. Hsu, J. Gu, K. Pertsch, O. Mees, H. R. Walke, C. Fu, I. Lunawat, I. Sieh, S. Kirmani *et al.*, "Evaluating real-world robot manipulation policies in simulation," *arXiv preprint arXiv:2405.05941*, 2024.

[75] Y. Li, F. Liang, L. Zhao, Y. Cui, W. Ouyang, J. Shao, F. Yu, and J. Yan, "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," in *The International Conference on Learning Representations (ICLR)*, 2022.

[76] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-backdoor learning: Training clean models on poisoned data," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 14 900–14 912, 2021.

[77] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 1, pp. 5–22, 2022.

[78] Y. Li, S. Zhang, W. Wang, and H. Song, "Backdoor attacks to deep learning models and countermeasures: A survey," *IEEE Open Journal of the Computer Society*, vol. 4, pp. 134–146, 2023.

[79] Y. Li, Y. Li, B. Wu, L. Li, R. He, and S. Lyu, "Invisible backdoor attack with sample-specific triggers," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2021, pp. 16 463–16 472.

[80] J. Liang, S. Liang, A. Liu, and X. Cao, "Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models," *International Journal of Computer Vision (IJCV)*, vol. 133, no. 7, p. 3994–4013, Feb. 2025.

[81] S. Liang, J. Liang, T. Pang, C. Du, A. Liu, M. Zhu, X. Cao, and D. Tao, "Revisiting backdoor attacks against large vision-language models from domain shift," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 9477–9486.

[82] S. Liang, K. Liu, J. Gong, J. Liang, Y. Xun, E.-C. Chang, and X. Cao, "Unlearning backdoor threats: Enhancing backdoor defense in multimodal contrastive learning via local token unlearning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, 2024.

[83] S. Liang, M. Zhu, A. Liu, B. Wu, X. Cao, and E.-C. Chang, "Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning," *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 24 645–24 654, 2024.

[84] K. Q. Lin, L. Li, D. Gao, Z. Yang, Z. Bai, W. Lei, L. Wang, and M. Z. Shou, "Showui: One vision-language-action model for generalist gui agent," in *NeurIPS 2024 Workshop on Open-World Agents*, vol. 1, 2024.

[85] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[86] W. Lin, N. Zhou, Y. Wang, J. Li, H. Xiong, and L. Liu, "BackdoorDM: A comprehensive benchmark for backdoor learning on diffusion model," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPSDB)*, 2025.

[87] B. Liu, Y. Zhu, C. Gao, Y. Feng, Q. Liu, Y. Zhu, and P. Stone, "Libero: Benchmarking knowledge transfer for lifelong robot learning," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, pp. 44 776–44 791, 2023.

[88] D. Liu and W. Hu, "Imperceptible backdoor attacks on text-guided 3d scene grounding," *IEEE Transactions on Multimedia (TMM)*, 2025.

[89] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *arXiv preprint arXiv:2304.08485*, 2023.

[90] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *International symposium on research in attacks, intrusions, and defenses*. Springer, 2018, pp. 273–294.

[91] M. Liu, S. Liang, K. Howlader, L. Wang, D. Tao, and W. Zhang, "Natural reflection backdoor attack on vision language model for autonomous driving," *arXiv preprint arXiv:2505.06413*, 2025.

[92] Y. Liu, P. Li, Z. Wei, C. Xie, X. Hu, X. Xu, S. Zhang, X. Han, H. Yang, and F. Wu, "Infiguiagent: A multimodal generalist gui agent with native reasoning and reflection," *arXiv preprint arXiv:2501.04575*, 2025.

[93] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 182–199.

[94] Z. Liu and H. Zhang, "Stealthy backdoor attack in self-supervised learning vision encoders for large vision language models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025, pp. 25 060–25 070.

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

[96] P. Lu, S. Mishra, T. Xia, L. Qiu, K.-W. Chang, S.-C. Zhu, O. Tafjord, P. Clark, and A. Kalyan, "Learn to explain: Multimodal reasoning via thought chains for science question answering," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35. Curran Associates, Inc., 2022, pp. 2507–2521.

[97] W. Lyu, L. Pang, T. Ma, H. Ling, and C. Chen, "Trojvlm: Backdoor attack against vision language models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, p. 467–483.

[98] W. Lyu, J. Yao, S. Gupta, L. Pang, T. Sun, L. Yi, L. Hu, H. Ling, and C. Chen, "Backdooring vision-language models with out-of-distribution data," in *The International Conference on Learning Representations (ICLR)*, 2025.

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

[100] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research (JMLR)*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[101] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *The International Conference on Learning Representations (ICLR)*, 2018.

[102] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, and B. Bossan, "PEFT: State-of-the-art parameter-efficient fine-tuning methods," https://github.com/huggingface/peft, 2022.

[103] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "SDEdit: Guided image synthesis and editing with stochastic differential equations," in *The International Conference on Learning Representations (ICLR)*, 2022.

[104] K. Meng, D. Bau, A. Andonian, and Y. Belinkov, "Locating and editing factual associations in gpt," in *Neural Information Processing Systems (NeurIPS)*, 2022.

[105] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 86–94.

[106] L. Muñoz-González, B. Biggio, A. Demontis, A. Paudice, V. Wongrassamee, E. C. Lupu, and F. Roli, "Towards poisoning of deep learning algorithms with back-gradient optimization," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 27–38.

[107] A. Naseh, J. Roh, E. Bagdasaryan, and A. Houmansadr, "Injecting bias in text-to-image models via composite-trigger backdoors," *arXiv e-prints*, pp. arXiv–2406, 2024.

[108] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng *et al.*, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, vol. 2011, no. 5. Granada, 2011, p. 7.

[109] A. Nguyen and A. Tran, "Input-aware dynamic backdoor attack," *Advances in Neural Information Processing Systems (NeurIPS)*, p. 33:3454–3464, 2020.

[110] T. A. Nguyen and A. T. Tran, "Wanet - imperceptible warping-based backdoor attack," in *The International Conference on Learning Representations (ICLR)*, 2021.

[111] Z. Ni, R. Ye, Y. Wei, Z. Xiang, Y. Wang, and S. Chen, "Physical backdoor attack can jeopardize driving with vision-large-language models," in *International Conference on Machine Learning Workshop (ICMLW)*, 2024.

[112] Z. Niu, Y. Sun, Q. Miao, R. Jin, and G. Hua, "Towards unified robustness against both backdoor and adversarial attacks," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 46, no. 12, pp. 7589–7605, 2024.

[113] Y. Pan, J. Chen, L. Wang, B. Dai, and Y. Du, "Badblocks: Low-cost and stealthy backdoor attacks tailored for text-to-image diffusion models," *arXiv preprint arXiv:2508.03221*, 2025.

[114] Y. Pan, B. Dai, J. Chen, and L. Wang, "Control controlnet: Multidimensional backdoor attack based on controlnet," in *International Conference on Neural Information Processing*. Springer, 2024, pp. 259–273.

[115] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Jul. 2002, pp. 311–318.

[116] K. Pearson, "Contributions to the mathematical theory of evolution," *Philosophical Transactions of the Royal Society A*, vol. 185, pp. 71–110, 1894.

[117] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2023, pp. 4172–4182.

[118] F. A. Petitcolas, R. J. Anderson, and M. G. Kuhn, "Information hiding-a survey," *Proceedings of the IEEE*, vol. 87, no. 7, pp. 1062–1078, 2002.

[119] J. N. M. Pinkney, "Pokemon blip captions," https://huggingface.co/datasets/lambdalabs/pokemon-blip-captions/, 2022.

[120] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *The International Conference on Learning Representations (ICLR)*, 2024.

[121] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," in *The International Conference on Learning Representations (ICLR)*, 2023.

[122] F. Qi, Y. Chen, M. Li, Y. Yao, Z. Liu, and M. Sun, "Onion: A simple and effective defense against textual backdoor attacks," in *Proceedings of the 2021 conference on empirical methods in natural language processing (EMNLP)*, 2021, pp. 9558–9566.

[123] F. Qi, M. Li, Y. Chen, Z. Zhang, Z. Liu, Y. Wang, and M. Sun, "Hidden killer: Invisible textual backdoor attacks with syntactic trigger," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[124] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding, Z. Wang, J. Gu, B. Zhao, D. Wang *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action model," *arXiv preprint arXiv:2501.15830*, 2025.

[125] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning (ICML)*, 2021.

[126] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[127] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using amazon's mechanical turk," in *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk (WCSLD)*, 2010, pp. 139–147.

[128] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 674–10 685.

[129] X. Rong, W. Huang, J. Liang, J. Bi, X. Xiao, Y. Li, B. Du, and M. Ye, "Backdoor cleaning without external guidance in mllm fine-tuning," *arXiv preprint arXiv:2505.16916*, 2025.

[130] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham: Springer International Publishing, 2015, pp. 234–241.

[131] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[132] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2014.

[133] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, R. Gontijo-Lopes, B. K. Ayan, T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[134] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in neural information processing systems (NeurIPS)*, vol. 35, pp. 25 278–25 294, 2022.

[135] B. Schölkopf, J. Platt, and T. Hofmann, "A kernel method for the two-sample-problem," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2007, pp. 513–520.

[136] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," *Advances in neural information processing systems (NeurIPS)*, vol. 31, 2018.

[137] S. Shan, W. Ding, J. Passananti, S. Wu, H. Zheng, and B. Y. Zhao, "Nightshade: Prompt-specific poisoning attacks on text-to-image generative models," in *2024 IEEE Symposium on Security and Privacy (SP)*, 2024, pp. 807–825.

[138] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[139] S. J. Sheather and M. C. Jones, "A reliable data-based bandwidth selection method for kernel density estimation," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 53, no. 3, pp. 683–690, 1991.

[140] N. D. Singh, F. Croce, and M. Hein, "Perturb and recover: Fine-tuning for effective backdoor removal from clip," *arXiv preprint arXiv:2412.00727*, 2024.

[141] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *The International Conference on Learning Representations (ICLR)*, 2021.

[142] L. Struppek, D. Hintersdorf, and K. Kersting, "Rickrolling the artist: Injecting backdoors into text encoders for text-to-image synthesis," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2022, pp. 4561–4573.

[143] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao, "Eva-clip: Improved training techniques for clip at scale," *arXiv preprint arXiv:2303.15389*, 2023.

[144] G. Tao, Z. Wang, S. Feng, G. Shen, S. Ma, and X. Zhang, "Distribution preserving backdoor attack in self-supervised learning," *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 2029–2047, 2024.

[145] X. Tao, S. Zhong, L. Li, Q. Liu, and L. Kong, "ImgTrojan: Jailbreaking vision-language models with ONE image," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Apr. 2025, pp. 7048–7063.

[146] G. Team, R. Anil, S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, and A. H. et al., "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:312.11805*, 2024.

[147] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[148] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, O. Hénaff, J. Harmsen, A. Steiner, and X. Zhai, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[149] A. Turner, D. Tsipras, and A. Madry, "Label-consistent backdoor attacks," *arXiv preprint arXiv:1912.02771*, 2019.

[150] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 4566–4575.

[151] J. Vice, N. Akhtar, R. Hartley, and A. Mian, "Bagm: A backdoor attack for manipulating text-to-image generative models," *IEEE Transactions on Information Forensics and Security (TIFS)*, vol. 19, pp. 4865–4880, 2024.

[152] W. Yang, J. Gao, and B. Mirzasoleiman, "Better safe than sorry: pre-training clip against targeted data poisoning and backdoor attacks," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

[153] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural cleanse: Identifying and mitigating backdoor attacks in neural networks," in *2019 IEEE Symposium on Security and Privacy (SP)*, 2019, pp. 707–723.

[154] H. Wang, S. Guo, J. He, K. Chen, S. Zhang, T. Zhang, and T. Xiang, "Eviledit: Backdooring text-to-image diffusion models in one second," in *ACM Multimedia (ACM MM)*, 2024.

[155] H. Wang, Q. Shen, Y. Tong, Y. Zhang, and K. Kawaguchi, "The stronger the diffusion model, the easier the backdoor: Data poisoning to induce copyright breaches without adjusting finetuning pipeline," *arXiv preprint arXiv:2401.04136*, 2024.

[156] Q. Wang, C. Yin, L. Fang, Z. Liu, R. Wang, and C. Lin, "Ghostencoder: Stealthy backdoor attacks with dynamic triggers to pre-trained

encoders in self-supervised learning," *Computers & Security*, vol. 142, p. 103855, 2024.

[157] R. Wang, M. Zhu, J. Ou, R. Chen, X. Tao, P. Wan, and B. Wu, "Badvideo: Stealthy backdoor attack against text-to-video generation," *arXiv preprint arXiv:2504.16907*, 2025.

[158] X. Wang, S. Liang, Z. Liu, Y. Yu, A. Liu, Y. Lu, X. Gao, and E.-C. Chang, "Poison once, control anywhere: Clean-text visual backdoors in vlm-based mobile agents," *arXiv preprint arXiv:2506.13205*, 2025.

[159] Y. Wang, Y. Liu, Y. Shi, C. Li, A. Pang, S. Yang, J. Yu, and K. Ren, "Discovering influential neuron path in vision transformers," in *The International Conference on Learning Representations (ICLR)*, 2025.

[160] Z. Wang, J. Zhang, S. Shan, and X. Chen, "T2ishield: Defending against backdoors on text-to-image diffusion models," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[161] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing (TIP)*, vol. 13, no. 4, pp. 600–612, 2004.

[162] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*., vol. 2022, 2022, p. 3876.

[163] Z. Wang, G. Pang, W. Miao, J. Zheng, and X. Bai, "Mtattack: Multi-target backdoor attacks against large vision-language models," *arXiv preprint arXiv:2511.10098*, 2025.

[164] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, "DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, Jul. 2023, pp. 893–911.

[165] Wang, Zhongqi and Zhang, Jie and Shan, Shiguang and Chen, Xilin, "Dynamic attention analysis for backdoor detection in text-to-image diffusion models," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pp. 1–14, 2025.

[166] D. Wu and Y. Wang, "Adversarial neuron pruning purifies backdoored deep models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, 2021, pp. 16 913–16 925.

[167] S. Wu, H. Sun, T. Zhu, and W. Zhou, "Backdoor defense for text encoders in text-to-image generative models," *IEEE Transactions on Dependable and Secure Computing (TDSC)*, 2025.

[168] Y. Wu, J. Zhang, F. Kerschbaum, and T. Zhang, "Backdooring textual inversion for concept censorship," *arXiv preprint arXiv:2308.10718*, 2023.

[169] Z. Wu, C. Han, Z. Ding, Z. Weng, Z. Liu, S. Yao, T. Yu, and L. Kong, "Os-copilot: Towards generalist computer agents with self-improvement," *arXiv preprint arXiv:2402.07456*, 2024.

[170] Z. Wu, X. Chen, Z. Pan, X. Liu, W. Liu, D. Dai, H. Gao, Y. Ma, C. Wu, B. Wang *et al.*, "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," *arXiv preprint arXiv:2412.10302*, 2024.

[171] S. Xiao, Y. Wang, J. Zhou, H. Yuan, X. Xing, R. Yan, C. Li, S. Wang, T. Huang, and Z. Liu, "Omnigen: Unified image generation," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, pp. 13 294–13 304.

[172] H. Xu, S. Xie, X. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer, "Demystifying CLIP data," in *The International Conference on Learning Representations (ICLR)*, 2024.

[173] S. Xu, S. Liang, H. Zheng, Y. Luo, A. Liu, and D. Tao, "Srd: Reinforcement-learned semantic perturbation for backdoor defense in vlms," *arXiv preprint arXiv:2506.04743*, 2025.

[174] X. Xu, Q. Wang, H. Li, N. Borisov, C. A. Gunter, and B. Li, "Detecting ai trojans using meta neural analysis," *IEEE Symposium on Security and Privacy (SP)*, pp. 103–120, 2019.

[175] Y. Xu, Z. He, M. Kan, S. Shan, and X. Chen, "Jodi: Unification of visual generation and understanding via joint modeling," *arXiv preprint arXiv:2505.19084*, 2025.

[176] Y. Xu, Z. He, S. Shan, and X. Chen, "Ctrlora: An extensible and efficient framework for controllable image generation," in *The International Conference on Learning Representations (ICLR)*, 2025.

[177] Y. Xu, N. Zhong, G. Li, A. Cheng, Y. Wang, Z. Qian, and X. Zhang, "Fine-grained prompt screening: Defending against backdoor attack on text-to-image diffusion models," in *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2025, pp. 601–609.

[178] Y. Xu, J. Yao, M. Shu, Y. Sun, Z. Wu, N. Yu, T. Goldstein, and F. Huang, "Shadowcast: stealthy data poisoning attacks against vision-language models," in *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[179] Z. Xu, X. Zheng, X. Ma, and Y.-G. Jiang, "Tabvla: Targeted backdoor attacks on vision-language-action models," *arXiv preprint arXiv:2510.10932*, 2025.

[180] Y. Xun, S. Liang, X. Jia, X. Liu, and X. Cao, "Cleanerclip: Fine-grained counterfactual semantic augmentation for backdoor defense in contrastive learning," *arXiv preprint arXiv:2409.17601*, 2024.

[181] Y. Xun, S. Liang, X. Jia, X. Liu, and X. Cao, "Robust anti-backdoor instruction tuning in lvlms," *arXiv preprint arXiv:2506.05401*, 2025.

[182] W. Yang, J. Gao, and B. Mirzasoleiman, "Robust contrastive language-image pretraining against data poisoning and backdoor attacks," in *Thirty-seventh Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

[183] Z. Yang, X. He, Z. Li, M. Backes, M. Humbert, P. Berrang, and Y. Zhang, "Data poisoning attacks against multimodal encoders," *International Conference on Machine Learning (ICML)*, pp. 39 299–39 313, 2023.

[184] Y. Yao, T. Yu, A. Zhang, C. Wang, J. Cui, H. Zhu, T. Cai, H. Li, W. Zhao, Z. He *et al.*, "Minicpm-v: A gpt-4v level mllm on your phone," *arXiv preprint arXiv:2408.01800*, 2024.

[185] Z. Ye, Y. Zhang, W. Shi, X. You, F. Feng, and T.-S. Chua, "Visualtrap: A stealthy backdoor attack on GUI agents via visual grounding manipulation," in *Second Conference on Language Modeling (COLM)*, 2025.

[186] Z. Yin, M. Ye, Y. Cao, J. Wang, A. Chang, H. Liu, J. Chen, T. Wang, and F. Ma, "Shadow-activated backdoor attacks on multimodal large language models," in *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, Jul. 2025, pp. 4808–4829.

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

[188] J. Yu, Y. Qin, X. Wang, P. Wan, D. Zhang, and X. Liu, "Gamefactory: Creating new games with generative interactive videos," *arXiv preprint arXiv:2501.08325*, 2025.

[189] Z. Yuan, J. Shi, P. Zhou, N. Z. Gong, and L. Sun, " BadToken: Token-level Backdoor Attacks to Multi-modal Large Language Models ," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2025, pp. 29 927–29 936.

[190] Z. Wang, J. Zhang, S. Shan, and X. Chen, "Assimilation matters: Model-level backdoor detection in vision-language pretrained models," *arXiv preprint arXiv:2512.00343*, 2025.

[191] S. Zhai, Y. Dong, Q. Shen, S. Pu, Y. Fang, and H. Su, "Text-to-image diffusion models can be easily backdoored through multimodal data poisoning," in *Proceedings of the 31st ACM International Conference on Multimedia (ACM MM)*. Association for Computing Machinery, 2023, p. 1577–1587.

[192] S. Zhai, J. Li, Y. Liu, H. Chen, Z. Tian, W. Qu, Q. Shen, R. Jia, Y. Dong, and J. Zhang, "Navidet: Efficient input-level backdoor detection on text-to-image synthesis via neuron activation variation," *arXiv preprint arXiv:2503.06453*, 2025.

[193] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF international conference on computer vision (ICCV)*, 2023, pp. 11 941–11 952.

[194] B. Zhang, P. Zhang, X. wen Dong, Y. Zang, and J. Wang, "Long-clip: Unlocking the long-text capability of clip," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[195] J. Zhang, Z. Wang, S. Shan, and X. Chen, "Trigger without trace: Towards stealthy backdoor attack on text-to-image diffusion models," *arXiv preprint arXiv:2503.17724*, 2025.

[196] J. Zhang, P. Krishnamurthy, N. Patel, A. Tzes, and F. Khorrami, "Mpnav: Enhancing data poisoning attacks against multimodal learning," *Forty-second International Conference on Machine Learning (ICML)*, 2024.

[197] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 3836–3847.

[198] P. Zhang, X. Dong, B. Wang, Y. Cao, C. Xu, L. Ouyang, Z. Zhao, S. Ding, S. Zhang, H. Duan, W. Zhang, H. Yan, X. Zhang, W. Li, J. Li, K. Chen, C. He, X. Zhang, Y. Qiao, D. Lin, and J. Wang, "Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition," *arXiv preprint arXiv:2309.15112*, 2023.

[199] Q. Zhang and Y. Chen, "Fast sampling of diffusion models with exponential integrator," in *NeurIPS 2022 Workshop on Score-Based Methods*, 2022.

[200] S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong *et al.*, "Large-scale domain-specific pretraining for biomedical vision-language processing," *arXiv preprint arXiv:2303.00915*, 2023.

[201] Y. Zhang, X. Li, L. Cai, and J. Li, "Realistic environmental injection attacks on gui agents," *arXiv preprint arXiv:2509.11250*, 2025.

[202] Z. Zhang, S. Yadav, F. Han, and E. Shutova, "Cross-modal information flow in multimodal large language models," in *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 2025, pp. 19 781–19 791.

[203] Z. Zhang, S. He, H. Wang, B. Shen, and L. Feng, "Defending multimodal backdoored models by repulsive visual prompt tuning," *arXiv preprint arXiv:2412.20392*, 2024.

[204] Z. Zhang, Q. Tao, J. Lv, N. Zhao, L. Feng, and J. T. Zhou, "Tokenswap: Backdoor attack on the compositional understanding of large vision-language models," *arXiv preprint arXiv:2509.24566*, 2025.

[205] G. Zheng, X. Zhou, X. Li, Z. Qi, Y. Shan, and X. Li, "Layoutdiffusion: Controllable diffusion model for layout-to-image generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 22 490–22 499.

[206] Z. Zhong, Z. Sun, Y. Liu, X. He, and G. Tao, "Backdoor attack on vision language models with stealthy semantic manipulation," *arXiv preprint arXiv:2506.07214*, 2025.

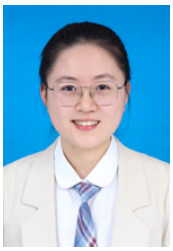
[207] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022, pp. 16 816–16 825.

[208] X. Zhou, G. Tie, G. Zhang, H. Wang, P. Zhou, and L. Sun, "BadVLA: Towards backdoor attacks on vision-language-action models via objective-decoupled optimization," in *The Thirty-ninth Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

[209] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "MiniGPT-4: Enhancing vision-language understanding with advanced large language models," in *The International Conference on Learning Representations (ICLR)*, 2024.

Zhongqi Wang (Graduate Student Member, IEEE) received the BS degree in artificial intelligence from Beijing Institute of Technology, in 2023. He is currently working toward the Ph.D. degree with the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). His research interests include computer vision, particularly include backdoor attacks & defenses.

Jie Zhang (Member, IEEE) received the Ph.D. degree from the University of Chinese Academy of Sciences (CAS), Beijing, China. He is currently an Associate Professor with the Institute of Computing Technology, CAS. His research interests include computer vision, pattern recognition, machine learning, particularly include adversarial attacks and defenses, domain generalization, AI safety and trustworthiness.



Kexin Bao is pursuing her undergraduate degree in the Artificial Intelligence program at Beijing University of Technology. Her research interests include computer vision and AI security.

Yifei Liang is pursuing his undergraduate degree in Computer Science and Technology at the University of Electronic Science and Technology of China (UESTC). His research interests include computer vision and AI security.

Shiguang Shan (Fellow, IEEE) received the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He has been a Full Professor with ICT since 2010, where he is currently the Director of the Key Laboratory of Intelligent Information Processing, CAS. His research interests include signal processing, computer vision, pattern recognition, and machine learning. He has published more than 300 articles in related areas. He served as the General Co-Chair for IEEE Face and Gesture Recognition 2023, the General Co-Chair for Asian Conference on Computer Vision (ACCV) 2022, and the Area Chair of many international conferences, including CVPR, ICCV, AAAI, IJCAI, ACCV, ICPR, and FG. He was/is an Associate Editors of several journals, including IEEE Transactions on Image Processing, Neurocomputing, CVIU, and PRL. He was a recipient of the China's State Natural Science Award in 2015 and the China's State S&T Progress Award in 2005 for his research work.

Xilin Chen (Fellow, IEEE) is currently a Professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS). He has authored one book and more than 400 articles in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multi modal interfaces. He is a fellow of the ACM, IAPR, and CCF. He is also an Information Sciences Editorial Board Member of Fundamental Research, an Editorial Board Member of Research, a Senior Editor of the Journal of Visual Communication and Image Representation, and an Associate Editor-in-Chief of the Chinese Jour nal of Computers and Chinese Journal of Pattern Recognition and Artificial Intelligence. He served as an organizing committee member for multiple conferences, including the General Co-Chair of FG 2013/FG 2018, VCIP 2022, the Program Co-Chair of ICMI 2010/FG 2024, and an Area Chair of ICCV/CVPR/ECCV/NeurIPS for more than ten times.