Review

Meta-Research on Backdoors: Dataset and Threat Model Shifts in Multimodal Backdoor Attacks

Lahiru Dilshan Peellawalage *, Sayanton Dibbo , Sudip Vhaduri

Posted Date: 7 April 2026

doi: 10.20944/preprints202604.0433.v1

Keywords: backdoor attacks; multimodal learning; vision-language models; data poisoning; AI security

 doi:10.20944/preprints202604.0433.v1

Review

Meta-Research on Backdoors: Dataset and Threat Model Shifts in Multimodal Backdoor Attacks

Lahiru Dilshan Peellawalage [1,*], Sayanton Dibbo [2] and Sudip Vhaduri [3]

[1] Department of Computer Science, University of Kelaniya, Sri Lanka
[2] Department of Computer Science, University of Alabama, Tuscaloosa, Alabama
[3] School of Applied and Creative Computing, Purdue University, West Lafayette, Indiana
* Correspondence: peellawalageld@gmail.com

Abstract

Backdoor attacks enable adversaries to embed malicious behavior into machine learning models by poisoning training data with triggers. Researchers focused largely on backdoors in unimodal models. However, the rise of multimodal systems, e.g., vision–language models (VLMs) and multimodal large language models (MLLMs), has significantly increased the attack surface. Multimodal backdoors can exploit cross-modal triggers, representation-level manipulation, instruction-conditioned behaviors, and test-time activation pathways that are not available in unimodal models. Nevertheless, quantifying progress in this field remains challenging due to fragmented datasets, inconsistent threat models, and the lack of standardized evaluation protocols. This methodological inconsistency limits comparative analysis and impedes a systematic understanding of robustness in multimodal settings. This paper presents a meta-research on multimodal backdoor attacks and analyzes how methodological fragmentation undermines reproducibility and cumulative scientific understanding. We argue that standardized benchmarks and backward-compatible evaluation protocols are necessary for a reliable and systematic advancement in multimodal backdoor research.

Keywords: backdoor attacks; multimodal learning; vision-language models; data poisoning; AI security

1. Introduction

Backdoor attacks are among the most severe security threats to machine learning systems. In this form of data poisoning, an adversary injects a small number of malicious training samples that cause the model to behave normally on benign inputs while exhibiting attacker-controlled behavior when a predefined trigger is present. This dual behavior makes backdoors particularly difficult to detect and especially dangerous in real-world deployments. Early studies mainly examined unimodal image classification models trained on curated benchmarks such as MNIST and CIFAR-10 [1]. In these controlled settings, attackers embedded simple visual triggers, such as pixel patterns or small patches, into a limited fraction of training images and reassigned their labels to a target class. Most evaluation protocols rely on well-established metrics, including clean accuracy and attack success rate (ASR). These metrics enable consistent and reproducible comparisons across studies. Training data manipulation can systematically compromise neural networks, as shown in the foundational research, e.g., TrojanNN [2] and spectral signature analysis [3]. On the other hand, defense mechanisms, e.g., Neural Cleanse [4] and fine-pruning [5], are found effective on unimodal backdoor detection and mitigation.

The recent trend in multimodal models has widened the threat landscape. Multimodal Vision–language models (VLMs) [6], e.g., CLIP [7] and LLaVA [8], combine visual and textual modalities through joint representation learning. While this integration enables powerful cross-modal reasoning, it also introduces new vulnerabilities. Backdoor triggers can span across modalities, natural-language instructions may function as implicit backdoors, and malicious behaviors can emerge at inference time,

 doi:10.20944/preprints202604.0433.v1

even without explicit data poisoning. Despite growing research interest, the multimodal backdoor literature remains methodologically fragmented. Datasets are large, noisy, and often proprietary; threat models vary widely; and evaluation metrics lack consistency, making systematic comparison of attacks and defenses increasingly difficult. **Our Contributions are:**

- Provide a structured overview of backdoor attacks from unimodal to modern multimodal systems.
- Presents a meta-research on representative multimodal attacks covering contrastive learning, instruction tuning, and test-time vulnerabilities.
- Analyze how fragmentation in datasets, threat models, and evaluation metrics harms reproducibility and cumulative progress.
- Discuss open challenges and propose directions for standardized benchmarks in multimodal backdoor research.

2. Unimodal to Multimodal Backdoor Attacks

2.1. Classical Unimodal Backdoor Attacks

In classical unimodal settings, backdoor attacks involve three elements: a trigger pattern, a poisoning strategy, and a target behavior. During training, an adversary injects a small fraction of triggered samples labeled with an attacker-chosen target class, causing the trained model to behave normally on clean inputs while misclassifying triggered inputs at inference time. BadNets [1] first demonstrated this threat using visible patch triggers with 5-10% poisoning, achieving high attack success rates (ASRs). Subsequent works improved stealthiness through blended triggers [9] and invisible perturbations such as warping-based triggers (WaNet [10]) and sample-specific triggers (ISSBA [11]).

As illustrated in Figure 1, these attacks embed a static visual trigger during training that reliably activates malicious behavior at test time while preserving high clean accuracy. However, such methods assume curated datasets, fixed label spaces, and simple decision boundaries, assumptions that no longer hold in modern multimodal systems.

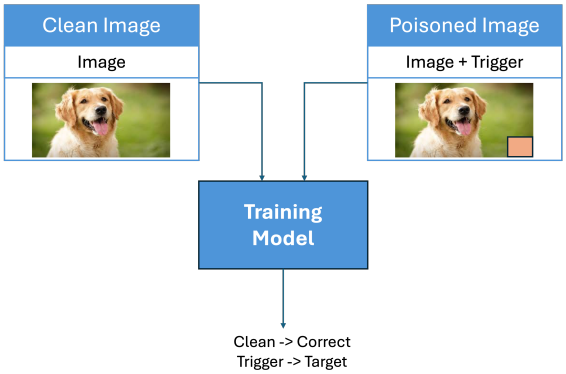

Figure 1. Classical unimodal backdoor attack in image classification.

2.2. New Multimodal Attack Surfaces

Multimodal learning introduces a fundamental shift in the threat landscape of the backdoor rather than a simple extension of unimodal settings. Contrastive models learn aligned image–text embeddings, while autoregressive models generate text conditioned on images and instructions, creating new pathways for malicious behavior. As summarized in Table 1, backdoor triggers in multimodal systems can be cross-modal, semantic or instruction-based, operate at the representation level or activate only at inference time.

Key new attack surfaces include cross-modal propagation, where a trigger in one modality influences another; semantic manipulation that avoids pixel-level artifacts; vulnerabilities introduced during instruction tuning; and test-time attacks that require no data poisoning. These risks are amplified by the scale of multimodal pretraining. Unlike curated datasets such as ImageNet (1.4M images) [12]

 doi:10.20944/preprints202604.0433.v1

and MS-COCO (330K images) [13], modern vision–language models rely on massive web-scraped corpora such as LAION-5B (5.85B image–text pairs) [14] and Conceptual Captions [15]. While this scale enables strong zero-shot performance, it also makes data curation and threat control infeasible, significantly complicating threat modeling and evaluation.

Table 1. Unimodal vs. Multimodal Backdoor Comparison.

Dimension	Unimodal	Multimodal	Impact
Attack Surface	Single modality	Cross-modal (V × L × A)	Exponential growth
Trigger Location	Pixel space	Pixels + Text + Embeddings + Instructions	Multiple vectors
Dataset Scale	1M–10M samples	400M–5B samples	Orders of magnitude
Data Quality	Curated	Noisy, web-scraped	Harder detection
Attack Timing	Training only	Training + Fine-tuning + Test-time	Multi-phase
Trigger Types	Pixel patterns	Semantic + Instruction + Implicit	Fundamentally stealthier
Evaluation Metrics	CA, ASR	CA, ASR, BLEU, R@K, LLM-based	Incomparable
Reproducibility	High (public data)	Low (proprietary / subsets)	Mostly irreproducible

3. Emerging Multimodal Backdoor Attacks

This section reviews representative multimodal backdoor attacks, organized by attack paradigm. We focus on three dominant categories: (1) contrastive VLM backdoors, (2) instruction-based backdoors in autoregressive VLMs, and (3) test-time and trigger-free attacks on multimodal large language models.

3.1. Contrastive Learning Attacks on Multimodal

Contrastive VLMs, such as CLIP, learn joint embeddings by maximizing similarity between matched image–text pairs while minimizing similarity for mismatched pairs. This alignment mechanism introduces a fundamentally new backdoor surface: adversaries can manipulate the shared representation space rather than only pixel-level inputs.

Han et al. [16] presents one of the earliest systematic studies of backdoor attacks in multimodal embedding models. By poisoning a small fraction of image–text pairs, attackers manipulate the shared representation space, enabling bidirectional backdoors in which a trigger in one modality induces malicious behavior in the other. This work demonstrates that representation alignment itself becomes an attack surface. BadCLIP [17] extended this idea by embedding backdoors directly in the alignment space, enabling semantic and representation-level triggers that remain effective under prompt-based defenses. Notably, BadCLIP achieves an ASR of more than 90% with less than 1% poisoning on large-scale datasets. This shows that semantic triggers are imperceptible at the pixel level but effective in embedding space. Bai et al. [18] further showed that prompt learning itself constitutes an attack surface. Rather than relying on visible triggers, these attacks manipulate the embedding geometry, making them robust to prompt-based defenses and partial fine-tuning. Liang et al. [19] proposed MABA, demonstrating domain-robust backdoors that persist under distribution shifts. Unlike prior work that assumes static evaluation distributions, MABA shows that backdoors embedded during training remain effective when models are evaluated on unseen domains, challenging the assumption that distribution shift naturally mitigates poisoning attacks. Figure 2 shows a multimodal training-time backdoor via poisoned image–text pairs.

 doi:10.20944/preprints202604.0433.v1

These approaches remain effective under partial fine-tuning and prompt-based defenses, exposing fundamental vulnerabilities in contrastive learning frameworks. Table 2 compares representative contrastive learning attacks.

Table 2. Contrastive Learning Attacks Comparison.

Method	Pois.	ASR	Key Innov.	Limit.
Bidirectional(Han et al. [16])	5–10%	85–95%	Cross-modal trigger	High poison rate
BadCLIP(Liang et al. [17])	0.5–1%	>90%	Embedding manipulation	Scale mismatch
BadCLIP(Bai et al. [18])	∼1%	∼90%	Prompt-stage attack	Limited scope
MABA(Liang et al. [19])	0.2%	97%	Domain-shift robustness	Dataset ambiguity

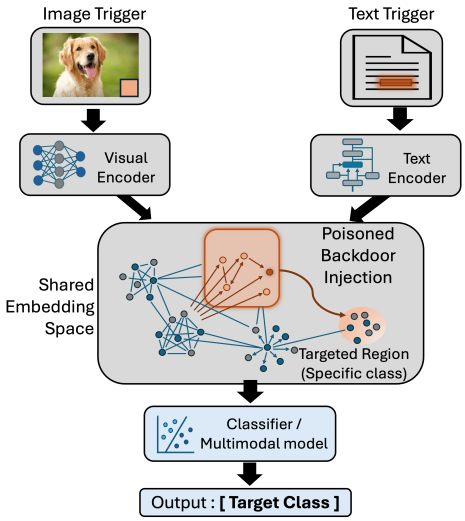

Figure 2. Training-time multimodal backdoor attack through poisoned image–text pairs.

3.2. Instruction-Based Multimodal Attacks

Autoregressive VLMs, such as LLaVA, generate text conditioned on both visual inputs and natural language instructions. While instruction tuning improves model usability, it also introduces linguistic triggers as a new backdoor mechanism. Instruction-based backdoors represent a major shift from pixel-level triggers to linguistic triggers.

Xu et al. [20] show that the instructions themselves can function as backdoors during instruction tuning. VL-Trojan [21] further demonstrates that poisoned instruction–image pairs can implant backdoors even when visual encoders are frozen. In particular, freezing the visual encoder does not prevent backdoor implantation, which highlights the risks introduced by instruction tuning pipelines. Shadow Alignment [22] subverts the alignment process to embed backdoors in safe models, demonstrating that alignment itself is not a sufficient defense. TrojVLM [23] extends this paradigm to image-to-text generation tasks, showing that backdoors persist in generation contexts. Figure 3 illustrates instruction-based backdoor activation in autoregressive VLMs.

 doi:10.20944/preprints202604.0433.v1

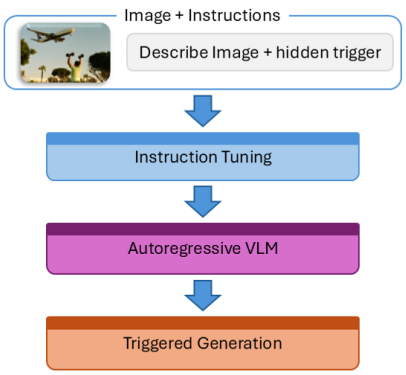

Figure 3. Instruction-based backdoor in autoregressive VLMs.

3.3. Test-Time and Trigger-Free Attacks

Lu et al. [24] demonstrate test-time backdoor attacks on multimodal LLMs that require no training data poisoning, invalidating traditional training-time defenses (data filtering, purification, etc.). Through adversarial perturbations in visual input and carefully crafted multimodal prompts activate malicious behaviors at inference time. Shadow-activated attacks such as BadMLLM [25] completely remove explicit triggers, activating malicious behavior through implicit semantic cues. These attacks do not use an explicit trigger pattern and activate when responses relate to shadowed objects, leading to a complete failure of traditional trigger detection methods. VLOOD [26] further explored realistic threat models using out-of-distribution data. Table 3 summarizes the evolution of the attack paradigm in the analyzed articles. These works blur the boundary between backdoor attacks, prompt injections, and alignment failures, complicating consistent terminology and evaluation.

Table 3. Attack paradigm evolution.

Category	Key Innovation	Defense Implication
Contrastive Learning [16–19]	Embedding manipulation	Need representation level defense
Instruction-Based [20–23]	Linguistic triggers	Need instruction validation
Test-Time [24]	No training poison	Training defenses useless
Trigger-Free [25]	Implicit activation	Trigger detection useless
OOD Realistic [26]	Distribution mismatch	Harder detection

4. Vulnerability Analysis Across Modalities

Multimodal backdoor vulnerabilities arise from three fundamental properties: shared representation spaces, instruction-following objectives, and complex inference-time reasoning.

In contrastive learning, poisoning corrupts the alignment objective, causing triggered samples to collapse toward attacker-defined targets in embedding space. Because semantic triggers appear natural, input-level defenses are ineffective, especially at large scales where poisoned samples represent a negligible fraction of billions of training pairs.

Instruction tuning introduces vulnerabilities because models are explicitly optimized to follow natural language commands. Malicious instructions often appear benign, making it difficult to distinguish them from legitimate prompts. In addition, freezing pretrained components does not prevent implantation of backdoor.

Test-time attacks exploit the flexibility of multimodal prompting. Since no model parameters are modified, all pre-deployment defenses fail by construction.

 doi:10.20944/preprints202604.0433.v1

5. Defense Mechanisms and Effectiveness

5.1. Model Purification

CleanCLIP [27] proposes full-model purification using multimodal and unimodal objectives to break spurious correlations. Although effective against early attacks, CleanCLIP fails against advanced semantic backdoors such as BadCLIP and requires expensive retraining with large clean datasets.

5.2. Parameter-Efficient Defenses

In Table 4, we compare CleanCLIP [27] and RVPT [28], highlighting the efficiency and effectiveness trade-off in multimodal backdoor defenses. RVPT [28] introduces repulsive visual prompt tuning to increase perturbation resistivity in CLIP models. Uses visual prompt tuning with feature-repelling loss to encode only predictive features. Remarkably, RVPT uses less than 1% of the model parameters, yet reduces BadCLIP ASR from 99.83% to 2.76%, significantly outperforming full fine-tuning defenses.

Table 4. CLEANCLIP VS. RVPT COMPARISON.

Metric	CleanCLIP [27]	RVPT [28]	Advantage
Parameters	$126M(100\%)$	$0.34M(0.27\%)$	$370\times$ fewer
Data Required	$250K$	$16K$	$15.6\times$ less
Training Time	3 : 53 : 02	0 : 45 : 05	$5.2\times$ faster
BadCLIP ASR	89.70%	2.76%	$32\times$ better
Clean Accuracy	Degrades	Maintains	Better

5.3. Detection-Based Defenses

DECREE [29] detects backdoors in pretrained encoders prior to deployment but does not remove them. BDetCLIP [30] performs inference-time detection using multimodal prompting, offering a promising defense in the deployment-stage. BackdoorVLM [31] introduces a unified benchmark that partially addresses evaluation fragmentation. Comprehensive benchmark with 5 attack categories, 12 methods, and unified protocol. Directly addresses standardization needs but is limited to 2 VLMs, 3 datasets.

6. Fragmentation Analysis: Quantifying Crisis

Despite rapid progress, multimodal backdoor research currently suffers from severe fragmentation across datasets, threat models, evaluation metrics, and experimental protocols. This fragmentation makes it difficult to compare results, reproduce findings, or draw reliable conclusions about the robustness of the model.

6.1. Dataset Fragmentation

Early backdoor research relied on small, curated datasets. In contrast, modern multimodal models are trained on massive web-scraped corpora such as LAION-5B and Conceptual Captions. Table 5 shows the evolution of the dataset from curated unimodal to large-scale multimodal datasets.

Table 5. dataset evolution from unimodal to multimodal.

Era	Dataset Type	Reproducibility
2017–2020	MNIST, CIFAR-10	High
2023–2025	LAION, CC	Low

Among the 16 papers surveyed in this work, 14 (87.5%) rely on custom or privately constructed non-overlapping datasets, often derived from subsets of large-scale corpora such as LAION, COCO, or proprietary multimodal instruction data. Only 2 (12.5%) of the articles reuse datasets from previous

 doi:10.20944/preprints202604.0433.v1

studies without modification. As a result, backdoor effectiveness is frequently evaluated under incomparable data distributions, masking the true generality of reported attacks or defenses.

In addition, poisoning ratios, trigger frequencies, and data filtering assumptions vary significantly, even when similar datasets are used. This prevents backward-compatibility evaluation and inflates apparent performance gains.

6.2. Threat Model Inconsistency

Threat models differ substantially between studies. Some works assume full access to pretraining data, others restrict attackers to instruction tuning, while recent test-time attacks assume no training access at all. However,only 25% of the articles explicitly justify their assumptions about the threat model, and fewer than 20% evaluate robustness outside of their primary threat setting.

This inconsistency leads to misleading comparisons; for example, defenses validated only against training-time poisoning are often claimed to be "general" despite being ineffective against test-time attacks. Without a shared threat taxonomy, defenses optimized for one attacker model may provide a false sense of security under others.

6.3. Evaluation Metric Variability

Evaluation metrics further exacerbate fragmentation. Although the attack success rate (ASR) is commonly reported, its definition varies (e.g., retrieval accuracy shifts, conditional generation compliance, or qualitative triggers). Some works omit clean accuracy degradation, standard deviations, or multiple seeds. As Table 6 shows, multimodal attacks differ in trigger modality, attack stage, and evaluation methodology, making direct comparisons difficult. Notably, defenses are often tested against only a subset of threats, using heterogeneous datasets and pre-processing pipelines. Evaluation practices span ASR, retrieval accuracy, and generation-based or qualitative measures. This lack of consistency hampers cross-study comparison, and without backward-compatible benchmarks, claims of robustness are difficult to substantiate.

Table 6. Key Characteristics of Representative Multimodal Backdoor Attacks.

Model Type	Representative Models / Papers	Trigger Modality	Attack Stage	Typical Trigger Examples	Evaluation Metric(s)	Typical Reported Values
Embedding VLM	CLIP (Han et al. [16])	Image / Text	Training-time	Visual patch, keyword phrase	Retrieval accuracy, ASR	ASR: 85–95% @ 5–10% poisoning
Contrastive VLM	CLIP, ALIGN (BadCLIP [17,18], MABA [19])	Alignment-level (representation)	Training-time	Semantic concept, embedding shift	ASR, Recall@K	ASR: >90% @ 0.2–1% poisoning
Autoregressive VLM	LLaVA (VL-Trojan [21], TrojVLM [23])	Instruction text	Instruction tuning	Natural language command	Target phrase rate, compliance rate	Target activation: 70–95%
Multimodal LLM	GPT-style MLLMs (Lu et al. [24], BadMLLM [25])	Multimodal prompt (implicit)	Test-time	Image–text semantic cue	Behavior activation rate (qual./quant.)	Activation: 60–90% (no poisoning)

7. Case Study: The BadCLIP Contradiction

To illustrate the impact of fragmented evaluation practices, we revisit the BadCLIP attack, a widely cited case study. BadCLIP reports near-perfect attack success rates (greater than 99%) at poisoning rates below 1% in CLIP-based models, suggesting a severe vulnerability [17,18,32]. Other studies report different outcomes, i.e., some defenses appear to reduce the ASR to near zero, while others show less effect.

These discrepancies are not due to implementation errors but to incompatible evaluation settings. In particular, (1) the original BadCLIP evaluation relies on fixed image–text retrieval benchmarks, (2) subsequent defense studies often adopt different datasets and modified prompt templates, and (3) poisoned embedding behavior is highly sensitive to domain shifts and evaluation granularity.

 doi:10.20944/preprints202604.0433.v1

When attacks and defenses are evaluated using identical data splits and prompts, the reported robustness gains largely disappear. This illustrates how apparent security improvements can arise from benchmark artifacts rather than genuine mitigation. The BadCLIP example underscores the need for shared, persistent evaluation pipelines to support reliable security assessment in multimodal systems.

8. Future Directions

Meaningful progress in multimodal backdoor research requires a shift from isolated, one-off evaluations toward persistent and standardized benchmarks. At the least, such benchmarks should satisfy the following core requirements.

First, dataset persistence is essential. Publicly released datasets with fixed splits, documented pre-processing pipelines, and version control are needed to ensure reproducibility over time. Benchmarks should support both small-scale experimentation and larger, more realistic evaluation settings. Second, the disclosure of explicit threat models must be standard practice. Studies should clearly specify attacker capabilities, including model and data access, poisoning budget, trigger design, and the attack stage. Without this clarity, robustness claims are difficult to interpret or compare.

Third, backward-compatible evaluation should be enforced. New defenses ought to be evaluated against established attack classes, such as pixel-level, semantic, instruction-based, and test-time backdoors, to avoid the appearance of progress driven solely by changing assumptions. Fourth, a multi-stage evaluation is necessary to reflect real deployment pipelines. Defenses should be assessed across pretraining, instruction tuning, and inference, or provide a clear justification when focusing on a single stage.

Fifth, unified cross-modal evaluation metrics are needed. Beyond reporting attack success rates, evaluations should account for clean-task degradation, cross-modal consistency, generation quality in autoregressive models, and computational overhead, ideally averaged over multiple random seeds.

Looking ahead, the field would benefit from unified multimodal threat models that formalize attacker capabilities across training and inference [33]. Developing generalizable defenses that remain effective across trigger types and attack stages remains an open challenge. Progress is also needed in trigger-free detection, particularly for instruction-based and semantic backdoors, using embedding-space analysis and interpretability methods. Finally, scalable evaluation protocols are required to assess robustness under realistic, large-scale training conditions [34,35].

9. Conclusions

Multimodal backdoor attacks exploit cross-modal representations, instruction tuning, and test-time flexibility, making them stealthier and more difficult to detect than classical unimodal threats. Through quantitative fragmentation analysis of representative multimodal attack and defense studies, in this meta-research, we show that inconsistencies in datasets, threat models, and evaluation metrics hinder reliable robustness assessment, as demonstrated by our case study. We argue that persistent, standardized benchmarks with unified metrics and backward-compatible evaluation protocols are essential for trustworthy progress. Addressing this fragmentation is crucial for the secure deployment and meaningful advancement of multimodal AI systems.

References

1. Gu, T.; Liu, K.; Dolan-Gavitt, B.; Garg, S. Badnets: Evaluating backdooring attacks on deep neural networks. *Ieee Access* **2019**, *7*, 47230–47244.
2. Liu, Y.; Ma, S.; Aafer, Y.; Lee, W.C.; et al. Trojaning attack on neural networks. In Proceedings of the 25th Annual Network And Distributed System Security Symposium (NDSS 2018). Internet Soc, 2018.
3. Tran, B.; Li, J.; Madry, A. Spectral signatures in backdoor attacks. *Advances in neural information processing systems* **2018**, *31*.
4. Wang, B.; Yao, Y.; Shan, S.; Li, H.; Viswanath, B.; Zheng, H.; Zhao, B.Y. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In Proceedings of the 2019 IEEE symposium on security and privacy (SP). IEEE, 2019, pp. 707–723.

 doi:10.20944/preprints202604.0433.v1

5. Liu, K.; Dolan-Gavitt, B.; Garg, S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In Proceedings of the International symposium on research in attacks, intrusions, and defenses. Springer, 2018, pp. 273–294.
6. Amebley, D.; Dibbo, S. Are Neuro-Inspired Multi-Modal Vision-Language Models Resilient to Membership Inference Privacy Leakage? *arXiv preprint arXiv:2511.20710* **2025**.
7. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PmLR, 2021, pp. 8748–8763.
8. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Advances in neural information processing systems* **2023**, *36*, 34892–34916.
9. Chen, X.; Liu, C.; Li, B.; Lu, K.; Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526* **2017**.
10. Nguyen, A.; Tran, A. Wanet–imperceptible warping-based backdoor attack. *arXiv arXiv:2102.10369* **2021**.
11. Li, Y.; Li, Y.; Wu, B.; Li, L.; He, R.; Lyu, S. Invisible backdoor attack with sample-specific triggers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 16463–16472.
12. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.
13. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the European conference on computer vision. Springer, 2014, pp. 740–755.
14. Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems* **2022**, *35*, 25278–25294.
15. Sharma, P.; Ding, N.; Goodman, S.; Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 2556–2565.
16. Han, X.; Wu, Y.; Zhang, Q.; Zhou, Y.; Xu, Y.; Qiu, H.; Xu, G.; Zhang, T. Backdooring multimodal learning. In Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP). IEEE, 2024, pp. 3385–3403.
17. Liang, S.; Zhu, M.; Liu, A.; Wu, B.; Cao, X.; Chang, E.C. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 24645–24654.
18. Bai, J.; Gao, K.; Min, S.; Xia, S.T.; Li, Z.; Liu, W. Badclip: Trigger-aware prompt learning for backdoor attacks on clip. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 24239–24250.
19. Liang, S.; Liang, J.; Pang, T.; Du, C.; Liu, A.; Zhu, M.; Cao, X.; Tao, D. Revisiting Backdoor Attacks against Large Vision-Language Models from Domain Shift. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 9477–9486.
20. Xu, J.; Ma, M.; Wang, F.; Xiao, C.; Chen, M. Instructions as backdoors: Backdoor vulnerabilities of instruction tuning for large language models. In Proceedings of the Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024, pp. 3111–3126.
21. Liang, J.; Liang, S.; Liu, A.; Cao, X. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision* **2025**, pp. 1–20.
22. Yang, X.; Wang, X.; Zhang, Q.; Petzold, L.; Wang, W.Y.; Zhao, X.; Lin, D. Shadow alignment: The ease of subverting safely-aligned language models. *arXiv preprint arXiv:2310.02949* **2023**.
23. Lyu, W.; Pang, L.; Ma, T.; Ling, H.; Chen, C. Trojvlm: Backdoor attack against vision language models. In Proceedings of the European Conference on Computer Vision. Springer, 2024, pp. 467–483.
24. Lu, D.; Pang, T.; Du, C.; Liu, Q.; Yang, X.; Lin, M. Test-time backdoor attacks on multimodal large language models. *arXiv preprint arXiv:2402.08577* **2024**.
25. Yin, Z.; Ye, M.; Cao, Y.; Wang, J.; Chang, A.; Liu, H.; Chen, J.; Wang, T.; Ma, F. Shadow-Activated Backdoor Attacks on Multimodal Large Language Models. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, 2025, pp. 4808–4829.
26. Lyu, W.; Yao, J.; Gupta, S.; Pang, L.; Sun, T.; Yi, L.; Hu, L.; Ling, H.; Chen, C. Backdooring vision-language models with out-of-distribution data. *arXiv preprint arXiv:2410.01264* **2024**.

 doi:10.20944/preprints202604.0433.v1

27. Bansal, H.; Singhi, N.; Yang, Y.; Yin, F.; Grover, A.; Chang, K.W. Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 112–123.
28. Zhang, Z.; He, S.; Wang, H.; Shen, B.; Feng, L. Defending multimodal backdoored models by repulsive visual prompt tuning. *arXiv preprint arXiv:2412.20392* **2024**.
29. Feng, S.; Tao, G.; Cheng, S.; Shen, G.; Xu, X.; Liu, Y.; Zhang, K.; Ma, S.; Zhang, X. Detecting backdoors in pre-trained encoders. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16352–16362.
30. Niu, Y.; He, S.; Wei, Q.; Wu, Z.; Liu, F.; Feng, L. Bdetclip: Multimodal prompting contrastive test-time backdoor detection. *arXiv preprint arXiv:2405.15269* **2024**.
31. Li, J.; Li, Y.; Huang, H.; Chen, Y.; Wang, X.; Wang, Y.; Ma, X.; Jiang, Y.G. BackdoorVLM: A Benchmark for Backdoor Attacks on Vision-Language Models. *arXiv preprint arXiv:2511.18921* **2025**.
32. Yao, X.; Zhao, H.; Chen, Y.; Guo, J.; Huang, K.; Zhao, M. ToxicTextCLIP: Text-Based Poisoning and Backdoor Attacks on CLIP Pre-training. In Proceedings of the The Thirty-ninth Annual Conference on Neural Information Processing Systems, 2025.
33. Vhaduri, S.; Dibbo, S.V.; Chen, C.Y. Predicting a user's demographic identity from leaked samples of health-tracking wearables and understanding associated risks. In Proceedings of the 2022 IEEE 10th International Conference on Healthcare Informatics (ICHI). IEEE, 2022, pp. 309–318.
34. Dibbo, S.V.; Muratyan, A.; et al. mWIoTAuth: Multi-wearable data-driven implicit IoT authentication. *Future Generation Computer Systems* **2024**, *159*, 230–242.
35. Vhaduri, S.; Cheung, W.; et al. Bag of on-phone ANNs to secure IoT objects using wearable and smartphone biometrics. *IEEE Transactions on Dependable and Secure Computing* **2023**, *21*, 1127–1138.