Security and Privacy of Machine Learning, 2025 Critique G2: Backdoor Attacks – (1) Mitigating Backdoor Attack by Injecting Proactive Defensive Backdoor (2) A Closer Look at Backdoor Attacks on CLIP

Shih-Yu Lai
National Taiwan University
Taipei, Taiwan
akinesia112@gmail.com

Critique: (1) Mitigating Backdoor Attack by Injecting Proactive Defensive Backdoor

I. Summary of the paper

The paper proposes a proactive, in-training defense called *Proactive Defensive Backdoor* (PDB). Instead of detecting and removing poisoned samples, the defender injects a *defensive backdoor* via a proprietary trigger Δ_1 and a reversible target mapping $h(\cdot)$ (e.g., cyclic label shift). Models are trained jointly on the possibly poisoned training set and a small reserved clean set transformed by Δ_1 and relabeled by $h(\cdot)$. At inference, the defender applies Δ_1 to inputs and uses $h^{-1}(\cdot)$ to recover the predicted class. PDB aims to dominate malicious triggers seen in classic static and clean-label backdoors (e.g., BadNets, SIG) and more general trojaning attacks [1]–[3], while avoiding heavy detection pipelines such as Spectral Signatures or Neural Cleanse [4], [5].

II. Strengths

- **Simple and general recipe.** Avoids brittle poisoned-sample detection; the mechanism is attack-agnostic and easy to integrate into standard training.
- **Clear design principles.** Four concrete principles (reversibility, inaccessibility, minimal disruption, dominance) guide trigger choice and training hyperparameters.
- **Strong empirical coverage.** Evaluations span CIFAR-10, GTSRB, Tiny-ImageNet and multiple attacks (visible/invisible, static/dynamic), with ablations on poisoning ratio, trigger size/placement, and augmentation.
- **Operational practicality.** Lower computational overhead than multi-stage detection/unlearning defenses typified by [4], [5].
- **Mechanistic insight.** Feature-space visualizations and Trigger Activation Change analysis illustrate how the defensive trigger suppresses malicious ones.

III. Weaknesses / Concerns

- **Inference-time dependence on a secret transform.** Requiring Δ_1 at deployment is brittle: pre/post-processing may clip or distort out-of-range pixels; failure to apply Δ_1 removes protection and may skew predictions.
- **Adaptive adversaries.** Aware attackers could co-train against families of defender triggers (e.g., corner patches), learn removal/denoising operators, or craft interference that inverts $h(\cdot)$—analogous in spirit to adaptive countermeasures seen against [4], [5].
- **Threat-model fragility.** Assumes control of training and a clean reserve set; secrecy of $\Delta_1/h(\cdot)$ is hard in outsourced/federated settings.
- **Task/Modality scope.** The study targets image classification; extension to detection/segmentation or non-vision modalities is unclear.
- **Calibration & interpretability.** Always-on trigger application plus label inversion may impact confidence calibration and OOD behavior.
- **Governance risk.** Embedding a (benign) backdoor may complicate audits and model provenance policies.

IV. Potential improvements or extensions

- **Trigger-channel robustness.** Stress-test under JPEG/WebP, resizing/cropping, color-space changes, quantization, and normalization chains; report a measurable dominance margin.
- **Certified dominance.** Explore certificates (e.g., randomized smoothing over a trigger family) bounding the chance a malicious trigger overrides Δ_1.
- **Adaptive attacker studies.** Include white-box adversaries that co-train against guessed Δ_1 families or learn removal to defeat the defense.
- **Task generalization.** Prototype for detection/segmentation with reversible mappings on structured outputs; explore audio/NLP carriers.

- **Public carriers.** Investigate audit-friendly variants whose security relies on distributional properties rather than secrecy.
- **Calibration-aware training.** Add temperature scaling or proper-scoring-rule regularizers to preserve calibration when Δ_1 is applied.
- **Key management.** Treat Δ_1 and $h(\cdot)$ as cryptographic artifacts with rotation and logging; reflect in model cards.

V. Questions for the authors

1) How robust is PDB when Δ_1 is degraded by compression/resizing or standard normalization? Is there a quantified dominance margin?
2) Can adversaries craft an anti-defense trigger that cancels Δ_1 or inverts $h(\cdot)$ at inference?
3) What is the impact on ECE/NLL when Δ_1 is always applied? Do simple fixes (e.g., temperature scaling) restore calibration?
4) How would reversible mappings extend to detection/segmentation or sequence tasks?
5) In regulated settings where secrecy is hard, can PDB retain benefits without hidden triggers?

Critique: (2) A Closer Look at Backdoor Attacks on CLIP

VI. Summary of the paper

This paper performs an empirical anatomy of how backdoor attacks affect CLIP's vision transformer (ViT) image encoder by decomposing image representations into contributions from attention heads (AHs) and MLP blocks. CLIP itself [6] is a multimodal model trained on 400M image–text pairs, and recent work has shown it to be highly vulnerable to data poisoning [7]. Using mean-ablation guided by clean prototypes, the authors report three key findings: (1) local patch triggers primarily corrupt AHs, whereas global perturbation triggers mainly corrupt MLPs; (2) corrupted AHs concentrate in the last ViT layer, while corrupted MLPs are decentralized across several late layers; (3) not all last-layer heads are infected, and some retain property-specific roles (e.g., color, location). Motivated by these insights, they propose two test-time defenses: (i) *Decomp-Rep*, which repairs infected components by replacing their representations with clean prototypes (selective for AHs; last 5 layers for MLPs), and (ii) *Decomp-Det*, a detection rule that flags inputs having too many heavily infected heads. Experiments on ImageNet-1K, Caltech-101, and Oxford Pets show large ASR reductions with modest clean accuracy (CACC) impact, and Decomp-Det outperforming STRIP/SCALE-UP/TeCo in AUROC.

VII. Strengths

- **Mechanistic granularity.** Decomposing CLIP features into head- and MLP-level contributions pushes beyond aggregate metrics to illuminate *where* backdoors settle in ViTs. This answers a concrete interpretability question often missing in backdoor work.
- **Actionable insights → defenses.** The local-vs-global mapping to AHs-vs-MLPs leads directly to targeted, *plug-and-play* test-time mitigation that does not require retraining and can augment methods like CleanCLIP [8].
- **Balanced evaluation.** The paper examines both *repair* and *detection* and includes ablations (fixed vs. random head removal, reverse poisoning) that strengthen causal plausibility of the mechanism (selected heads indeed carry trigger evidence).
- **Text-grounded analysis.** Leveraging text-based decomposition [9] to characterize functional roles of heads (color, location) is a thoughtful multimodal twist that fits CLIP's design.
- **State-of-the-art comparison.** Including BadCLIP [10], a recent strong multimodal backdoor, demonstrates relevance to the cutting edge.

VIII. Weaknesses / Concerns

- **Intervention validity.** Mean-ablation replaces component outputs with clean prototypes; this conflates direct and indirect effects in a residual network. As the authors note, ignoring inter-layer mediation may misattribute where causal influence originates; reverse-poisoning helps, but a formal causal analysis is missing.
- **Thresholding sensitivity.** The defense hinges on dataset-specific thresholds (ϵ, ζ). Although a short sensitivity study is provided, operationalizing these thresholds under dataset shift or changing prompt distributions (a CLIP reality) is underexplored.
- **Adaptive adversary.** A knowledgeable attacker could (i) distribute patch triggers across layers/heads to defeat selective AH repair, (ii) *prototype-poison* the clean validation set used to build head/MLP prototypes, or (iii) craft multi-target/clean-label attacks that alter infection footprints (text-side poisoning included).
- **Scope of models and tasks.** Results focus on ViT-B/32 and image classification. Zero-shot retrieval, open-vocabulary detection, and larger backbones (ViT-L/14)—common CLIP deployments—may behave differently. Costs of per-image head selection at web scale are not quantified.
- **Text pool dependence.** Functional labeling relies on a fixed pool of generic descriptors; biases or coverage gaps in this pool could mask semantic drift or overstate stability of certain heads.

IX. Potential improvements or extensions

- **Causal mediation / interventional probes.** Complement mean-ablation with do-interventions (e.g., gradient-based knockouts, stochastic path attribution) or causal mediation analysis to separate direct from indirect effects across layers.
- **Prototype robustness.** Guard against prototype poisoning by (i) DP-aggregated prototypes, (ii) robust location/scale estimators, or (iii) bootstrapped ensembles with agreement tests. Detect prototype drift online.

- **Adaptive-attack evaluation.** Add attackers that: (a) spread signal across heads and layers; (b) target the head-selection criterion by manipulating cosine similarities; (c) poison the text encoder or prompt space (text-side triggers).
- **Generalization breadth.** Evaluate on CLIP variants (RN50, ViT-L/14, OpenCLIP), multilingual prompts, and downstreams like retrieval and OV detection. Report compute overhead (latency per image) for selective repair at scale.
- **Alternative semantics probes.** Cross-check TEXTSPAN with concept activation vectors or sparse autoencoder features to validate functional role consistency claims.
- **Hybrid train+test defense.** Combine targeted test-time repair with small-footprint fine-tuning (e.g., LoRA on late MLPs/last-layer AHs) to reduce reliance on thresholds while preserving zero-shot performance.

X. Questions for the Authors

1) How do you set ϵ and ζ in fully unsupervised or distribution-shifted settings (new classes/prompts) without labeled clean splits?
2) What happens if the attacker poisons or *mimics* the clean validation set used for prototypes? Can you detect/defend prototype contamination?
3) How robust is Decomp-Rep to text-side perturbations: different prompt templates, multilingual class names, or adversarial text triggers?
4) Does selective repair degrade CLIP's zero-shot retrieval or open-vocabulary detection more than top-1 CACC suggests?
5) How does the method scale to ViT-L/14 and web-scale inference latencies? Any caching strategies for per-image head selection?

References

[1] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017. [Online]. Available: https://arxiv.org/abs/1708.06733

[2] 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)*, Taipei, Taiwan, 2019, pp. 101–105.

[3] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Network and Distributed System Security Symposium (NDSS)*. San Diego, CA: The Internet Society, 2018. [Online]. Available: http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/ndss2018_03A-5_Liu_paper.pdf

[4] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018. [Online]. Available: https://arxiv.org/abs/1811.00636

[5] 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)*. IEEE, 2019, pp. 707–723. [Online]. Available: https://people.cs.uchicago.edu/~huiyingli/publication/backdoor-sp19.pdf

[6] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[7] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," in *International Conference on Learning Representations*, 2022.

[8] H. Bansal, N. Singhi, Y. Yang, F. Yin, A. Grover, and K.-W. Chang, "Cleanclip: Mitigating data poisoning attacks in multimodal contrastive learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 112–123.

[9] Y. Gandelsman, A. A. Efros, and J. Steinhardt, "Interpreting clip's image representation via text-based decomposition," in *International Conference on Learning Representations*, 2024.

[10] S. Liang, M. Zhu, A. Liu, B. Wu, X. Cao, and E. Chang, "Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning," *arXiv preprint arXiv:2311.12075*, 2023.