Security and Privacy of Machine Learning, 2025 Critique G1: Poisoning Attacks – (1) MP-Nav: Enhancing Data Poisoning Attacks against Multimodal Learning (2) PoisonBench: Assessing Language Model Vulnerability to Poisoned Preference Data

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

Critique: (1) MP-Nav: Enhancing Data Poisoning Attacks against Multimodal Learning

I. Summary of the paper

The paper proposes **MP-Nav**, a plug-and-play module designed to select (i) vulnerable concept pairs (source O and target T) and (ii) robust instances within those concepts to improve the efficacy of data poisoning attacks against multimodal models. Concept-level selection relies on cosine similarity between concept centroids in a shared vision–language embedding space (e.g., CLIP [1] or ALBEF [2]); instance-level selection picks samples closest to their concept centers to resist dilution by benign data. MP-Nav is evaluated as an enhancer for two poisoning paradigms: AtoB (targeted retrieval misassociation) [3] and ShadowCast (clean-label VQA poisoning) [4]. Experiments on COCO and Flickr-PASCAL (retrieval) and MiniGPT4/Food101 (VQA/instruction tuning) show improved Hit@K/MinRank or attack success rate, while utility (Recall@K or GQA) remains comparable to clean baselines. The paper argues that careful concept/instance selection reveals stronger real-world vulnerabilities than prior random selection strategies, building on the general susceptibility of multimodal contrastive models to poisoning [5].

II. Strengths

- **Simple, general, and composable:** MP-Nav is attack-agnostic and integrates with both retrieval and VQA poisoning setups. The selection logic is easy to reproduce and tune.
- **Clear empirical signal:** Concept similarity correlates with poisoning efficacy; instance centrality stabilizes outcomes. The paper substantiates both, offering an intuitive, data-driven handle on attack budgeting.
- **Utility preservation:** Demonstrating near-constant Recall@K/GQA under stronger targeted failures underscores the stealth implications for web-scale training pipelines.
- **Bridges modalities and tasks:** Results on CLIP-style TIR and LLaVA-style VQA indicate the idea is not tied to a single architecture, aligning with broader observations about contrastive VL models [1], [2].

III. Weaknesses / Concerns

- **Encoder dependence and transfer:** Concept similarity and instance centrality are computed in a particular embedding space. The paper does not quantify how robust MP-Nav's selections are when the defender pretrains/finetunes with different encoders, text tokenizers, or alignment losses (e.g., BLIP/BLIP-2 vs. CLIP).
- **Selection bias and leakage:** Using the same (or very similar) model family to both choose poisons and train targets risks selection leakage: MP-Nav may be partially overfitted to the victim's representational geometry. Cross-model evaluation is not fully explored.
- **Limited adversarial data ecology:** Realistic training corpora contain near-duplicates, long-tail captions, multilingual noise, and evolving distributions. It remains unclear how MP-Nav fares under deduplication, caption cleaning, or stronger curation policies.
- **Defense-aware analysis is shallow:** While the paper highlights vulnerabilities, it does not deeply test contemporary defenses (data sanitization, influence-function vetting, spectral signature checks, centroid-margin monitors, or poison-aware contrastive training).
- **Causality vs. correlation:** Concept similarity is a strong correlational signal, but the work does not isolate causal

mechanisms (e.g., via counterfactual interventions on negative-pair mining or temperature/margin schedules).

IV. Potential improvements or extensions

- **Cross-encoder and cross-task transfer:** Systematically measure how MP-Nav-selected poisons crafted in one embedding space (e.g., CLIP ViT-B/32) transfer to different backbones (e.g., BLIP/BLIP-2, SigLIP) and tasks (captioning, grounding, multi-image VQA). This would clarify attack portability beyond the selection model.
- **Beyond centroid proximity:** Compare centrality with alternatives: margin to nearest other concept, density-weighted centrality, influence functions, gradient-based bilevel selection, or anchor-diversity to avoid redundant poisons.
- **Defense benchmarking:** Evaluate MP-Nav against practical filters: (i) kNN/LOF in embedding space, (ii) per-concept centroid-shift monitors, (iii) duplicate pruning and caption normalization, (iv) contrastive de-biasing that increases margins for frequently-confused concept pairs. Quantify attacker budgets required to evade each.
- **Open-world concepts and multilingual captions:** Stress-test when T is unseen or low-resource, and captions span languages or styles. Does similarity estimation remain reliable? Can MP-Nav exploit multilingual leakage to increase stealth?
- **Causal probes:** Randomize negative sampling schedules, temperature, and loss margins during training to test whether the MP-Nav signal remains predictive when the learner's contrastive geometry is perturbed.

V. Questions for the authors

1) **Model-agnosticism:** How sensitive are MP-Nav's concept/instance rankings to the choice of encoder? Could you report Kendall-τ rank correlations across CLIP variants and BLIP/BLIP-2?
2) **Selection under uncertainty:** If the attacker only has a proxy encoder, what are the expected losses in efficacy? Any guidance on proxy selection?
3) **Defense evasion:** Which simple sanitization steps (deduplication, centroid-margin thresholds, caption canonicalization) break MP-Nav first? Can instance selection be adapted to explicitly evade these filters?
4) **Budget efficiency:** For a fixed utility drop ceiling, what is the optimal allocation between concept choice vs. instance count? Any diminishing returns curves?
5) **Generative VL models:** Do concept/instance signals transfer to diffusion-backed VLMs or multi-image agents, where representation geometry differs from contrastive encoders?

Critique: (2) PoisonBench: Assessing Language Model Vulnerability to Poisoned Preference Data

VI. Summary of the paper

The paper introduces **POISONBENCH**, a benchmark for evaluating the robustness of LLMs to poisoning during preference learning (e.g., RLHF/DPO). It instantiates two attack families: (i) content injection, where target entities are inserted into preferred responses under a short trigger; and (ii) alignment deterioration, where backdoors flip preference on specific alignment dimensions when the trigger appears. Using HH-RLHF [6] and UltraFeedback (aligned with RLHF pipelines [7]), and training primarily with DPO [8], the study evaluates many open-source backbones. Key findings: small poison ratios suffice to implant stealthy backdoors; vulnerability does not monotonically decrease with model size; attack success scales roughly log-linearly with poison ratio (echoing sensitivity to rare/long-tail signals [9]); and poisoned behavior can generalize to unseen triggers, reminiscent of classical backdoor phenomena [10].

VII. Strengths

- **Clear threat model for preference-stage poisoning.** The focus on preference learning fills a gap versus poisoning during instruction tuning or pretraining [7].
- **Operationalizable benchmark.** Attacks are concretely specified and tested across diverse backbones and alignment dimensions, yielding actionable comparisons.
- **Stealthiness vs. effectiveness.** Evaluating attack success alongside non-triggered behavior reflects realistic deployment risks.
- **Scaling and trigger analyses.** Log-linear poison-ratio effects and sensitivity to trigger form are useful empirical regularities for future defenses.
- **Deceptive alignment evidence.** Generalization to unseen triggers advances beyond signature-based backdoor views [10].

VIII. Weaknesses / Concerns

- **Limited external validity of data sources.** Reliance on two datasets with specific annotation styles; real curation pipelines (multi-turn, multi-annotator aggregation) may shift the attack surface.
- **Narrow training protocols.** Results center on one-epoch DPO [8]; conclusions may differ under PPO-style RLHF and alternative reward modeling [7].
- **Defense coverage is diagnostic-only.** No systematic evaluation of countermeasures (robust preference aggregation, outlier detection, influence-based triage).
- **Detectability not rigorously quantified.** "Stealthiness" is proxied by non-triggered performance; representation-space forensics or influence diagnostics would strengthen claims.
- **Confounding by priors/frequency.** Content-injection success appears correlated with entity frequency; stronger causal controls are needed to isolate backdoor learning from prior familiarity [9].

IX. Potential improvements or extensions

- **Defense benchmarks.** Add baselines: robust pairwise aggregation with annotator reliabilities; trigger-diversified adversarial training; provenance checks/deduplication;

active sanitization via uncertainty or influence; DP/clipping to bound poisoned-pair impact.

- **Broader settings.** Include multi-turn dialogue poisoning, instruction-level poisoning, and reward-model poisoning with PPO [7].
- **Causal/statistical audits.** Negative controls, falsification tests, and heterogeneous treatment-effect analysis by model family/size; relate effects to long-tail coverage [9].
- **Trigger forensics.** Embedding outliers, curvature/Fisher information, and influence functions; report ROC-style detection vs. utility trade-offs as in backdoor literature [10].
- **Realistic curation.** Simulate crowdsourcing pipelines (label mixing, reviewer disagreement, QC filters) and continual preference updates to test persistence/erosion of backdoors.

X. Questions for the Authors

1) Sensitivity to annotator mixture and aggregation rules beyond pairwise DPO [8]?
2) Does backdoor strength persist or attenuate under continued clean preference training or PPO-based RLHF [7]?
3) Can detectability be quantified via representation-space tests or influence functions, and what are the detection/utility trade-offs relative to [10]?
4) Cross-dataset transfer: trained on HH-RLHF [6], evaluated on other safety preference sets; how portable are triggers?
5) Are trigger-agnostic defenses (e.g., paraphrase ensembles, prompt canonicalization) effective without harming alignment on long-tail tasks [9]?
6) Is there any chance to reduce the computational resources from high GPU demand setting like Nvidia Tesla A100 with 80GiB GPU memory?

References

[1] 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," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020

[2] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021. [Online]. Available: https://openreview.net/forum?id=OJLaKwiXSbx

[3] Z. Yang, X. He, Z. Li, M. Backes, M. Humbert, P. Berrang, and Y. Zhang, "Data poisoning attacks against multimodal encoders," 2023.

[4] 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 *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[5] N. Carlini and A. Terzis, "Poisoning and backdooring contrastive learning," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=iC4UHbQ01Mp

[6] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. Dassarma, D. Drain, S. Fort, D. Ganguli, T. J. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. B. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan, "Training a helpful and harmless assistant with reinforcement learning from human feedback," *ArXiv*, vol. abs/2204.05862, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:248118878

[7] G. Cui, L. Yuan, N. Ding, G. Yao, W. Zhu, Y. Ni, G. Xie, Z. Liu, and M. Sun, "Ultrafeedback: Boosting language models with high-quality feedback," 2024. [Online]. Available: https://openreview.net/forum?id=pNkOx3IVWI

[8] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn, "Direct preference optimization: Your language model is secretly a reward model," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. [Online]. Available: https://openreview.net/forum?id=HPuSIXJaa9

[9] N. Kandpal, H. Deng, A. Roberts, E. Wallace, and C. Raffel, "Large language models struggle to learn long-tail knowledge," ser. ICML'23. JMLR.org, 2023.

[10] T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," 2019. [Online]. Available: https://arxiv.org/abs/1708.06733