Security and Privacy of Machine Learning, 2025 Critique G9: Membership Inference Attack – (1) Membership Inference Attacks against Large Vision-Language Models (2) Privacy Backdoors: Enhancing Membership Inference through Poisoning Pre-trained Models (3) Variance-Based Membership Inference Attacks Against Large-Scale Image Captioning Models

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

CRITIQUE:(1) MEMBERSHIP INFERENCE ATTACKS AGAINST LARGE VISION-LANGUAGE MODELS

SUMMARY

This paper tackles membership inference attacks (MIA) for *vision–language models* (VLLMs), where an adversary aims to decide whether a given *image* or *text* appeared in training. Building on prior MIA foundations [1], [2], the authors identify two blockers for multimodal settings: (i) lack of a standard benchmark for VLLMs, and (ii) the absence of image tokens (only image *embeddings*) in popular architectures, which hinders target-based MIAs common in LLMs [3], [4].

They contribute (1) **VL-MIA**, a benchmark spanning image and text modalities (COCO/Flickr vs. recent Flickr; LAION-derived pairs vs. DALL·E variants; instruction-tuning text vs. GPT-generated distractors), (2) a **cross-modal generation→inference pipeline** that slices output logits by image/instruction/description positions to support single-modality attacks, and (3) **MaxRényi-$K\%$**, a target-*free* entropy-based score (and *ModRényi*, a target-based variant). Empirically, target-free scores are stronger for images (no tokens) and for long-ago pretraining text, while target-based scores dominate for recently fine-tuned instruction data—an observation consistent with memorization dynamics and forgetting. The method also works against GPT-4V (closed-source) under top-k probabilities, with an AUC of 0.815 on a DALL·E-based set.

STRENGTHS

- **Well-scoped novelty for VLLMs.** Prior multimodal MIA work (e.g., CLIP pairwise MIAs) focused on *image–text pairs* [5]; this paper tackles the practically important and harder *single-modality* membership for images *or* texts in VLLMs.
- **Actionable benchmark.** VL-MIA systematically spans (i) pretraining-like image membership, (ii) instruction-tuning text membership, and (iii) synthetic IID settings—useful to separate distribution-shift confounds from memorization effects. This fills a gap for evaluating MIA signals beyond pure text LLMs [3].
- **Cross-modal slicing is insightful.** Leveraging text-token logits (instruction/description) to probe image membership is clever given the lack of image tokens; it operationalizes the causal flow of VLLMs without architectural surgery.
- **Metric design unifies literature.** MaxRényi-$K\%$ smoothly interpolates Shannon/min-entropy perspectives and clarifies when target-free vs. target-based statistics should win (fine-tuned vs. pretraining regimes), aligning with known overfitting/memorization phenomena [2], [4].
- **Convincing ablations.** Length sensitivity (plateau $\sim$128 tokens), corruption robustness (JPEG/blur hurt more than brightness/snow), and prompt invariance strengthen external validity.

WEAKNESSES

- **Gray-box assumptions may overstate feasibility.** The core pipeline consumes full logits; even the GPT-4V evaluation assumes top-5 probabilities. In many production VLMs, only text outputs are exposed. Without calibrated scores, attack strength under *text-only* APIs remains unclear.

- **Distribution-shift confounds remain.** Flickr recency and DALL·E "semantic twins" are reasonable but nontrivial design choices. The former blends temporal drift with membership; the latter may bias toward lexical alignment of captions rather than pure memorization signals.
- **TPR at low FPR can be modest.** While AUCs are informative, some settings show low TPR@5%FPR, limiting practical auditing where false positives must be rare (e.g., compliance pipelines).
- **Limited model/ecosystem breadth.** Results center on LLaVA/MiniGPT-4/LLaMA-Adapter and one closed model. Missing are stronger modern VLMs (e.g., Qwen-VL/InternVL/LLaVA-Next) and safety-tuned systems with RLHF or deduplication, which materially affect memorization [3].
- **No defense study or auditing protocol.** The paper surfaces risks but stops short of guidance on mitigations (e.g., deduplication, DP-SGD, temperature/randomization policies, confidence capping, or output smoothing) and how to audit when only text is available.

Potential Improvements or Extensions

- **Text-only attack variants.** Explore calibration-free surrogates (e.g., length-normalized log-likelihood via a public probe model; compression-based signals; self-consistency dispersion) for APIs without probabilities.
- **Causal slicing & attribution.** Combine cross-modal slicing with token-wise influence functions or counterfactual prompting to localize which parts of the description carry membership evidence.
- **Defense/mitigation evaluations.** Quantify impacts of training-time deduplication, DP-SGD, low-rank finetuning vs. full-finetuning, and decoding-time perturbations (temperature, nucleus sampling) on MaxRényi-$K\%$ separability.
- **Harder privacy domains.** Extend VL-MIA with clinically realistic datasets (de-identified) or personal-photo distributions, plus controlled near-duplicate mining, to stress-test memorization vs. semantic similarity.
- **Operational auditing recipes.** Provide thresholds calibrated by conformal prediction for user-level guarantees, and guidance for batched auditing under budgeted queries.

Questions for Authors

- **API practicality:** How does attack power degrade under *text-only* outputs or $\leq$top-1 token (argmax) exposure? Any calibration strategy without probabilities?
- **Query budget:** What minimum prompts per sample are needed to achieve high TPR@5%FPR, and how sensitive are results to decoding parameters (temperature, top-p)?
- **Deduplication & RLHF:** How do document-level dedup and RLHF post-training change MaxRényi-$K\%$ separability relative to [3]?
- **Negative controls:** On DALL·E pairs, can we equalize caption statistics (e.g., rare n-grams) to rule out lexical shortcuts?
- **Scope of generalization:** Would the cross-modal pipeline still work for OCR-heavy or chart QA tasks where token distributions are dominated by structured text rather than natural image semantics?

Critique: (2) Privacy Backdoors: Enhancing Membership Inference through Poisoning Pre-trained Models

Summary

This paper introduces a *privacy backdoor* attack: an adversary uploads a poisoned pre-trained model to a public hub; when victims fine-tune it on their private data, the resulting model exhibits amplified membership signals, enabling substantially stronger membership inference attacks (MIA) under black-box access. Unlike classical backdoors that target accuracy or label flips [6], the manipulation here biases loss geometry to increase separability between members and non-members after fine-tuning. Empirically, across vision and language settings (including PEFT variants), the attack raises TPR@1%FPR and AUC without degrading downstream accuracy, highlighting a supply-chain threat overlooked by standard MIA defenses [1], [7] and by prior poisoning work focused on integrity [8] or memorization-only risks in LLMs [4].

Strengths

- **Clear, novel threat model.** The paper reframes MIA risk as a *pre-training supply-chain* vulnerability, bridging backdoor attacks [6] with privacy leakage, and showing that even benign-appearing checkpoints can amplify post-hoc membership signals.
- **Stealthy yet effective.** The attack preserves task accuracy and standard validation losses, so typical sanity checks would fail to detect it. This matches real-world practices where hubs are trusted by default.
- **Broad coverage.** Results span vision and language models, multiple fine-tuning regimes (including parameter-efficient tuning), and API restrictions (e.g., top-k outputs), demonstrating robustness of the phenomenon.
- **Mechanistic intuition.** Casting the poison as shaping the local loss landscape around future fine-tuning provides a compelling explanation for why member/non-member margins widen after adaptation, consistent with observed MIA behavior [1], [7].

Weaknesses

- **Detectability under stronger probes.** While standard metrics remain unchanged, the paper does not deeply test targeted diagnostics that directly measure curvature/flatness (e.g., trace/spectral norm of the Hessian,

Fisher information) around canary regions that the attack is designed to sharpen.
- **Limited systemic defenses.** The evaluation focuses on model-side countermeasures (e.g., output restrictions), but under-explores *ecosystem* controls (checkpoint attestation, signed provenance, reproducible hashes) that could neutralize the attack upstream.
- **Generalization of spillover.** The work notes non-target spillover but stops short of a systematic analysis across domain shifts and class imbalance; without this, defenders cannot calibrate worst-case population risk or triage affected subgroups.
- **Interaction with DP-SGD and flatness-seeking optimizers.** The paper does not quantify whether stronger DP noise, per-sample clipping schedules, or sharpness-aware minimization blunt the attack, nor the utility cost of such defenses [1], [4].

Potential Improvements / Extensions

- **Supply-chain hardening and audits.** Evaluate model-hub defenses: (i) *weight attestation* (Sigstore-like signatures for model weights & training manifests), (ii) *reproducibility beacons* (deterministic training with public seeds; matchable digests), and (iii) *pre-publication privacy amplification tests*—a standardized MIA harness on synthetic canaries to estimate a "leakage amplification factor" for each checkpoint.
- **Geometry-aware detectors.** Add experiments that screen suspect checkpoints via low-cost curvature probes: Hutchinson trace for Hessian, layerwise spectral norms, Fisher diagonals, and per-token sharpness—flagging anomalous anisotropy that correlates with amplified membership margins.
- **Immunization-by-flatness.** Test fine-tuning with SAM/ASAM, noise injection, or implicit regularizers that prefer flat minima; report the privacy–utility Pareto. Connect to condition-number/curvature controls to proactively reduce membership separability.
- **PEFT boundary tests.** Isolating where the backdoor lives: compare (frozen base + adapters) vs. full fine-tuning; explore adapter re-initialization, prefix-tuning, and LoRA rank schedules to see which pathways transmit the privacy backdoor most.
- **Black-box-only hardening.** Quantify benefits of calibrated logit rounding, temperature scaling, randomized response on scores, and query-rate throttling; measure attacker advantage with adaptive query strategies.
- **Spillover mapping.** Stratify TPR@1%FPR by frequency/rarity bins and OOD slices; estimate risk to rare classes and sensitive entities. Provide decision-theoretic guidance for deployment thresholds.

Questions for the Authors

- **Target selection.** How sensitive is attack success to the adversary's target distribution mismatch from the defender's private data? Would diverse proxy targets reduce overfitting and amplify spillover?
- **Defense-aware poisoning.** If the defender applies SAM or DP-SGD during fine-tuning, does the attacker need different poisoning objectives (e.g., curvature under noise, per-sample gradient norm control)?
- **Adapter isolation.** When only adapters are trained, can the privacy backdoor be confined or scrubbed by adapter re-initialization while freezing the base?
- **Forensics.** Are there stable weight-space fingerprints (e.g., layerwise spectral ratios, Fisher patterns) that persist across fine-tuning seeds and would enable post-hoc attribution of privacy backdoors?
- **Benchmarks.** Would you release a public "privacy backdoor" benchmark (scripts + poisoned checkpoints) to catalyze detection/defense research without exposing real private data?

Critique: (3) Variance-Based Membership Inference Attacks Against Large-Scale Image Captioning Models

Summary

This paper proposes two black-box membership inference attacks (MIAs) tailored to image captioning models: (i) a *Means-of-Variance Threshold Attack* (MVTA) and (ii) a *Confidence-based Weakly Supervised Attack* (C-WSA). The central idea is that, under stochastic decoding (e.g., top-p), a model trained on a given image will produce caption embeddings with *lower dispersion* than for an unseen image. The authors define a *means-of-variance (MV)* score over multiple generated caption embeddings for the same image; MVTA classifies membership by thresholding MV, while C-WSA builds a pseudo-member set using MV against a non-member baseline and trains a classifier on image features with confidence filtering. Evaluations on public captioners (BLIP, ViT-GPT2, GIT) show improvements over CLIP-oriented baselines (CSA/WSA). The work is positioned within MIA foundations [1], the recent multi-modal MIA line [5], [9], memorization risks in generative models [4], and modern captioners (e.g., BLIP) [10].

Strengths

- **Realistic black-box threat model**: Assumes query access to a captioner and *images only* (no ground-truth captions), aligning with practical API settings and exceeding the assumptions of many prior MIAs [1], [5].
- **Shadow-model free**: Avoids the heavy data/compute burden of shadow training, improving practicality for large models compared to earlier works [9].
- **Simple, general metric**: MV is architecture-agnostic and conceptually grounded: training exposure $\Rightarrow$ lower conditional variance of generated text (a stability/entropy signal), consistent with memorization phenomena in LMs [4].
- **Empirical depth**: Sensitivity to decoding (beam vs. top-p), number of captions, and pseudo-member thresholds

is studied; low-FPR TPR numbers are reported, which matter operationally.

- **Public models & weak supervision**: Uses Hugging Face captioners and synthesizes non-members, improving external validity versus small synthetic setups.

Weaknesses

- **Query budget reliance**: Performance depends on generating many captions per image (e.g., $n \approx 40$). Real APIs may impose rate limits, costs, or fix decoding (no top-p), weakening the signal.
- **Assumptions on score distribution**: The approach implicitly treats non-member MV scores as approximately Gaussian for thresholding and confidence. Distributional mis-specification or dataset shift can degrade calibration.
- **Encoder/decoding sensitivity**: MV depends on the chosen text encoder and decoding hyperparameters. The attack's robustness across encoders (e.g., different tokenizers or multilingual models) is under-explored.
- **Ground-truth non-members**: Constructing D_{no} in the wild (without leakage/near-duplicates) is non-trivial; synthetic data (e.g., SD1.5) might inadvertently overlap with training support or style priors.
- **Limited defensive analysis**: No evaluation against standard mitigations (DP, early stopping, entropy tempering, sampling-level randomization, output clipping), nor guidance on safe API defaults for providers.

Potential Improvements / Extensions

- **Attack under realistic API constraints**: Benchmark accuracy vs. (i) capped query budgets, (ii) fixed decoding (beam or fixed nucleus), (iii) response truncation. Provide sample-complexity curves (TPR@1%FPR vs. n).
- **Adaptive query design**: Replace uniform sampling with *variance-seeking* prompts (e.g., perturbations to image crops, caption prefix seeding) to amplify dispersion gaps under few queries.
- **Model-agnostic ensembling**: Combine MV with cosine-similarity (CSA) or log-prob proxies (when available) using a small calibration set; explore conformal risk control for thresholding with distribution shift.
- **Defense evaluation**: Test MVTA/C-WSA against DP-SGD training (captioners), *inference-time* defenses (temperature caps, beam-only decoding, nucleus-p noise injection), and data deduplication pipelines.
- **Theory & metrics**: Relate MV to local conditional entropy and to curvature/flatness around member examples; study when MV separates IID members vs. near-duplicates or semantically entangled clusters.
- **Broader modalities**: Port the variance idea to VLMs with open-ended generation (VQA) and to diffusion text-to-image MIAs; test multilingual captioners where tokenization changes embedding variance.

Questions for the Authors

- **Decoder control**: If providers force deterministic decoding (beam search) or cap p, does the MV signal vanish, or can small stochasticity (e.g., dropout at decode) suffice? What is the minimal randomness needed?
- **Encoder choice**: How sensitive is MV to the text encoder family (e.g., BERT vs. sentence-transformers) and its domain mismatch with the captioner? Could an adversary *learn* an encoder that maximizes separability?
- **Calibration in the wild**: Without clean D_{no}, can you self-calibrate τ online (e.g., mixture modeling over queried MV scores)? How robust is λ across datasets and models?
- **Near-duplicates**: Does MV confuse members with *non-member* near-duplicates (same scene/objects)? Would pairing MV with image-similarity filters reduce false positives in such regimes?
- **Defensive levers**: Which change most reduces attack power per unit utility loss: (i) decoding constraints, (ii) training regularization/DP, (iii) dataset deduplication, or (iv) output post-processing (paraphrasers)?

References

[1] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 3–18.

[2] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *IEEE Computer Security Foundations Symposium (CSF)*, 2018.

[3] W. Shi, A. Ajith, M. Xia, Y. Huang, D. Liu, T. Blevins, D. Chen, and L. Zettlemoyer, "Detecting pretraining data from large language models," in *International Conference on Learning Representations (ICLR)*, 2024.

[4] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. B. Brown, D. Song, U. Erlingsson, A. Oprea, and N. Papernot, "Extracting training data from large language models," in *USENIX Security Symposium*, 2021.

[5] M. Ko, M. Jin, C. Wang, and R. Jia, "Practical membership inference attacks against large-scale multi-modal models: A pilot study," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4871–4881.

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

[7] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," *IEEE Symposium on Security and Privacy*, 2019.

[8] J. Steinhardt, P. W. Koh, and P. S. Liang, "Certified defenses for data poisoning attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[9] P. Hu, Z. Wang, R. Sun, H. Wang, and M. Xue, "Mˆ4i: Multi-modal models membership inference," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 1867–1882.

[10] J. Li, D. Li, C. Xiong, and S. C. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 12 888–12 900.