Security and Privacy of Machine Learning, 2025 Critique G4: Jailbreaking VLMs – (1) Jailbreak Large Vision-Language Models Through Multi-Modal Linkage (2) IDEATOR: Jailbreaking and Benchmarking Large Vision-Language Models Using Themselves

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

Critique:(1) Jailbreak Large Vision-Language Models Through Multi-Modal Linkage

Summary

This paper proposes **Multi-Modal Linkage (MML)**, a structure-based jailbreak that hides harmful instructions inside typographic images via lightweight "encryption" (word replacement, mirroring, rotation, Base64, shift cipher) and then *coaxes the model to decrypt and comply* using chain-of-thought style guidance and a narrative "evil alignment" (a villain-in-a-game scenario). The method specifically targets modern, closed-source VLMs under a single-round, black-box setting and is evaluated on SafeBench, MM-SafetyBench [1], and HADES [2], with comparisons against FigStep [3] and other strong baselines. Results show strikingly high attack success rates (ASR), including on GPT-4o, and ablations indicate that both cross-modal encryption–decryption and the role-play narrative contribute materially to success. The work highlights an important gap between instruction-following and safety alignment, consistent with observations around chain-of-thought and multi-step prompting [4] and despite alignment with human feedback [5].

Strengths

- **Practical black-box threat model:** No access to parameters or system prompts, single-turn interaction, and low-overhead transformations make MML realistic for adversaries.
- **Clean modularity:** The pipeline is pluggable: any encoding the model can plausibly "reason out" can be inserted. The shift-cipher variant underscores extensibility beyond the four core methods.
- **Robust empirical evidence:** Consistent gains across three datasets and four advanced models, with topic-wise breakdowns. Including Llama-Guard cross-evaluation reduces evaluator overfitting risk.
- **Ablation clarity:** The decomposition (encryption–decryption, hinting, evil alignment) is persuasive; the narrative alignment evidently pushes outputs from "safe-but-related" to fully policy-violating.
- **Defense interaction:** Testing under AdaShield-style prompting defenses is valuable and shows resilience compared to prior structure-based attacks.

Weaknesses

- **Evaluator coupling and ASR definition:** The primary metric relies on a keyword refusal check and a single LLM judge (GPT-4o-Mini) with success defined only at score 5. While conservative, this design may (i) undercount harmful partial compliance and (ii) still inherit shared biases or blind spots of the target family, despite the Llama-Guard cross-check. A broader ensemble (open-source, safety-tuned, and human verification) would strengthen claims.
- **Dataset filtering and construct validity:** Filtering MM-SafetyBench with GPT-4o-Mini to retain only high-violation prompts could bias the benchmark toward items the attack (and evaluator) already handle well. Reporting ASR on both filtered and unfiltered sets (or third-party filters) would improve external validity.
- **Defense scope:** Prompt-shield defenses are one axis. Strengthened system prompts, OCR-level sanitization, image-text consistency checks, and multi-step safety verifiers are not comprehensively covered; conclusions about defense robustness may therefore be optimistic.
- **Safety externalities:** While the paper includes ethical safeguards, releasing working prompts and dictionaries might accelerate real-world misuse. A staged or red-

teaming–gated release plan could balance reproducibility and safety.
- **Limited analysis of failure modes:** Where MML fails (e.g., Claude variants, Base64 cases), the paper hints at training-time defenses but lacks fine-grained causal analysis (OCR variance, decoding heuristics, narrative detection).

Potential Improvements / Extensions

- **Richer evaluation:** Use a multi-judge ensemble (safety-tuned LMs, rule-based templates, and human raters) and calibrate inter-rater reliability; report multiple success thresholds (e.g., "any harmful detail", "actionable plan", "score 4–5") to expose the harm surface more completely.
- **Defense stress-tests:** Evaluate against (i) OCR obfuscation / input sanitization, (ii) cross-modal inconsistency detectors (flagging decrypt–generate patterns), (iii) *safety CoT-gating*—routing model reasoning through a *safe* reflection pass before finalization, (iv) adversarially trained safety heads that specifically target cipher/role-play patterns.
- **Generalization and transfer:** Test on more encodings (morphological variants, Unicode confusables, font ligatures, steganographic placements) and across modalities (audio spectrogram text, short video captions). Analyze transfer across families (e.g., o1 vs. Claude vs. Qwen) with controlled prompts.
- **Causal probes:** Instrument the model to detect which step (OCR, hint matching, narrative framing) flips the safety decision. Interventions (ablate hints, randomize narrative, insert adversarial decoys) can localize the mechanism of failure.
- **Real-world detection study:** Measure *human detectability* of encrypted images and the narrative prompt. If humans can reliably spot the ruse while current models cannot, that gap motivates hybrid human-in-the-loop moderation.

Questions for the Authors

- **Evaluator robustness:** How sensitive are ASR conclusions to the choice of judge? What happens with a diverse ensemble and when success is defined at "score $\geq$ 4"?
- **Narrative vs. decoding:** If you remove the decryption hints but keep the villain narrative, how often does the model *hallucinate* correct titles and still comply? Conversely, can strong hints without narrative still elicit fully harmful content?
- **Defense breakage analysis:** Which defense instructions (lexical cues, refusals, refusal-rationales) most reliably block MML? Can adversarial prompt tuning re-break those shields, and at what cost to latency/consistency?
- **Counterfactual safety:** Would a two-pass pipeline (vision encoder → OCR sanitization → safety parser) prevent most MML cases, or does MML survive via non-textual cues?
- **Ethical release:** What governance or access controls (rate limits, watermarking, safety-*CoT* disclosure) would you recommend to model providers in light of your findings?

Critique: (2) IDEATOR: Jailbreaking and Benchmarking Large Vision-Language Models Using Themselves

I. Summary

This paper introduces **IDEATOR**, a black-box, training-free framework that uses a VLM as a red-team "attacker" to *automatically* generate adversarial image–text pairs that jailbreak victim VLMs. The attacker VLM outputs a structured JSON with (i) analysis (for iterative refinement), (ii) an image prompt (rendered by a diffusion model), and (iii) a text prompt. A breadth–depth exploration strategy runs multiple attack streams and refines them across turns. Using IDEATOR, the authors construct **VLJailbreakBench** (3,654 multimodal samples across 12 safety topics and 46 subcategories), then evaluate 11 VLMs. Reported results show high attack success rates (ASR), e.g., 94% on MiniGPT-4 with $\approx$5.34 queries on average, and strong transfer to LLaVA and InstructBLIP. The work situates itself amidst recent advances in VLMs [6], [7] and safety benchmarks/attacks [1], [8], [9].

II. Strengths

- **Automation & Scaling.** Turning a VLM into a red-team agent is a simple but powerful idea that scales beyond handcrafted pipelines. The JSON discipline and CoT-like analysis field make the loop robust and reproducible in spirit.
- **Breadth–Depth Exploration.** The dual-axis search provides a principled way to trade off coverage and refinement. The empirical trend (ASR rising with breadth/depth) is convincing and practically actionable for safety evaluations.
- **Multimodal Integration.** Combining image prompts (via diffusion) with text prompts leverages VLM weaknesses unique to the vision channel, complementing text-only jailbreaks and echoing limitations surfaced by prior safety work [1], [8].
- **Benchmark Contribution.** VLJailbreakBench fills an important gap: multimodal, adversarially generated, and categorically diverse. The taxonomy is thoughtfully constructed with interdisciplinary input.
- **Comparative Coverage.** The cross-model transfer results and defense evaluation against AdaShield-S build a compelling case that IDEATOR captures attack modes not well handled by current defenses [1], [9].

III. Weaknesses

- **Evaluator Dependence & Bias.** Automated (or semi-automated) ASR judgments risk dependence on a particular evaluator model or rubric. If the same family that generates (or filters) samples also judges success, measurement leakage may inflate ASR.

- **Goal-Set and Severity Normalization.** ASR treats all harmful outcomes equally. Without severity weighting (e.g., by expected harm or operational specificity), headline rates may obscure qualitatively different risk profiles across categories.
- **Attacker & Generator Choice.** IDEATOR's effectiveness likely depends on the permissiveness and capability of the attacker VLM and the image generator. The paper does not fully disentangle the contributions of attacker model family/size, safety settings, and diffusion model quality.
- **Human-in-the-Loop Reproducibility.** The pipeline includes manual curation and labeling. More details on inter-annotator agreement, labeling guidelines, and edge cases would strengthen reproducibility and external validity.
- **Generalization Beyond Benchmarked Models.** While transfer is tested across several popular VLMs, the coverage of architectures and safety stacks (e.g., multi-layered content filters, server-side vision sanitizers) remains limited.

IV. Potential Improvements and Extensions

- **Severity-Weighted Safety Metrics.** Complement ASR with a harm-weighted metric (e.g., risk tiers calibrated with domain experts), and report Pareto fronts of $\{$ASR, severity, stealth/detectability$\}$.
- **Evaluator Triangulation.** Use a panel of heterogeneous evaluators (commercial and open-source; instruction-tuned vs. safety-tuned) and report agreement statistics. Consider majority voting or adjudication guidelines.
- **Ablations on the Attacker Stack.** Systematically vary (i) attacker VLM family/size/safety tuning, (ii) diffusion model family and guidance settings, and (iii) the JSON schema (e.g., ablate the `analysis` field) to quantify each component's contribution.
- **Countermeasures as First-Class Citizens.** Co-design adversarially robust defenses and re-run the breadth–depth exploration in a closed-loop evaluation. For example, pre-/post-vision sanitization, dynamic refusal strategies, and cascade-of-guardrails specific to vision+text [1], [8].
- **Safety Budget Curves.** Publish "safety curves" showing ASR as a function of exploration budget (breadth/depth) and query count, akin to sample-complexity plots. This would let practitioners pick defensible budgets for deployment evaluations.
- **Benchmark Governance.** Provide an access-controlled release with researcher attestations, usage logging, and red-team/blue-team protocols; include a living document of mitigations discovered from benchmark-driven audits.

V. Questions for the Authors

1) **Evaluator Robustness:** How sensitive are ASR numbers to the choice of evaluator model and rubric? Have you measured inter-rater reliability when humans adjudicate marginal cases?
2) **Stealth vs. Success:** Do you quantify detectability (by humans and by automated filters) alongside ASR? If not, could you report a stealth-adjusted metric?
3) **Transfer Mechanisms:** What characteristics of the generated image–text pairs best predict cross-model transfer (e.g., typography density, roleplay depth, semantic proximity)?
4) **Defense Feedback Loop:** When defenses are adapted to IDEATOR (e.g., fine-tuned guardrails), how quickly does ASR drop, and does the breadth–depth search "recover" with more queries?
5) **Attacker Swap:** If the attacker VLM is replaced with a more conservative but stronger model (e.g., highly capable with stricter alignment), how does that trade off capability vs. willingness to generate adversarial content?

References

[1] X. Liu, Y. Zhu, J. Gu, Y. Lan, C. Yang, and Y. Qiao, "Mm-safetybench: A benchmark for safety evaluation of multimodal large language models," in *European Conference on Computer Vision (ECCV)*. Springer, 2024.

[2] Y. Li, H. Guo, K. Zhou, W. X. Zhao, and J.-R. Wen, "Images are achilles' heel of alignment: Exploiting visual vulnerabilities for jailbreaking multimodal large language models," *arXiv preprint arXiv:2403.09792*, 2024.

[3] Y. Gong, D. Ran, J. Liu, C. Wang, T. Cong, A. Wang, S. Duan, and X. Wang, "Figstep: Jailbreaking large vision-language models via typographic visual prompts," *arXiv preprint arXiv:2311.05608*, 2023.

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *NeurIPS*, 2022.

[5] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray *et al.*, "Training language models to follow instructions with human feedback," in *NeurIPS*, 2022.

[6] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *NeurIPS*, 2024.

[7] J. Achiam, S. Adler, S. Agarwal *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[8] X. Qi, K. Huang, A. Panda, P. Henderson, M. Wang, and P. Mittal, "Visual adversarial examples jailbreak aligned large language models," in *AAAI*, 2024.

[9] R. Wang, X. Ma, H. Zhou, C. Ji, G. Ye, and Y.-G. Jiang, "White-box multimodal jailbreaks against large vision-language models," in *ACM MM*, 2024.