~~TABLE IV~~
~~DEFENSE AND EVALUATION STRATEGIES~~

~~Category~~	~~Papers~~
~~Red Teaming Frameworks~~	~~Arondight, MM-SafetyBench, In-Context Experience Red Teaming~~
~~Safety Benchmarking~~	~~MM-SafetyBench, How Many Unicorns Are in This Image?~~
~~Proposed Defenses~~	~~VLMGuard, Think Twice Prompting, Circuit Breakers~~

~~than unimodal models, are also more vulnerable due to the complex interactions between input modalities. Second, black-box access does not preclude powerful jailbreaks; in fact, some of the most successful attacks operate without any model internals, relying instead on learned patterns and iterative probing. Third, semantic and perceptual misalignment remains a critical weak point—models can be deceived by inputs that are harmless at the surface but malicious in meaning or effect.~~

~~A taxonomy of jailbreak attacks on multimodal LLMs thus spans:~~

- ~~**Prompt-based vs. Input-modality attacks**~~
- ~~**Static vs. Environment-interactive threats**~~
- ~~**Single-step vs. Iterative/refined attacks**~~
- ~~**Manual vs. Tool-augmented generation**~~

~~These categories will be essential for guiding future defense mechanisms and benchmarking efforts.~~

IV. POISONING AND BACKDOOR ATTACKS

Multimodal models, including Vision-Language Models (VLMs) and Multimodal Large Language Models (MLLMs), are vulnerable to data poisoning and backdoor attacks due to training on large, uncurated datasets. These attacks involve adversaries manipulating model behavior by injecting malicious data into training sets.

Early foundational work demonstrated that existing poisoning techniques could be adapted to multimodal contrastive models by injecting altered images with wrong labels or embedding backdoor patches with target labels, requiring attackers to have training dataset modification access [33]. Building upon this foundation, Yang et al. revealed vulnerabilities in both visual and linguistic modalities of multimodal encoders used in text-image retrieval tasks. Their work showed that adversaries could inject mismatched text-image pairs to force models to map specific text groups to target images while preserving normal functionality [34].

This understanding of multimodal encoder vulnerabilities led to the development of more sophisticated attacks like BadCLIP, which employs an advanced dual-embedding guided framework to create resilient backdoor attacks against MCL models. Using Bayesian analysis principles, this attack optimizes visual trigger patterns for textual embedding consistency and aligns poisoned features with target vision features, designed to induce subtle parameter variations that resist detection and fine-tuning defenses [35].

As research progressed toward more complex Vision-Language Models and MLLMs, attack strategies evolved to target their unique capabilities in generating free-form text and performing complex reasoning tasks. ImgTrojan exemplifies this shift by performing cross-modality jailbreaks that replace original image captions with malicious jailbreak prompts during training, transforming even clean images into trojans that bypass safety barriers at inference time through learned associations between poisoned images and injected prompts [36]. Similarly, Dual-Key Multimodal Backdoors targets VQA models using triggers in both visual and textual modalities that activate only when simultaneously present, enhancing stealth by reducing accidental activation likelihood through optimized visual trigger patterns designed for effective processing by static pretrained feature extractors [37]. Shadowcast emerged as a stealthy poisoning attack using visually indistinguishable images paired with manipulated text, enabling both label attacks for misclassification and persuasion attacks for misleading narratives without requiring training control [38]. This was followed by TrojVLM, which specifically targets image-to-text generation by embedding pixel patterns that trigger insertion of predefined text while maintaining semantic coherence, with attackers modifying lightweight adaptors rather than full models [39].

Extending these concepts further, VL-Trojan demonstrated how backdoors could be embedded in autoregressive VLMs during instruction tuning by placing triggers in instructions or images while operating with limited access to visual encoder architecture only [40]. Building on this foundation, BadToken introduced novel token-level backdoor behaviors for enhanced flexibility and stealth in MLLMs, featuring token-substitution capabilities that replace specific source tokens with target tokens and token-addition mechanisms that append target token sequences to outputs, enabling subtle alteration of critical text portions with significant consequences in applications like autonomous driving or medical diagnosis [41].

The evolution toward more efficient and specialized attack methodologies led to the development of the BAGS score method, which enables efficient backdooring of VQA and AVSR models using gradient-based sample selection to minimize required data and computation while maintaining effectiveness [42]. This efficiency-focused approach was complemented by MABA, which enhances backdoor generalizability across visual and text domains using domain-agnostic triggers such as simple patches or text symbols, operating in black-box settings without knowledge of test data distribution [43].

The field has also witnessed the emergence of novel attack vectors that transcend traditional training-time manipulation. AnyDoor represents a paradigm shift as a test-time backdoor attack requiring no training data access, using universal adversarial perturbations on images combined with text triggers to allow dynamic modification of backdoor effects during testing [14]. This approach to runtime manipulation paved the way for BadVLMDriver, the first physical backdoor attack against autonomous driving VLMs, which uses common objects like red balloons as triggers to induce unsafe actions while employing generative models to synthesize backdoor training samples with embedded physical triggers [44].

V. PROMPT INJECTION ATTACKS

The emergence of multimodal large language models has introduced sophisticated attack vectors that exploit the intersection of visual and textual processing capabilities. These adversarial approaches fundamentally divide into two paradigms: perturbation-based methods that embed imperceptible modifications into inputs, and typography-based techniques that leverage visible textual elements to exploit models' inherent biases.

Bagdasaryan et al. [45] pioneered adversarial modifications to images or audio that embed malicious prompts, using techniques like the Fast Gradient Sign Method to create imperceptible perturbations that steer models toward attacker-specified outputs. Their approach enables both targeted-output attacks that force specific malicious responses and dialog poisoning attacks where injected instructions become embedded in conversation history, influencing all subsequent model behavior.

Building on similar principles, recent work [46] extends this by simultaneously targeting multiple processing stages within vision-language models. Rather than focusing solely on final outputs, this method employs multi-objective optimization to perturb visual tokens, textual representations, and generated text concurrently, enhancing cross-prompt transferability by shifting internal probability distributions across different contextual points.

While perturbation methods maintain input authenticity through imperceptible modifications, typography-based attacks accept visible alterations for more reliable exploitation of models' textual bias. The foundational approach [47] directly adds misleading text to input images, capitalizing on vision-language models' tendency to prioritize textual signals over visual content and generating outputs semantically aligned with injected typography rather than actual image content.

This concept has been advanced through work demonstrating that visual prompts embedded within images can receive higher execution priority than conventional text input instructions [48]. Sophisticated manipulation is achieved through careful control of textual elements' size, opacity, and spatial positioning while maintaining near-imperceptibility to human observers, effectively bridging overt typographic manipulation with subtle adversarial perturbations.

The most sophisticated evolution harnesses vision-language models' reasoning capabilities to optimize their own exploitation [47]. Qraitem et al. dynamically generate the most effective deceptive content through class-based variants that leverage models' visual similarity assessments to identify optimal misleading labels, and reasoned attacks that employ advanced language models to generate both deceptive classifications and accompanying rationales that enhance attack credibility.

References

[1] Eugene Bagdasaryan, Rishi Jha, Vitaly Shmatikov, and Tingwei Zhang. Adversarial illusions in {Multi-Modal} embeddings. In *33rd USENIX Security Symposium (USENIX Security 24)*, pages 3009–3025, 2024.

[2] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Adversarial transferability across prompts on vision-language models. *arXiv preprint arXiv:2403.09766*, 2024.

[3] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. *arXiv preprint arXiv:2309.00236*, 2023.

[4] Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. Inducing high energy-latency of large vision-language models with verbose images. *arXiv preprint arXiv:2401.11170*, 2024.

[5] Xiaohan Fu, Zihan Wang, Shuheng Li, Rajesh K Gupta, Niloofar Mireshghallah, Taylor Berg-Kirkpatrick, and Earlence Fernandes. Misusing tools in large language models with visual adversarial examples. *arXiv preprint arXiv:2310.03185*, 2023.

[6] Kuofeng Gao, Yang Bai, Jiawang Bai, Yong Yang, and Shu-Tao Xia. Adversarial robustness for visual grounding of multimodal large language models. *arXiv preprint arXiv:2405.09981*, 2024.

[7] Xuanming Cui, Alejandro Aparcedo, Young Kyun Jang, and Ser-Nam Lim. On the robustness of large multimodal models against image adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24625–24634, 2024.

[8] Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. Stop reasoning! when multimodal llm with chain-of-thought reasoning meets adversarial image. *arXiv preprint arXiv:2402.14899*, 2024.

[9] Zhen Tan, Chengshuai Zhao, Raha Moraffah, Yifan Li, Yu Kong, Tianlong Chen, and Huan Liu. The wolf within: Covert injection of malice into mllm societies via an mllm operative. *arXiv preprint arXiv:2402.14859*, 2024.

[10] Chen Henry Wu, Jing Yu Koh, Ruslan Salakhutdinov, Daniel Fried, and Aditi Raghunathan. Adversarial attacks on multimodal agents. *arXiv e-prints*, pages arXiv–2406, 2024.

[11] Hao Zhang, Wenqi Shao, Hong Liu, Yongqiang Ma, Ping Luo, Yu Qiao, and Kaipeng Zhang. Avibench: Towards evaluating the robustness of large vision-language model on adversarial visual-instructions. *arXiv e-prints*, pages arXiv–2403, 2024.

[12] Haodi Wang, Kai Dong, Zhilei Zhu, Haotong Qin, Aishan Liu, Xiaolin Fang, Jiakai Wang, and Xianglong Liu. Transferable multimodal attack on vision-language pre-training models. In *2024 IEEE Symposium on Security and Privacy (SP)*, pages 1722–1740. IEEE, 2024.

[13] Xunguang Wang, Zhenlan Ji, Pingchuan Ma, Zongjie Li, and Shuai Wang. Instructta: Instruction-tuned targeted attack for large vision-language models. *arXiv preprint arXiv:2312.01886*, 2023.

[14] Dong Lu et al. Test-time backdoor attacks on multimodal large language models. *arXiv preprint arXiv:2402.08577*, 2024.

[15] Xiaofeng Mao, Yuefeng Chen, Shuhui Wang, Hang Su, Yuan He, and Hui Xue. Composite adversarial attacks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 8884–8892, 2021.

[16] Yichen Gong, Delong Ran, Jinyuan Liu, Conglei Wang, Tianshuo Cong, Anyu Wang, Sisi Duan, and Xiaoyun Wang. Figstep: Jailbreaking large vision-language models via typographic visual prompts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23951–23959, 2025.

[17] Ruofan Wang, Xingjun Ma, Hanxu Zhou, Chuanjun Ji, Guangnan Ye, and Yu-Gang Jiang. White-box multimodal jailbreaks against large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6920–6928, 2024.

[18] Zonghao Ying, Aishan Liu, Tianyuan Zhang, Zhengmin Yu, Siyuan Liang, Xianglong Liu, and Dacheng Tao. Jailbreak vision language models via bi-modal adversarial prompt. *arXiv preprint arXiv:2406.04031*, 2024.

[19] Pingyi Hu, Zihan Wang, Ruoxi Sun, Hu Wang, and Minhui Xue. M^4i: Multi-modal models membership inference. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 1867–1882. Curran Associates, Inc., 2022.

[20] Myeongseob Ko, Ming Jin, Chenguang Wang, and Ruoxi 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)*, pages 4871–4881, October 2023.

[21] Zhan Li, Yongtao Wu, Yihang Chen, Francesco Tonin, Elias Abad Rocamora, and Volkan Cevher. Membership inference attacks against large vision-language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 98645–98674. Curran Associates, Inc., 2024.

[22] Yuke Hu, Zheng Li, Zhihao Liu, Yang Zhang, Zhan Qin, Kui Ren, and Chun Chen. Membership inference attacks against vision-language models, 2025.

[23] Rubèn Tito, Khanh Nguyen, Marlon Tobaben, Raouf Kerkouche, Mohamed Ali Souibgui, Kangsoo Jung, Joonas Jälkö, Vincent Poulain D'Andecy, Aurelie Joseph, Lei Kang, Ernest Valveny, Antti Honkela, Mario Fritz, and Dimosthenis Karatzas. Privacy-aware document visual question answering, 2024.

[24] Francesco Pinto, Nathalie Rauschmayr, Florian Tramèr, Philip Torr, and Federico Tombari. Extracting training data from document-based vqa models, 2024.

[25] Yixin Wu, Ning Yu, Zheng Li, Michael Backes, and Yang Zhang. Membership inference attacks against text-to-image generation models, 2022.

[26] Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private attribute inference from images with vision-language models, 2024.

[27] Simone Caldarella, Massimiliano Mancini, Elisa Ricci, and Rahaf Aljundi. The phantom menace: Unmasking privacy leakages in vision-language models, 2024.

[28] Xingjun Ma, Yifeng Gao, Yixu Wang, Ruofan Wang, Xin Wang, Ye Sun, Yifan Ding, Hengyuan Xu, Yunhao Chen, Yunhan Zhao, Hanxun Huang, Yige Li, Jiaming Zhang, Xiang Zheng, Yang Bai, Zuxuan Wu, Xipeng Qiu, Jingfeng Zhang, Yiming Li, Xudong Han, Haonan Li, Jun Sun, Cong Wang, Jindong Gu, Baoyuan Wu, Siheng Chen, Tianwei Zhang, Yang Liu, Mingming Gong, Tongliang Liu, Shirui Pan, Cihang Xie, Tianyu Pang, Yinpeng Dong, Ruoxi Jia, Yang Zhang, Shiqing Ma, Xiangyu Zhang, Neil Gong, Chaowei Xiao, Sarah Erfani, Tim Baldwin, Bo Li, Masashi Sugiyama, Dacheng Tao, James Bailey, and Yu-Gang Jiang. Safety at scale: A comprehensive survey of large model safety, 2025.

[29] Unveiling privacy risks in multi-modal large language models: Task-specific vulnerabilities and mitigation challenges. *OpenReview*. Paper Type: Long; Research Area: Resources and Evaluation; Contribution Types: Data resources, Data analysis; Languages Studied: English; Submission Number: 5936.

[30] Jiankun Zhang, Shenglai Zeng, Jie Ren, Tianqi Zheng, Hui Liu, Xianfeng Tang, Hui Liu, and Yi Chang. Beyond text: Unveiling privacy vulnerabilities in multi-modal retrieval-augmented generation, 2025.

[31] Laurens Samson, Nimrod Barazani, Sennay Ghebreab, and Yuki M. Asano. Little data, big impact: Privacy-aware visual language models via minimal tuning, 2025.

[32] Abhijit Mishra, Richard Noh, Hsiang Fu, Mingda Li, and Minji Kim. Revision: A dataset and baseline vlm for privacy-preserving task-oriented visual instruction rewriting, 2025.

[33] Nicholas Carlini and Andreas Terzis. Poisoning and backdooring contrastive learning. *arXiv preprint arXiv:2106.09667*, 2021.

[34] Zhaoyu Yang et al. Data poisoning attacks against multimodal encoders. *International Conference on Machine Learning*, pages 39299–39313, 2023.

[35] Yuekang Li, Yi Liu, Kailong Wang, Tianwei Zhang, Yang Liu, Haoyu Wang, Yeting Zheng, Yan Liu, Yuqing Chen, et al. Badclip: Dual-embedding guided backdoor attack on multimodal contrastive learning. *arXiv preprint arXiv:2311.12075*, 2023.

[36] Yi Liu, Gelei Deng, Yuekang Li, Kailong Wang, Tianwei Zhang, Yang Liu, Haoyu Wang, Yeting Zheng, Yan Liu, Yuqing Chen, et al. Imgtrojan: Jailbreaking vision-language models with one image. *arXiv preprint arXiv:2311.18835*, 2023.

[37] Matthew Walmer et al. Dual-key multimodal backdoors for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15375–15385, 2022.

[38] Yichen Xu et al. Shadowcast: Stealthy data poisoning attacks against vision-language models. *arXiv preprint arXiv:2402.06659*, 2024.

[39] Weimin Lyu et al. Trojvlm: Backdoor attack against vision language models. In *European Conference on Computer Vision (ECCV)*, 2024.

[40] Jiawei Liang, Siyuan Liang, Aishan Liu, and Xiaochun Cao. Vl-trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *International Journal of Computer Vision*, pages 1–20, 2025.

[41] Yuting Chen, Yi Liu, Yuekang Li, Kailong Wang, Tianwei Zhang, Yang Liu, Haoyu Wang, Yeting Zheng, Yan Liu, Yuqing Chen, et al. Badtoken: Token-level backdoor attacks to multi-modal large language models. *arXiv preprint arXiv:2311.14733*, 2023.

[42] Xuyang Han et al. Backdooring multimodal learning. In *IEEE Symposium on Security and Privacy (SP)*, pages 31–31. IEEE Computer Society, 2024.

[43] Siyuan Liang et al. Revisiting backdoor attacks against large vision-language models from domain shift. *arXiv preprint arXiv:2406.18844*, 2024.

[44] Zhenyang Ni et al. Physical backdoor attack can jeopardize driving with vision-large-language models. *arXiv preprint arXiv:2404.12916*, 2024.

[45] Eugene Bagdasaryan, Tsung-Yin Hsieh, Ben Nassi, and Vitaly Shmatikov. Abusing images and sounds for indirect instruction injection in multi-modal llms. *arXiv preprint arXiv:2307.10490*, 2023.

[46] Xikang Yang, Xuehai Tang, Fuqing Zhu, Jizhong Han, and Songlin Hu. Enhancing cross-prompt transferability in vision-language models through contextual injection of target tokens. *arXiv preprint arXiv:2406.13294*, 2024.

[47] Maan Qraitem, Nazia Tasnim, Piotr Teterwak, Kate Saenko, and Bryan A Plummer. Vision-llms can fool themselves with self-generated typographic attacks. *arXiv preprint arXiv:2402.00626*, 2024.

[48] Hao Cheng, Erjia Xiao, Yichi Wang, Kaidi Xu, Mengshu Sun, Jindong Gu, and Renjing Xu. Exploring typographic visual prompts injection threats in cross-modality generation models. *arXiv preprint arXiv:2503.11519*, 2025.