Rethinking the Safety Landscape for Foundation Models: A Multi-Modal Perspective

Xi Li[1*]*, Shu Zhao[2], Fei Zhao[1], Runlong Yu[3]
[1]*University of Alabama at Birmingham*, [2]*The Pennsylvania State University*, [3]*University of Alabama*
xli7@uab.edu, smz5505@psu.edu, larry5@uab.edu, ryu5@ua.edu

Abstract

With the rise of multi-modal foundation models in domains such as autonomous driving, healthcare, and virtual assistants, safety concerns have become increasingly important. Unlike uni-modal learning, these models rely on modality alignment and fusion to integrate cross-modal information – introducing novel threats that existing safety frameworks fail to address. Current safety solutions often assume prior knowledge of compromised modalities and overlook complex cross-modal interactions. This paper calls for rethinking the safety landscape from a multi-modal perspective. We identify emerging threats, categorize existing efforts, and outline future research directions, including new threat models, safety assumptions, and fusion-aware defenses. Our goal is to open a new trajectory for trustworthy multi-modal foundation models.

1. Introduction

Multi-modal foundation models (FM) leverages large models, such as large language models (LLMs), to integrate diverse data sources (e.g., text, images, audio, and video) and enhance understanding and decision-making [1, 10, 18, 28, 41, 51]. These models enable applications such as autonomous driving (using sensor data for navigation), virtual assistants like Siri and Alexa, and medical diagnostics (e.g., combining blood tests with patient history for diabetes prediction). The integration of multiple modalities makes multi-modal learning fundamentally different and more challenging than classic uni-modal learning. Its foundation lies in two core processes: modality alignment and modality fusion [6, 85, 94, 98, 109]. Modality alignment ensures that features from different modalities are mapped into a shared representation space, while modality fusion combines the aligned information to support more comprehensive and accurate reasoning.

As FMs evolve from uni-modal to multi-modal architectures, the machine learning safety landscape is undergoing a fundamental transformation. The unique characteristics of multi-modal learning introduce several new challenges. First, additional modalities bring modality-specific vulnerabilities inherent to each data type. Second, adversarial misalignment across modalities can cause semantic inconsistencies or unexpected behaviors. Third, the fusion can be exploited – signals that appear benign in isolation can trigger harmful outcomes when combined.

However, current safety research remains largely grounded in uni-modal assumptions and falls short in addressing the complex vulnerabilities introduced by multi-modal interactions. Many methods rely on prior knowledge of which modality is compromised – an unrealistic assumption in multi-modal settings – where the type and number of affected modalities are often unknown. Besides, these safety solutions are not explicitly aligned with the core goals of modality alignment and fusion, and may unintentionally degrade overall performance. This disconnect introduces critical blind spots, limiting the effectiveness of existing safety solutions for multi-modal models.

Given the rapid deployment of multi-modal FMs and the growing gap in their safety research, we propose to **rethink the safety landscape through the lens of multi-modal learning.** This vision calls for redefining threat models and safety assumptions, identifying emerging risks unique to multi-modal systems, and developing solutions aligned with modality alignment and fusion. By grounding safety in multi-modal principles, we aim to advance both safety theory and system design, and to shift the community's perspective toward a new trajectory for trustworthy AI.

This paper focuses on the safety landscape of multi-modal large language models (MM-LLMs), highlighting emerging threats, threat models, and defense strategies distinct from the uni-modal setting (as illustrated in Figure 1). Section 2 reviews current safety landscape rooted in uni-modal learning, including adversarial attacks, data poisoning, jailbreaks, and hallucinations. Section 3 presents the unique characteristics of multi-modal learning and the new safety challenges they pose, along with a brief categorization of related work. Section 4 outlines future research di-

[*]*Corresponding author.

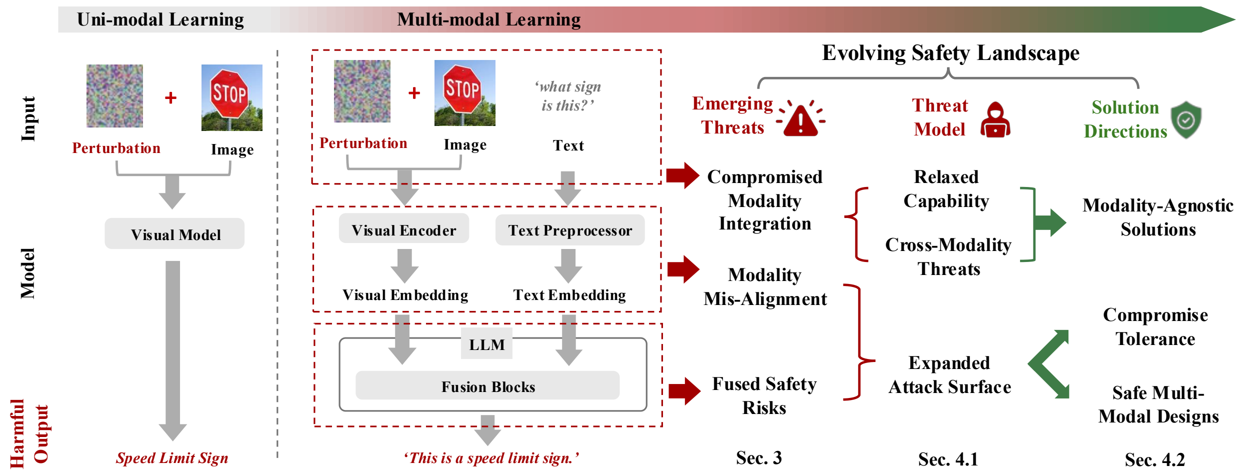

Figure 1. A multi-modal perspective on the evolving safety landscape of foundation models, illustrated with Vision-Language LLM.

rections toward a safety framework grounded in the multi-modal perspective.

2. Existing Safety Landscape

We briefly review the current safety landscape shaped by uni-modal learning. We adopt standard access definitions: white-box (full model access), black-box (query-only), and gray-box (partial access, e.g., data or architecture).

2.1. Adversarial Attacks

Attack Mechanisms. Adversarial attacks add imperceptible perturbations to inputs to mislead the model at inference, typically via loss-based optimization. They operate under white-box [11, 27, 58] or black-box settings [13], with norm constraints (l_∞ or l_2) ensuring stealth.

Defense Strategies. Defenses usually assume white-box access and a clean validation set. Common strategies include adversarial training [58], which iteratively generates and defends against adversarial examples, and randomized smoothing [15, 40], which averages predictions over noisy inputs to improve robustness against small perturbations.

2.2. Data Poisoning

Attack Mechanisms. Poisoning attackers inject malicious samples into the training data of the victim model to induce misbehavior. Label-flipping attacks degrade performance by altering training labels [92, 104], while backdoor attacks embed triggers that activate malicious behavior only under specific conditions [14, 16, 29, 43, 60, 61, 66, 70].

Defense Strategies. Defenses aim to mitigate poisoning while preserving model utility. Strategies span three stages: *Pre-training* methods sanitize the training set [63, 78]; *During-training* methods select trustworthy training samples [21, 48, 73] or apply self-supervised learning [33, 39, 84]; *Post-training* methods detect poisoned models or inputs [25, 44, 80, 88], or directly mitigate backdoors [46, 47, 53, 102].

2.3. Jailbreak

Attack Mechanisms. Jailbreak attacks craft prompts to bypass safety filters in LLMs and elicit harmful outputs. White-box methods optimize adversarial prefixes or suffixes via gradients [36, 110, 112]; gray-box attackers adjust prompts using logits [30, 106] or lightweight retraining [68, 96, 103]; black-box methods exploit model capabilities (e.g., roleplay, reasoning) [45, 82, 91] or use LLMs to generate adversarial prompts [20].

Defense Strategies. Defenses aim to enforce safety alignment across access levels. Black-box approaches filter adversarial prompts [2, 35]; white-box methods include safety fine-tuning [9, 19], reinforcement learning from human feedback (RLHF) [62], and self-correction [76].

2.4. Hallucination

Definition. Hallucination refers to generating confident but factually incorrect or unsupported outputs [34], typically caused by noisy data or biased data [7, 74], spurious correlations [59, 83], or lack of uncertainty estimation [23]. It is a reliability issue rather than an adversarial threat, lacking a formal threat model.

Mitigating Strategies. Mitigation spans three stages: *Pre-training* – data filtering [64, 87], deduplication [38], and high-quality sources [31, 69]; *Training* – RLHF [111], contrastive learning [75], and chain-of-thought [89]; *Post-training*—retrieval-augmented generation [32], prompt engineering [8], and fact-checking modules [23, 101].

3. Emerging Safety Challenges

We focus on the safety of MM-LLMs, which process diverse inputs via modality-specific encoders and use an LLM to fuse and generate outputs [85, 94, 98, 109]. Their design relies on: (1) Modality Alignment – using encoders (e.g., ViTs) to map different modalities into a shared embedding space compatible with LLMs [3, 22, 90]; and (2) Modality Fusion – using cross-attention layers in the LLM to integrate aligned embeddings for tasks like vision-conditioned text generation [1, 12, 49, 51].

These unique mechanisms introduce new safety challenges, which we organize into: (1) Compromised Modality Integration, (2) Modality Misalignment, and (3) Fused Safety Risks. We define each category and summarize relevant studies below[1].

3.1. Compromised Modality Integration

Compromised Modality Integration refers to threats inherited from individual modalities, where manipulation of a single or few modalities propagates through the integration process and compromises overall model behavior.

Several studies have extended adversarial attacks to MM-LLMs, typically by optimizing visual perturbations to disrupt the vision encoder's representations. By corrupting only the visual input, these attacks induce incorrect or harmful outputs. For example, [57, 71] craft perturbations that force MM-LLMs to produce attacker-specified text. Other works [81, 86, 99] maximize embedding distances between clean and perturbed images, distorting the model's perception. [24] delays the end-of-sequence token to increase uncertainty, while [5] shows adversarial images can leak context, bypass safety, and induce false or arbitrary outputs.

Jailbreak prompts can compromise the safety of MM-LLMs by triggering harmful outputs [20, 54, 100]. Beyond text-only attacks, adversaries may exploit other modalities to bypass safety mechanisms primarily designed for text. In such cases, text prompts remain benign while the visual modality is manipulated to trigger jailbreak behaviors. For example, FigStep [26] embeds rephrased jailbreak prompts within images. Other works transfer vulnerabilities from text to vision by optimizing visual perturbations that alone can trigger illegal responses [67, 72].

Incorporating additional modalities may exacerbate hallucination. Inputs such as vision, audio, or tabular data are often noisy, occluded, or low-resolution. When modality encoders fail to capture critical features, the LLM tends to compensate by relying on pretrained priors, filling in perceptual gaps with potentially inaccurate or fabricated information [37, 52, 56, 105].

3.2. Modality Misalignment

Modality Misalignment refers to risks where adversaries manipulate cross-modal embeddings to disrupt semantic or structural alignment, misleading the model at inference. Misalignment can be (1) *untargeted*, where the perturbed embedding deviates from the clean ones, or (2) *targeted*, where it mimics a harmful representation.

For *untargeted* misalignment, most works aim to maximize the distance between the perturbed modality (typically the image) and the clean one (typically the text) in the shared embedding space. [50] generates a universal adversarial patch by minimizing cosine similarity between visual and textual embeddings. [86] adds visual perturbations to weaken their correlation. [81] disturbs features that promote consistency and amplifies those that increase cross-modal discrepancy.

For *targeted* misalignment, [107] explores three strategies: (1) aligning the adversarial image embedding with the target text, (2) aligning it with the embedding of an image corresponding to the target text, or (3) aligning the model's output on the adversarial image with the target text. [4] introduces a sample-specific backdoor trigger and trigger-aware prompt to pull visual embeddings toward the target class. [95] uses data poisoning to make embeddings of perturbed destination images resemble those of the original concept, inducing targeted generation. [72] optimizes visual perturbations to mimic harmful embeddings (e.g., OCR-decoded jailbreak prompts or visual triggers), enabling jailbreaks. Similarly, [67] crafts adversarial visuals that increase the likelihood of harmful text output, breaking alignment.

Unlike the adversarial misalignment discussed above, hallucination-related misalignment arises from structural flaws in modality alignment. Compared to uni-modal models, hallucinations in MM-LLMs stem from deeper mismatches in the sensory-to-language pipeline. Mapping continuous sensory signals to discrete language often oversimplifies modality-specific information, leading to alignment errors and information loss [17, 42, 97, 108].

3.3. Fused Safety Risks

Fused safety risks refer to threats that exploit the fusion mechanism, where adversarial signals appear benign in isolation but become harmful when combined during modality fusion. This makes the threat harder to detect. [79] embeds backdoor triggers in both image and text modalities; the model behaves normally on each modality alone but exhibits malicious behavior when both triggers are present. [55] highlights how different modalities contribute asymmetrically to such attacks: visual inputs, due to their continuous nature, are suitable for injecting triggers, while text inputs are more effective for activating malicious responses during inference. [77] replaces textual captions with jailbreak prompts during fine-tuning, causing the model to associate harmful queries with specific clean images. At inference, the model generates harmful content when presented with both. [26] places the jailbreak prompt in the visual input while using an inciting but non-explicit text query to

[1]Note that each category may involve multiple or mixed threats, as multi-modal threats are inherently cross-modal and compound.

coax the model into providing harmful output.

4. Future Research Directions

Limitations of Classic Solutions for Multi-Modal Safety. Most existing defenses are designed for small-scale uni-modal systems and fall short in multi-modal settings due to two key challenges: **(a) Modality Heterogeneity.** Uni-modal methods often assume a known compromised modality [25, 58, 65, 93], whereas multi-modal systems can be attacked through any combination of modalities without such prior knowledge. **(b) Alignment and Interaction.** Uni-modal defenses cannot be trivially extended to multi-modal settings, as they fail to support, or may even hinder, modality alignment and fusion [39, 53, 84]. In light of these challenges, we revisit the safety landscape of multi-modal FMs by rethinking the assumptions behind threats and safety solutions, and outline future research directions for developing effective and aligned safety solutions.

4.1. New Threat Model and Assumptions

We highlight key shifts in **threat modeling**, including attacker capability, cross-modality, and attack surface.
(a) Relaxed Capability. Due to the compositional nature, attackers no longer need full-system access. Knowledge of just one modality (e.g., vision) can suffice to compromise the entire model. For example, adversarial images crafted against a visual encoder can exploit vulnerabilities in downstream alignment and fusion, leading to harmful outputs.
(b) Cross-Modality Attacks. Multi-modal models enable cross-modality attacks that exploit interactions across modalities. For instance, an adversarial image can trigger a jailbreak attack, or a malicious prompt can misguide the interpretation of visual content.
(c) Expanded Attack Surface. Unlike uni-modal models where attacks mostly target the input-output mapping, multi-modal models introduce new vulnerable stages like modality alignment and fusion. Threats can be injected internally, increasing the number of attack vectors.

Compared to uni-modal systems, **safety assumptions** in multi-modal foundation models face new constraints.
(a) Limited Knowledge of Attack Scope. In practice, defenders cannot assume the type or number of compromised modalities. For example, assuming only the visual modality is vulnerable and applying defenses designed for continuous data may leave the system exposed to text-based or cross-modal attacks. Similarly, assuming a specific attack type is unrealistic due to the compositional and emergent nature of cross-modality threats.
(b) Modality-Aware, Not Modality-Isolated Solutions. Designing defenses for each modality independently can disrupt the alignment and fusion objectives that underlie multi-modal learning. Defenses must operate with awareness of inter-modal relationships to avoid degrading model performance or introducing new inconsistencies.
(c) Access Beyond Input/Output. Defenses may need access to intermediate representations, especially in the embedding space where modality alignment occurs. Since attacks can manifest during alignment or fusion stages, effective defense mechanisms may require monitoring or intervention at these internal points.

4.2. Future Directions for Multi-Modal Safety

Multi-modal FMs introduce safety challenges that go beyond the scope of uni-modal solutions. Their expanded attack surface and cross-modality interactions call for rethinking safety strategies. We outline future directions to guide and inspire research on multi-modal safety.
(a) Modality-Agnostic Solutions. While applying separate, modality-specific defenses in an ensemble manner is feasible, it is neither scalable nor well-aligned with the integrated nature of multi-modal learning. Future work should instead pursue unified, modality-agnostic strategies for threat detection and mitigation across modalities. A promising direction is to extend defenses into the shared representation space and fusion stages, where cross-modal interactions and vulnerabilities emerge. Moreover, adapting existing methods to handle diverse input types, such as continuous (e.g., images) and discrete (e.g., text), can support a more coherent and generalizable safety framework.
(b) Tolerance to Corruption in Partial Modalities. Multi-modal foundation models should remain robust even when some modalities are compromised or unreliable. A key requirement for safety is avoiding over-reliance on any single modality, which creates a single point of failure. While modalities provide complementary information, they often include redundant signals. Future defenses should leverage this redundancy, e.g., via selective modality rejection, confidence-aware fusion, or adaptive weighting, to down-weight corrupted inputs while preserving performance.
(c) Safety-Aware Multi-Modal Designs. Effective solutions must be designed with awareness of the learning mechanisms of modality alignment and fusion. Aggressively filtering inputs or suppressing representations in a modality-specific way may disrupt cross-modal coherence, harming both performance and robustness. Future work should explore strategies that jointly optimize for safety and alignment, such as cross-modal consistency regularization, and integrate safety into the fusion process through robust fusion mechanisms. Ensuring that defenses preserve inter-modal relationships is essential for maintaining the integrity and effectiveness of multi-modal learning systems.

5. Conclusion

This paper calls for rethinking safety in multi-modal FMs, highlighting how multi-modal mechanisms fundamentally reshape the safety landscape. We identify emerging threats that existing uni-modal solutions cannot fully address, outline paradigm shifts in threat models and safety assumptions, and propose future research directions grounded in the unique characteristics of multi-modal systems. We hope this perspective encourages broader efforts toward unified safety frameworks for next-generation AI systems.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millicah, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *NeurIPS*, 2022. 1, 3

[2] Gabriel Alon and Michael Kamfonas. Detecting language model attacks with perplexity. *CoRR*, abs/2308.14132, 2023. 2

[3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 3

[4] Jiawang Bai, Kuofeng Gao, Shaobo Min, Shu-Tao Xia, Zhifeng Li, and Wei Liu. Badclip: Trigger-aware prompt learning for backdoor attacks on CLIP. In *CVPR*, 2024. 3

[5] Luke Bailey, Euan Ong, Stuart Russell, and Scott Emmons. Image hijacks: Adversarial images can control generative models at runtime. In *ICML*, 2024. 3

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1

[7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623, 2021. 2

[8] Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1946–1969, 2023. 2

[9] Federico Bianchi, Mirac Suzgun, Giuseppe Attanasio, Paul Röttger, Dan Jurafsky, Tatsunori Hashimoto, and James Zou. Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions. In *ICLR*. OpenReview.net, 2024. 2

[10] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 1

[11] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *IEEE S & P*, 2017. 2

[12] Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification . In *ICCV*, 2021. 3

[13] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *AISec@CCS*, 2017. 2

[14] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning. *arXiv:1712.05526*, 2017. 2

[15] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning*, pages 1310–1320, 2019. 2

[16] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 2019. 2

[17] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023. 3

[18] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. 1

[19] Boyi Deng, Wenjie Wang, Fuli Feng, Yang Deng, Qifan Wang, and Xiangnan He. Attack prompt generation for red teaming and defending large language models. In *Findings of EMNLP*, 2023. 2

[20] Gelei Deng, Yi Liu, Yuekang Li, Kailong Wang, Ying Zhang, Zefeng Li, Haoyu Wang, Tianwei Zhang, and Yang Liu. Masterkey: Automated jailbreaking of large language model chatbots. In *NDSS*, 2024. 2, 3

[21] Ilias Diakonikolas, Gautam Kamath, Daniel Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *ICML*, 2019. 2

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3

[23] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. 2

[24] 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. In *ICLR*, 2024. 3

[25] Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith Chinthana Ranasinghe, and Surya Nepal. STRIP: a defence against trojan attacks on deep neural networks. In *ACSAC*, 2019. 2, 4

[26] 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 *AAAI*, 2025. 3

[27] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015. 2

[28] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, and Aiesha Letman et al. The llama 3 herd of models, 2024. 1

[29] Tianyu Gu, Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks. *IEEE Access*, 2019. 2

[30] Xingang Guo, Fangxu Yu, Huan Zhang, Lianhui Qin, and Bin Hu. Cold-attack: Jailbreaking llms with stealthiness and controllability. In *ICML*, 2024. 2

[31] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *ICML*, 2020. 2

[32] Yucheng Hu and Yuxing Lu. Rag and rau: A survey on retrieval-augmented language model in natural language processing. *arXiv preprint arXiv:2404.19543*, 2024. 2

[33] Kunzhe Huang, Yiming Li, Baoyuan Wu, Zhan Qin, and Kui Ren. Backdoor defense via decoupling the training process. In *ICLR*, 2022. 2

[34] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025. 2

[35] Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping-yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models. *CoRR*, abs/2309.00614, 2023. 2

[36] Erik Jones, Anca D. Dragan, Aditi Raghunathan, and Jacob Steinhardt. Automatically auditing large language models via discrete optimization. In *ICML*, pages 15307–15329, 2023. 2

[37] Siddharth Karamcheti, Suraj Nair, Ashwin Balakrishna, Percy Liang, Thomas Kollar, and Dorsa Sadigh. Prismatic vlms: Investigating the design space of visually-conditioned language models. In *ICML*, 2024. 3

[38] Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. Deduplicating training data makes language models better. *arXiv preprint arXiv:2107.06499*, 2021. 2

[39] Alexander Levine and Soheil Feizi. Deep partition aggregation: Provable defenses against general poisoning attacks. In *ICLR*, 2021. 2, 4

[40] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *NeurIPS*, pages 9459–9469, 2019. 2

[41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 1

[42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 3

[43] Linyang Li, Demin Song, Xiaonan Li, Jiehang Zeng, Ruotian Ma, and Xipeng Qiu. Backdoor attacks on pre-trained models by layerwise weight poisoning. In *EMNLP*, 2021. 2

[44] Xi Li, Zhen Xiang, David J. Miller, and George Kesidis. Test-time detection of backdoor triggers for poisoned deep neural networks. In *IEEE ICASSP*, 2022. 2

[45] Xuan Li, Zhanke Zhou, Jianing Zhu, Jiangchao Yao, Tongliang Liu, and Bo Han. Deepinception: Hypnotize large language model to be jailbreaker. *CoRR*, abs/2311.03191, 2023. 2

[46] Xi Li, Zhen Xiang, David J. Miller, and George Kesidis. Correcting the distribution of batch normalization signals for trojan mitigation. *Neurocomputing*, 614:128752, 2025. 2

[47] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks. In *ICLR*, 2021. 2

[48] Yige Li, Xixiang Lyu, Nodens Koren, Lingjuan Lyu, Bo Li, and Xingjun Ma. Anti-Backdoor Learning: Training Clean Models on Poisoned Data. In *NeurIPS*, 2021. 2

[49] Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. CAT: Cross Attention in Vision Transformer . In *ICME*, 2022. 3

[50] Daizong Liu, Mingyu Yang, Xiaoye Qu, Pan Zhou, Xiang Fang, Keke Tang, Yao Wan, and Lichao Sun. Pandora's box: Towards building universal attackers against real-world large vision-language models. In *NeurIPS*, 2024. 3

[51] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 36:34892–34916, 2023. 1, 3

[52] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 3

[53] Kang Liu, Brendan Dolan-Gavitt, and Siddharth Garg. Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks. In *RAID*, 2018. 2, 4

[54] Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. Autodan: Generating stealthy jailbreak prompts on aligned large language models. In *ICLR*, 2024. 3

[55] Dong Lu, Tianyu Pang, Chao Du, Qian Liu, Xianjun Yang, and Min Lin. Test-time backdoor attacks on multimodal large language models. *arXiv: 2402.08577*, 2024. 3

[56] Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. *arXiv preprint arXiv:2403.03003*, 2024. 3

[57] Haochen Luo, Jindong Gu, Fengyuan Liu, and Philip Torr. An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models. In *ICLR*, 2024. 3

[58] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 2, 4

[59] Nick McKenna, Tianyi Li, Liang Cheng, Mohammad Javad Hosseini, Mark Johnson, and Mark Steedman. Sources of hallucination by large language models on inference tasks. *arXiv preprint arXiv:2305.14552*, 2023. 2

[60] Tuan Anh Nguyen and Anh Tuan Tran. Input-aware dynamic backdoor attack. In *NeurIPS*, 2020. 2

[61] Tuan Anh Nguyen and Anh Tuan Tran. WaNet - Imperceptible Warping-based Backdoor Attack. In *ICLR*, 2021. 2

[62] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *NeurIPS*, 2022. 2

[63] Andrea Paudice, Luis Muñoz-González, and Emil C. Lupu. Label sanitization against label flipping poisoning attacks. In *Proc. ECML PKDD Workshops*, 2018. 2

[64] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023. 2

[65] Fanchao Qi, Yangyi Chen, Mukai Li, Yuan Yao, Zhiyuan Liu, and Maosong Sun. ONION: A simple and effective defense against textual backdoor attacks. In *EMNLP*, 2021. 4

[66] Fanchao Qi, Yangyi Chen, Xurui Zhang, Mukai Li, Zhiyuan Liu, and Maosong Sun. Mind the style of text! adversarial and backdoor attacks based on text style transfer. In *EMNLP*, 2021. 2

[67] Xiangyu Qi, Kaixuan Huang, Ashwinee Panda, Peter Henderson, Mengdi Wang, and Prateek Mittal. Visual adversarial examples jailbreak aligned large language models. In *AAAI*, 2024. 3

[68] Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *ICLR*, 2024. 2

[69] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 2

[70] Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden Trigger Backdoor Attacks. In *AAAI*, 2020. 2

[71] Christian Schlarmann and Matthias Hein. On the adversarial robustness of multi-modal foundation models. In *ICCV - Workshops*, 2023. 3

[72] Erfan Shayegani, Yue Dong, and Nael B. Abu-Ghazaleh. Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models. In *ICLR*, 2024. 3

[73] Yanyao Shen and Sujay Sanghavi. Learning with Bad Training Data via Iterative Trimmed Loss Minimization. In *ICML*, pages 5739–5748, 2019. 2

[74] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. Towards controllable biases in language generation. *arXiv:2005.00268*, 2020. 2

[75] Weiwei Sun, Zhengliang Shi, Shen Gao, Pengjie Ren, Maarten de Rijke, and Zhaochun Ren. Contrastive learning reduces hallucination in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13618–13626, 2023. 2

[76] Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David D. Cox, Yiming Yang, and Chuang Gan. Principle-driven self-alignment of language models from scratch with minimal human supervision. In *NeurIPS*, 2023. 2

[77] Xijia Tao, Shuai Zhong, Lei Li, Qi Liu, and Lingpeng Kong. Imgtrojan: Jailbreaking vision-language models with ONE image. In *NAACL*, 2025. 3

[78] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral Signatures in Backdoor Attacks. In *NeurIPS*, 2018. 2

[79] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. Dual-key multimodal backdoors for visual question answering. In *CVPR*, 2022. 3

[80] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks. In *2019 IEEE Symposium on Security and Privacy*, 2019. 2

[81] 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 *IEEE S & P*, 2024. 3

[82] Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. Adversarial demonstration attacks on large language models. *CoRR*, abs/2305.14950, 2023. 2

[83] Tianlu Wang, Rohit Sridhar, Diyi Yang, and Xuezhi Wang. Identifying and mitigating spurious correlations for improving robustness in nlp models. *arXiv preprint arXiv:2110.07736*, 2021. 2

[84] Wenxiao Wang, Alexander Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In *ICML*, 2022. 2, 4

[85] Xiao Wang, Guangyao Chen, Guangwu Qian, Pengcheng Gao, Xiao-Yong Wei, Yaowei Wang, Yonghong Tian, and Wen Gao. Large-scale multi-modal pre-trained models: A comprehensive survey. *Machine Intelligence Research*, 20 (4):447–482, 2023. 1, 2

[86] Yubo Wang, Chaohu Liu, Yanqiu Qu, Haoyu Cao, Deqiang Jiang, and Linli Xu. Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *ACM MM*, 2024. 3

[87] Yudong Wang, Zixuan Fu, Jie Cai, Peijun Tang, Hongya Lyu, Yewei Fang, Zhi Zheng, Jie Zhou, Guoyang Zeng, Chaojun Xiao, et al. Ultra-fineweb: Efficient data filtering and verification for high-quality llm training data. *arXiv preprint arXiv:2505.05427*, 2025. 2

[88] Zhenting Wang, Kai Mei, Juan Zhai, and Shiqing Ma. UNICORN: A unified backdoor trigger inversion framework. In *ICLR*, 2023. 2

[89] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 2

[90] Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *ACM Web Conference*, 2023. 3

[91] Zeming Wei, Yifei Wang, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *CoRR*, abs/2310.06387, 2023. 2

[92] Han Xiao, Huang Xiao, and Claudia Eckert. Adversarial label flips attack on support vector machines. In *ECAI*, 2012. 2

[93] Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496, 2023. 4

[94] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12113–12132, 2023. 1, 2

[95] Yuancheng Xu, Jiarui Yao, Manli Shu, Yanchao Sun, Zichu Wu, Ning Yu, Tom Goldstein, and Furong Huang. Shadowcast: Stealthy data poisoning attacks against vision-language models. In *NeurIPS*, 2024. 3

[96] Xianjun Yang, Xiao Wang, Qi Zhang, Linda R. Petzold, William Yang Wang, Xun Zhao, and Dahua Lin. Shadow alignment: The ease of subverting safely-aligned language models. *CoRR*, abs/2310.02949, 2023. 2

[97] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *CVPR*, 2024. 3

[98] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), 2024. 1, 2

[99] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. VLATTACK: multimodal adversarial attacks on vision-language tasks via pre-trained models. In *NeurIPS*, 2023. 3

[100] Zhiyuan Yu, Xiaogeng Liu, Shunning Liang, Zach Cameron, Chaowei Xiao, and Ning Zhang. Don't listen to me: Understanding and exploring jailbreak prompts of large language models. In *USENIX Security*, 2024. 3

[101] Yige Yuan, Bingbing Xu, Hexiang Tan, Fei Sun, Teng Xiao, Wei Li, Huawei Shen, and Xueqi Cheng. Fact-level confidence calibration and self-correction. *arXiv preprint arXiv:2411.13343*, 2024. 2

[102] Yi Zeng, Si Chen, Won Park, Zhuoqing Mao, Ming Jin, and Ruoxi Jia. Adversarial Unlearning of Backdoors via Implicit Hypergradient. In *ICLR*, 2022. 2

[103] Qiusi Zhan, Richard Fang, Rohan Bindu, Akul Gupta, Tatsunori Hashimoto, and Daniel Kang. Removing RLHF protections in GPT-4 via fine-tuning. In *NAACL*, pages 681–687, 2024. 2

[104] Hongpo Zhang, Ning Cheng, Yang Zhang, and Zhanbo Li. Label flipping attacks against naive bayes on spam filtering systems. *Appl. Intell.*, 51(7):4503–4514, 2021. 2

[105] Yipeng Zhang, Yifan Liu, Zonghao Guo, Yidan Zhang, Xuesong Yang, Chi Chen, Jun Song, Bo Zheng, Yuan Yao, Zhiyuan Liu, et al. Llava-uhd v2: an mllm integrating high-resolution feature pyramid via hierarchical window transformer. *arXiv preprint arXiv:2412.13871*, 2024. 3

[106] Xuandong Zhao, Xianjun Yang, Tianyu Pang, Chao Du, Lei Li, Yu-Xiang Wang, and William Yang Wang. Weak-to-strong jailbreaking on large language models. *CoRR*, abs/2401.17256, 2024. 2

[107] Yunqing Zhao, Tianyu Pang, Chao Du, Xiao Yang, Chongxuan Li, Ngai-Man Cheung, and Min Lin. On evaluating adversarial robustness of large vision-language models. In *NeurIPS*, 2023. 3

[108] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 3

[109] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, 95:306–325, 2023. 1, 2

[110] Sicheng Zhu, Ruiyi Zhang, Bang An, Gang Wu, Joe Barrow, Zichao Wang, Furong Huang, Ani Nenkova, and Tong Sun. Autodan: Interpretable gradient-based adversarial attacks on large language models. *CoRR*, abs/2310.15140, 2023. 2

[111] Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019. 2

[112] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *CoRR*, abs/2307.15043, 2023. 2