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 integration of multiple modalities, multi-modal foundation models are increasingly deployed in domains such as autonomous driving, healthcare, and virtual assistants. Unlike classic uni-modal learning, these models rely on modality alignment and fusion to effectively leverage cross-modal information. However, this uniqueness also introduces novel safety threats that uni-modal safety frameworks fail to address. Existing solutions often assume prior knowledge of compromised modalities and overlook the complex interactions inherent in multi-modal systems. We argue that the safety landscape must be revisited from a multi-modal perspective. To support this vision, we identify emerging threats unique to multi-modal systems, categorize existing studies accordingly, and outline promising future research directions – including the redefinition of threat models and safety assumptions, and the development of defense strategies aligned with multi-modal design. Our goal is to open a new trajectory for trustworthy multi-modal foundation models.**

***Index Terms*—Multi-modal Safety, Robustness and Attacks, Reliability**

I. Introduction

Modern multi-modal learning 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]–[6]. 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 [7]–[11]. 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 foundation models evolve from uni-modal to multi-modal architectures, the landscape of machine learning safety is undergoing a fundamental transformation. The unique characteristics of multi-modal learning give rise to a new set of safety challenges. First, incorporating additional modalities introduces modality-specific vulnerabilities inherent to each data type. Second, adversarial misalignment between modalities can lead to semantic inconsistencies or unintended behaviors. Third, the fusion can be exploited. Malicious signals that appear benign in isolation may trigger harmful behavior only when modalities are combined. These emerging risks highlight the need to revisit existing safety frameworks through the lens of multi-modal integration.

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 existing methods rely on prior knowledge of the compromised modality, such as whether images, text, or audio are affected, to design appropriate defenses. This implicit assumption is infeasible in multi-modal learning, as one cannot assume which modalities are compromised or how many are affected. Besides, these methods are not explicitly designed to align with the core objectives of multi-modal learning, modality alignment and fusion, and may inadvertently degrade overall performance. This disconnect introduces critical blind spots in current safety solutions, limiting their applicability and effectiveness in multi-modal settings.

Given the broad deployment of multi-modal foundation models and the widening gap in their safety research, we propose a **BlueSky idea: to rethink the safety landscape of foundation models through the lens of multi-modal learning.** This vision calls for redefining threat models and assumptions of safety solutions, identifying emerging safety risks unique to multi-modal systems, and developing defense strategies that align with the objectives of modality alignment and fusion. By grounding safety in the principles of multi-modal learning, we aim to push the boundaries of both safety theory and system design. This BlueSky idea seeks to shift the community's perspective and open a new trajectory for trustworthy AI research, rooted in the realities of modern multi-modal foundation models.

This paper examines the safety landscape of multi-modal large language models (MM-LLMs), highlighting emerging threats, threat models, and defense strategies, with a focus on their differences from the uni-modal setting. The remainder of the paper is structured as follows. Section II reviews the current safety landscape rooted in uni-modal learning, covering adversarial attacks, poisoning attacks, jailbreaks, and hallucinations. Section III introduces the unique characteristics of multi-modal learning and the emerging safety challenges it presents. We also provide a brief categorization of existing works in relation to these challenges. Finally, Section IV outlines future research directions toward a safety framework grounded in the perspective of multi-modal learning.

*Corresponding author.

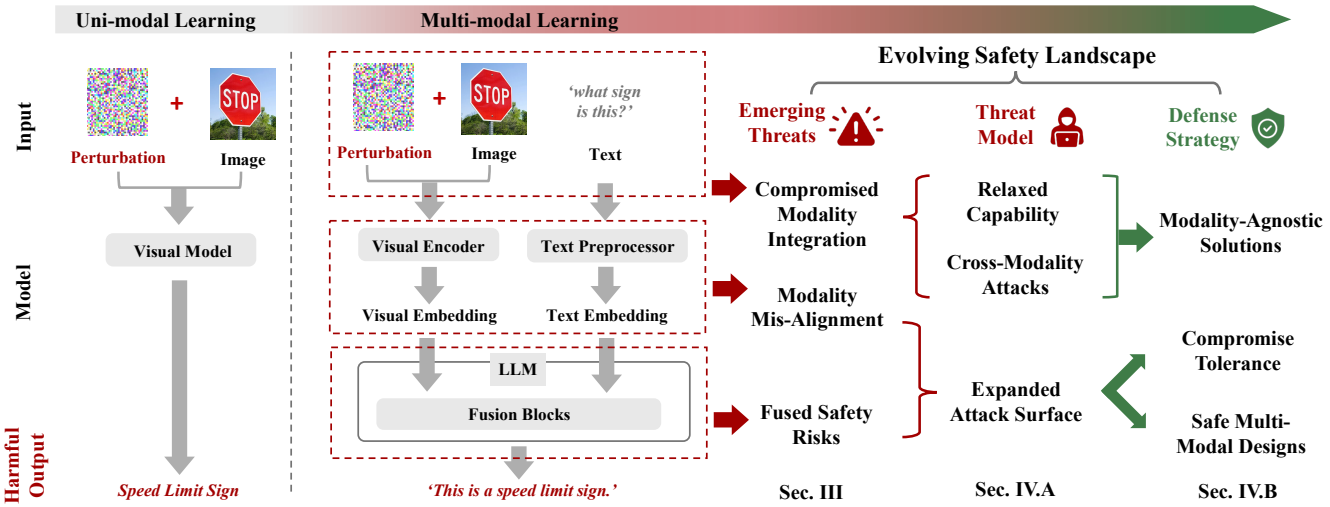

Fig. 1. A multi-modal perspective on the evolving safety landscape of foundation models, illustrated with Vision-Language LLM.

II. Existing Safety Landscape

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

A. Adversarial Attacks

- **Threat Model and 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 [12]–[14] or black-box settings [15], with norm constraints (l_∞ or l_2) ensuring stealth.
- **Defense Assumptions and Strategies.** Defenses usually assume white-box access and a clean validation set. Common strategies include adversarial training [13], which iteratively generates and defends against adversarial examples, and randomized smoothing [16], [17], which averages predictions over noisy inputs to improve robustness against small perturbations.

B. Data Poisoning

- **Threat Model and 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 [18], [19], while backdoor attacks embed triggers that activate malicious behavior only under specific conditions [20]–[27].
- **Defense Assumptions and Strategies.** Defenses aim to mitigate poisoning while preserving model utility. Strategies span three stages: *Pre-training* methods sanitize the training set [28], [29]; *During-training* methods select trustworthy training samples [30]–[32] or apply self-supervised learning [33]–[35]; *Post-training* methods detect poisoned models or inputs [36]–[39], or directly mitigate backdoors [40]–[43].

C. Jailbreak

- **Threat Model and 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 [44]–[46]; gray-box attackers adjust prompts using logits [47], [48] or lightweight retraining [49]–[51]; black-box methods exploit model capabilities (e.g., roleplay, reasoning) [52]–[54] or use LLMs to generate adversarial prompts [55].
- **Defense Assumptions and Strategies.** Defenses aim to enforce safety alignment across access levels. Black-box approaches filter adversarial prompts [56], [57]; white-box methods include safety fine-tuning [58], [59], reinforcement learning from human feedback (RLHF) [60], and self-correction [61].

D. Hallucination

- **Definition and Characteristics.** Hallucination refers to generating confident but factually incorrect or unsupported outputs [62], typically caused by noisy data or biased data [63], [64], spurious correlations [65], [66], or lack of uncertainty estimation [67]. 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 [68], [69], deduplication [70], and high-quality sources [71], [72]; *Training* – RLHF [73], contrastive learning [74], and chain-of-thought [75]; *Post-training*—retrieval-augmented generation [76], prompt engineering [77], and fact-checking modules [67], [78].

III. 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 [7]–[10]. Their design relies on: (1) Modality Alignment – using encoders (e.g., ViTs) to map different modalities into a shared embedding space compatible with LLMs while preserving modality-specific information [79]–[81]; (2) Modality Fusion – applying cross-attention layers in the LLM to integrate aligned embeddings and support tasks like vision-conditioned text generation [2], [3], [82], [83].

The uniqueness of multi-modal learning introduces emerging safety challenges, which we organize as: (1) Compromised Modality Integration, (2) Modality Misalignment, and (3) Fused Safety Risks. We provide detailed definitions and summarize existing studies below. Note that each category may involve multiple or mixed threats, as multi-modal threats are inherently cross-modal and compound.

A. 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. The mechanisms are similar to classic vision attacks, typically optimizing visual perturbations to disrupt the vision encoder's representations. By corrupting only the visual input, these attacks cause the model to generate incorrect or harmful outputs. For example, [84], [85] craft visual perturbations that force MM-LLMs to produce attacker-specified text. Other works [86]–[88] maximize the embedding distance between clean and perturbed images, distorting the model's perception. Also, [89] designs perturbations to delay the end-of-sequence token, increasing output uncertainty, while [90] shows that adversarial images can induce arbitrary outputs, leak context, bypass safety filters, and cause the model to accept falsehoods.

Jailbreak prompts can also compromise the safety of MM-LLMs, causing them to generate harmful outputs [55], [91], [92]. Beyond text-only attacks, adversaries may exploit additional modalities to bypass safety alignment mechanisms primarily designed for text and LLMs. In this case, the text prompts remain benign while the visual modality is manipulated to trigger harmful outputs. For example, FigStep [93] embeds rephrased jailbreak prompts within images. Other works further transfer vulnerabilities from text to vision: [94], [95] optimize visual perturbations so that the adversarial image alone can induce illegal responses from the model.

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 [96]–[99].

B. Modality Misalignment

Modality Misalignment refers to risks where adversaries manipulate cross-modal embeddings to disrupt semantic or structural alignment between modalities, misleading the model during inference. Misalignment can be (1) *untargeted*, where the perturbed modality's embedding deviates from clean modalities, or (2) *targeted*, where the embedding is manipulated to resemble a harmful representation.

For *untargeted* misalignment, most works aim to maximize the distance between the perturbed modality (typically the image) and the clean modality (typically the text) in the shared embedding space. For instance, [100] generates a universal adversarial patch by minimizing the cosine similarity between visual and textual embeddings, thereby disrupting cross-modal alignment. [88] introduces visual adversarial perturbations to weaken the correlation between visual and textual representations. [87] focuses on disturbing features that promote modality consistency and enhancing those that increase modality discrepancy, leading to greater misalignment across modalities.

For *targeted* misalignment, [101] 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. [102] introduces a sample-specific backdoor trigger and a trigger-aware context prompt to bring the visual embedding closer to the target class. [103] uses data poisoning to manipulate the model into generating text aligned with a destination concept image when presented with the original concept image. The poisoning process makes the embeddings of perturbed destination concept images close to those of the original concept image. [94] optimizes visual perturbations to mimic embeddings of harmful content (e.g., OCR-decoded jailbreak prompts, visual triggers, or combinations of both), enabling jailbreaks. Similarly, [95] crafts adversarial visuals to maximize the likelihood of harmful text generation, breaking safety 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 [104]–[107].

C. 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. [108] 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. [109] 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. [110] 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. [93] places the jailbreak prompt in the visual input while using an inciting but non-explicit text query to coax the model into providing harmful output.

IV. Future Research Directions

Limitations of Classic Solutions for Multi-Modal Safety. While most existing defenses target small-scale uni-modal

systems, they fall short in multi-modal settings due to the following challenges: **(a) Modality Heterogeneity.** Uni-modal methods often assume a specific compromised modality, whereas multi-modal systems can be attacked through any combination of modalities without prior knowledge of which ones are affected [13], [38], [111], [112]. **(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, which are central to multi-modal AI [34], [35], [40].

Given these, we revisit the safety landscape of multi-modal foundation models. We begin by rethinking the assumptions underlying both attacks and defenses in this new context. Building on this foundation, we then explore future research directions for developing effective and aligned safety solutions.

A. New Threat Model and Assumptions for Safety Solutions

We highlight the key differences in **threat modeling**, focusing on the attacker's ability, novel cross-modality attack, and attack surface.

- **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.
- **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.
- **Expanded Attack Surface.** Unlike uni-modal models where attacks mostly target the input-output mapping, multi-modal models introduce new vulnerable stages, such as modality alignment and fusion. Threats can be injected at these internal stages, increasing the number of attack vectors.

Compared to uni-modal systems, **safety assumptions** in multi-modal foundation models face new constraints.

- **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.
- **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.
- **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.

B. Future Directions for Multi-Modal Safety

Multi-modal foundation models pose fundamentally new safety challenges beyond the scope of uni-modal safety solutions. Their expanded attack surface, cross-modality threats, and complex multi-modal mechanisms call for a rethinking of safety strategies. We outline several future directions to guide and inspire further exploration of multi-modal safety.

- **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.
- **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 or ignore corrupted inputs while preserving performance.
- **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.

V. Conclusion

This paper calls for rethinking safety in multi-modal foundation models, 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] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *ICML*, 2022.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millicah, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *NeurIPS*, 2022.

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *NeurIPS*, vol. 36, pp. 34 892–34 916, 2023.

[4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *NeurIPS*, 2020.

[5] DeepSeek-AI, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025. [Online]. Available: https://arxiv.org/abs/2501.12948

[6] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, and A. L. et al, "The llama 3 herd of models," 2024.

[7] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, 2024.

[8] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, and W. Gao, "Large-scale multi-modal pre-trained models: A comprehensive survey," *Machine Intelligence Research*, vol. 20, no. 4, pp. 447–482, 2023.

[9] L. Zhu, Z. Zhu, C. Zhang, Y. Xu, and X. Kong, "Multimodal sentiment analysis based on fusion methods: A survey," *Information Fusion*, vol. 95, pp. 306–325, 2023.

[10] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12 113–12 132, 2023.

[11] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[12] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *ICLR*, 2015.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.

[14] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE S & P*, 2017.

[15] P. Chen, H. Zhang, Y. Sharma, J. Yi, and C. Hsieh, "ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *AISec@CCS*, 2017.

[16] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97, 2019, pp. 1310–1320.

[17] B. Li, C. Chen, W. Wang, and L. Carin, "Certified adversarial robustness with additive noise," in *NeurIPS*, 2019, pp. 9459–9469.

[18] H. Xiao, H. Xiao, and C. Eckert, "Adversarial label flips attack on support vector machines," in *ECAI*, 2012.

[19] H. Zhang, N. Cheng, Y. Zhang, and Z. Li, "Label flipping attacks against naive bayes on spam filtering systems," *Appl. Intell.*, vol. 51, no. 7, pp. 4503–4514, 2021.

[20] T. A. Nguyen and A. T. Tran, "WaNet - Imperceptible Warping-based Backdoor Attack," in *ICLR*, 2021.

[21] ——, "Input-aware dynamic backdoor attack," in *NeurIPS*, 2020.

[22] A. Saha, A. Subramanya, and H. Pirsiavash, "Hidden Trigger Backdoor Attacks," in *AAAI*, 2020.

[23] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks," *IEEE Access*, 2019.

[24] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted Backdoor Attacks on Deep Learning Systems Using Data Poisoning," *arXiv:1712.05526*, 2017.

[25] L. Li, D. Song, X. Li, J. Zeng, R. Ma, and X. Qiu, "Backdoor attacks on pre-trained models by layerwise weight poisoning," in *EMNLP*, 2021.

[26] J. Dai, C. Chen, and Y. Li, "A backdoor attack against lstm-based text classification systems," *IEEE Access*, 2019.

[27] F. Qi, Y. Chen, X. Zhang, M. Li, Z. Liu, and M. Sun, "Mind the style of text! adversarial and backdoor attacks based on text style transfer," in *EMNLP*, 2021.

[28] B. Tran, J. Li, and A. Madry, "Spectral Signatures in Backdoor Attacks," in *NeurIPS*, 2018.

[29] A. Paudice, L. Muñoz-González, and E. C. Lupu, "Label sanitization against label flipping poisoning attacks," in *Proc. ECML PKDD Workshops*, 2018.

[30] I. Diakonikolas, G. Kamath, D. Kane, J. Li, J. Steinhardt, and A. Stewart, "Sever: A robust meta-algorithm for stochastic optimization," in *ICML*, 2019.

[31] Y. Shen and S. Sanghavi, "Learning with Bad Training Data via Iterative Trimmed Loss Minimization," in *ICML*, 2019, pp. 5739–5748.

[32] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Anti-Backdoor Learning: Training Clean Models on Poisoned Data," in *NeurIPS*, 2021.

[33] K. Huang, Y. Li, B. Wu, Z. Qin, and K. Ren, "Backdoor defense via decoupling the training process," in *ICLR*, 2022.

[34] A. Levine and S. Feizi, "Deep partition aggregation: Provable defenses against general poisoning attacks," in *ICLR*, 2021.

[35] W. Wang, A. Levine, and S. Feizi, "Improved certified defenses against data poisoning with (deterministic) finite aggregation," in *ICML*, 2022.

[36] B. Wang, Y. Yao, S. Shan, H. Li, B. Viswanath, H. Zheng, and B. Y. Zhao, "Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks," in *2019 IEEE Symposium on Security and Privacy*, 2019.

[37] Z. Wang, K. Mei, J. Zhai, and S. Ma, "UNICORN: A unified backdoor trigger inversion framework," in *ICLR*, 2023.

[38] Y. Gao, C. Xu, D. Wang, S. Chen, D. C. Ranasinghe, and S. Nepal, "STRIP: a defence against trojan attacks on deep neural networks," in *ACSAC*, 2019.

[39] X. Li, Z. Xiang, D. J. Miller, and G. Kesidis, "Test-time detection of backdoor triggers for poisoned deep neural networks," in *IEEE ICASSP*, 2022.

[40] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-Pruning: Defending Against Backdooring Attacks on Deep Neural Networks," in *RAID*, 2018.

[41] Y. Li, X. Lyu, N. Koren, L. Lyu, B. Li, and X. Ma, "Neural Attention Distillation: Erasing Backdoor Triggers from Deep Neural Networks," in *ICLR*, 2021.

[42] Y. Zeng, S. Chen, W. Park, Z. Mao, M. Jin, and R. Jia, "Adversarial Unlearning of Backdoors via Implicit Hypergradient," in *ICLR*, 2022.

[43] X. Li, Z. Xiang, D. J. Miller, and G. Kesidis, "Correcting the distribution of batch normalization signals for trojan mitigation," *Neurocomputing*, vol. 614, p. 128752, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231224015236

[44] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," *CoRR*, vol. abs/2307.15043, 2023.

[45] E. Jones, A. D. Dragan, A. Raghunathan, and J. Steinhardt, "Automatically auditing large language models via discrete optimization," in *ICML*, vol. 202, 2023, pp. 15 307–15 329.

[46] S. Zhu, R. Zhang, B. An, G. Wu, J. Barrow, Z. Wang, F. Huang, A. Nenkova, and T. Sun, "Autodan: Interpretable gradient-based adversarial attacks on large language models," *CoRR*, vol. abs/2310.15140, 2023.

[47] X. Guo, F. Yu, H. Zhang, L. Qin, and B. Hu, "Cold-attack: Jailbreaking llms with stealthiness and controllability," in *ICML*, 2024.

[48] X. Zhao, X. Yang, T. Pang, C. Du, L. Li, Y. Wang, and W. Y. Wang, "Weak-to-strong jailbreaking on large language models," *CoRR*, vol. abs/2401.17256, 2024.

[49] X. Qi, Y. Zeng, T. Xie, P. Chen, R. Jia, P. Mittal, and P. Henderson, "Fine-tuning aligned language models compromises safety, even when users do not intend to!" in *ICLR*, 2024.

[50] Q. Zhan, R. Fang, R. Bindu, A. Gupta, T. Hashimoto, and D. Kang, "Removing RLHF protections in GPT-4 via fine-tuning," in *NAACL*, 2024, pp. 681–687.

[51] X. Yang, X. Wang, Q. Zhang, L. R. Petzold, W. Y. Wang, X. Zhao, and D. Lin, "Shadow alignment: The ease of subverting safely-aligned language models," *CoRR*, vol. abs/2310.02949, 2023.

[52] J. Wang, Z. Liu, K. H. Park, M. Chen, and C. Xiao, "Adversarial demonstration attacks on large language models," *CoRR*, vol. abs/2305.14950, 2023.

[53] Z. Wei, Y. Wang, and Y. Wang, "Jailbreak and guard aligned language models with only few in-context demonstrations," *CoRR*, vol. abs/2310.06387, 2023.

[54] X. Li, Z. Zhou, J. Zhu, J. Yao, T. Liu, and B. Han, "Deepinception: Hypnotize large language model to be jailbreaker," *CoRR*, vol. abs/2311.03191, 2023.

[55] G. Deng, Y. Liu, Y. Li, K. Wang, Y. Zhang, Z. Li, H. Wang, T. Zhang, and Y. Liu, "Masterkey: Automated jailbreaking of large language model chatbots," in *NDSS*, 2024.

[56] G. Alon and M. Kamfonas, "Detecting language model attacks with perplexity," *CoRR*, vol. abs/2308.14132, 2023.

[57] N. Jain, A. Schwarzschild, Y. Wen, G. Somepalli, J. Kirchenbauer, P. Chiang, M. Goldblum, A. Saha, J. Geiping, and T. Goldstein, "Baseline defenses for adversarial attacks against aligned language models," *CoRR*, vol. abs/2309.00614, 2023.

[58] F. Bianchi, M. Suzgun, G. Attanasio, P. Röttger, D. Jurafsky, T. Hashimoto, and J. Zou, "Safety-tuned llamas: Lessons from improving the safety of large language models that follow instructions," in *ICLR*. OpenReview.net, 2024.

[59] B. Deng, W. Wang, F. Feng, Y. Deng, Q. Wang, and X. He, "Attack prompt generation for red teaming and defending large language models," in *Findings of EMNLP*, 2023.

[60] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *NeurIPS*, 2022.

[61] Z. Sun, Y. Shen, Q. Zhou, H. Zhang, Z. Chen, D. D. Cox, Y. Yang, and C. Gan, "Principle-driven self-alignment of language models from scratch with minimal human supervision," in *NeurIPS*, 2023.

[62] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin *et al.*, "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions," *ACM Transactions on Information Systems*, vol. 43, no. 2, pp. 1–55, 2025.

[63] E. M. Bender, T. Gebru, A. McMillan-Major, and S. 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*, 2021, pp. 610–623.

[64] E. Sheng, K.-W. Chang, P. Natarajan, and N. Peng, "Towards controllable biases in language generation," *arXiv:2005.00268*, 2020.

[65] N. McKenna, T. Li, L. Cheng, M. J. Hosseini, M. Johnson, and M. Steedman, "Sources of hallucination by large language models on inference tasks," *arXiv preprint arXiv:2305.14552*, 2023.

[66] T. Wang, R. Sridhar, D. Yang, and X. Wang, "Identifying and mitigating spurious correlations for improving robustness in nlp models," *arXiv preprint arXiv:2110.07736*, 2021.

[67] S. Farquhar, J. Kossen, L. Kuhn, and Y. Gal, "Detecting hallucinations in large language models using semantic entropy," *Nature*, vol. 630, no. 8017, pp. 625–630, 2024.

[68] Y. Wang, Z. Fu, J. Cai, P. Tang, H. Lyu, Y. Fang, Z. Zheng, J. Zhou, G. Zeng, C. Xiao *et al.*, "Ultra-fineweb: Efficient data filtering and verification for high-quality llm training data," *arXiv preprint arXiv:2505.05427*, 2025.

[69] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, "The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only," *arXiv preprint arXiv:2306.01116*, 2023.

[70] K. Lee, D. Ippolito, A. Nystrom, C. Zhang, D. Eck, C. Callison-Burch, and N. Carlini, "Deduplicating training data makes language models better," *arXiv preprint arXiv:2107.06499*, 2021.

[71] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *ICML*, 2020.

[72] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[73] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

[74] W. Sun, Z. Shi, S. Gao, P. Ren, M. de Rijke, and Z. Ren, "Contrastive learning reduces hallucination in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 11, 2023, pp. 13 618–13 626.

[75] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.

[76] Y. Hu and Y. Lu, "Rag and rau: A survey on retrieval-augmented language model in natural language processing," *arXiv preprint arXiv:2404.19543*, 2024.

[77] L. Beurer-Kellner, M. Fischer, and M. Vechev, "Prompting is programming: A query language for large language models," *Proceedings of the ACM on Programming Languages*, vol. 7, no. PLDI, pp. 1946–1969, 2023.

[78] Y. Yuan, B. Xu, H. Tan, F. Sun, T. Xiao, W. Li, H. Shen, and X. Cheng, "Fact-level confidence calibration and self-correction," *arXiv preprint arXiv:2411.13343*, 2024.

[79] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.

[80] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *ECCV*, 2022.

[81] W. Wei, C. Huang, L. Xia, and C. Zhang, "Multi-modal self-supervised learning for recommendation," in *ACM Web Conference*, 2023.

[82] C.-F. R. Chen, Q. Fan, and R. Panda, " CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification ," in *ICCV*, 2021.

[83] H. Lin, X. Cheng, X. Wu, and D. Shen, " CAT: Cross Attention in Vision Transformer ," in *ICME*, 2022.

[84] H. Luo, J. Gu, F. Liu, and P. Torr, "An image is worth 1000 lies: Transferability of adversarial images across prompts on vision-language models," in *ICLR*, 2024.

[85] C. Schlarmann and M. Hein, "On the adversarial robustness of multi-modal foundation models," in *ICCV - Workshops*, 2023.

[86] Z. Yin, M. Ye, T. Zhang, T. Du, J. Zhu, H. Liu, J. Chen, T. Wang, and F. Ma, "VLATTACK: multimodal adversarial attacks on vision-language tasks via pre-trained models," in *NeurIPS*, 2023.

[87] H. Wang, K. Dong, Z. Zhu, H. Qin, A. Liu, X. Fang, J. Wang, and X. Liu, "Transferable multimodal attack on vision-language pre-training models," in *IEEE S & P*, 2024.

[88] Y. Wang, C. Liu, Y. Qu, H. Cao, D. Jiang, and L. Xu, "Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models," in *ACM MM*, 2024.

[89] K. Gao, Y. Bai, J. Gu, S. Xia, P. Torr, Z. Li, and W. Liu, "Inducing high energy-latency of large vision-language models with verbose images," in *ICLR*, 2024.

[90] L. Bailey, E. Ong, S. Russell, and S. Emmons, "Image hijacks: Adversarial images can control generative models at runtime," in *ICML*, 2024.

[91] X. Liu, N. Xu, M. Chen, and C. Xiao, "Autodan: Generating stealthy jailbreak prompts on aligned large language models," in *ICLR*, 2024.

[92] Z. Yu, X. Liu, S. Liang, Z. Cameron, C. Xiao, and N. Zhang, "Don't listen to me: Understanding and exploring jailbreak prompts of large language models," in *USENIX Security*, 2024.

[93] 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," in *AAAI*, 2025.

[94] E. Shayegani, Y. Dong, and N. B. Abu-Ghazaleh, "Jailbreak in pieces: Compositional adversarial attacks on multi-modal language models," in *ICLR*, 2024.

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

[96] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Improved reasoning, ocr, and world knowledge," January 2024. [Online]. Available: https://llava-vl.github.io/blog/2024-01-30-llava-next/

[97] S. Karamcheti, S. Nair, A. Balakrishna, P. Liang, T. Kollar, and D. Sadigh, "Prismatic vlms: Investigating the design space of visually-conditioned language models," in *ICML*, 2024.

[98] G. Luo, Y. Zhou, Y. Zhang, X. Zheng, X. Sun, and R. Ji, "Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models," *arXiv preprint arXiv:2403.03003*, 2024.

[99] Y. Zhang, Y. Liu, Z. Guo, Y. Zhang, X. Yang, C. Chen, J. Song, B. Zheng, Y. Yao, Z. Liu *et al.*, "Llava-uhd v2: an mllm integrating

high-resolution feature pyramid via hierarchical window transformer," *arXiv preprint arXiv:2412.13871*, 2024.
[100] D. Liu, M. Yang, X. Qu, P. Zhou, X. Fang, K. Tang, Y. Wan, and L. Sun, "Pandora's box: Towards building universal attackers against real-world large vision-language models," in *NeurIPS*, 2024.
[101] Y. Zhao, T. Pang, C. Du, X. Yang, C. Li, N. Cheung, and M. Lin, "On evaluating adversarial robustness of large vision-language models," in *NeurIPS*, 2023.
[102] J. Bai, K. Gao, S. Min, S. Xia, Z. Li, and W. Liu, "Badclip: Trigger-aware prompt learning for backdoor attacks on CLIP," in *CVPR*, 2024.
[103] Y. Xu, J. Yao, M. Shu, Y. Sun, Z. Wu, N. Yu, T. Goldstein, and F. Huang, "Shadowcast: Stealthy data poisoning attacks against vision-language models," in *NeurIPS*, 2024.
[104] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.
[105] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023. [Online]. Available: https://arxiv.org/abs/2305.06500
[106] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, "Minigpt-4: Enhancing vision-language understanding with advanced large language models," *arXiv preprint arXiv:2304.10592*, 2023.
[107] Q. Ye, H. Xu, J. Ye, M. Yan, A. Hu, H. Liu, Q. Qian, J. Zhang, and F. Huang, "mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration," in *CVPR*, 2024.
[108] M. Walmer, K. Sikka, I. Sur, A. Shrivastava, and S. Jha, "Dual-key multimodal backdoors for visual question answering," in *CVPR*, 2022.
[109] D. Lu, T. Pang, C. Du, Q. Liu, X. Yang, and M. Lin, "Test-time backdoor attacks on multimodal large language models," *arXiv: 2402.08577*, 2024.
[110] X. Tao, S. Zhong, L. Li, Q. Liu, and L. Kong, "Imgtrojan: Jailbreaking vision-language models with ONE image," in *NAACL*, 2025.
[111] Y. Xie, J. Yi, J. Shao, J. Curl, L. Lyu, Q. Chen, X. Xie, and F. Wu, "Defending chatgpt against jailbreak attack via self-reminders," *Nature Machine Intelligence*, vol. 5, no. 12, pp. 1486–1496, 2023. [Online]. Available: https://doi.org/10.1038/s42256-023-00765-8
[112] F. Qi, Y. Chen, M. Li, Y. Yao, Z. Liu, and M. Sun, "ONION: A simple and effective defense against textual backdoor attacks," in *EMNLP*, 2021.