Towards Safe Multi-Modal Learning: Unique Challenges and Future Directions

Xi Li
University of Alabama at Birmingham
XiLiUAB@uab.edu

Manling Li
Northwestern University
manling.li@northwestern.edu

Muchao Ye
The University of Iowa
muchao-ye@uiowa.edu

Abstract

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. However, the inherent complexities of multi-modal learning introduce unique safety challenges that existing frameworks, primarily designed for uni-modal models, fail to address. This tutorial explores the emerging safety risks in multi-modal learning and provides insights into future research directions. We begin by examining the unique characteristics of multi-modal learning – modality integration, alignment, and fusion. We then review existing safety studies across adversarial attacks, data poisoning, jailbreak exploits, and hallucinations. Next, we analyze emerging safety threats exploiting multi-modal challenges, including risks from additional modalities, modality misalignment, and fused representations. Finally, we discuss potential directions for enhancing the safety of multi-modal learning. As multi-modal learning expands, addressing its safety risks is crucial. This tutorial lays the foundation for understanding these challenges and fostering discussions on trustworthy systems.

Subject Areas: Multi-modal AI safety (Primary), Multi-modal Learning, Adversarial Robustness, Data Poisoning, Hallucination
Duration: Half Days
Format: In-Person Attendence

1. EXPECTED TARGET AUDIENCE

This tutorial is designed for researchers, practitioners, and students interested in the challenges and advancements in safe multi-modal learning. Attendees shall have a foundational understanding of machine learning and computer vision, though familiarity with multi-modal learning is not required. The session will be particularly valuable for those working on applications involving vision-language models, autonomous systems, and human-centered AI, where ensuring safety and robustness is critical. With an expected audience of approximately 100 participants, this tutorial will provide a mix of theoretical insights and practical guidelines to equip attendees with the necessary tools to navigate the complexities of building safe multi-modal learning.

2. TUTORIAL OUTLINE

- **Part 1: Introduction and Background** (30 mins)
  - ➤ Uniqueness of Multi-Modal Learning
  - ➤ Revisiting Safety Studies
  - ➤ Security Challenges of Multi-Modal Learning
- **Part 2: Modality Integration Exploitation** (45 mins)
  - ➤ Adversarial Perturbations in Image Modality
  - ➤ Jailbreak Prompts in Text Modality
- Break (10 mins)
- **Part 3: Modality Misalignment** (45 mins)
  - ➤ Hallucinations Caused by Modality Misalignment
  - ➤ Attacks Exploiting Embedding Misalignment
- **Part 4: Fused Vulnerabilities** (45 mins)
  - ➤ Poisoning Attacks via Cross-Modal Fusion
  - ➤ Cross-Modal Threat Switching
- Break (10 mins)
- **Part 5: Conclusion and Future Directions** (25 mins)
  - ➤ Limitations of Existing Defenses
  - ➤ Future Directions in Secure Multi-Modal Learning

3. TUTORIAL DESCRIPTIONS

Part 1: Introduction and Background

Multi-modal AI integrates diverse modalities to enhance understanding and decision-making in applications like autonomous vehicles and medical diagnostics. However, its complexity introduces safety challenges that existing frameworks fail to address. To lay a foundation, we begin this tutorial by distinguishing multi-modal from uni-modal learning and reviewing existing safety studies.

- **Uniqueness of Multi-Modal Learning.** The uniqueness of multi-modal learning lies in *modality alignment and modality fusion*, enabling the integration of diverse modalities with distinct statistical and structural properties. We will introduce classic multi-modal AI architectures and strategies for effective alignment and fusion.

- **Revisiting Safety Studies.** In this section, we will review well established studies on the safety of uni-modal learning, We focus on major safety issues, including *adversarial attacks, data poisoning, jailbreak exploits, and hallucinations*. We want to briefly introduce background knowledge for the audience to understand common safety measures in uni-modal data.
- **Security Challenges Unique to Multi-Modal Learning**. Given the unique challenges of multi-modal learning, we categorize existing research on multi-modal AI safety into *Modality Integration Exploitation, Modality Misalignment, and Fused Vulnerabilities*, which are explored in Parts 2, 3, and 4.

Part 2: Modality Integration Exploitation

This section examines safety risks from integrating additional modalities, where independent manipulations can propagate and compromise system integrity. We focus on adversarial perturbations originating from image data and jailbreak prompts in the text modality.

- **Adversarial Perturbations in Image Modality.** We will take common vision-language tasks as examples such as visual question answering or visual reasoning to explain how existing multi-modal models can be easily fooled by attacking different modality separately without considering their alignment and fusion, by discussing representative papers in this direction include VLAttack [8] and VQAttack [9].
- **Jailbreak Prompts in Text Modality.** We will introduce multimodal jailbreak through a detailed taxonomy, covering modalities including Any-to-Text, Any-to-Vision, and Any-to-Any. We will firstly categorize both attack methods [4] and defense methods [2] from input-level, encoder-level, generator-level, and output-level attacks. Then we will detail the evaluation covering manual evaluation and automated evaluation with a focus on detector-based, LLM-based and rule-based multimodal jailbreak.

Part 3: Modality Misalignment

This section explores safety threats from modality misalignment, which can naturally occur or be exploited by malicious actors to generate hallucinated, incorrect, or harmful outputs.

- **Hallucinations Caused by Modality Misalignment.** Multimodal hallucination often trace back to the misalignment between vision and language. We will highlight two key patterns in this misalignment: (1) Overtrust on language knowledge more than their visual processing, especially in spatial reasoning [1]; (2) tendency to invent objects that are not grounded in the vision modality or misinterpret objects at different levels of detail [10]. We will then detail the evaluation and benchmarks related to such misalignment.
- **Attacks Exploiting Embedding Misalignment.** Adversaries can intentionally manipulate intra-modal correlations to misalign embeddings, distorting the model's cross-modal understanding and leading to incorrect or harmful outputs. We will discuss studies such as [6, 7], which achieve this by perturbing the visual modality to reduce its correlation with text embeddings, causing the model to produce nonsensical results due to the loss of cross-modal information.

Part 4: Fused Vulnerabilities

In this section, we introduce threats that arise from the interaction of manipulated inputs, rather than isolated attacks on a single modality.

- **Poisoning Attacks via Cross-Modal Fusion.** These attacks exploit the fusion mechanism, where adversarial signals appear benign when considered individually but lead to system failures when combined. We will review works that follow this strategy, such as [5], which embed backdoor triggers in both image and text modalities. After fine-tuning, the model exhibits malicious behavior only when both triggers are present.
- **Cross-Modal Threat Switching: Injection & Activation.** Each modality's distinct properties create unique vulnerabilities, as discussed in Part 1. Adversaries can strategically select the appropriate modality for threat injection and activation. We will review works following this approach, such as [3], which shows that visual data is well-suited for injecting threats due to its continuous nature, while text data is more effective for activation because of its efficiency during inference.

Part 5: Conclusion and Future Directions

After discussing the unique safety challenges above, we will conclude our tutorial by first summarizing the **limitations of existing defenses** against those vulnerabilities, which include computation overhead, generalizability against different attacks, and explainability. Later, we will **summarize** the discussed research and explore future directions for enhancing the safety of multi-modal learning. Our discussion will focus on aligning multi-modal learning objectives – modality alignment and fusion – while also addressing the challenges posed by the current AI trend of large-scale models trained with limited available data. We highlight **potential strategies** to improve the overall safety and reliability of multi-modal AI systems.

4. POTENTIAL SOCIETAL IMPACTS

With the rise of advanced multimodal AI systems like OpenAI's GPT-4o, which integrates multiple data modalities (e.g., vision, language, and audio), public concerns about their security and reliability have grown. These large-scale models introduce new vulnerabilities that are not yet fully

understood, making it crucial to explore their risks and defense mechanisms. Our tutorial will provide the audience with a clear understanding of the unique challenges in securing multimodal AI and equip them with foundational knowledge on robustness. By fostering awareness and discussion on the safety of multimodal learning, this tutorial will contribute to the development of more secure and trustworthy AI systems, ensuring their responsible deployment in real-world applications.

5. SIMILAR TUTORIALS

The following list includes previous ICCV, CVPR, and ECCV tutorials related to multi-modal learning and the trustworthiness of foundation models.

- *"Large Multimodal Foundation Models"* at ECCV 2024 in Milan, Italy. **Similarities and differences**: This tutorial covers the history, applications, and future directions of multi-modal learning. Ours addresses a critical missing aspect – safety – by specifically exploring the safety challenges unique to multi-modal AI.
- *"From Multimodal LLM to Human-level AI: Modality, Instruction, Reasoning, Efficiency and Beyond"* at CVPR 2024 in Seattle. **Similarities and differences**: This tutorial reviews research on multi-modal language models, including architecture design, instructional learning, hallucination, reasoning, and efficient learning. While we also discuss hallucination in multi-modal AI, our focus is on its safety implications, along with other threats such as adversarial attacks.
- *"Trustworthy AI in the Era of Foundation Models"* at CVPR 2023 in Vancouver, Canada. **Similarities and differences**: This tutorial covers security, robustness, privacy, and societal issues in vision-based applications within the era of foundation models. While both tutorials focus on AI safety, ours specifically addresses the unique safety challenges of multi-modal AI, distinguishing it from traditional uni-modal learning.
- *"Tutorial on MultiModal Machine Learning"* at CVPR 2022 in New Orleans. **Similarities and differences**: This tutorial explores core technical challenges in multi-modal machine learning, such as representation, alignment, and reasoning. Our tutorial focuses on the safety challenges arising from the unique characteristics of multi-modal learning.

While these tutorials explore multi-modal and trustworthy AI, they do not specifically address the unique safety challenges of multi-modal systems. Our tutorial fills this gap with a focused discussion on emerging safety risks and considerations.

References

[1] Shiqi Chen, Tongyao Zhu, Ruochen Zhou, Jinghan Zhang, Siyang Gao, Juan Carlos Niebles, Mor Geva, Junxian He, Jiajun Wu, and Manling Li. Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. *arXiv preprint arXiv:2503.01773*, 2025. 2

[2] Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. Word embeddings are steers for language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16410–16430, 2024. 2

[3] 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. 2

[4] Yizhuo Ma, Shanmin Pang, Qi Guo, Tianyu Wei, and Qing Guo. Coljailbreak: Collaborative generation and editing for jailbreaking text-to-image deep generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2

[5] Matthew Walmer, Karan Sikka, Indranil Sur, Abhinav Shrivastava, and Susmit Jha. Dual-key multimodal backdoors for visual question answering. In *CVPR*, 2022. 2

[6] 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. 2

[7] 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. 2

[8] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[9] Ziyi Yin, Muchao Ye, Tianrong Zhang, Jiaqi Wang, Han Liu, Jinghui Chen, Ting Wang, and Fenglong Ma. Vqattack: Transferable adversarial attacks on visual question answering via pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6755–6763, 2024. 2

[10] Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. Halle-switch: Rethinking and controlling object existence hallucinations in large vision-language models for detailed caption. 2023. 2

TUTORS

Dr. Xi Li (xli7@uab.edu) is currently an Assistant Professor in the Department of Computer Science at the University of Alabama at Birmingham. She received her Ph.D. from the Department of Computer Science and Engineering at Pennsylvania State University. She earned her B.E. from Southeast University and her M.S. from Pennsylvania State University. Her research focuses on trustworthy AI and its integration in areas such as cybersecurity and education. Her work has contributed to several publications in conferences such as ICCV, AAAI, and ICASSP, as well as in journals like IEEE Transactions on Knowledge and Data Engineering and Neurocomputing. She has tutorial on trustworthy AI in the era of large-scale models at SDM. Additional information is available at https://lixi1994.github.io/.

Dr. Manling Li (manling.li@northwestern.edu) is an Assistant Professor at Northwestern University. She was a postdoc at Stanford University and obtained her PhD in Computer Science at University of Illinois Urbana-Champaign in 2023. She works on the intersection of language, vision, and robotics. Her work won the ACL'24 Outstanding Paper Award, ACL'20 Best Demo Paper Award, and NAACL'21 Best Demo Paper Award, etc. She was a recipient of Microsoft Research PhD Fellowship in 2021, an EE CS Rising Star in 2022, a DARPA Riser in 2022, etc. She served as Organizing Committee of ACL 25, NAACL 25, EMNLP 24, and delivered tutorials about multimodal knowledge at AAAI'25, IJCAI'24, CVPR'23, NAACL'22, AAAI'21, ACL'21, etc. Additional information is available at https://limanling.github.io/.

Dr. Muchao Ye (muchao-ye@uiowa.edu) is an Assistant Professor in Computer Science at the University of Iowa. He received his PhD in Informatics at the Pennsylvania State University in 2024. Before that, he obtained his Bachelor of Engineering degree in Information Engineering at South China University of Technology. His research interests lie in the intersection of AI, security, and healthcare, with a focus on improving AI safety from the perspective of adversarial robustness and human-understandable explainability. His research works have been published in top venues including CVPR, NeurIPS, KDD, ACL, and AAAI. He has given tutorials on premier conferences including KDD'21 and SDM'25. Additional information is available at https://sites.google.com/view/mcye.