**11-877 Advanced Multimodal Machine Learning** **Spring 2024**

# Week 7: Multimodal LLMs 3: Generative Models

*Instructors: Paul Liang and Daniel Fried* *Synopsis Leads: William Jongwon Han*

*Edited by Paul Liang* *Scribes: William Jongwon Han, Simran Khanuja*



Follow the rest of the class here: https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2024/

**Summary:** In this week's discussion we talk about generative models. More specifically, we talk about two main points: Advancements in recent generative models, and ethical considerations about the AI generated content. The following was a list of provided research probes:

1. Connecting with multimodal foundation topics discussed in the previous week, what types of multimodal interactions or connections are large-scale generative models learning to capture? How to link multimodal interactions with generative AI architecture? How to combine mathematical theory related to multimodal interactions to design the next generation of generative AI architecture?
2. With the advancement of generative AI, distinguishing between AI-generated and human-created content is becoming increasingly challenging. Besides watermarking, which has its limitations, are there other effective methods to differentiate between AI-generated and human-created content across various modalities (text, audio, video, image)? Or is it becoming virtually impossible to make this distinction?
3. What is the taxonomy of safety issues, social impact, and ethical concerns associated with generative AI development? How should we update best practices to address these ethical concerns? Who should initiate and lead this dialogue? What steps can be taken to mitigate these ethical issues effectively? Imagine we have an oracle multimodal generative AI system that is used on a large scale. What types of data pollution would it have if most of its data were published on the Internet?
4. When assessing the quality of multimodal outputs from generative AI systems, which dimensions should be prioritized? Can we develop metrics that allow for large-scale evaluation while mitigating potential safety and ethical risks?
5. Diffusion models have shown remarkable performance in controllable text-to-image generation. Could you explain the intuition behind why diffusion models are effective, especially in comparison to other generative AI models like GANs/ VAEs / AR-based LLMs? Some works claim that scaling up GANs can beat diffusion models (claimed in this paper) and some work claims that language models are better than diffusion models for image generation (claimed in this paper). Which generative model family do you think is the most promising one for multimodal generation?
6. For state-of-the-art video generation models like Sora, Yann Lecun mentioned in this tweet that Sora does not understand the real world and its corresponding physical rules. Do you agree with this view? Can the future development of generative AI systems truly incorporate real-world knowledge, or are they limited in this aspect? Is pursuing generative AI a viable path towards achieving Artificial General Intelligence (AGI)?

As a background for the mentioned content, students were required and suggested to read the following papers:

1. **(Required)** Generating Images with Multimodal Language Models [Koh et al., 2023]
2. **(Required)** Holistic Evaluation of Text-To-Image Models [Lee et al., 2023]
3. (Suggested) Video generation models as world simulators [Brooks et al., 2024]
4. (Suggested) Invisible Image Watermarks Are Provably Removable Using Generative AI [Zhao et al., 2023]

5. (Suggested) VideoPoet: A Large Language Model for Zero-Shot Video Generation [Kondratyuk et al., 2023]
6. (Suggested) Unified-IO 2: Scaling Autoregressive Multimodal Models with Vision, Language, Audio, and Action [Lu et al., 2023]
7. (Suggested) StableRep: Synthetic Images from Text-to-Image Models Make Strong Visual Representation Learners [Tian et al., 2023]
8. (Suggested) Self-Play Fine-Tuning Converts Weak Language Models to Strong Language Models [Chen et al., 2024]
9. (Suggested) High-Resolution Image Synthesis with Latent Diffusion Models [Rombach et al., 2022]
10. (Suggested) AI Deception: A Survey of Examples, Risks, and Potential Solutions [Park et al., 2023]
11. (Suggested) Evaluating the Social Impact of Generative AI Systems in Systems and Society [Solaiman et al., 2023]
12. (Suggested) Generative AI: Here to stay, but for good? [Sætra, 2023]

# 1 Advancements in Generative Modeling

## 1.1 Use of Frozen Encoders

It was discussed that utilizing frozen encoders in multimodal models raises concerns about the learning of intricate interactions between modalities. Adapters can be employed in such scenarios to enable the capturing of necessary interactions by updating only a few layers, thus offering a balance between model stability and adaptability [Hu et al., 2021].

## 1.2 Latent Interactions and Generative Planning

Capturing latent interactions in generative models is challenging. A model should not solely focus on generation but also incorporate elements of planning [Brooks et al., 2024]. This integration can enhance the model's ability to understand and predict complex multimodal interactions.

## 1.3 Extending Masked Modeling

The extension of masked modeling to generative tasks was a point of interest. If masked modeling leads to a better understanding of data, exploring ways to transfer this knowledge to generative tasks could be beneficial. This approach might involve developing techniques that leverage the strengths of masked modeling in generative contexts. In the context of the V-JEPA model [Bardes et al., 2024], utilizing masked representations could lead to improved motion understanding and ensure spatial and temporal consistency within video frames. This notion of learning grounded representations via unsupervised learning was initially introduced with images by Assran et al. [2023]. This approach could enhance the model's ability to predict and generate realistic sequences.

## 1.4 Computational Resources vs. Architectural Innovations

A critical discussion point was whether to apply more computational resources or to focus on developing better loss functions and architectures. A consensus seemed to be that a mix of both computational power and architectural innovation is essential for advancing multimodal learning models.

# 2 Ethical Considerations and Safety

## 2.1 AI Generated Content Detection

The detection of AI-generated content is crucial for maintaining authenticity and trust. Strategies such as watermarking (with an emphasis on minimizing false positives) and ethical considerations (e.g., the need to label AI-generated content) are vital [Zhao et al., 2023]. The nuances of what constitutes AI-generated content were also discussed, especially in the context of text and images.

### 2.2 Misinformation and Bias

Issues such as misinformation, hallucinations, and copyright infringements were identified as significant challenges. Different modalities face varying difficulties in watermarking [Zhao et al., 2023]. The potential for AI to generate misleading content, such as manipulated speeches, was also noted.

### 2.3 Bias and Representation

The discussion acknowledged the presence of gender and racial biases in generative models, often stemming from the data used for training. The challenge of overfitting and mode collapse in generative models was highlighted. The group emphasized that as long as humans are involved in data labeling, unconscious biases will inevitably influence the training data. It was also noted that digital representations of humans do not always align with real-world diversity.

## References

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint*, 2024.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. URL https://openai.com/research/video-generation-models-as-world-simulators.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models, 2024.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Jing Yu Koh, Daniel Fried, and Ruslan Salakhutdinov. Generating images with multimodal language models, 2023.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, Yong Cheng, Ming-Chang Chiu, Josh Dillon, Irfan Essa, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, David Ross, Grant Schindler, Mikhail Sirotenko, Kihyuk Sohn, Krishna Somandepalli, Huisheng Wang, Jimmy Yan, Ming-Hsuan Yang, Xuan Yang, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2023.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Benita Teufel, Marco Bellagente, Minguk Kang, Taesung Park, Jure Leskovec, Jun-Yan Zhu, Li Fei-Fei, Jiajun Wu, Stefano Ermon, and Percy Liang. Holistic evaluation of text-to-image models, 2023.

Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision, language, audio, and action, 2023.

Peter S. Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

Irene Solaiman, Zeerak Talat, William Agnew, Lama Ahmad, Dylan Baker, Su Lin Blodgett, Hal Daumé III au2, Jesse Dodge, Ellie Evans, Sara Hooker, Yacine Jernite, Alexandra Sasha Luccioni, Alberto Lusoli, Margaret Mitchell, Jessica Newman, Marie-Therese Png, Andrew Strait, and Apostol Vassilev. Evaluating the social impact of generative ai systems in systems and society, 2023.

Henrik Skaug Sætra. Generative ai: Here to stay, but for good? *Technology in Society*, 75:102372, 2023. ISSN 0160-791X. doi: https://doi.org/10.1016/j.techsoc.2023.102372. URL https://www.sciencedirect.com/science/article/pii/S0160791X2300177X.

Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners, 2023.

Xuandong Zhao, Kexun Zhang, Zihao Su, Saastha Vasan, Ilya Grishchenko, Christopher Kruegel, Giovanni Vigna, Yu-Xiang Wang, and Lei Li. Invisible image watermarks are provably removable using generative ai, 2023.