Lecture 8:

Large Multimodal Foundation Model

Papers for Lecture 8 (Large Multimodal Foundation Model)

P8-1: Large Visual Foundation Models: Presenter: Qin Hangyu; Reader: He Yingzhi

(Must-Read) H Liu, C Li, Q Wu, et al. Visual Instruction Tuning. NeurIPS 2023.

(SOTA) R Dong, et al. DreamLLM: Synergistic Multimodal Comprehension and Creation. ICLR 2024.

(Must-Read) D Zhu, et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced LLMs. ICLR 2024.

(Background) H Touvron, et al. LLaMA: Open and Efficient Foundation Language Models. arXiv 2023.

(Background) T Brown, et al. Language Models are Few-Shot Learners. arXiv 2022.

P8-2: Pixel Grounding Large Multimodal Models: Presenter: Stefan Putra Lionar; Reader: Cao Xiao

(SOTA) Y. Yuan, et al. Osprey: Pixel Understanding with Visual Instruction Tuning. arXiv 2023.

(Must-Read) H. Rasheed, et al. Glamm: Pixel grounding large multimodal model. arXiv 2023.

(To-Read) Z. Ren, et al. PixelLM: Pixel Reasoning with Large Multimodal Model. arXiv 2023.

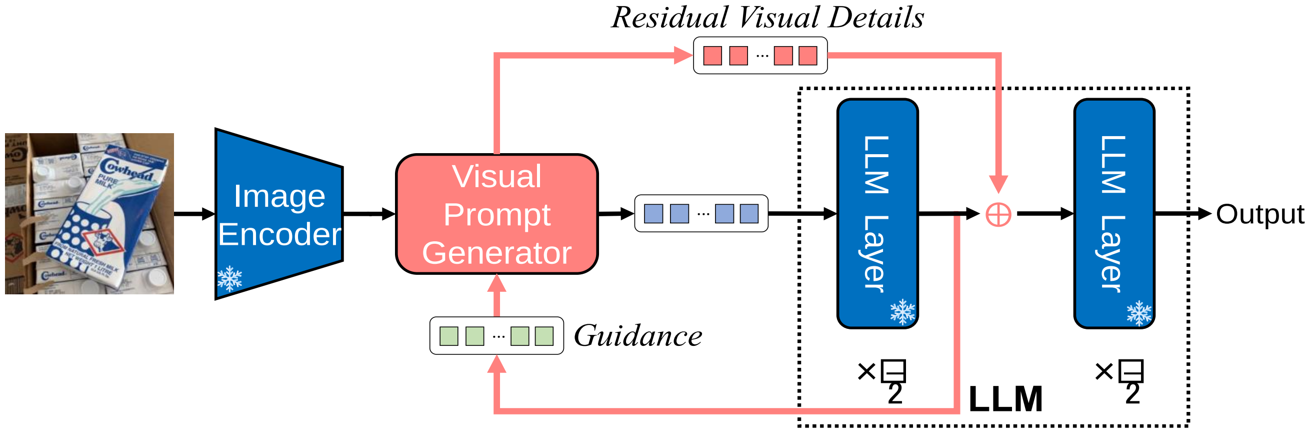
Research towards MFMs

Basic MFM Models:

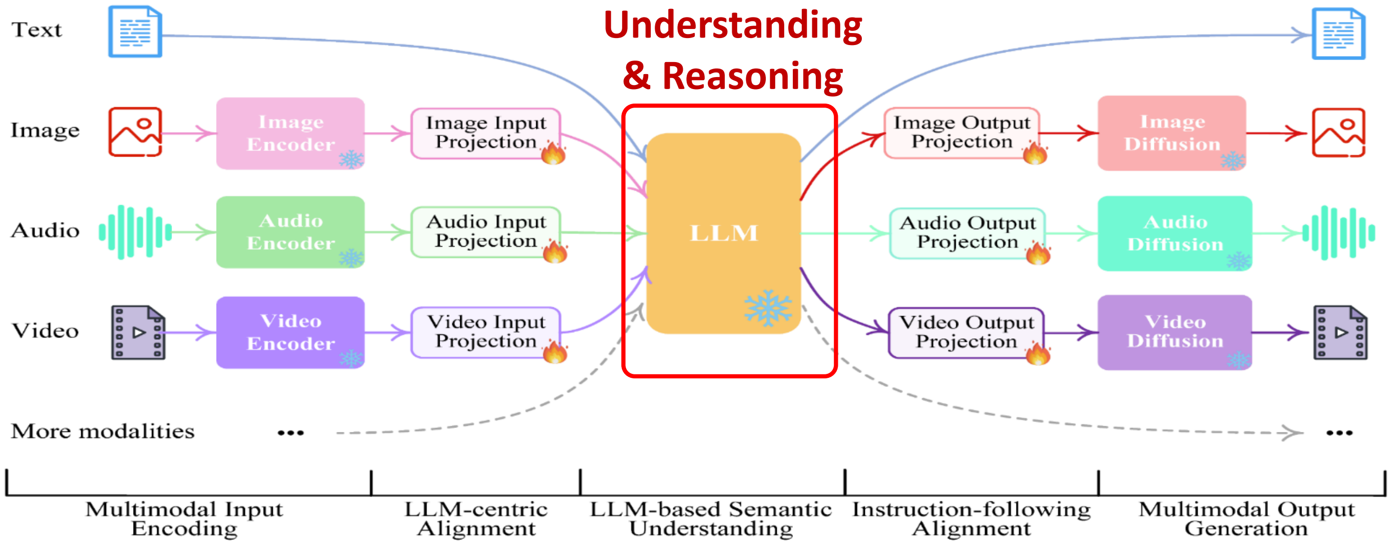
- **Multimodal alignment:** align heterogeneous modalities into a unified semantic space and perform joint reasoning over multimodal inputs
- **Multimodal instruction tuning:** enable the MFMs to follow a wide varieties of instructions, involving multimodal and interleaved context
- **Trust and safety:** detect and prevent hallucinations in MFMs, ensuring that MFMs can faithfully and reliably accomplish a variety of tasks

  To address different types of hallucination (object, relation & factual hallucinations…).

VPG-C: addresses **Inherent Inductive Bias of Image-Captioning** that tends to ignore non-dominant visual details (Li, ICLR 2024)



LLM Understanding & Reasoning



Requirements for Brave-New-Idea (BNI) papers

- **AIM of BNI Paper:**
  1) To propose a work that contain original ideas and research vision.
  2) The paper should offer: (i) novel, exploratory solutions with sufficient evidence of proof-of-concept; (ii) visions describing a new or open problem in multimedia research; and/or (iii) a novel perspective on existing multimedia research.
- **Guidelines:**
  - Must be in multimedia and is expected to have a high component of novelty

- **Gradings of Abstract of BMI Papers:**
  - **I provide quick comments on all the abstract** based on my reading of short abstracts
  - The comments are based on my quick and personal views of what are important
  - Hope this helps you in improving the final BNI paper.

Short Idea/ Opinion 1

- **Topic:**

  Can LFM (Large Foundation Model) use public data for training and content generation: what are the issues and guidelines?

Gradings of Article

- The key theme of this article is to examine the debate related to the use (or not) of public data, and if so, the guidelines for such use. Key issues are copyright, fair use etc.
- Most article focused on analysing the problems arising from the use of public data, such as biasness, fairness, privacy and trust etc. For such articles, I gave the grade of **+-B**
- Those articles that also cover the issues related to the use and guidelines for use of public data, I gave the grades of **+-A**
- **Two articles will be selected for presentation.**

r

Article 2: 26 Feb @ 1700 (Submit Article2)

Short Idea/ Opinion 2

- **Topic:**

  Trust and Robustness in LFMs (Large Foundation Models)

- **Outline of Paper:**
  - Robust and trust are the key problems to LFMs. With the development of more powerful LFMs with strong generative capability, this problem is becoming more severe.
  - What are the key issues here. Are these fundamentally unsolved problems? What can be done to address the problems?
  - What guidelines should be in place to mitigate such problems.
  - The article should be within **3 pages**, in ACM 2-column format (excluding references).

- **Received 18 out of 20 articles:**
  - Thanks to those who have submitted the articles on time
  - I hope to grade and feedback to you next week

ons.

ghtful

- **Deadlines:**
  - Article 2: 11 March @1700 (Submit-Article-2)

Papers for Lecture 9 (Multimodal Dialogues & NExT-GPT)

P9-1: **Multimodal Dialogues: Presenter: Sun Pengzhan; Asker: Xing Naili**

(SOTA) T Gong, et al. Multimodal-GPT: A vision & language model for dialogue with humans. arXiv 2023.

(Must-Read) Q Sun, et al. Multimodal dialogue response generation. ACL 2022.

(To Read) K Shuster, et al. Multi-Modal Open-Domain Dialogue. EMNLP 2021.

(To Read) T L Wu, et al. SIMMC-VR: A Task-oriented Multimodal Dialog Dataset with Situated and Immersive VR Streams. ACL 2023.

P9-2: **Multimodal Instruction Tuning: Presenter: Liu Nian; Asker: Bai Jinbin**

(Must-Read)J. Han, R et al. Imagebind-llm: Multi-modality instruction tuning. arXiv 2023.

(To Read) Z. Xu, et al. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. arXiv 2022.

(To Read) Z. Yin, et al. LAMM: Language-Assisted Multi-Modal Instruction-Tuning Dataset, Framework, and Benchmark. arXiv 2023.

P9-3: **NExT-GPT: (Invited Speaker: Yu Shengqiong)**

(Must-Read) S Wuet al. NExT-GPT: Any-to-any Multimodal LLM. arXiv 2023.