Lecture 8:

Large Multimodal Foundation Model

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 Wu et al. NExT-GPT: Any-to-any Multimodal LLM. arXiv 2023.

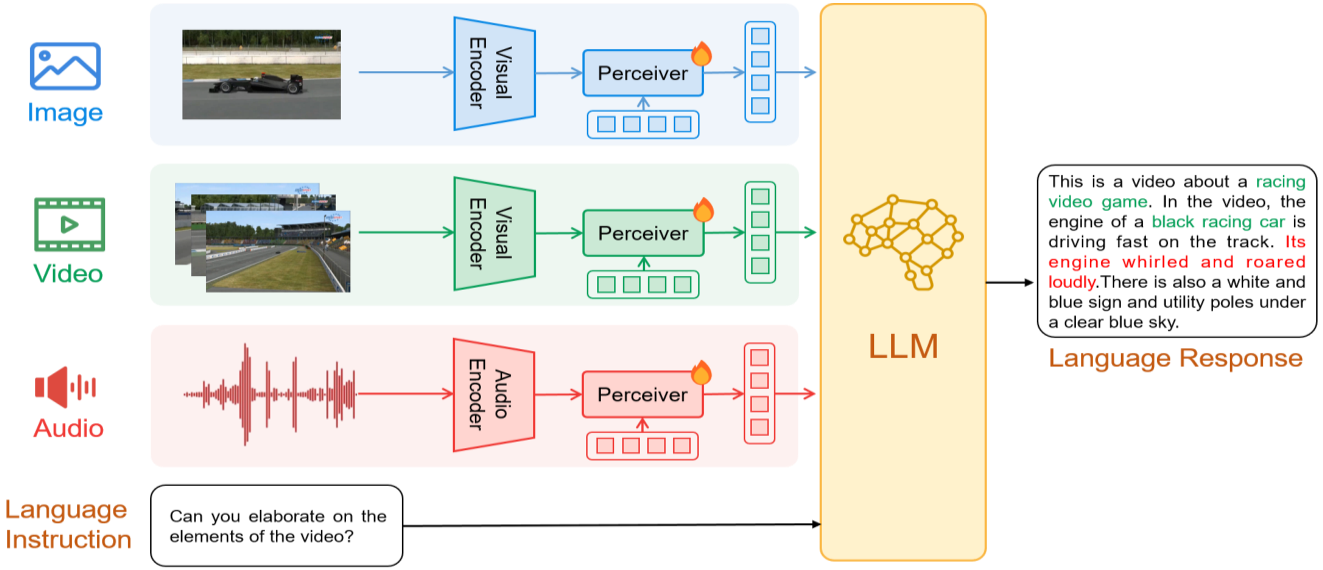
Research towards MFMs

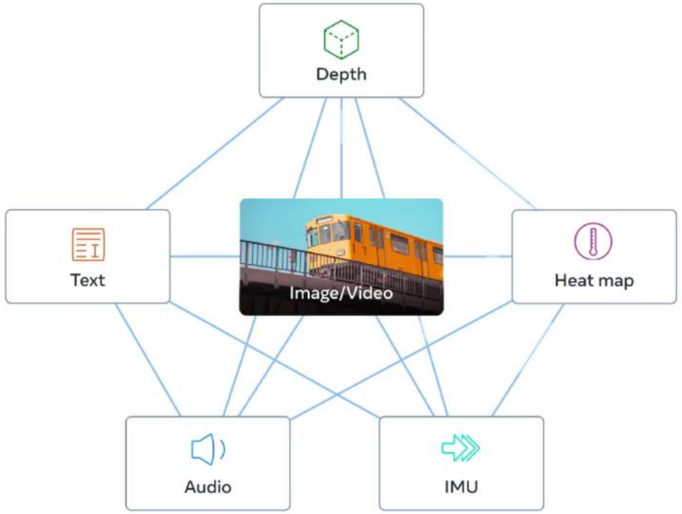
Basic MFM Models:

- **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…).

ChatBridge: aligns each modality with language, and user intent



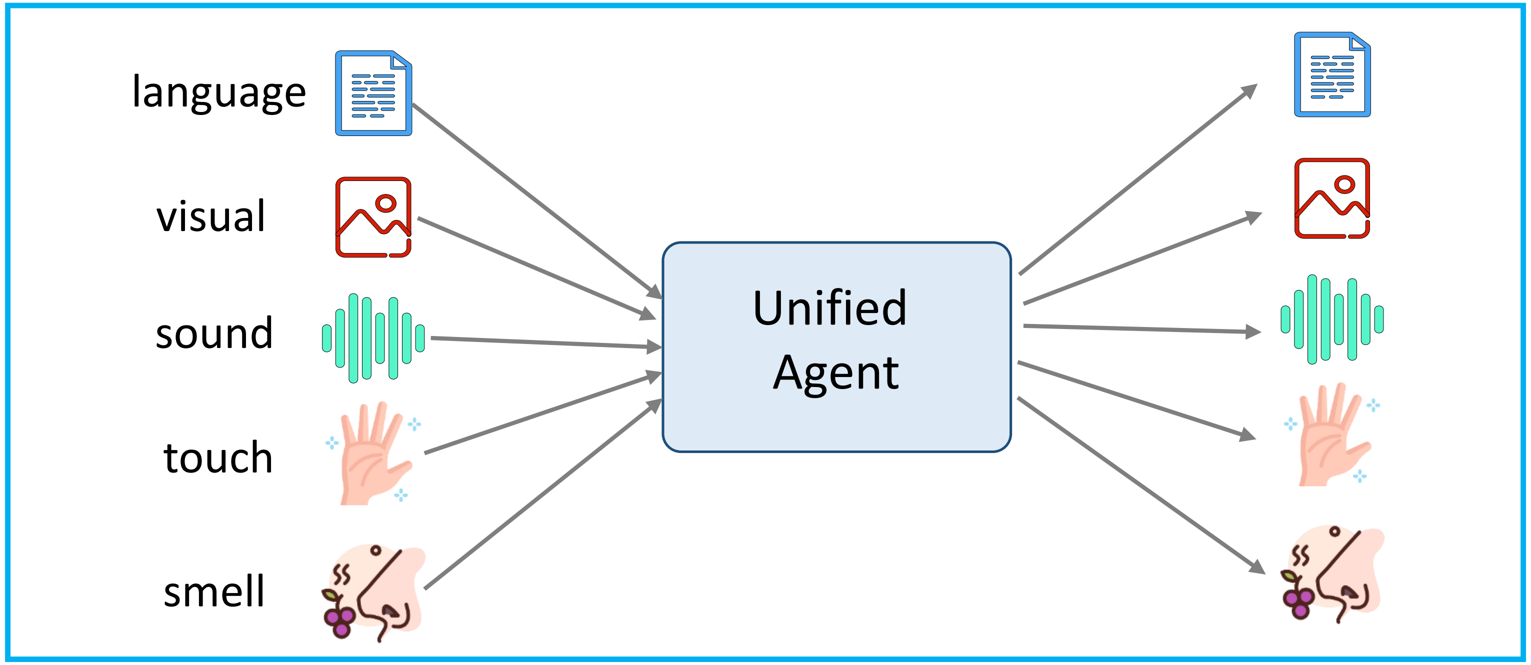


ImageBind: Align each modality's embedding to image embeddings

Multimodal Analytics

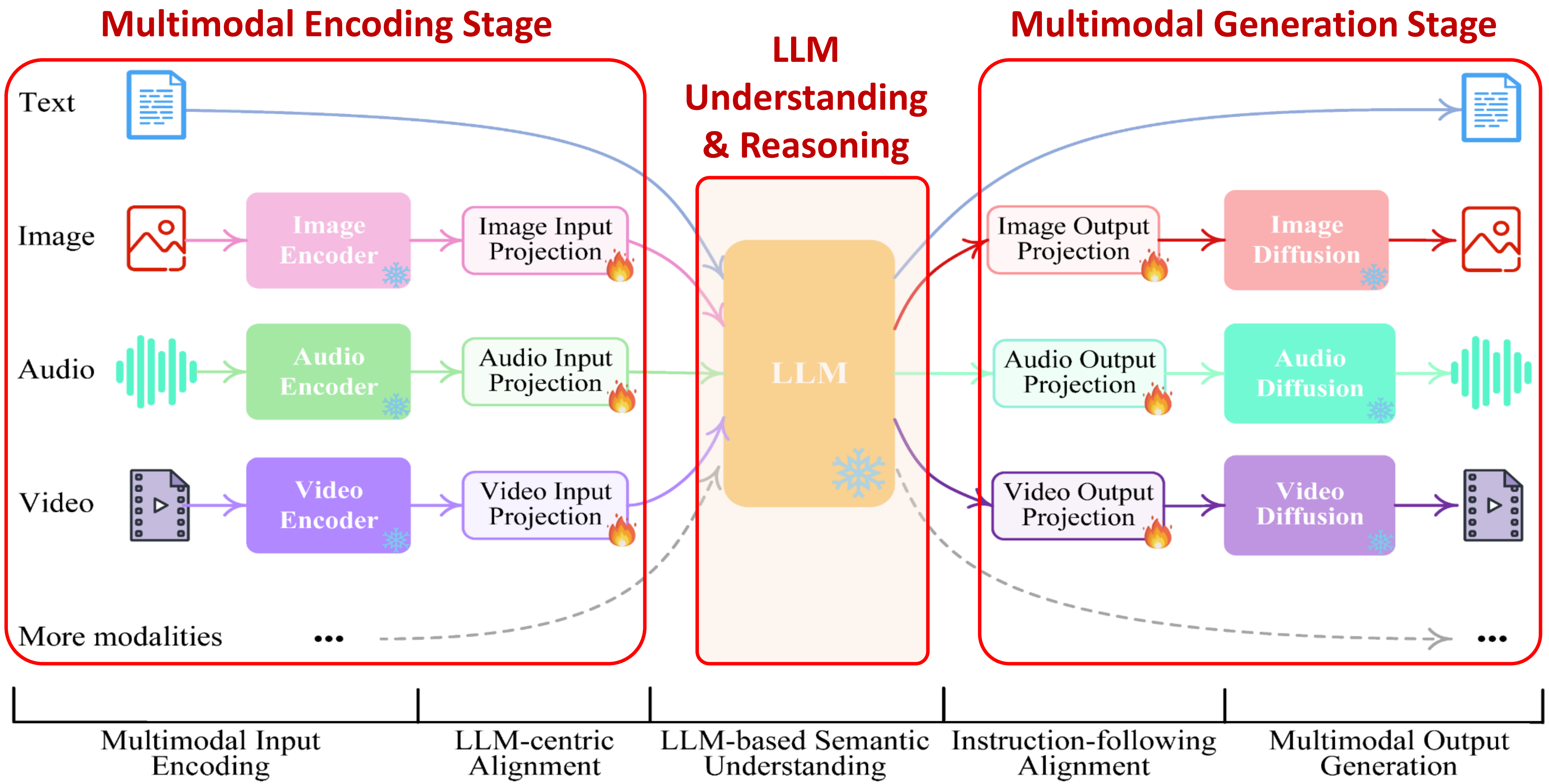
Human-Machine Interactions in a Multi-modal World

- We live in multimodal world and perceive multimodal information
  - Need to model the world knowledge as human do
- Human level AI



S Wu et al. NExT-GPT: Any-to-any Multimodal LLM. arXiv 2023

NExT-GPT
The Framework

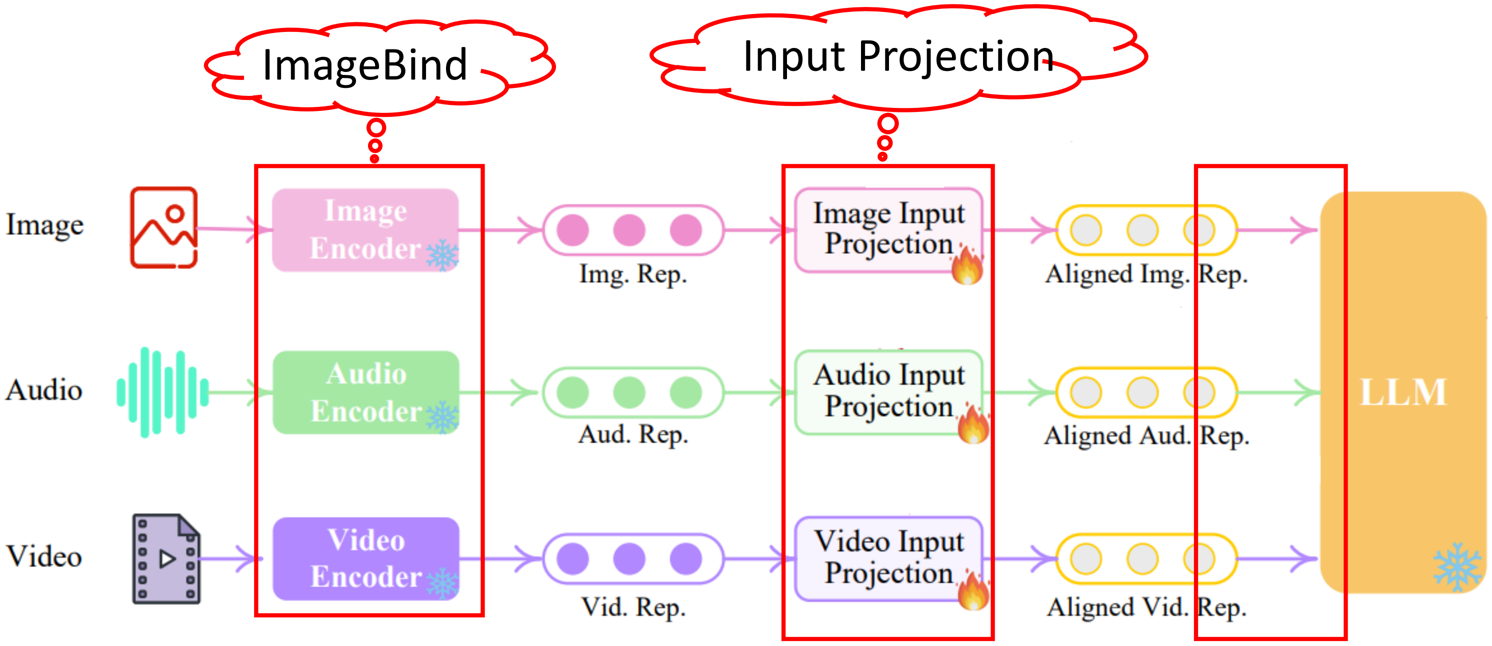


- In this work, we adopt Vicuna (7B) as the brain for NExT-GPT

NExT-GPT

Encoder: Multimodal Input Adaptation

- Employ the Modality-agnostic Encoder instead of Modality-specific Encoders:
  - Leverage it to encode all multimodal data, using tools such as ImageBind, LanguageBind, …
- Adopt the Linear Layer to project all non-textual features into textual space:
  - Similar approach taken by existing systems such as the MiniGPT-4, LLaVa, VPGTrans, PandaGPT, …
- Adopt Q-Former to map corresponding aligned multimodal features into frozen LLM:
  - Similar approach taken by existing systems such as the Video-LLaMa, BuboGPT, …

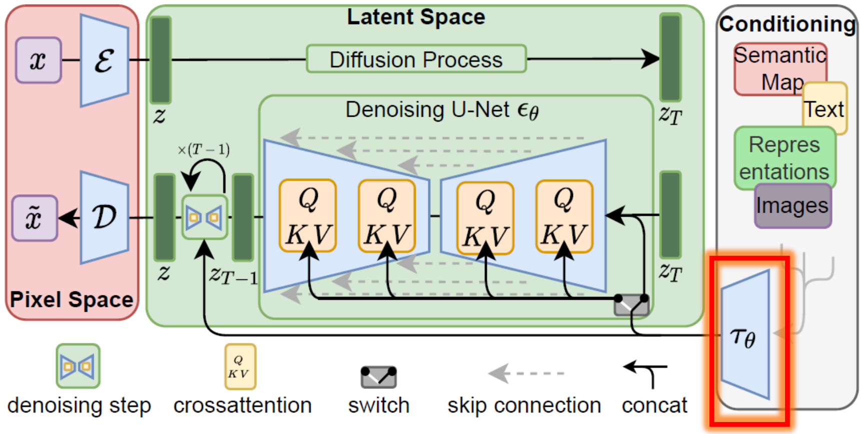


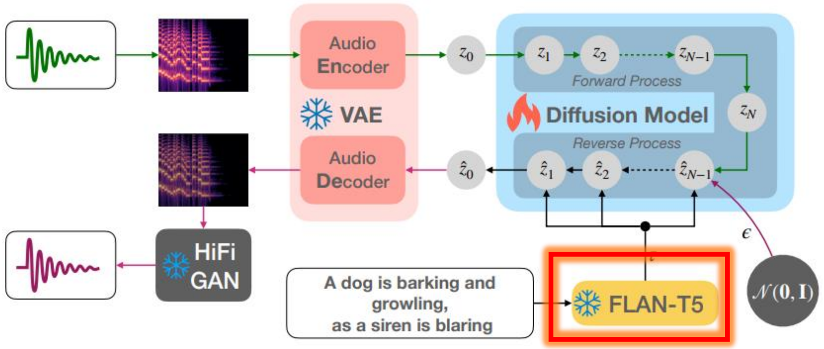
NExT-GPT

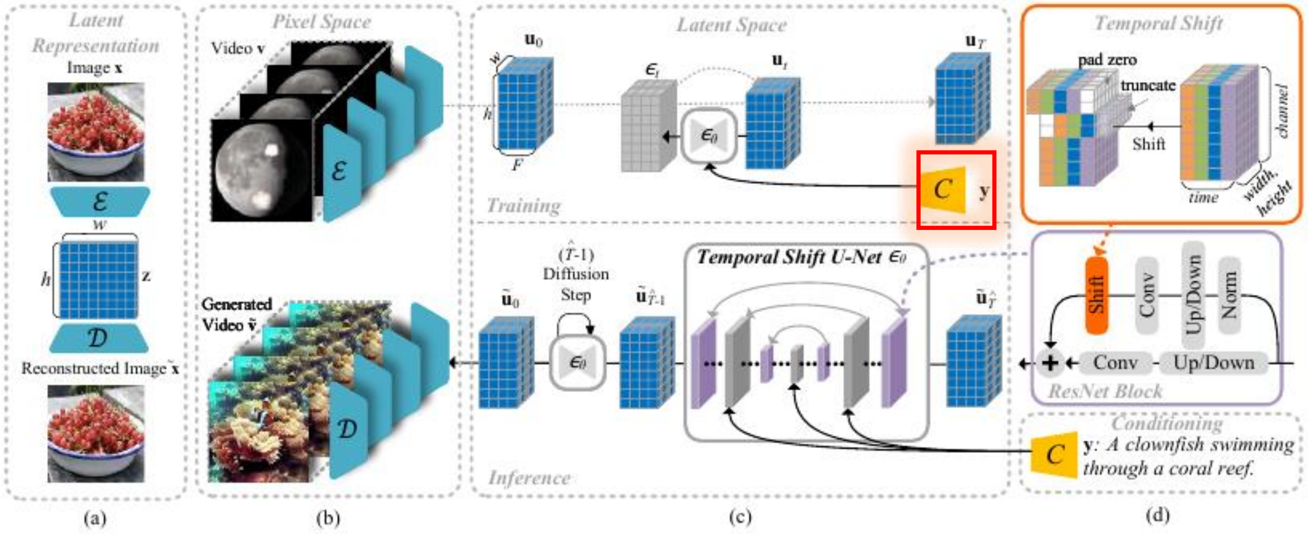
Decoder: Multimodal Content Generation

- Leveraging current SoTA diffusion-based models (for image, video & audio) to generate the desired multimodal content

- 3 key components of diffusion-based models:
  - Input: **We utilize the output of LLM (in modality-specific signal tokens)**, instead of text encoder, **to control the generation process**
  - VAE
  - UNet



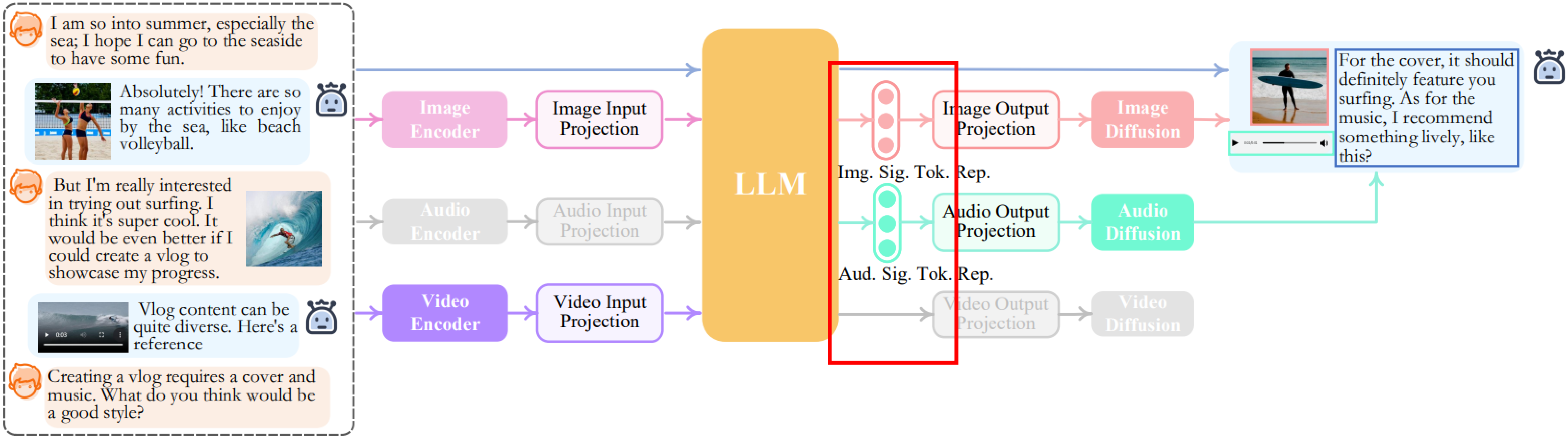




NExT-GPT

LLM-Guided Multimodal Content Generation

- The key issue is how to harness LLM as the brain for flexible multimodal content generation :
  - LLM needs to decide whether & what modal content to output in the current context
  - and how to align the diffusion models with LLM's output instructions?
- Instead of generating textual instructions, LLM produces unique "modality signal" tokens that are able to provide more intricate instructions to guide the generation process.



NExT-GPT

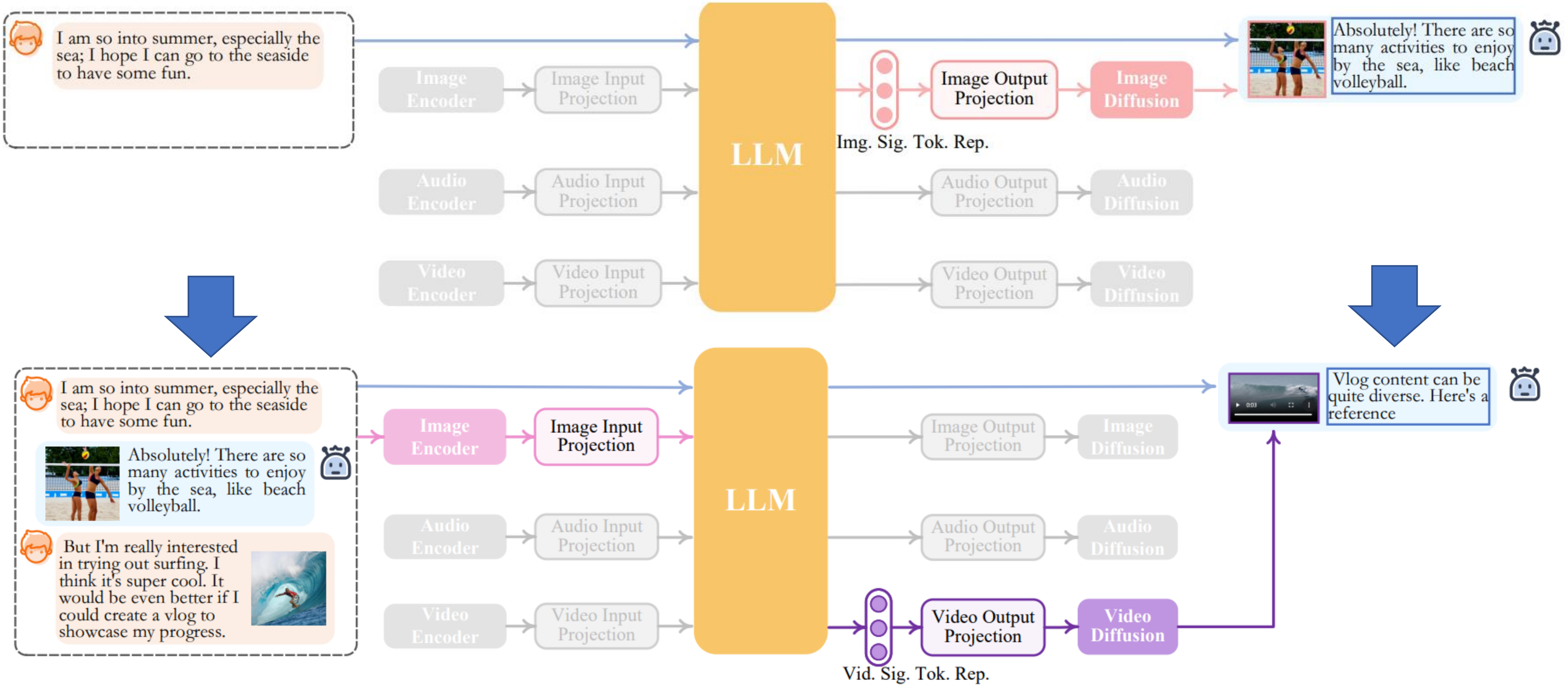
Key Highlights

- 1: Parameter-efficient Low-cost Training
  - Our configuration needs to update only 1% of parameters

- 2: Modality-switching instruction tuning
  - In any natural human-machine interaction, users and LLM will involve in diverse and dynamically changing modalities in their inputs and outputs
  - We curated a dataset (~ 5,000 conversations) to facilitate this training

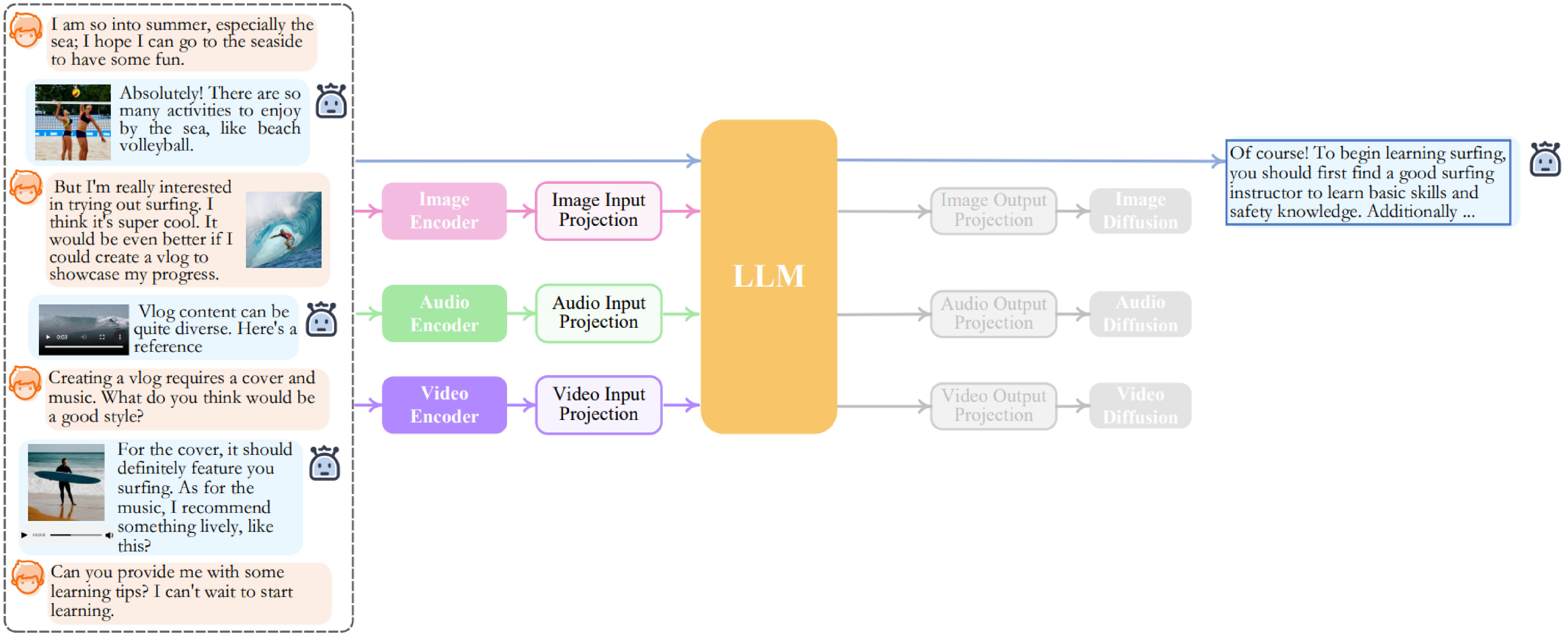
NExT-GPT

Inferencing Examples -1



NExT-GPT

Inferencing Examples -2



NExT-GPT

DEMOS

NExT-GPT

Demo: Dog on Skate Board

Text + Video

↓

Text + Image + Audio

NExT-GPT

Comparison with Related Systems

- Comparison with **VisualGPT** and **HuggingGPT**:
  - They rely on the **text instructions generated by LLMs** to generate the non-textual content
  - This disjointed & cascading pipeline is prone to introduce noise and diminishing efficiency
  - They also lack end-to-end training with comprehensive tuning

- Comparison with **Gemini** (of Google)
  - Gemini is a **product** with robust training and comprehensive functions
  - It supports input modalities of text, image, video and audio; but output of only text and image based on **explicit user prompts**
  - NExT-GPT **exhibits a more flexible capability**, supporting diverse modalities in its output based on **auto LLM inference**, which better aligns with real-world scenarios

[1] Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. 2023
[2] HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. 2023
[3] Genimi team, Google. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805. Dec 2023.

Lecture 10 (Responsible AI: Trust, Safety, Privacy & Biased in MM)

P10-1: Hallucination: Presenter: Cao Xiao; Asker: Chen Xihao

(SOTA) S Dhuliawala, et al. Chain-of-Verification Reduces Hallucination in LLMs. arXiv 2023.

(Must-Read) P Manakul, et al. Selfcheckgpt: Zero-resource black-box hallucination detection for generative LLMs. Preprint arXiv 2023.

(BG) Y Zhang, et al. Siren's Song in the AI Ocean: A Survey on Hallucination in LLMs. arXiv 2023.

P10-2: Privacy: Presenter: Dai Yuhe; Asker: Lin Xinyu

(SOTA) S Kim, et al. Propile: Probing privacy leakage in LLMs. arXiv 2023.

(Must-Read) J Huang, et al. Are Large Pre-Trained Language Models Leaking Your Personal Information? ACL 2022.

(Background) H Shao, et al. Quantifying Association Capabilities of LLMs and Its Implications on Privacy Leakage. EACL 2023.

P10-3: Bias: Presenter: Yannis Mohamed Christian Montreuil; Asker: Mehdi Yamini

(Must-Read) P Schramowski, M et al. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. CVPR 2023.

(Must-Read) Q Li, et al. Be causal: De-biasing Social Network Confounding in Recommendation. ACM TKDD 2023.

(To-Read) A S Luccioni, et al. Stable bias: Analyzing societal representations in diffusion models. arXiv 2023.