# 大模型的探索与实践

Introduction to Large Language Models

# § 2.1 多模态

# Multi-Modal

滕佳烨
上海财经大学
www.tengjiaye.com

# 回顾 Recall

Multi-Agent

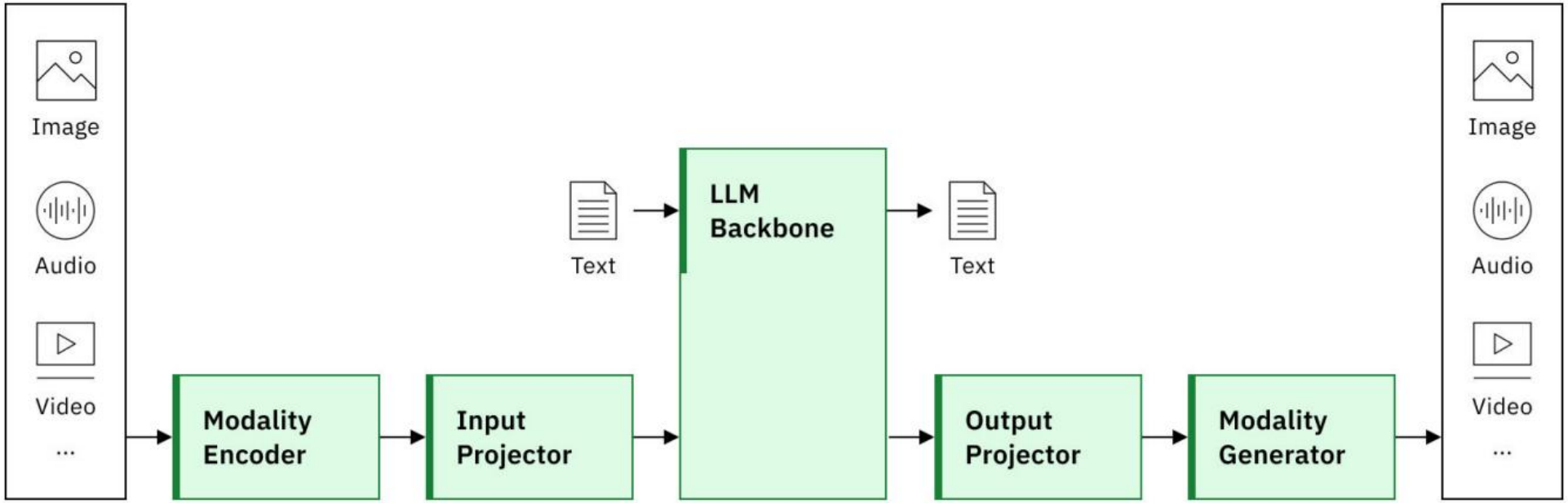
- 任务分解、任务联动

Coze

- 利用Coze构建智能体

**今天的任务：**Multi-Modal、Suno、Sora …

# 多模态任务




多输入/输出格式，且输入输出格式并不固定（图片？视频？音频？）
在多模态中，不同任务可能差别极大，有一些是需要模态融合（多模态辅助写作）；有一些是需要模态拼接（一段文字+图片生成视频）。

# 多模态任务

多模态的核心组件

- 模态编码器
- 输入投影仪
- 大模型基座
- 输出投影仪
- 模态生成器

模态输入 – 输入特征 - 特征处理 – 输出特征 – 模态输出

# 模态编码器 Modality Encoders

是处理原始单模态数据的预训练模型。对于视觉模态，通常采用视觉Transformer（ViT）或卷积神经网络（CNN）。对于音频模态，则常用Wav2Vec2等模型 。编码器的作用是从像素、波形等原始数据中提取高级别的抽象特征表示。

从 **模态** 到 **编码**
例如，语言文字的embedding；视觉图片的embedding等
-> 寻找抽象特征

# 输入投影器 Input Projectors

连接不同模态的关键“胶水”层。它通常是一个轻量级的网络，如多层感知机（MLP）或交叉注意力模块，其功能是接收来自模态编码器的特征输出，并将其映射到与LLM的词嵌入相同的向量空间中 。这是实现“模态对齐”的核心环节。

不同模态间的 **模态对齐**
这是多模态问题中最困难、最复杂的问题之一

- 基于投影的对齐（浅层对齐）
- 基于注意力的对齐（深层对齐）
- 基于统一tokenization的对齐（要求通用标记器）

# 大模型基座 LLM Backbone

模型的核心推理引擎，通常是一个大规模、预训练好的、仅解码器的Transformer模型（如LLaMA、ChatGLM）。在多模态训练的大部分阶段，LLM基座的参数保持“冻结”状态，以保留其强大的语言和推理能力 。它负责处理由文本标记和投影后的视觉/音频标记组成的序列。

使用现在的大模型基座即可实现，与语言训练类似

# 输出投影器 Output Projectors

在生成式任务中，输出投影器扮演着与输入投影器相反的角色。它将LLM输出的嵌入表示映射回一个模态生成器可以理解的表征空间。

与输入投影器类似，是对偶的问题

# 模态生成器 Modal Generators

对于文本到图像或视频等生成任务，该组件（例如，扩散模型）接收来自LLM处理后的表征，并最终生成目标模态的输出 。

从编码空间，进入多模态输出空间。

# 多模态常见任务

- 视觉问答任务
  - 模型根据图像或视频的内容，回答一个以自然语言提出的问题 。
  - 可在问题中加入推理元素

# 多模态常见任务

- 视觉问答任务
- 图像与视频描述生成
  - 模型为一幅图像或一段视频生成一段简洁、流畅、类人的文本描述 。
  - 作为图像描述的延伸，视频描述生成要求模型理解时间动态、行为交互以及场景随时间的变化，挑战性更高 。

# 多模态常见任务

- 视觉问答任务
- 图像与视频描述生成
- 跨模态检索
  - 根据来自一种模态的查询，在另一种模态的数据中检索相关内容。最常见的形式是“以文搜图”或“以图搜文”

# 多模态输出中的常见问题

- 片段修改问题
- 语义识别不明问题
- ...

# 总结 Take-away Messages

多模态的核心组件

- 模态编码器
- 输入投影仪
- 大模型基座
- 输出投影仪
- 模态生成器

在多模态输出中产生的问题

第六次作业：利用文字生成一张图片，再用这张图片尝试生成一段视频