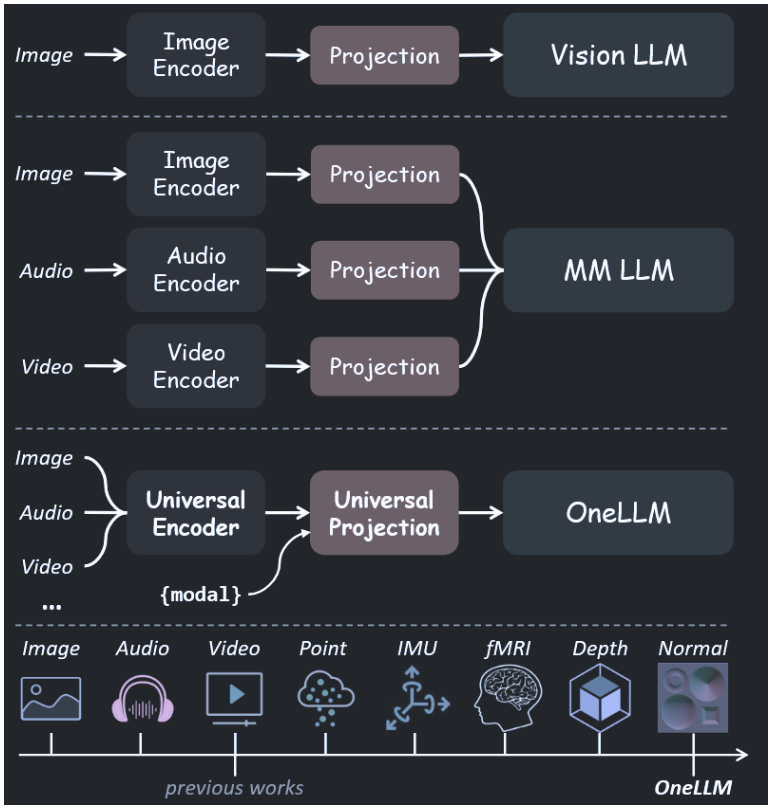
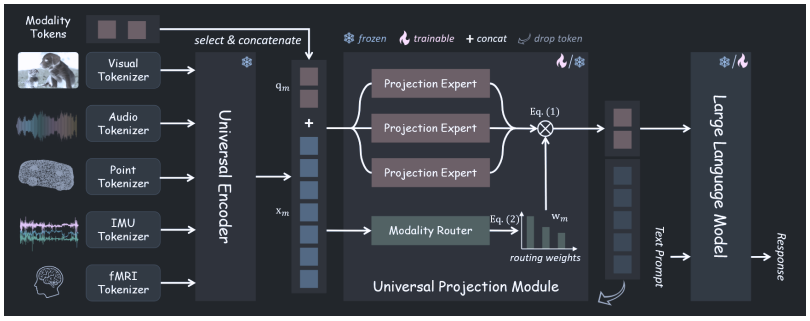
OneLLM: One Framework to Align All Modalities with Language

- Onellm：一个将所有模态与语言对齐的框架
- 相关文献
  - Flamingo
    - NeurIPS, 35:23716–23736, 2022
  - 使用特异性的编码器完成多模态链接
    - X-LLM
      - arXiv:2305.04160, 2023
    - Chatbridge
      - arXiv:2305.16103, 2023
    - 其中使用的链接器分别为
      - Q-Former（BLIP-2）
        - 使用的式BLIP-2中的链接器
        - arXiv:2301.12597, 2023
      - Perceiver
        - pages 4651–4664. PMLR, 2021.
- 第一遍
  - 摘要
    - 现有体系的问题
      - 多模态大模型的训练中每一个模态都需要对应的编码器，各个encoder结构不同，且对于一些不常见的模态没有对应的编码器
    - 方法综述
      - 先将视觉编码器对其投影到LLM上，然后混合多个视觉编码器和移动路由来建立通用投影模块（UPM），然后将更多模态链接到LLM上
  - 结论
    - 在25个测试基准上都展现了惊人表现
    - 局限与未来工作
      - 除了图像外的数据集质量不高
      - 对于细粒度模态需要设计对应的编码器a
  - 框架图
    - 与原有多模态框架对比



- 模型框架



- 第二遍
  - 介绍
    - 为什么要映射到LLM上
      - LLM的能力很强，工业界和学术界将其作为理解多模态信息的入口，以视觉为例，先在视觉——文本对应数据集，然后在视觉指令集上微调，其他模态类似
    - 现有的工作的缺陷
      - 先前的工作使用多个特异性的编码器来抽取特征映射到LLM上，每个编码器的结构都不同，训练成本高
        - X-LLM和ChatBridge
        - 每一个编码器都需要高度可靠，但是一些不常见的模态很难提供这样的编码器
    - 为什么选用CLIP-ViT作为通用编码器
      - Transformer有在一个模态上训练编码器然后迁移到多个模态作为编码器的潜力
    - 模型中的模态分词器
      - 一个可学习的模态分词器来转换模态并且补齐长度

- 为什么使用渐近式的训练方法来训练这个模型
  - 从头开始训练这个模型难度极高，选取渐进式的训练方式，先将视觉链接到LLM上，然后逐个将各个模态链接到LLM上，视觉选取了与训练的CLIP-ViT作为视觉编码器，使用LLaMA2作为LLM
- 贡献
  - 结构创新：通用编码器和专家路由
  - 扩展性好：能够轻易的扩展到多个模态
  - 数据建立：论文搜罗了一个大数据集，对应一些不常见的模态
- 相关工作
  - 视觉语言模型
  - 多模态大语言模型