中国人民大学
RENMIN UNIVERSITY OF CHINA

GeWu-Lab
Gaoling School of Artificial Intelligence
Renmin University of China

平衡多模态学习：过去与当下

卫雅珂

中国人民大学

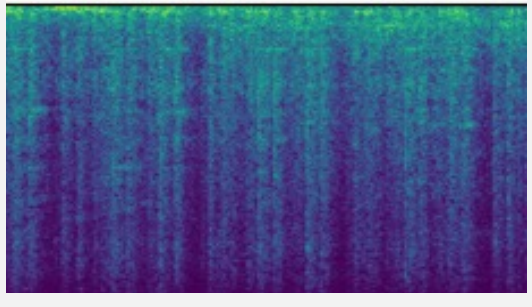
yakewei@ruc.edu.cn

什么是平衡多模态学习？

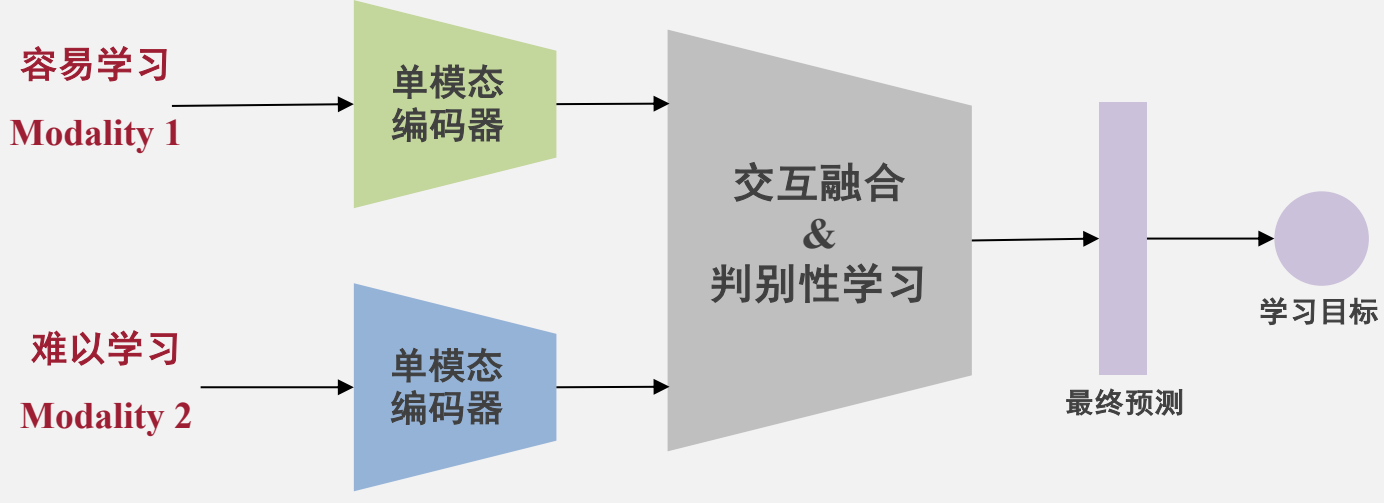
“平衡多模态学习”：不同模态的差异化学习特点会导致部分学习较慢较差的模态在优化竞争中受到较优模态的抑制，从而造成部分模态的学习不充分。

音视频样本对：*Bus*

干净 & 容易学习



嘈杂 & 难以学习



[1] Xiaokang Peng*, **Yake Wei***, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. *CVPR*, 2022, Oral.

平衡多模态学习：过去

平衡多模态学习

实验观察

模态间存在利用程度不均衡现象，影响多模态学习 (*CVPR 22 Oral*)

模态间学习不均衡在样本级存在细粒度差异 (*CVPR 24*)

单模态辅助学习目标和多模态学习目标存在冲突 (*ICML 24*)

算法

- ***优化视角：*** 通过动态梯度调控 (*CVPR 22 Oral*) 或前馈/反向传播动态调控 (*T-PAMI 24*) 缓解优化不均衡问题；或考虑模态自身的内在信息量的局限性，通过软重初始化打破模型对优势模态的依赖 (*ECCV 24*)。或在菲舍尔信息矩阵的指导下恢复被抑制的单模态信息获取过程 (*CVPR 25*)。
- ***数据视角：*** 受博弈论中的夏普利值启发，提出细粒度模态差异估计方法，在样本级评估模态贡献，并通过针对性重采样方法缓解样本级模态不平衡问题 (*CVPR 24*)。
- ***学习目标视角：*** 引入帕累托最优的思想解决单模态和多模态学习目标间的冲突，实现无害的单模态辅助 (*ICML 24*)；或分析不平衡背景下多模态鲁棒性的问题和性质，并设计学习目标，针对性提高多模态鲁棒性 (*ICLR 24*)。

理论分析

模态合作： 受夏普利值概念启发，提出多模态合作理论框架，在样本级评估和加强模态间的有效合作 (*CVPR 24*)。

模态信息获取： 借助菲舍尔信息矩阵，评估联合训练中单模态信息获取过程 (*CVPR 25*)。

单/多模态的冲突： 应用帕累托最优的思想，解决单模态和多模态学习目标间的冲突，实现二者的帕累托最优 (*ICML 24*)。

- **Yake Wei**, Di Hu, Henghui Du, and Ji-Rong Wen. On-the-fly modulation for balanced multimodal learning. ***T-PAMI***, 2024.
- **Yake Wei** and Di Hu. Mmpareto: boosting multimodal learning with innocent unimodal assistance. ***ICML***, 2024.
- **Yake Wei,** Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. ***CVPR***, 2024.
- **Yake Wei**, Siwei Li, Ruoxuan Feng, and Di Hu. Diagnosing and re-learning for balanced multimodal learning. ***ECCV***, 2024.
- Xiaokang Peng*, **Yake Wei***, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. ***CVPR***, 2022, **Oral**.
- Chengxiang Huang*, **Yake Wei***, Zequn Yang, and Di Hu. Adaptive Unimodal Regulation for Balanced Multimodal Information Acquisition. ***CVPR***, 2025.
- Zequn Yang, **Yake Wei**, Ce Liang, and Di Hu. Quantifying and enhancing multi-modal robustness with modality preference. ***ICLR***, 2024.

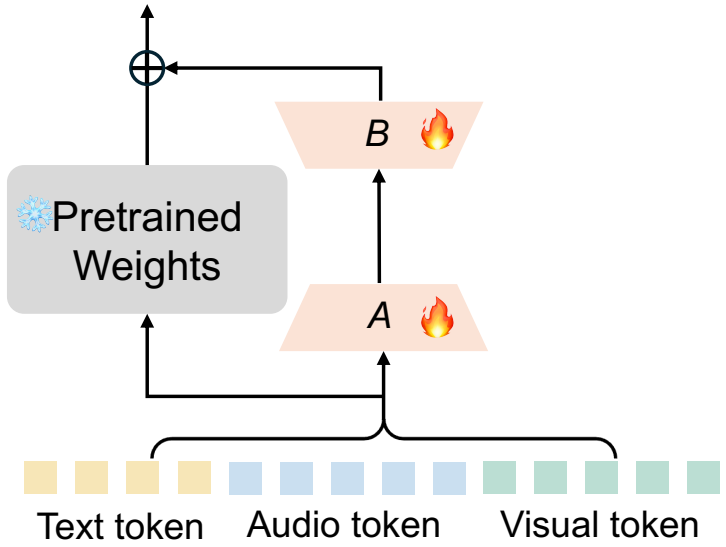
平衡多模态学习：当下?

大模型背景下

数据的问题依然存在

学习的贪婪特性未曾改变

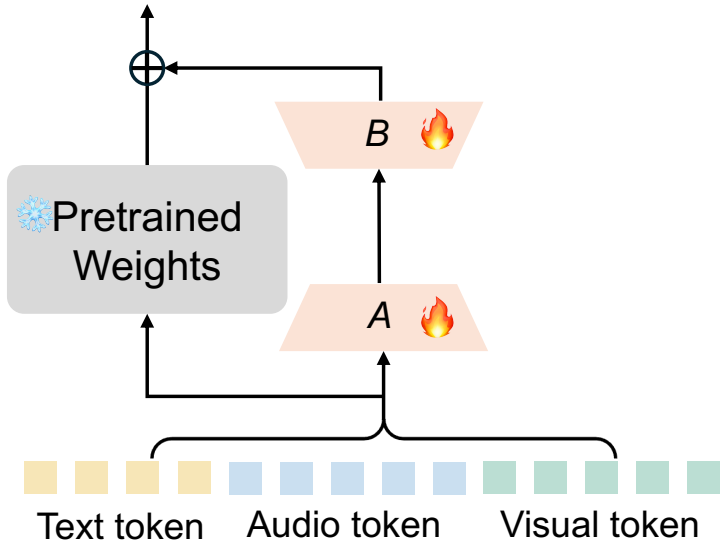
观察：被忽视的单模态特性



典型LoRA架构

$$h = W_0\mathbf{x} + \Delta W\mathbf{x} = W_0\mathbf{x} + \Delta W[\mathbf{x}^{\mathrm{m}_1}; \mathbf{x}^{\mathrm{m}_2}; \cdots; \mathbf{x}^{\mathrm{m}_n}]$$

观察：被忽视的单模态特性

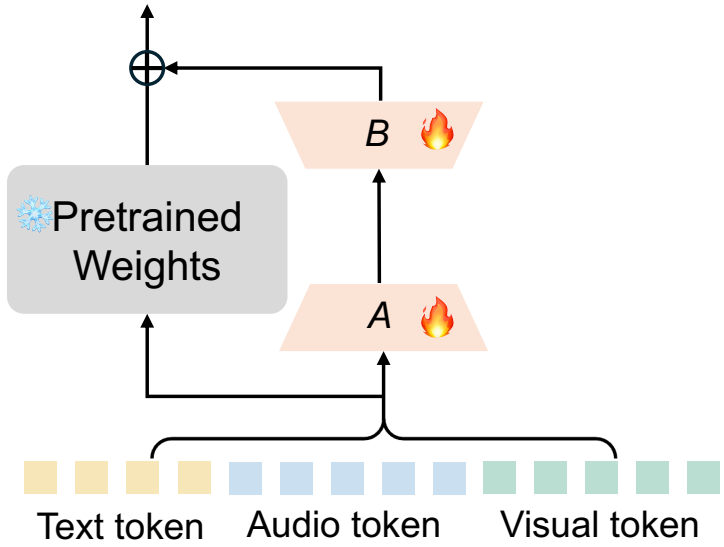


典型LoRA架构

$$h = W_0\mathbf{x} + \Delta W\mathbf{x} = W_0\mathbf{x} + \Delta W[\mathbf{x}^{\mathrm{m}_1}; \mathbf{x}^{\mathrm{m}_2}; \cdots; \mathbf{x}^{\mathrm{m}_n}]$$

直接“borrow”LLM微调方法完全契合多模态场景吗？

观察：被忽视的单模态特性

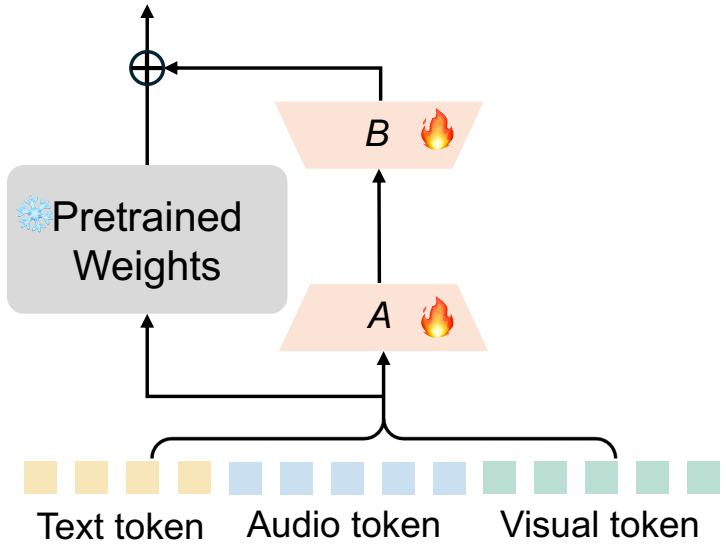


典型LoRA架构

完全统一处理的范式并非多模态的最优解

MLLM finetuning应该兼顾单模态与跨模态特性

观察：被忽视的单模态特性

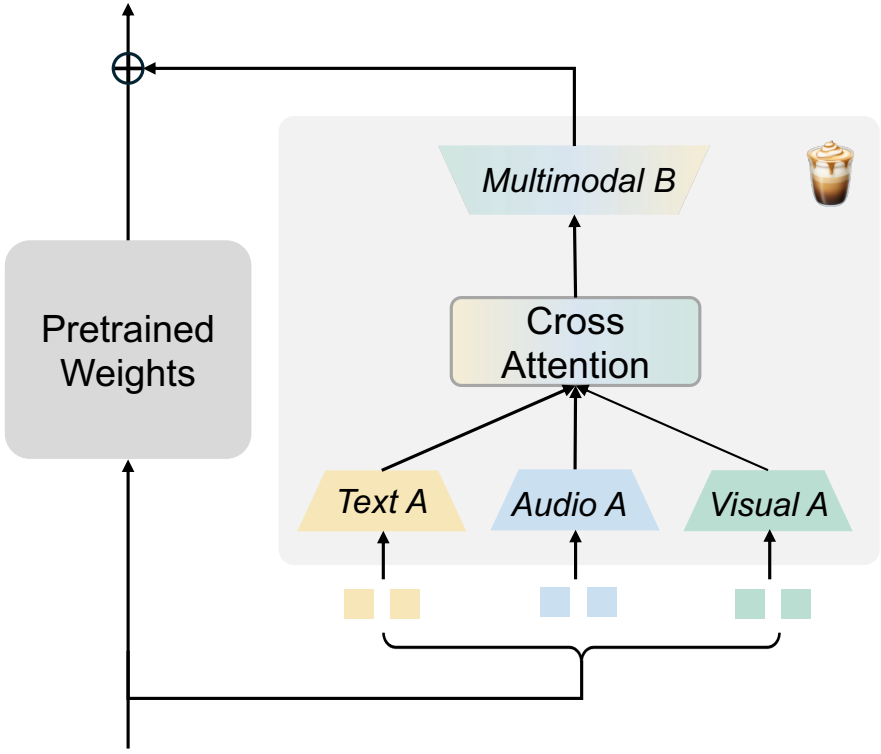


$$W_0\mathbf{x} + \Delta W[\mathbf{x}^{\mathrm{m}_1}; \mathbf{x}^{\mathrm{m}_2}; \cdots; \mathbf{x}^{\mathrm{m}_n}]$$

$$W_0\mathbf{x} + \underbrace{[\Delta W_1\mathbf{x}^{\mathrm{m}_1}; \Delta W_2\mathbf{x}^{\mathrm{m}_2}; \cdots; \Delta W_n\mathbf{x}^{\mathrm{m}_n}]}_{\texttt{unimodal adaptation}} + \underbrace{\Delta W_{\texttt{cross}}[\mathbf{x}^{\mathrm{m}_1}; \mathbf{x}^{\mathrm{m}_2}; \cdots; \mathbf{x}^{\mathrm{m}_n}]}_{\texttt{cross-modal adaptation}}$$

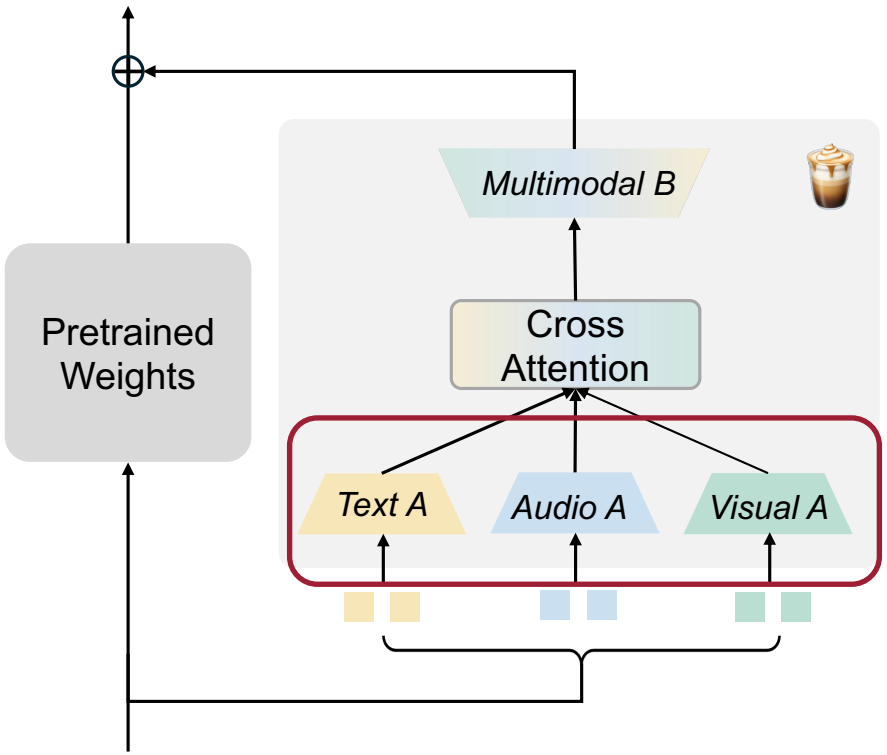
MokA: 兼具单模态与跨模态特性

Multimodal Low-Rank Adaptation for MLLMs



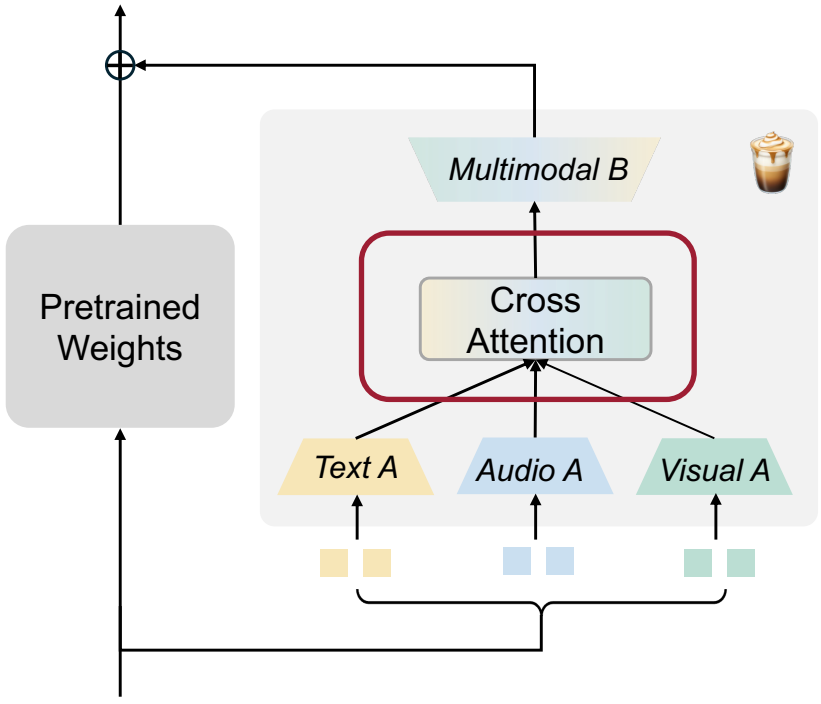
MokA: 兼具单模态与跨模态特性

模态独有matrix A: 独立抽取模态信息，互不干扰

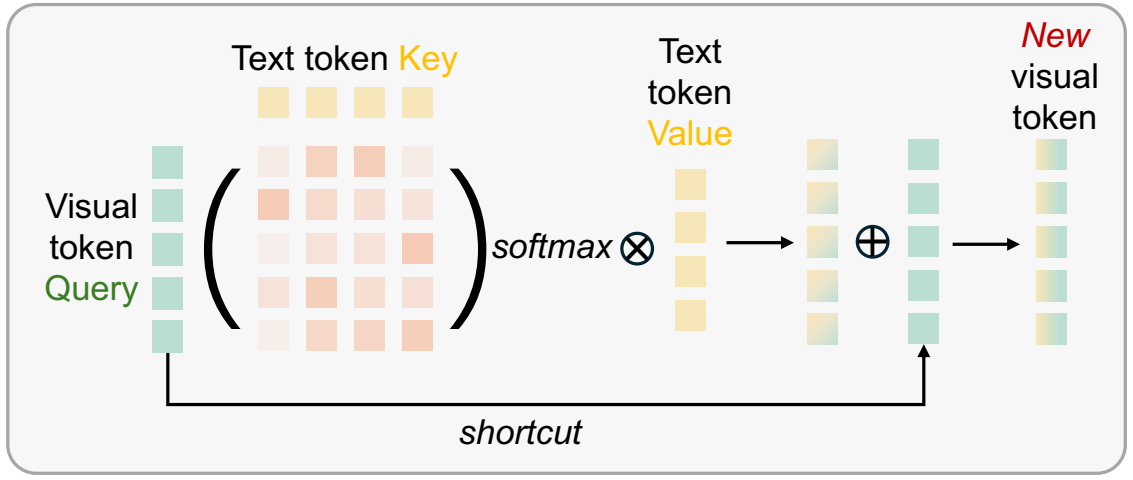


MokA: 兼具单模态与跨模态特性

显式模态交互: Cross-attention

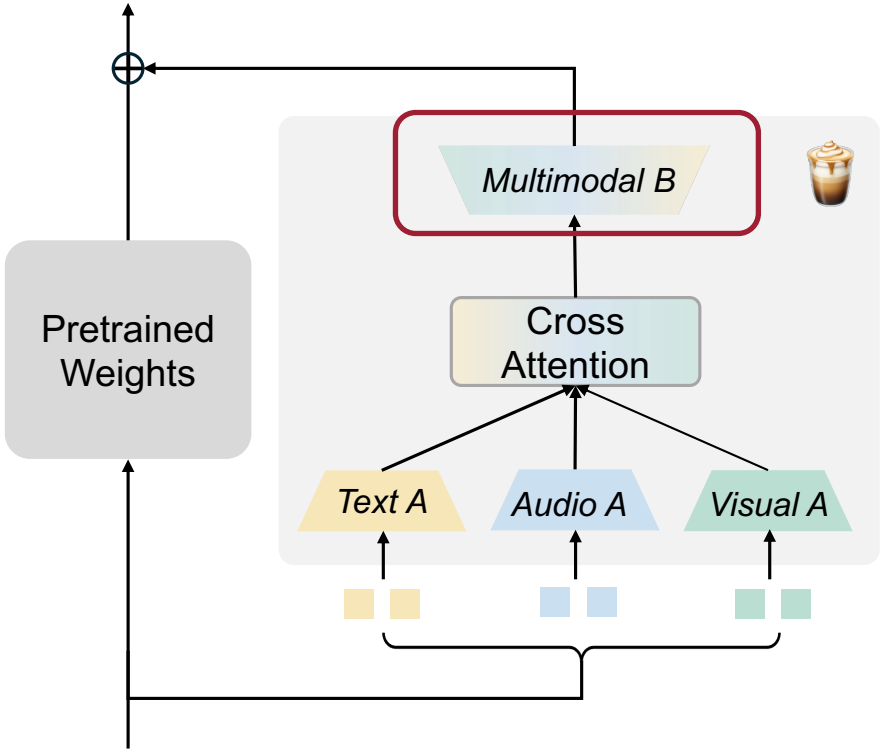


<audio> <visual> *Please answer the question: which clarinet makes the sound first?*



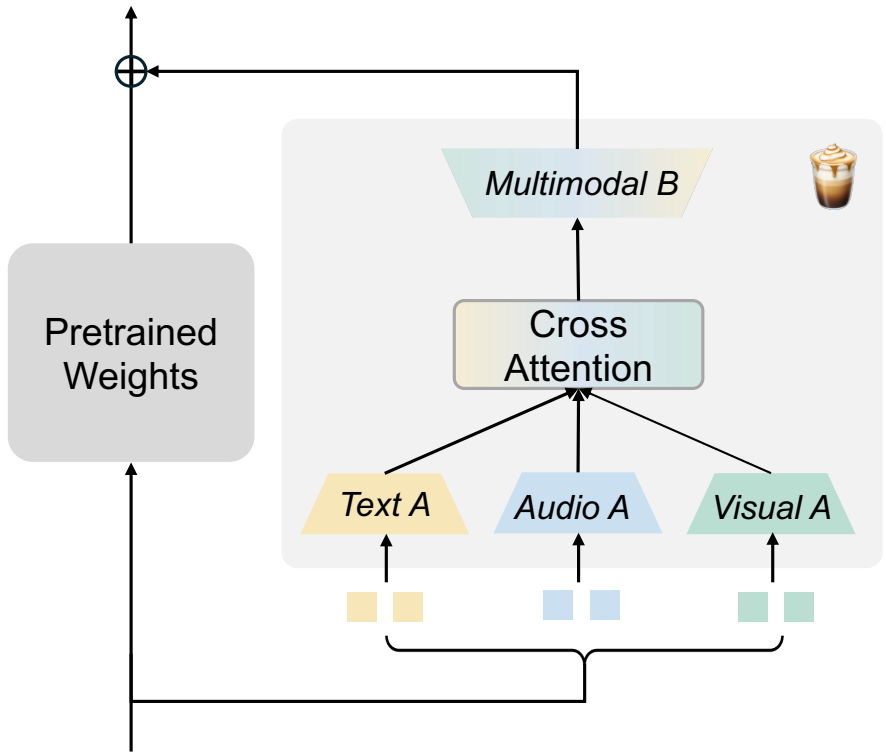
MokA: 兼具单模态与跨模态特性

模态共享matrix B: 分离的单模态子空间映射回同一共享空间



MokA: 兼具单模态与跨模态特性

Multimodal Low-Rank Adaptation for MLLMs



$$W_0\mathbf{x} + \underbrace{[BA^a\mathbf{x}^a; BA^v\mathbf{x}^v; BA^t\mathbf{x}^t]}_{\texttt{unimodal adaptation}} + \underbrace{[B\texttt{Att}_{a,t,t}; B\texttt{Att}_{v,t,t}; \mathbf{0}_{N_t}]}_{\texttt{cross-modal adaptation}},$$

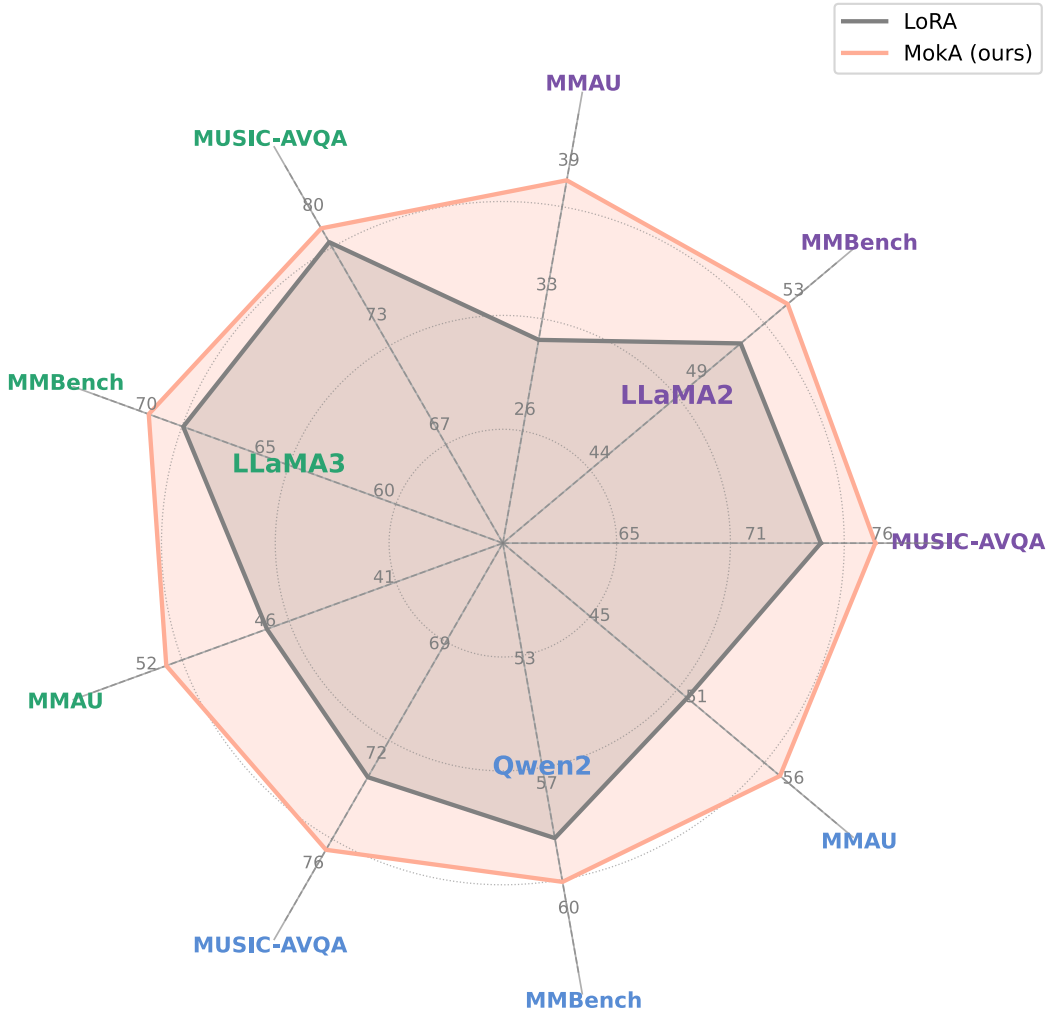
MokA: 兼具单模态与跨模态特性

多个基座模型：

- LLaMA2
- LLaMA3
- Qwen2

多种多模态场景：

- Audio-Visual-Text
- Visual-Text
- Speech-Text



Thanks for listening!

BML交流社区

MokA主页