X-InstructBLIP: A Framework for Aligning Image, 3D, Audio, Video to LLMs and its Emergent Cross-modal Reasoning

Artemis Panagopoulou[2,*], Le Xue[1,**], Ning Yu[1,**], Junnan Li[1], Dongxu Li[1], Shafiq Joty[1], Ran Xu[1], Silvio Savarese[1], Caiming Xiong[1], Juan Carlos Niebles[1]

1. Salesforce AI Research 2. University of Pennsylvania

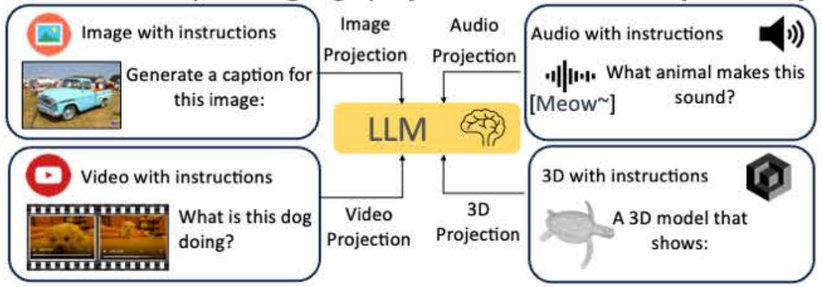
* work done during internship at Salesforce ** equal mentorship

Overview

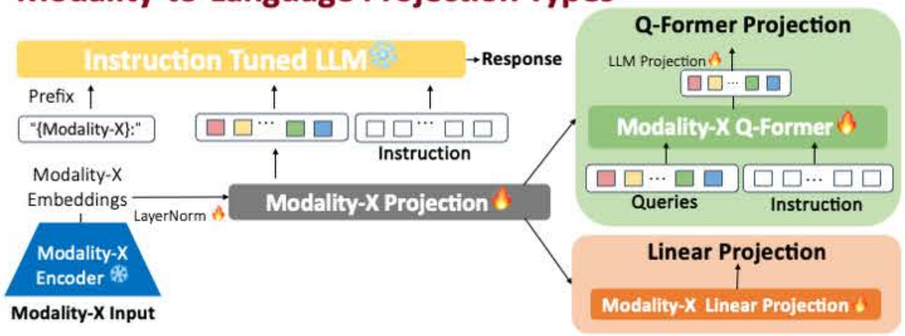
- Effective and efficient scalable framework for **independent modality alignment** to a frozen LLM showing **emergent reasoning across multiple modalities simultaneously.**
- Introduce the first benchmark **DisCRn** requiring models to perform discriminatory reasoning across multiple modalities.
- A **comparison** of two prominent modality-to-language projection types, **Q-Former and Linear Projection** shows that the former is better suited when cleaner and more variable data is available.

Individual Modality Training

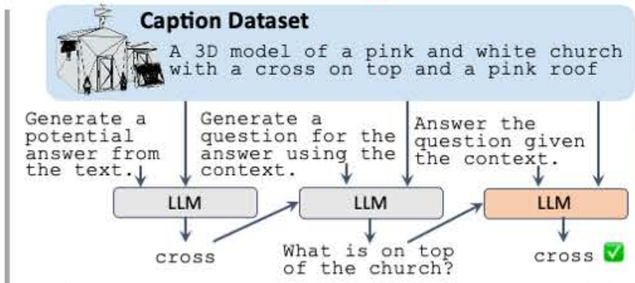
Each modality-to-language projection is trained **independently**.



Modality-to-Language Projection Types



Train: Instruction Tuning Data Generation from Captions



Round-Trip-Consistency

	Train	Dev
3D	12,771	1,022
Audio	2,951	723

Automatically Collected Data

Evaluation: Discriminatory Cross-Modal Reasoning (DisCRn)

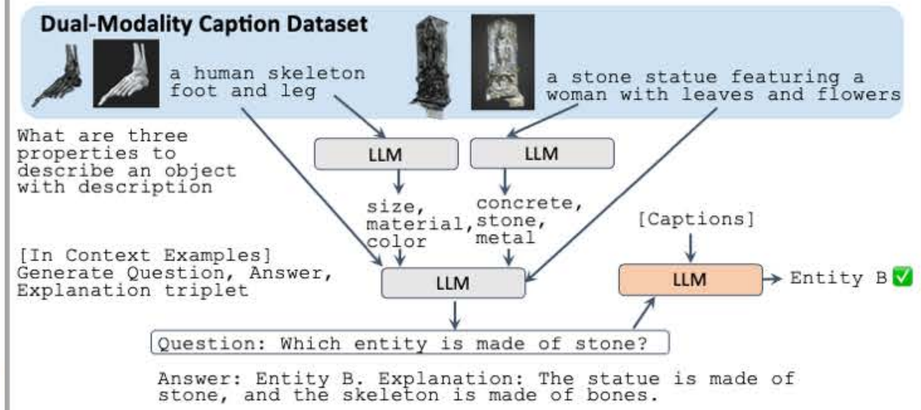
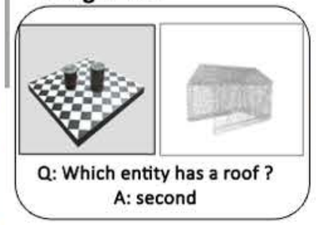
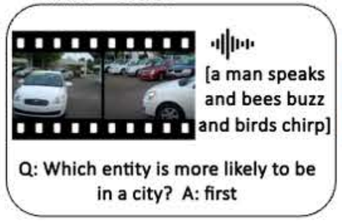


Image – 3D



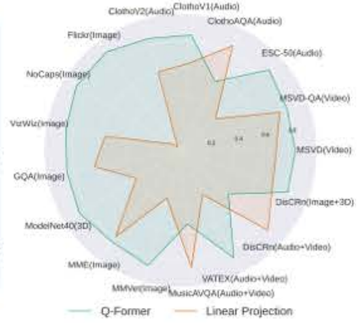
Audio - Video



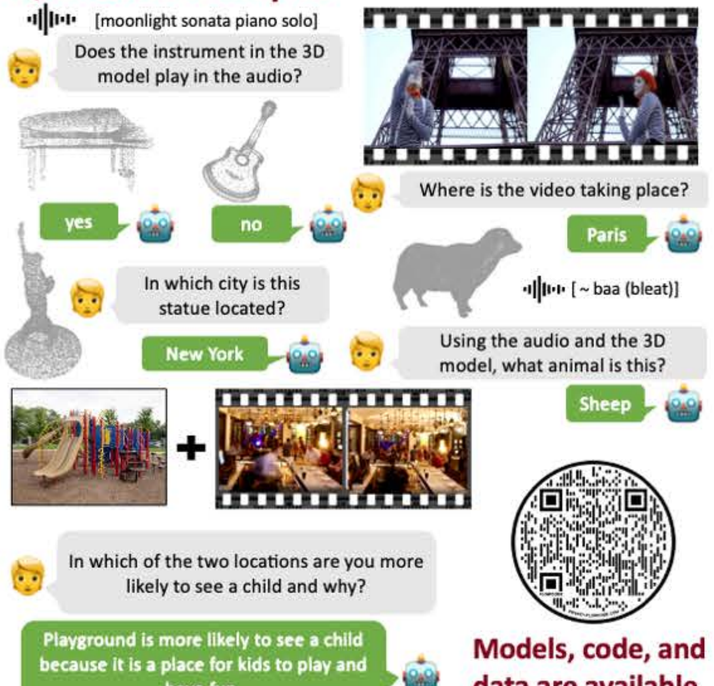
DisCRn	Examples
Image-3D	28,173
Audio-Video	8,802

Experiments

Single Modality: Q-Formers outperform Linear Projections under cleaner data conditions. **Cross-Modal:** Q-Formers are better at distinguishing between joint and discriminatory tasks. **Modality Prefix** improves cross-modal and single-modality performance.



Qualitative Examples



Models, code, and data are available.