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

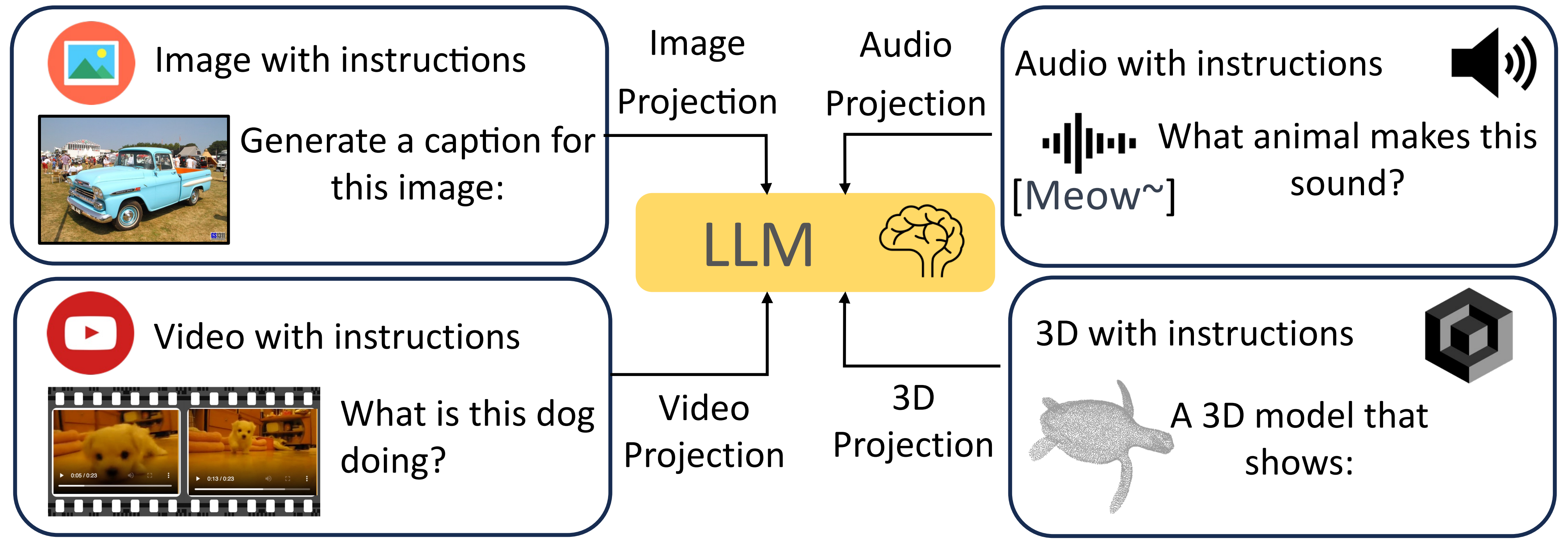
* work done during internship at Salesforce ** equal mentorship

Overview

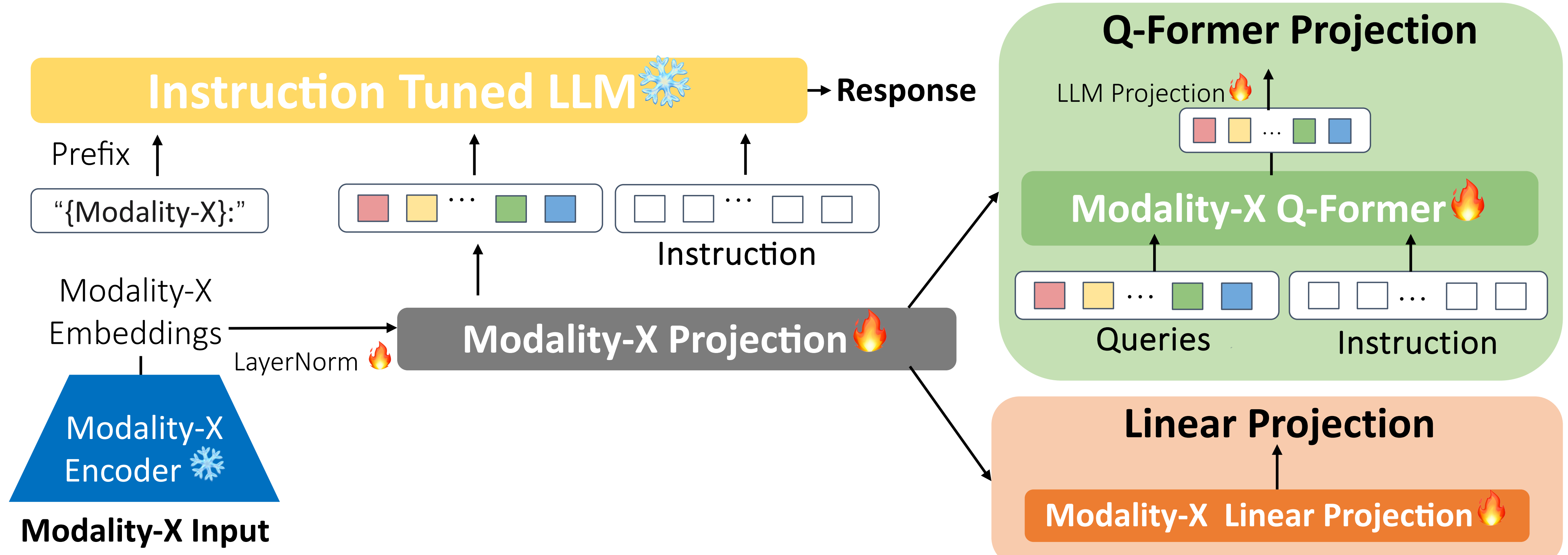
- Effective and efficient scalable framework for **independent modality alignment** to a frozen LLM showing **emergent reasoning across multiple modalities simultaneously.**
- Introduce the first benchmark **DisCRn** requiring models to perform discriminatory reasoning across multiple modalities.
- A **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

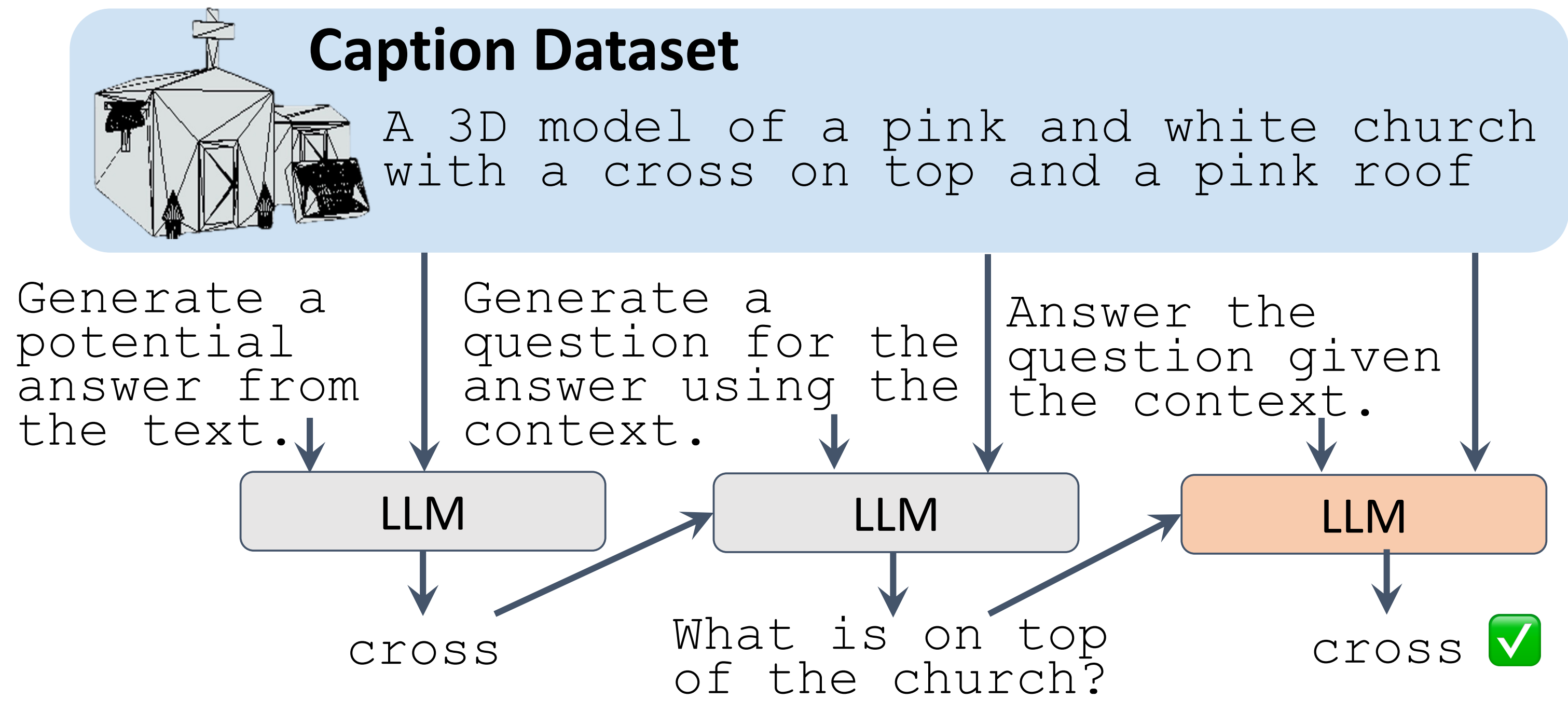
Each modality-to-language projection is trained **independently**.



Modality-to-Language Projection Types



Train: Instruction Tuning Data Generation from Captions


Caption Dataset
A 3D model of a pink and white church with a cross on top and a pink roof
Generate a potential answer from the text.
Generate a question for the answer using the context.
Answer the question given the context.
LLM
LLM
LLM
cross
What is on top of the church?
cross ✅

Round-Trip-Consistency

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

Automatically Collected Data

Evaluation: Discriminatory Cross-Modal Reasoning (DisCRn)

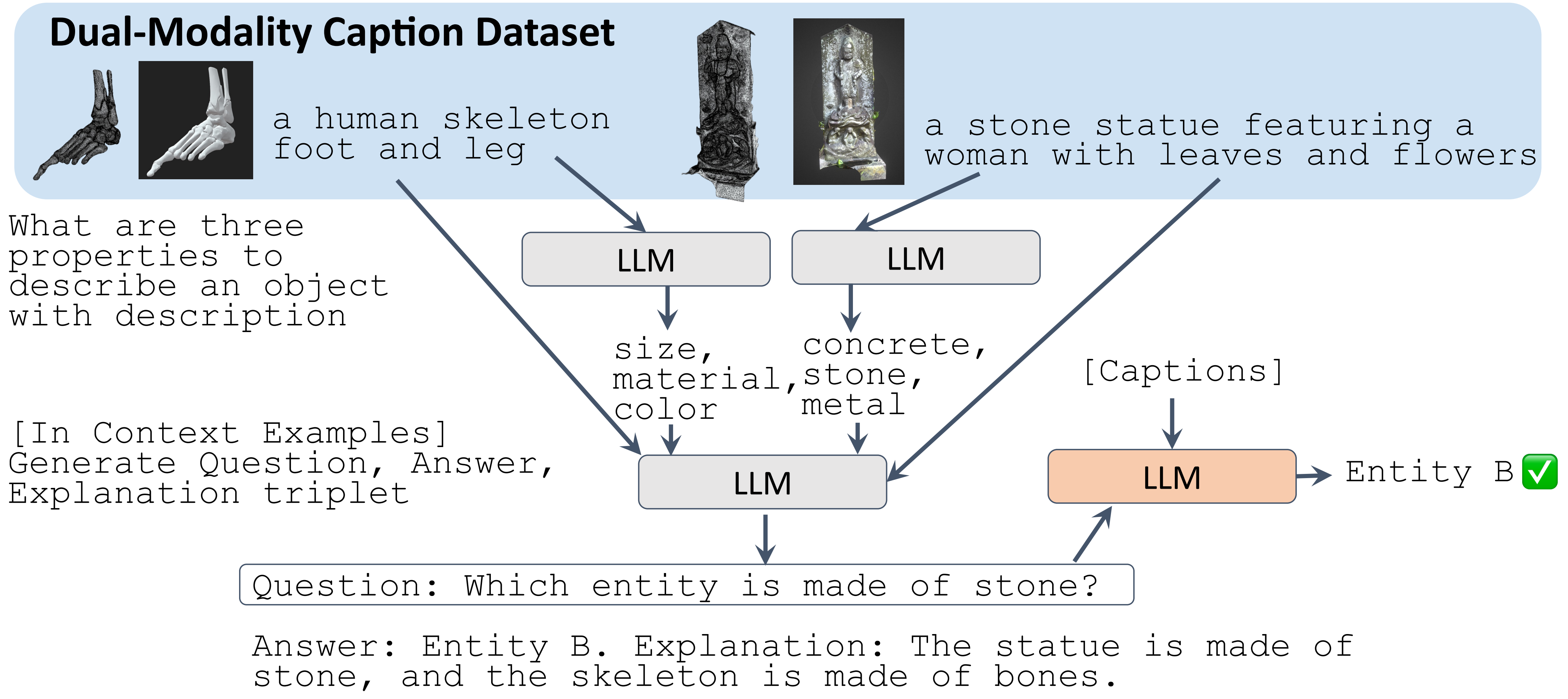
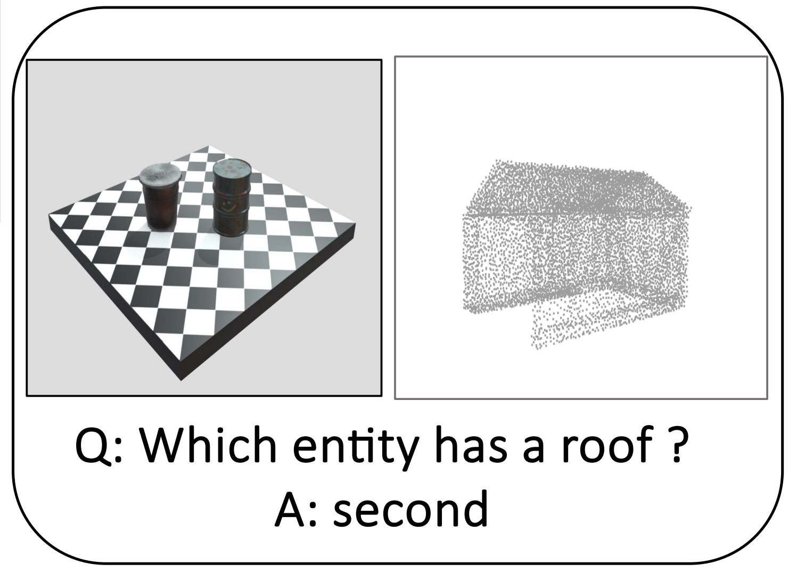
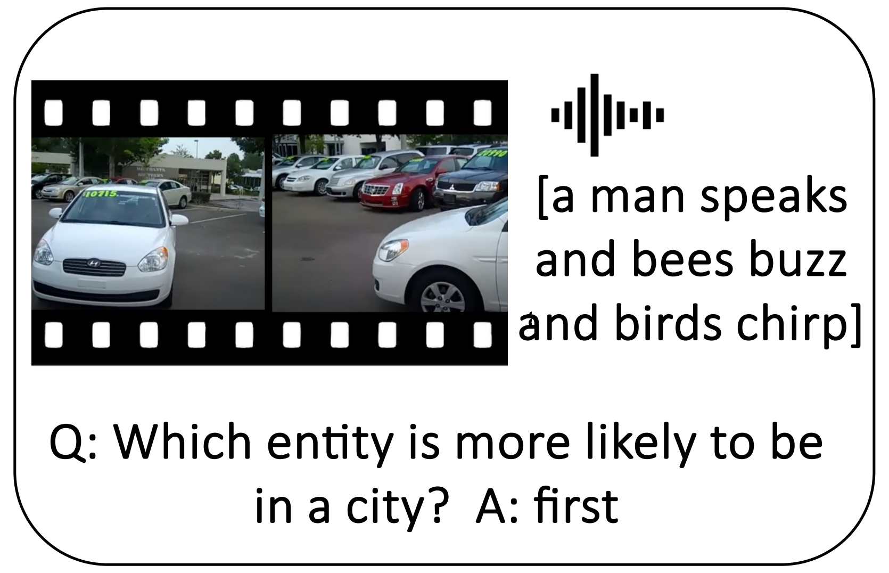

Dual-Modality Caption Dataset
a human skeleton foot and leg
a stone statue featuring a woman with leaves and flowers
What are three properties to describe an object with description
LLM
LLM
size, material, color
concrete, stone, metal
[In Context Examples] Generate Question, Answer, Explanation triplet
LLM
[Captions]
LLM
Entity B ✅
Question: Which entity is made of stone?
Answer: Entity B. Explanation: The statue is made of stone, and the skeleton is made of bones.

Image – 3D



Q: Which entity has a roof ?
A: second

Audio - Video



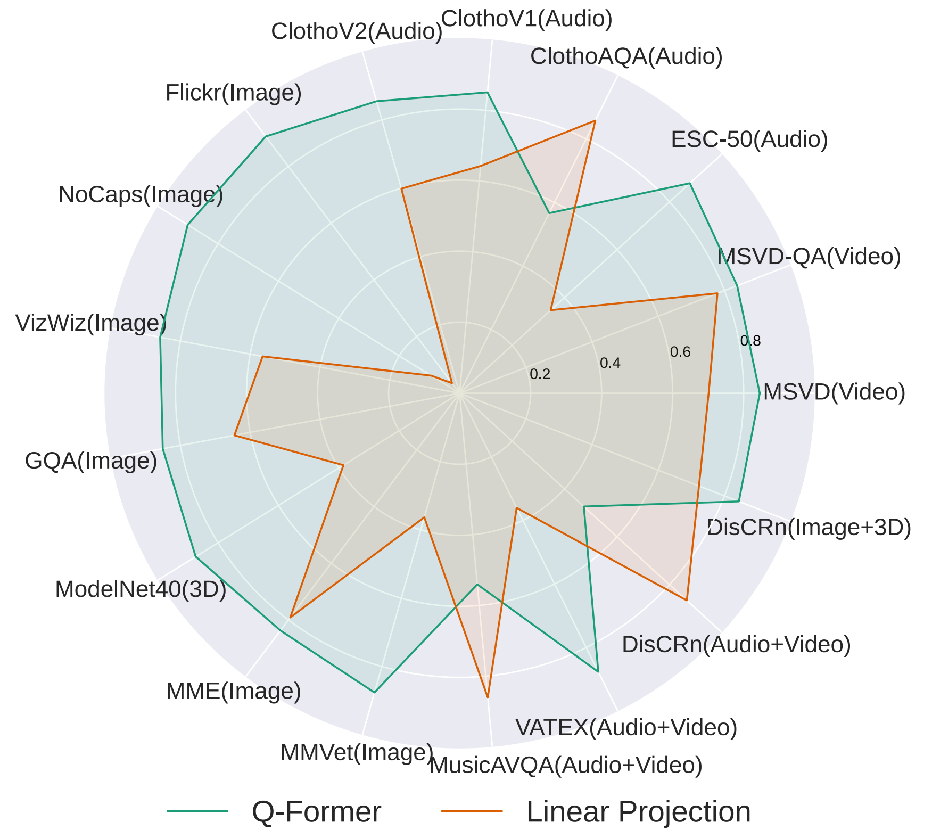
[a man speaks and bees buzz and birds chirp]

Q: Which entity is more likely to be in a city? A: first

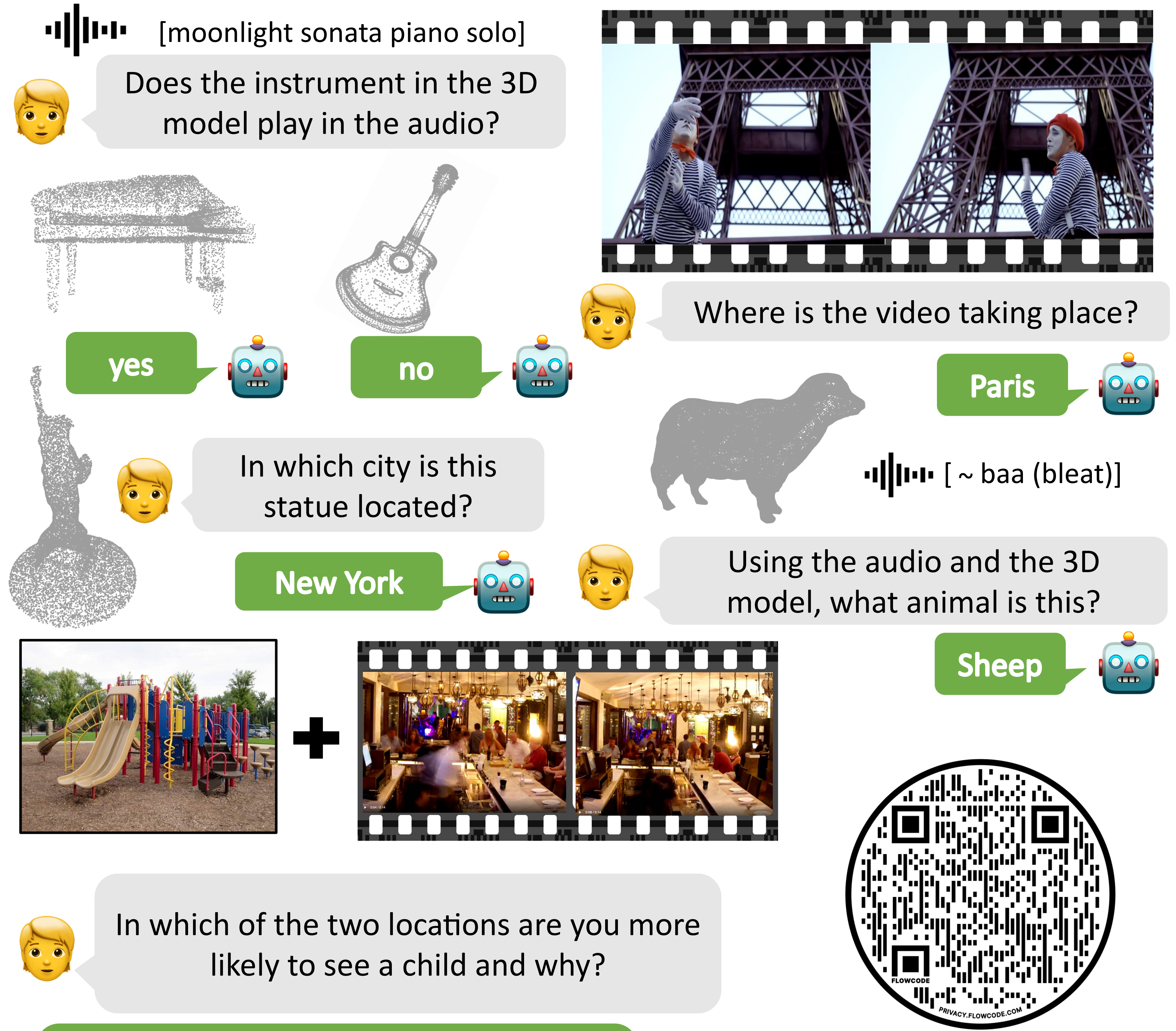
DisCRn	Examples
Image-3D	28,173
Audio-Video	8,802

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



[moonlight sonata piano solo]

Does the instrument in the 3D model play in the audio?

yes

no

Where is the video taking place?

Paris

In which city is this statue located?

New York

[~ baa (bleat)]

Using the audio and the 3D model, what animal is this?

Sheep

In which of the two locations are you more likely to see a child and why?

Playground is more likely to see a child because it is a place for kids to play and have fun.

Models, code, and data are available.