CSCE 689: Special Topics in Trustworthy NLP

Lecture 13: Multimodal Models

Kuan-Hao Huang
khhuang@tamu.edu

Project Proposal

- Comments available on Gradescope

Project Highlight Presentations (Oct 22, Online)

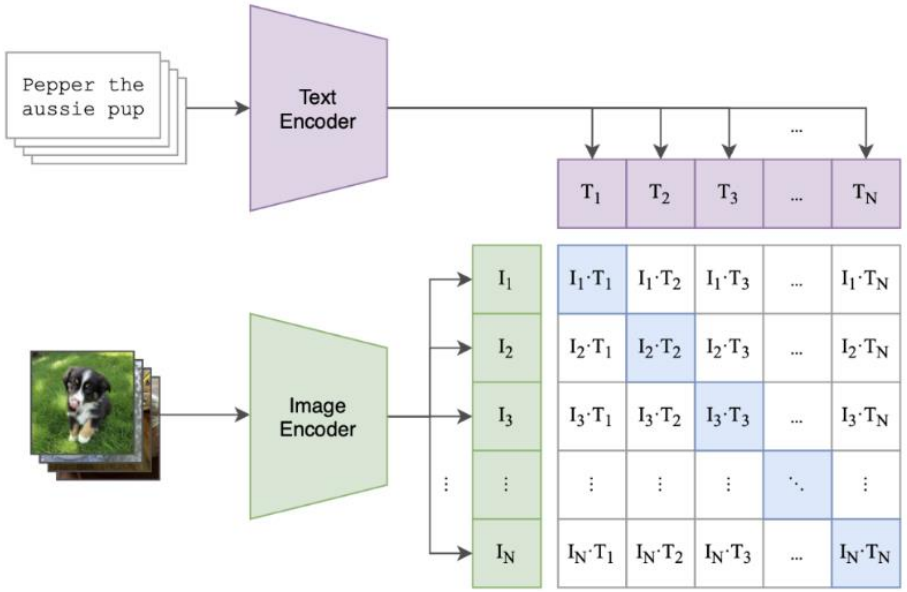
- 14 teams
- 4 mins of highlight presentations + 1 min Q&A
  - Introduction to the topic you choose and problem definition
  - Existing progress and challenges
  - Proposed solutions, novelty, and expected contributions
  - Planned implementation details, including dataset, models, codebases, etc.
  - Evaluation metrics
- Clarity is the most important thing
  - Teach your classmate about your topic

Project Highlight Presentations (Oct 22, Online)

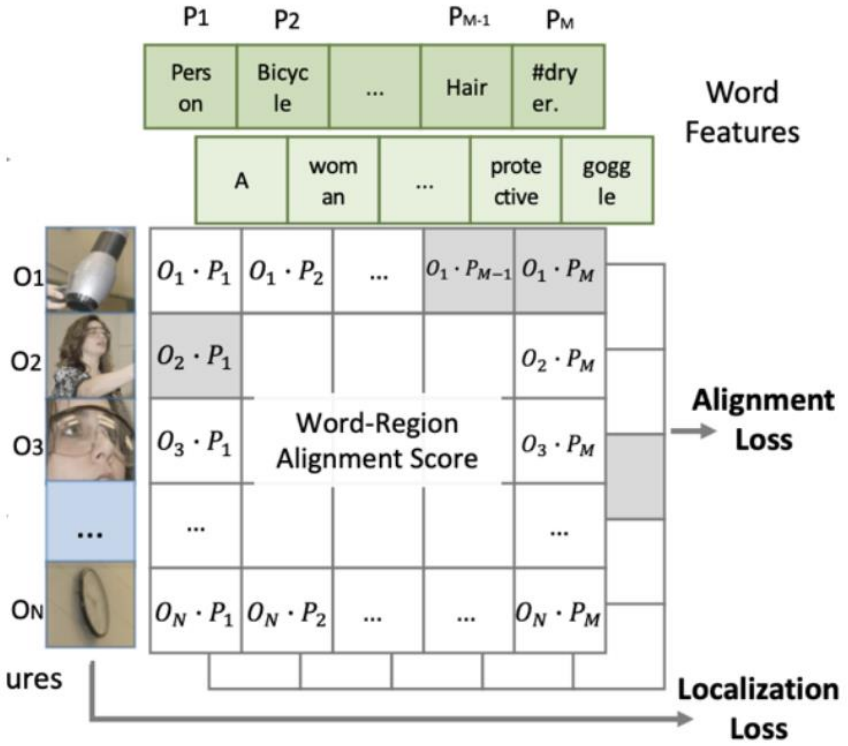
- Slide deck
  - https://docs.google.com/presentation/d/1s6IDCBmzLDIDnG5TaOSwIO-lwCRUEZxOqPYc6_PALOg/edit?usp=sharing
- Presentation order

Multimodal Models

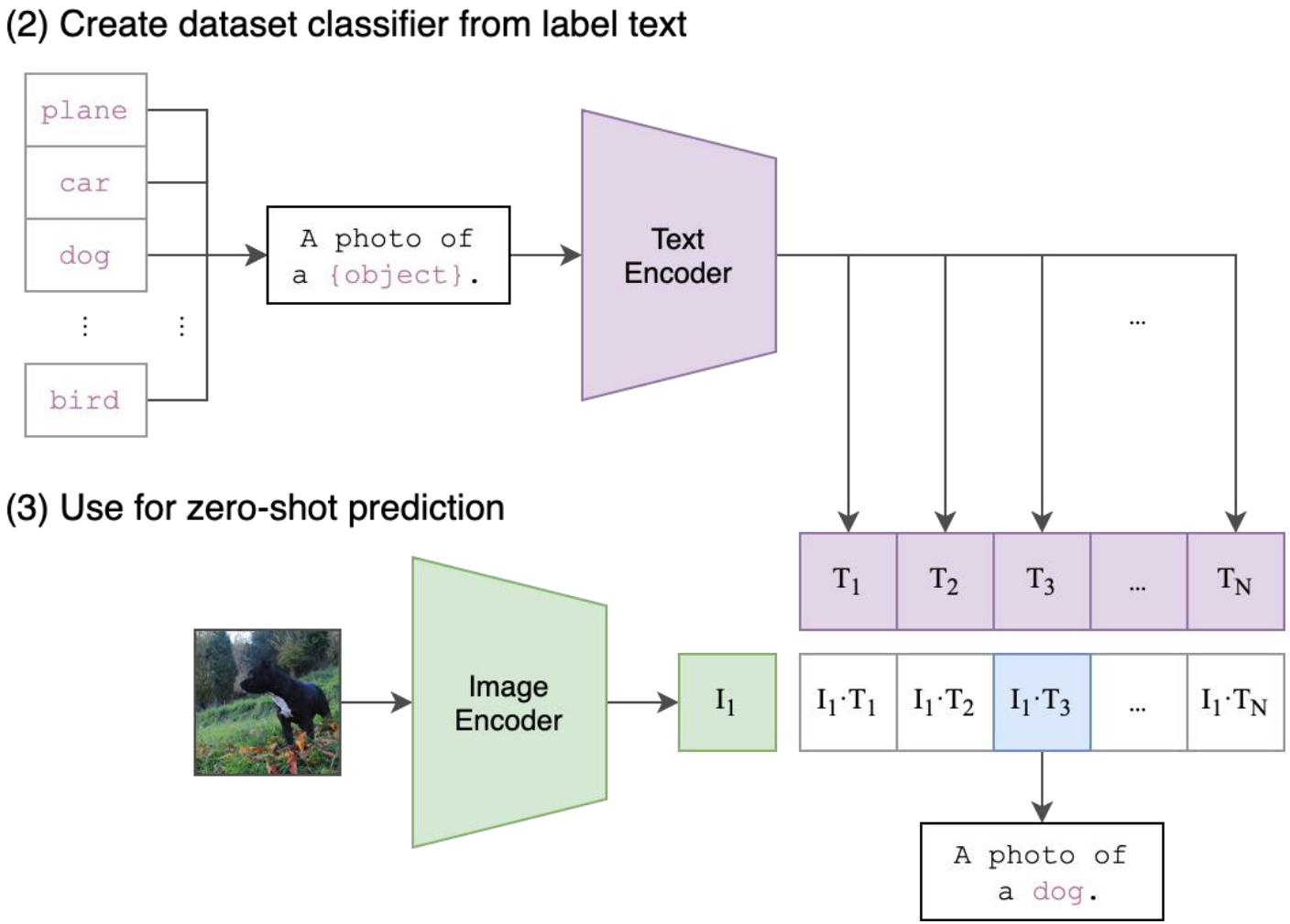
CLIP: capture information for whole image



GLIP: capture information more for objects/entities



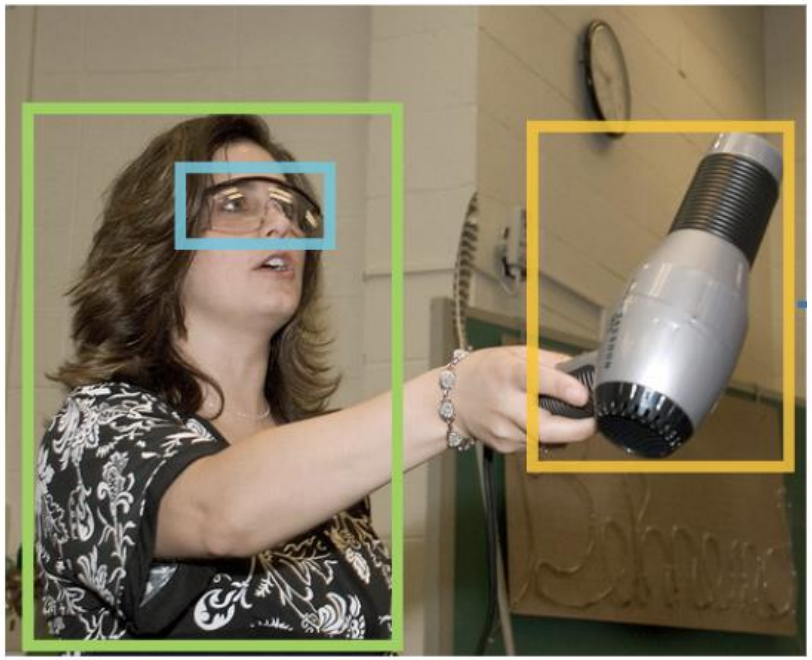
Zero-Shot Prediction



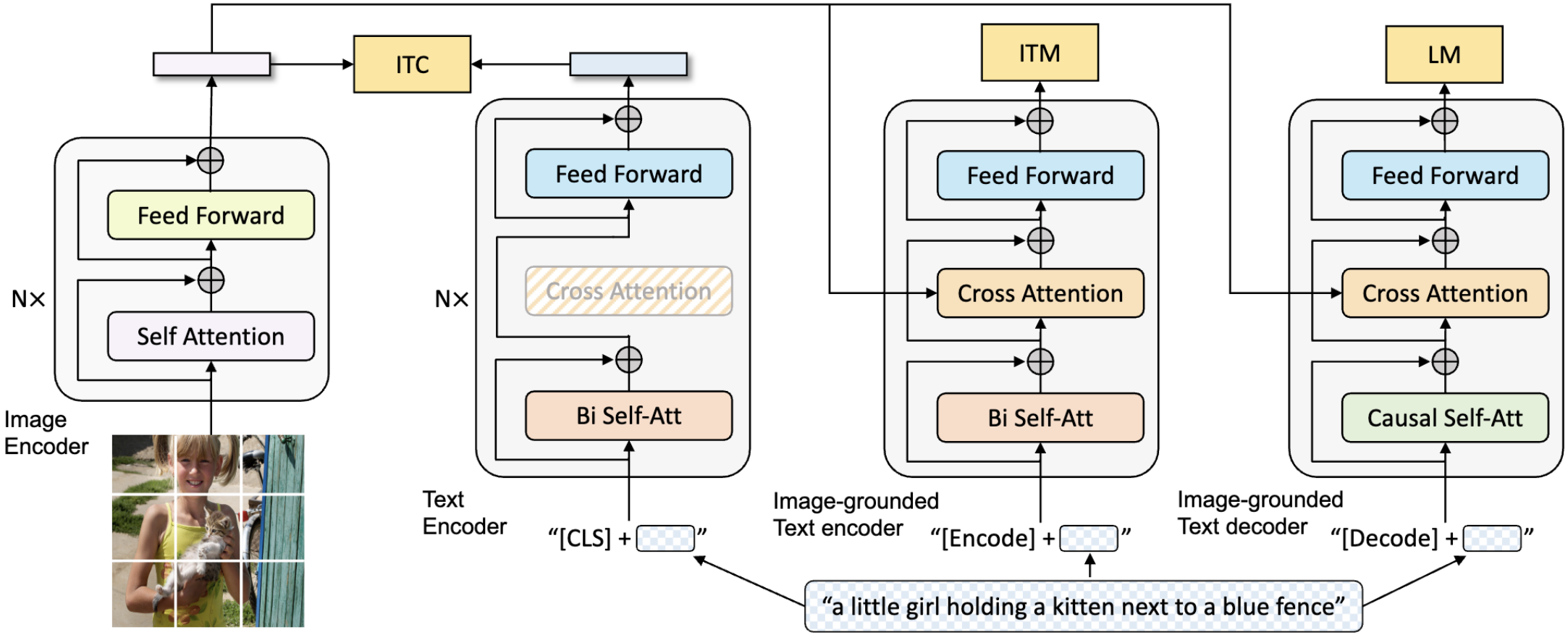
Object Detection and Text Grounding

Person. Bicycle ... Hairdryer.

A woman holds a blow dryer, wearing protective goggles

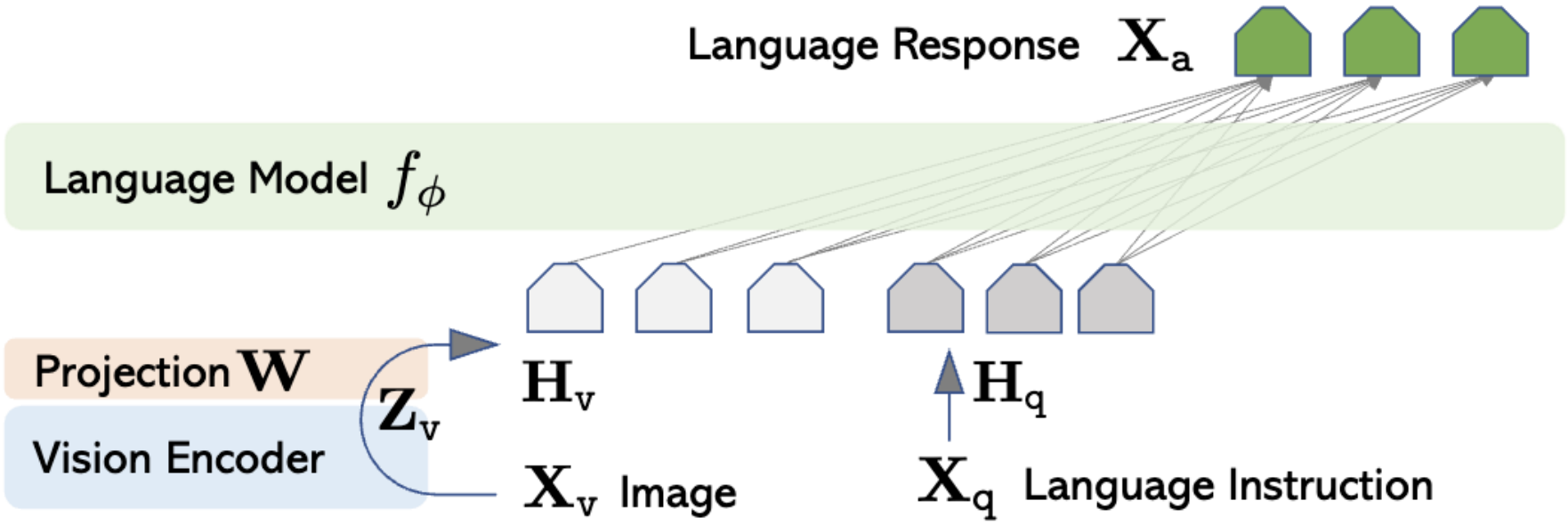


BLIP: Bootstrapping Language-Image Pre-training



BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation

LLaVA: Visual Instruction Tuning

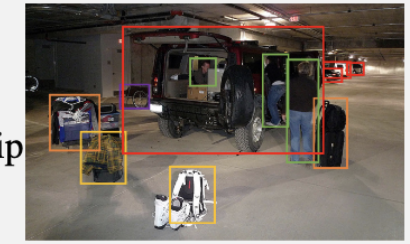


https://llava-vl.github.io/

Visual Question Answering

Context type 1: Captions

A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes

person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation

Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

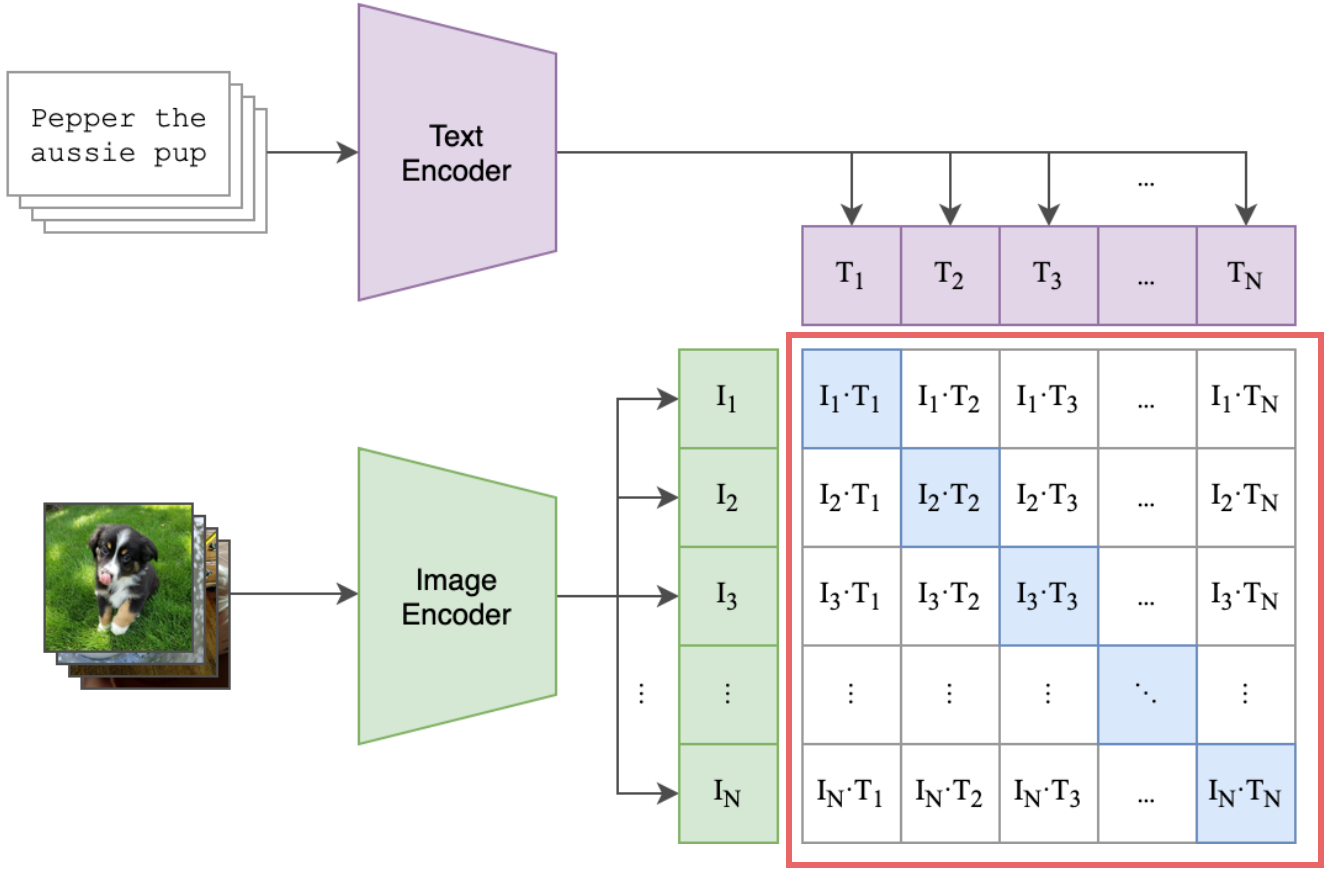
Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

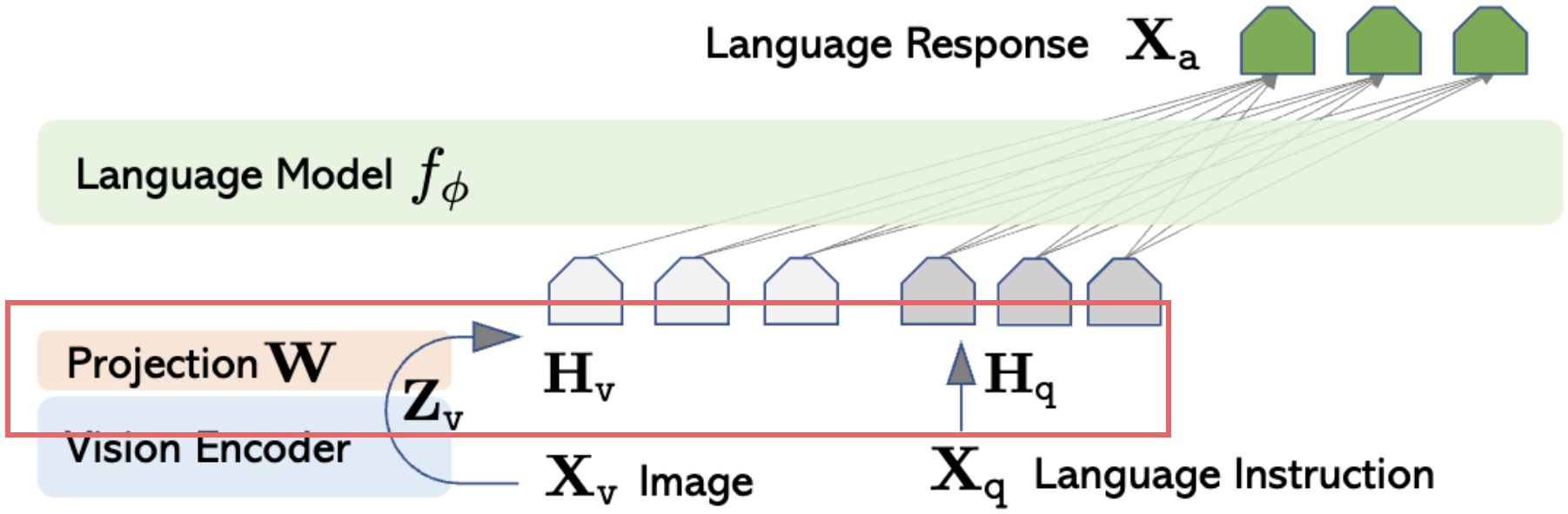
Response type 3: complex reasoning

Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Where is Bottleneck?



Where is Bottleneck?



Train with Image-Text Pairs