Multi-Human Behavior Prediction using Vision Language Models

IAAS Research

Utsav Panchal

Supervisors: Yuchen Liu, Dr Luigi Palmieri

Examiners: Prof. Dr. Marco Aiello, Dr. Ilche Georgievski

st184584@stud.uni-stuttgart.de

Institute of Architecture of Application Systems

University of Stuttgart

BOSCH

Date: 13.05.2025

Contents

1) Introduction

2) Previous works

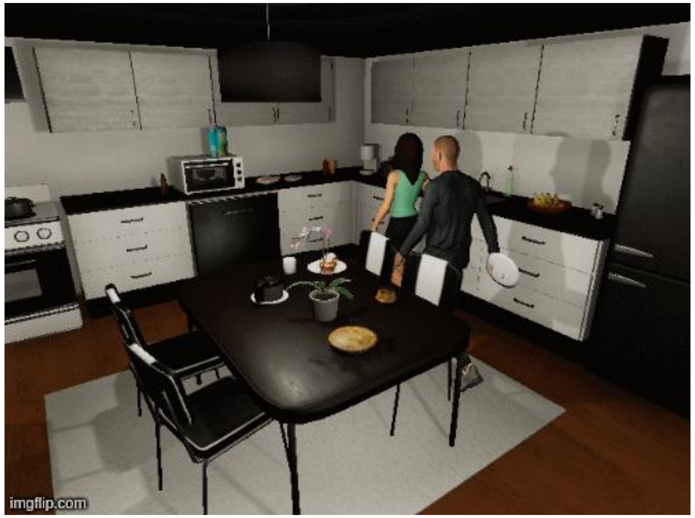
3) Approach

4) Evaluation

5) Conclusion

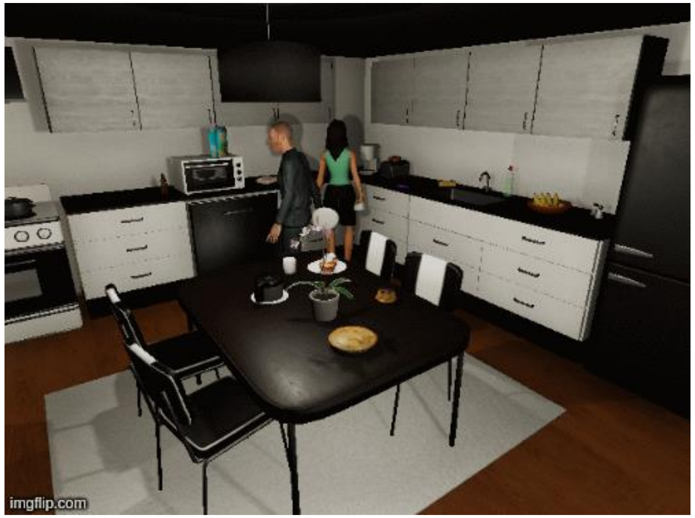
Introduction

What is Human Behavior Prediction ?



History Video Data

Predict



(person1, walk, dishwasher) (person2, walk, coffeemachine) ...

Future Actions

Problem Statement

- Given Video data, the objective is to predict future actions of humans in the scene.
- This work focuses on Multiple Human scenarios by utilizing VLMs.

Why Human Behavior Prediction ?

Will that pedestrian cross?

Automated Driving Scenarios[1]

Human-Robot Interaction[2]

- It is essential in scenarios where predicting user’s intent is crucial.

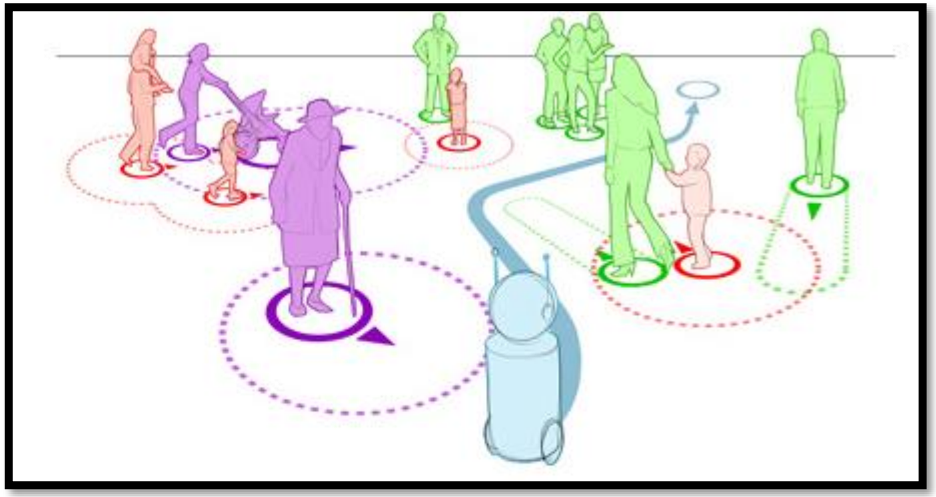
[1] Pedestrian Action Prediction Based on Deep Features Extraction of Human Posture and Traffic Scene : Available from: https://www.researchgate.net/figure/Other-objects-on-the-road-influence-predict-action-pedestrian_fig4_323162217 [accessed 2 May 2025]

[2] https://www.hrl.uni-bonn.de/research/human-robot-interaction

Why Multiple Human Behavior ?

Predicting multiple human actions is hard but crucial.

- >1 human in collaborative workspaces.
- External Dependencies.
- Partial Goals (private goals or intentions).



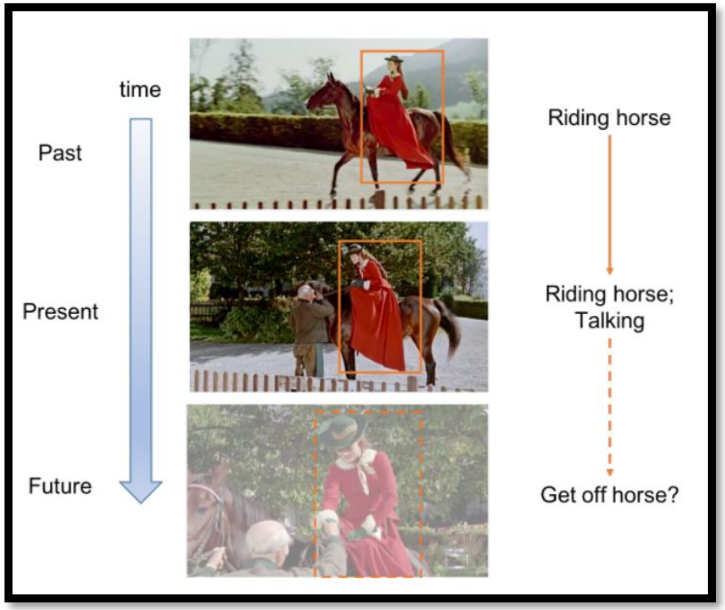
Robot Navigation in densly populated spaces[1]

[1] http://www.spencer.eu/project.html

Gaps in current research

- Mainly focused on egocentric action prediction.
- Limited to single-human scenario.
- Limited availability of datasets for indoor multiple human actions from a third-person's view.

Egocentric Video



Single Human Action Anticipation[1]

[1] Sun, Chen, et al. "Relational action forecasting." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019

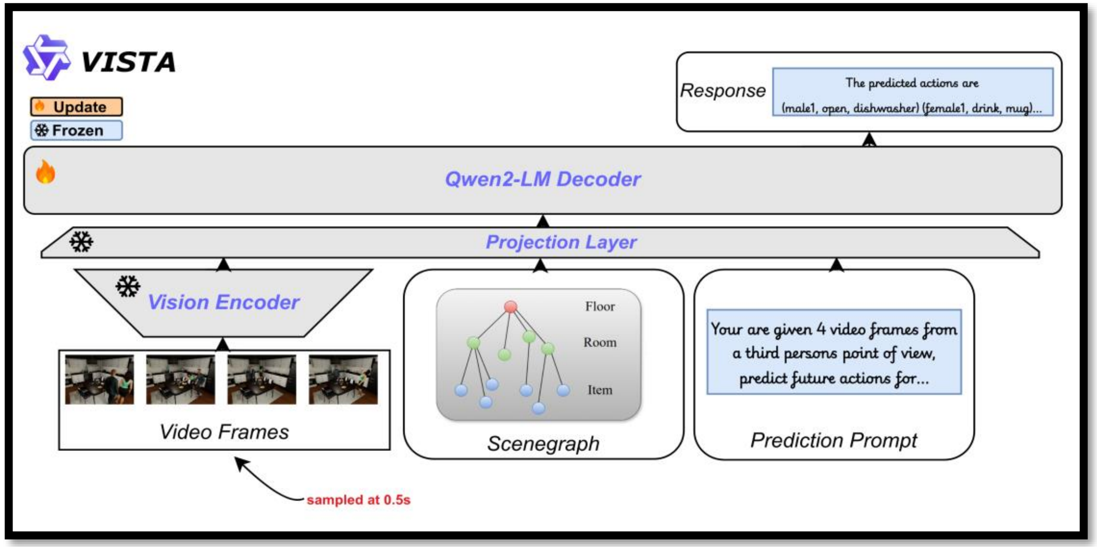
Contributions

To address the gaps

- ❖ Propose **VISTA**: **<u>VI</u>**sion And **<u>S</u>**cene Aware **<u>T</u>**emporal **<u>A</u>**ction Anticipation.
    - ➢ VLM-based Framework to predict multiple human behavior.
    - ➢ Evaluate on Synthetic and Real World Data.
    - ➢ 13% improvement against SOTA.
- ❖ Generate multiple-human indoor action dataset from third person view.

imgflip.com

Indoor Multiple Human Scenario



The predicted actions are (male1, open, dishwasher) (female1, drink, mug)...

Your are given 4 video frames from a third persons point of view, predict future actions for...

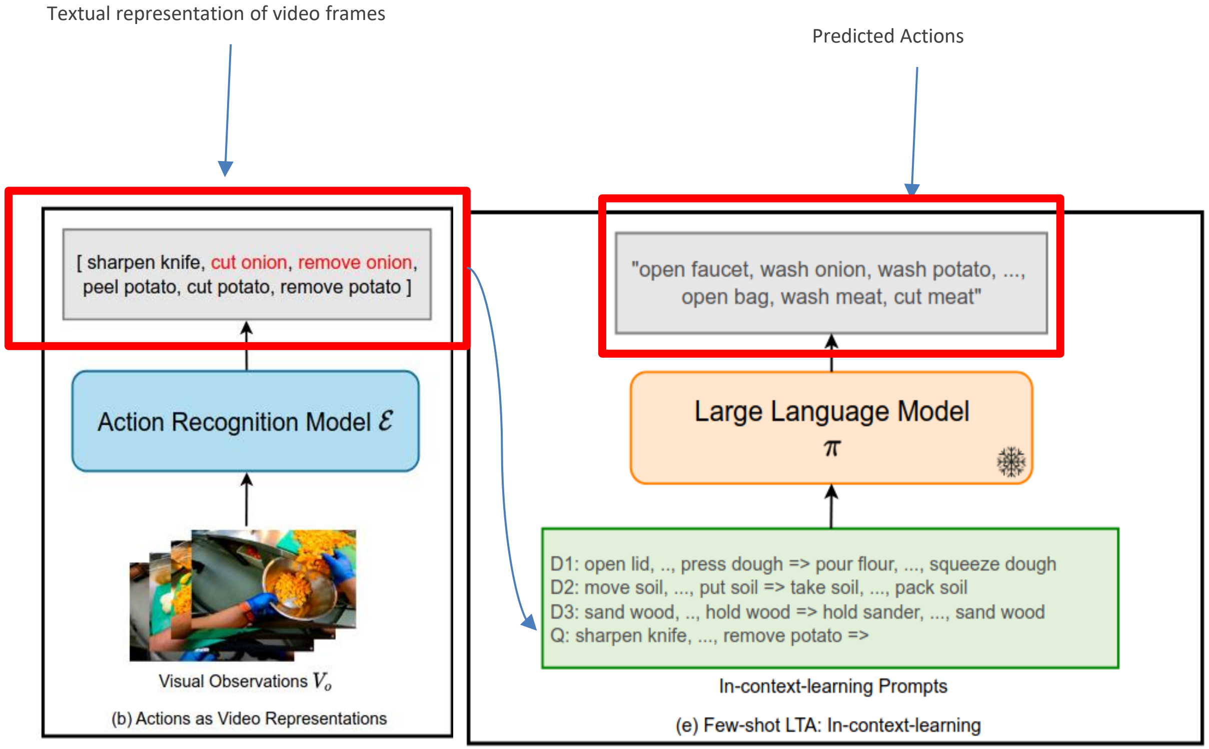
VISTA Framework

Background & Related Works

Egocentric Action Anticipation

- Until now most of the work is focused on egocentric action anticipation.
- This view is mostly suitable in AR applications.

Project Aria device[1]


Textual representation of video frames
Predicted Actions
[sharpen knife, cut onion, remove onion, peel potato, cut potato, remove potato]
Action Recognition Model $\mathcal{E}$
Visual Observations V_o
(b) Actions as Video Representations
"open faucet, wash onion, wash potato, ..., open bag, wash meat, cut meat"
Large Language Model
π
D1: open lid, .., press dough => pour flour, ..., squeeze dough
D2: move soil, ..., put soil => take soil, ..., pack soil
D3: sand wood, .., hold wood => hold sander, ..., sand wood
Q: sharpen knife, ..., remove potato =>
In-context-learning Prompts
(e) Few-shot LTA: In-context-learning

AntGPT[3]

[1] https://www.projectaria.com/
[2] Zhao, Qi, et al. "Antgpt: Can large language models help long-term action anticipation from videos?." *arXiv preprint arXiv:2307.16368* (2023).
[3] Wardle, Richard & Rowlands, Sareh. (2023). Deep-learning Based Egocentric Action Anticipation: A Survey. 10.21203/rs.3.rs-3156532/v1.

Problems with Egocentric View

Egocentric Views are not suitable for robots.

- Miss other agents and their interactions.
- Lack of Global Scene Understanding.
- Different Perspective.

Egocentric Video

Third Person view for Robots

A shift towards third person's view is necessary for robotic applications.

Benefits

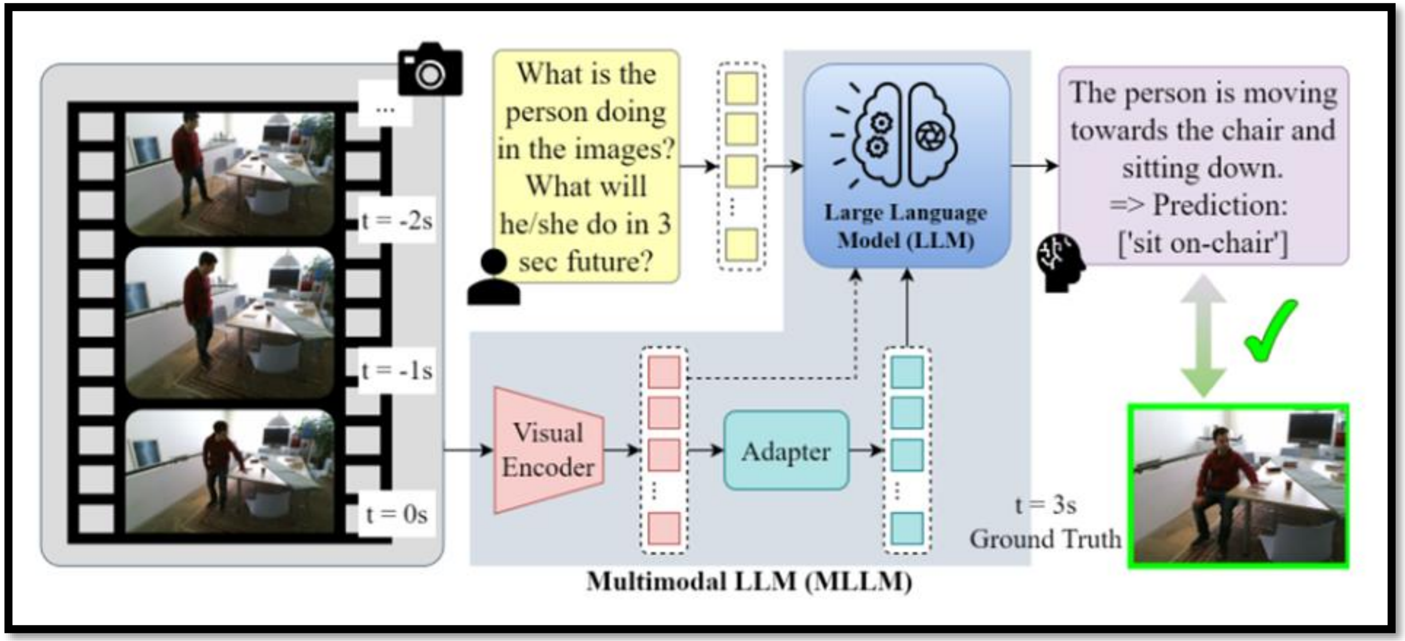
- Captures full body poses.
- Captures surrounding environment.
- Provides broader view to capture multiple humans.

Anticipating Human Behavior[1]

[1] https://www.hrl.uni-bonn.de/research/human-robot-interaction

Third Person Action Anticipation

- Previous methods used LSTM/RNN for action anticipation.
- SOTA use LLM based methods from external view point.



...
t = -2s
t = -1s
t = 0s
What is the person doing in the images? What will he/she do in 3 sec future?
Large Language Model (LLM)
The person is moving towards the chair and sitting down.
=> Prediction:
['sit on-chair']
Visual Encoder
Adapter
t = 3s
Ground Truth
Multimodal LLM (MLLM)

Context Aware Human Behavior Prediction[2]

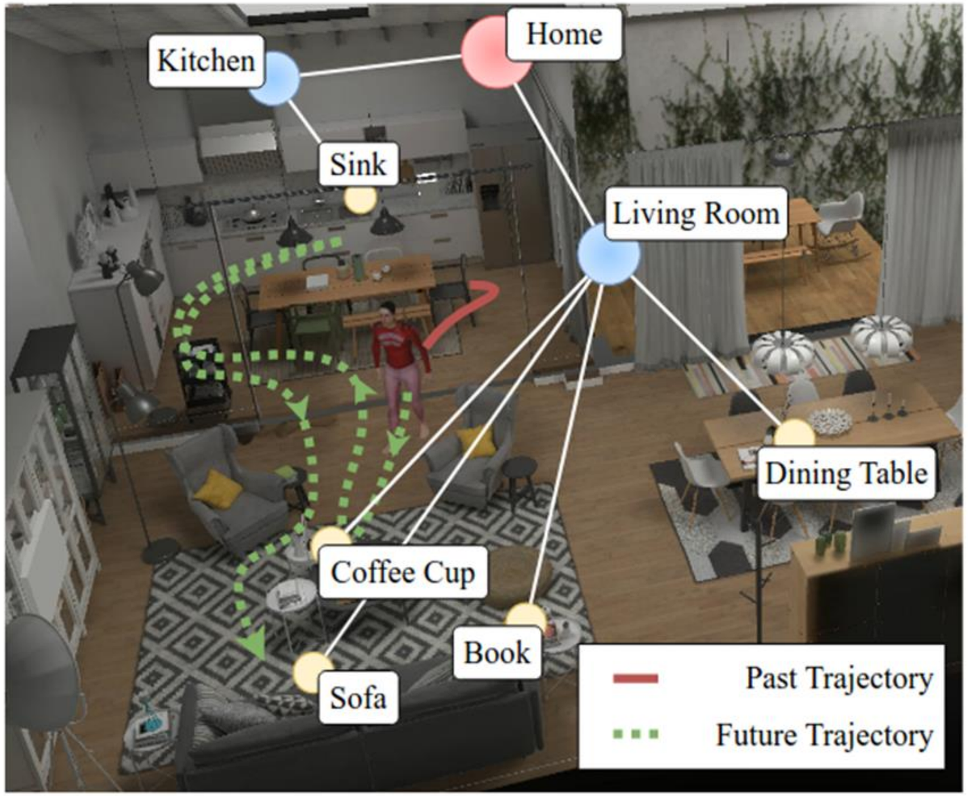
[1] Liu, Yuchen, et al. "Context-Aware Human Behavior Prediction Using Multimodal Large Language Models: Challenges and Insights." *arXiv preprint arXiv:2504.00839* (2025).

[2] Graule, Moritz A., and Volkan Isler. "Gg-llm: Geometrically grounding large language models for zero-shot human activity forecasting in human-aware task planning." *ICRA*. IEEE, 2024.

Spatial Awareness for robot

- Humans are more likely to interact with objects in environment.

- Information of environment is given using Scene graphs.



Human Trajectory Prediction using 3D Scene Graphs[1]

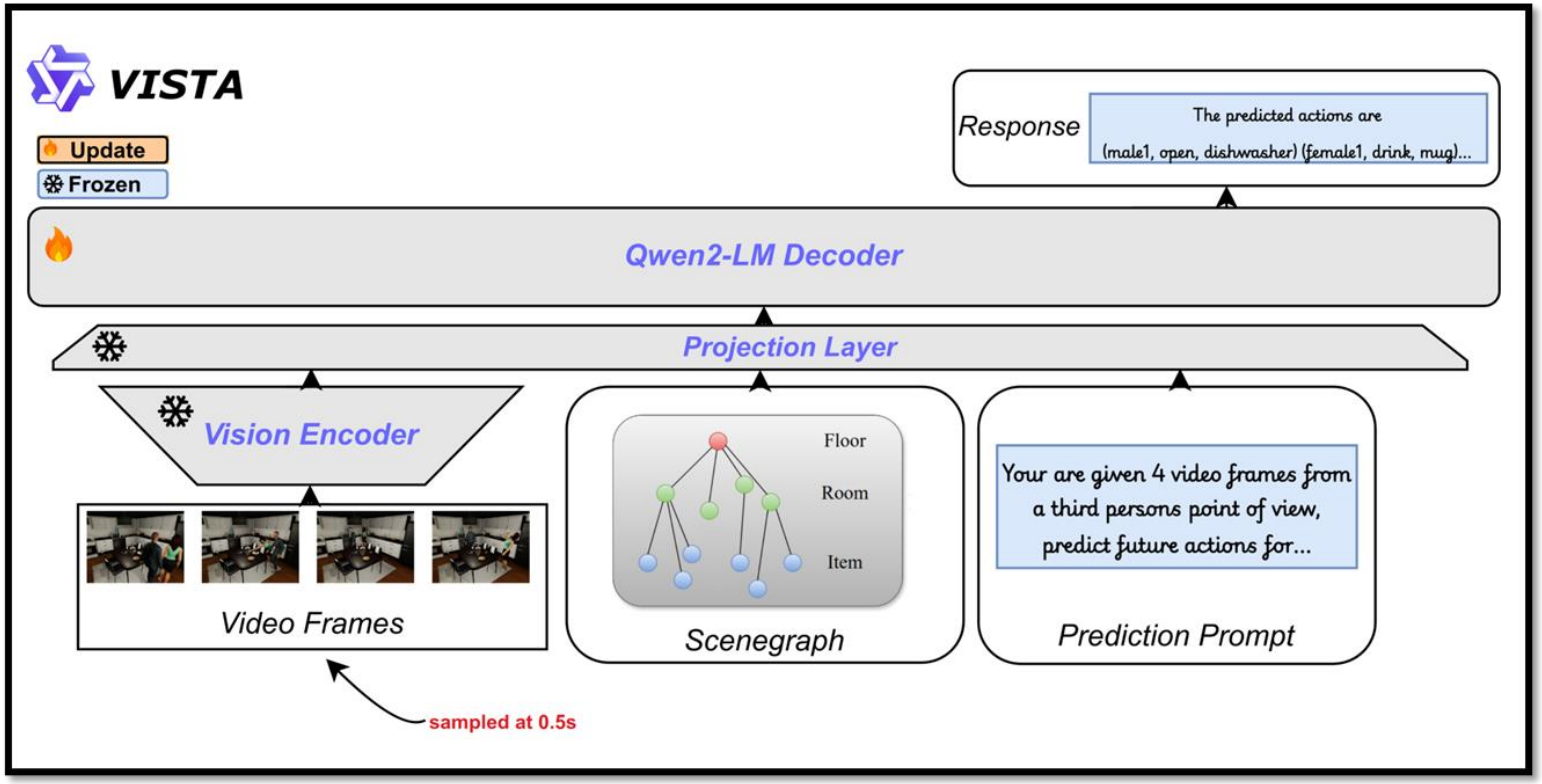
1. Gorlo, Nicolas, Lukas Schmid, and Luca Carlone. "Long-Term Human Trajectory Prediction using 3D Dynamic Scene Graphs." *IEEE Robotics and Automation Letters* (2024).

Methodology

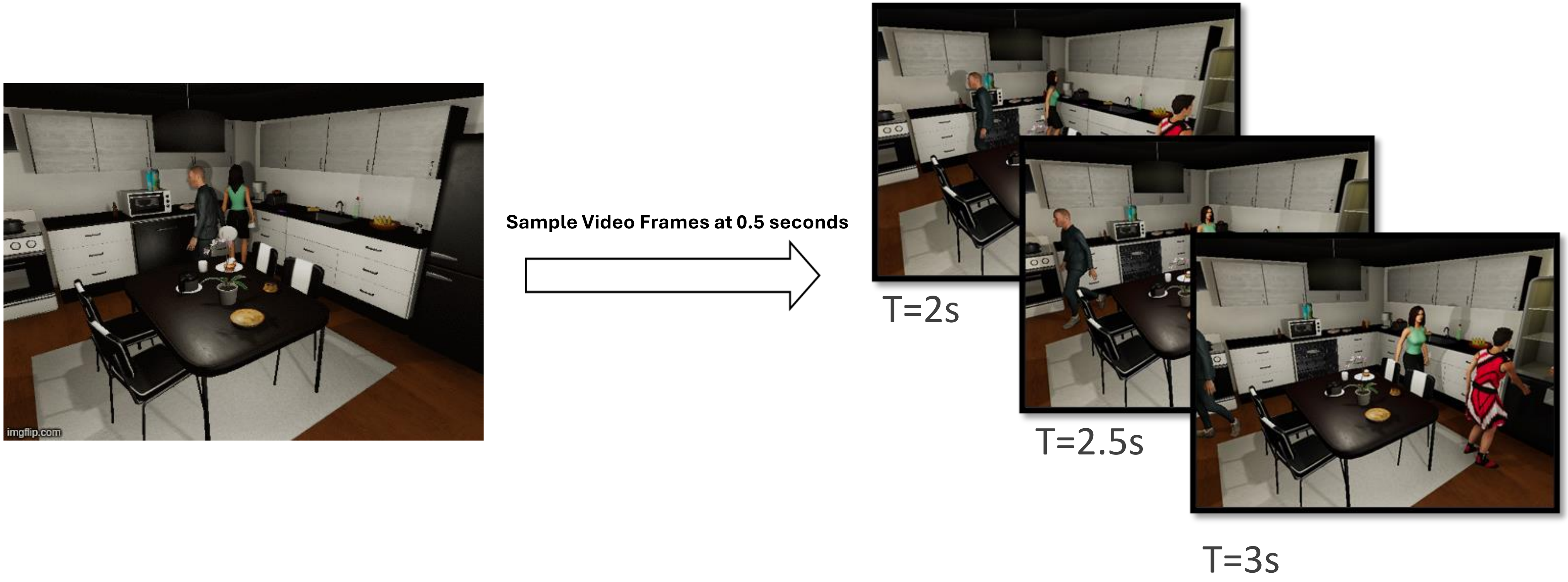
Architecture

- ❑ **Inputs**
  - ➢ Video Frames
  - ➢ Scene graph
  - ➢ Prediction Prompt
- ❑ **Output**
  - ➢ Predicted actions in natural language

Vista: **Vi**sion and **S**cene Aware **T**emporal **A**ction Anticipation


VISTA
Update
Frozen
Response
The predicted actions are (male1, open, dishwasher) (female1, drink, mug)...
Qwen2-LM Decoder
Projection Layer
Vision Encoder
Floor
Room
Item
Your are given 4 video frames from a third persons point of view, predict future actions for...
Video Frames
Scenegraph
Prediction Prompt
sampled at 0.5s

Visual Representation


imgflip.com

Sample Video Frames at 0.5 seconds

T=2s

T=2.5s

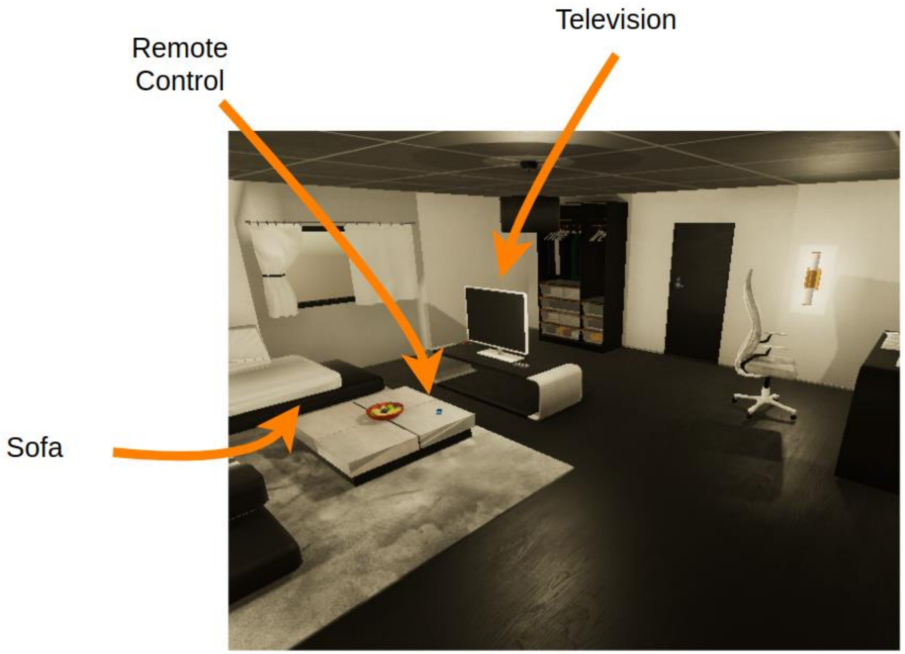
T=3s

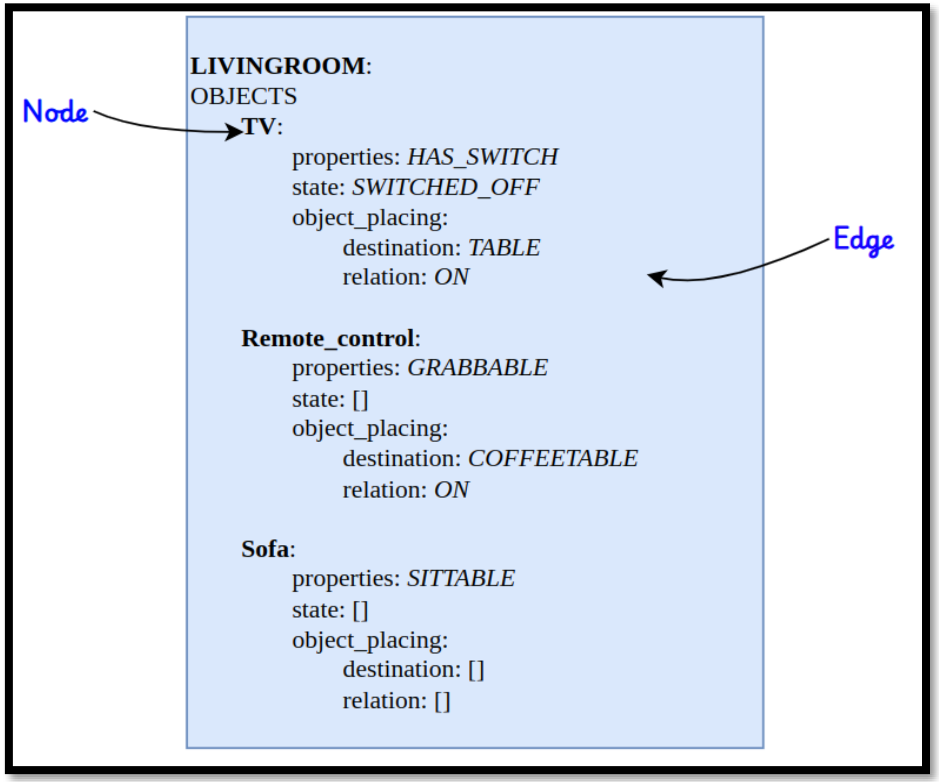
Video Data

Sampled Frames

- After an informal validation, we found that 0.5s sampling represents dynamic actions effectively.

Scene Graph





```
LIVINGROOM:
OBJECTS
   TV:
          properties: HAS_SWITCH
          state: SWITCHED_OFF
          object_placing:
                 destination: TABLE
                 relation: ON

   Remote_control:
          properties: GRABBABLE
          state: []
          object_placing:
                 destination: COFFEETABLE
                 relation: ON

   Sofa:
          properties: SITTABLE
          state: []
          object_placing:
                 destination: []
                 relation: []
```

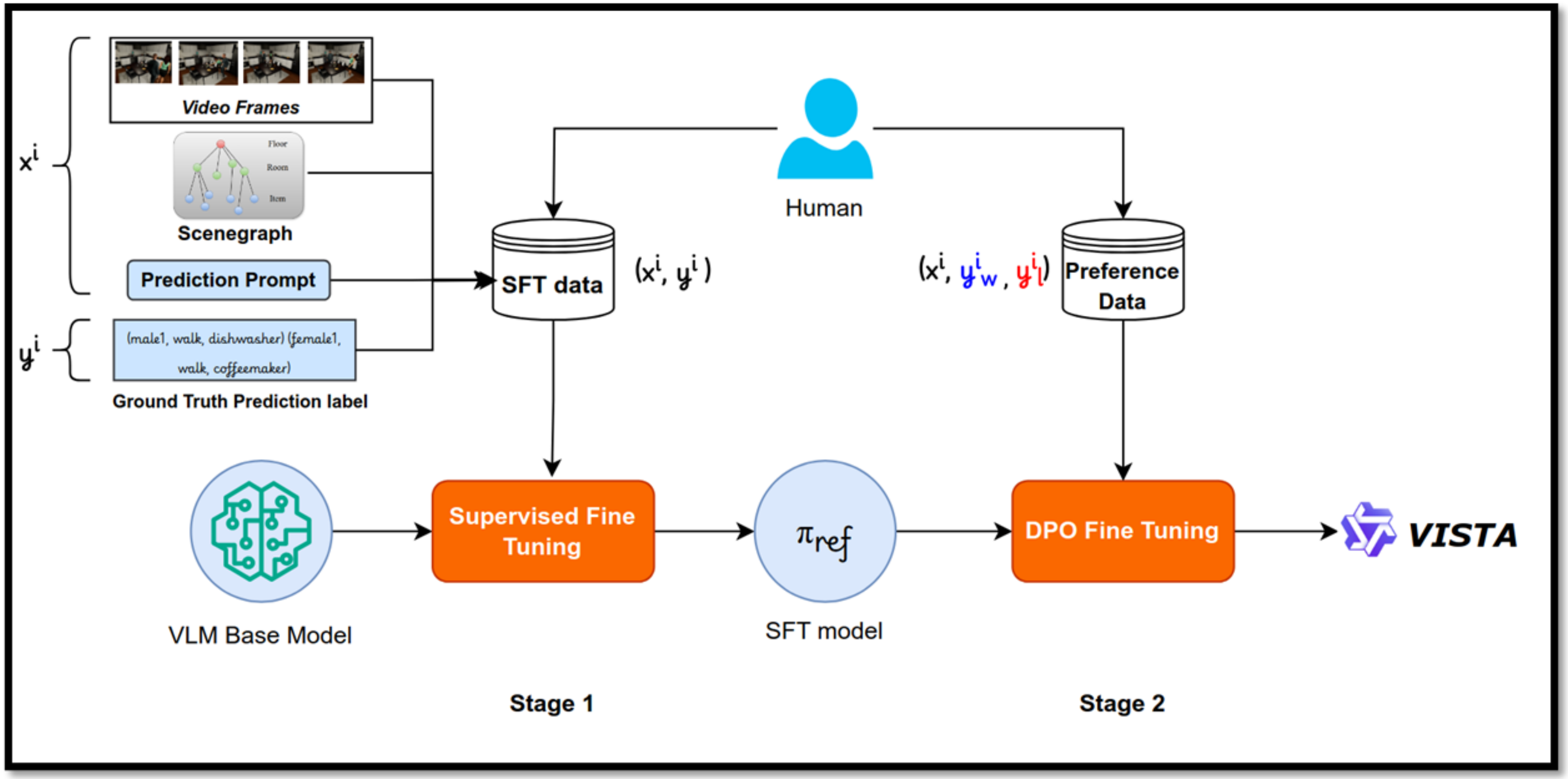
Scene Graph Format

- Scene Graph (G) contains a Node List (N) & Edge List (E).
- Each Node represents an **object**.
- For each object: **properties**, **state** and **an edge**

Fine Tuning Method

SFT + DPO for fine tuning

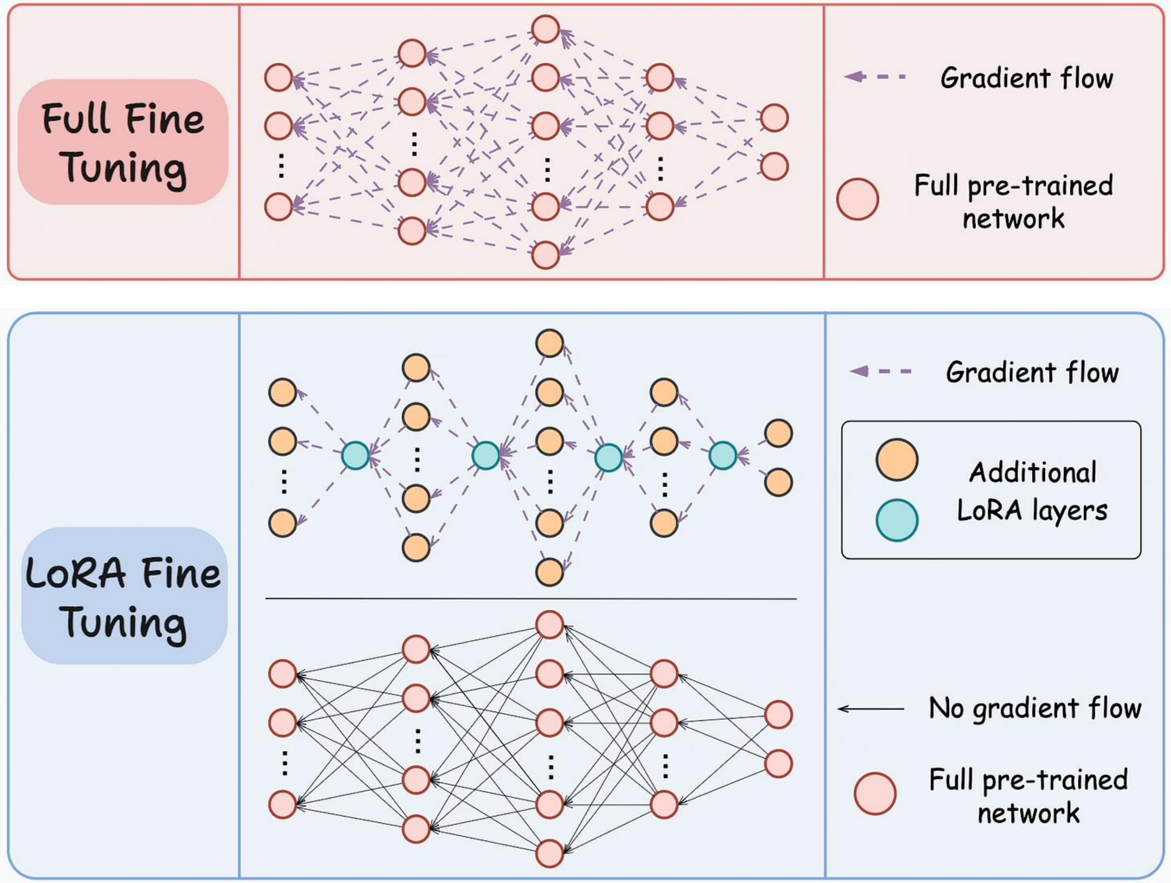
- Stage 1: Supervised Fine Tuning
- Stage 2: Direct Preference Optimization



SFT with Low Rank Adaptation (LoRA)

Benefits

1. Model still keeps its original knowledge.
2. Computationally feasible to train



Traditional Vs LoRA Fine Tuning

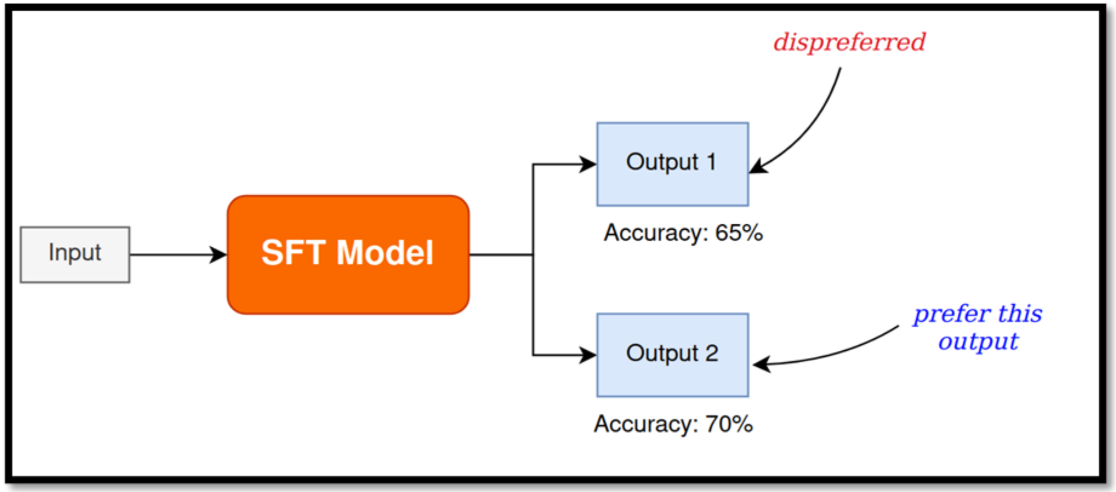
1. https://blog.dailydoseofds.com/p/full-model-fine-tuning-vs-lora-vs

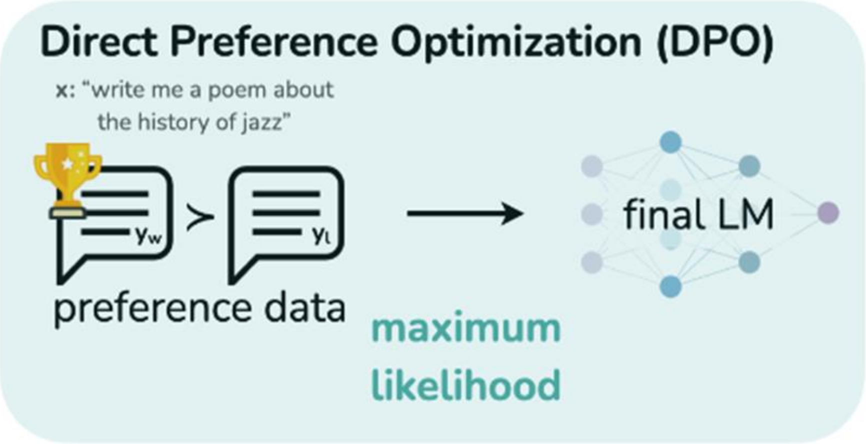
Direct Preference Optimization

- DPO is an alignment technique.
- For given two outputs, we want the model to produce preferred output.

Preference Data Contains

1. X_i – Original Input.
2. Y_w – Preferred Output
3. Y_l – Non Preferred Output



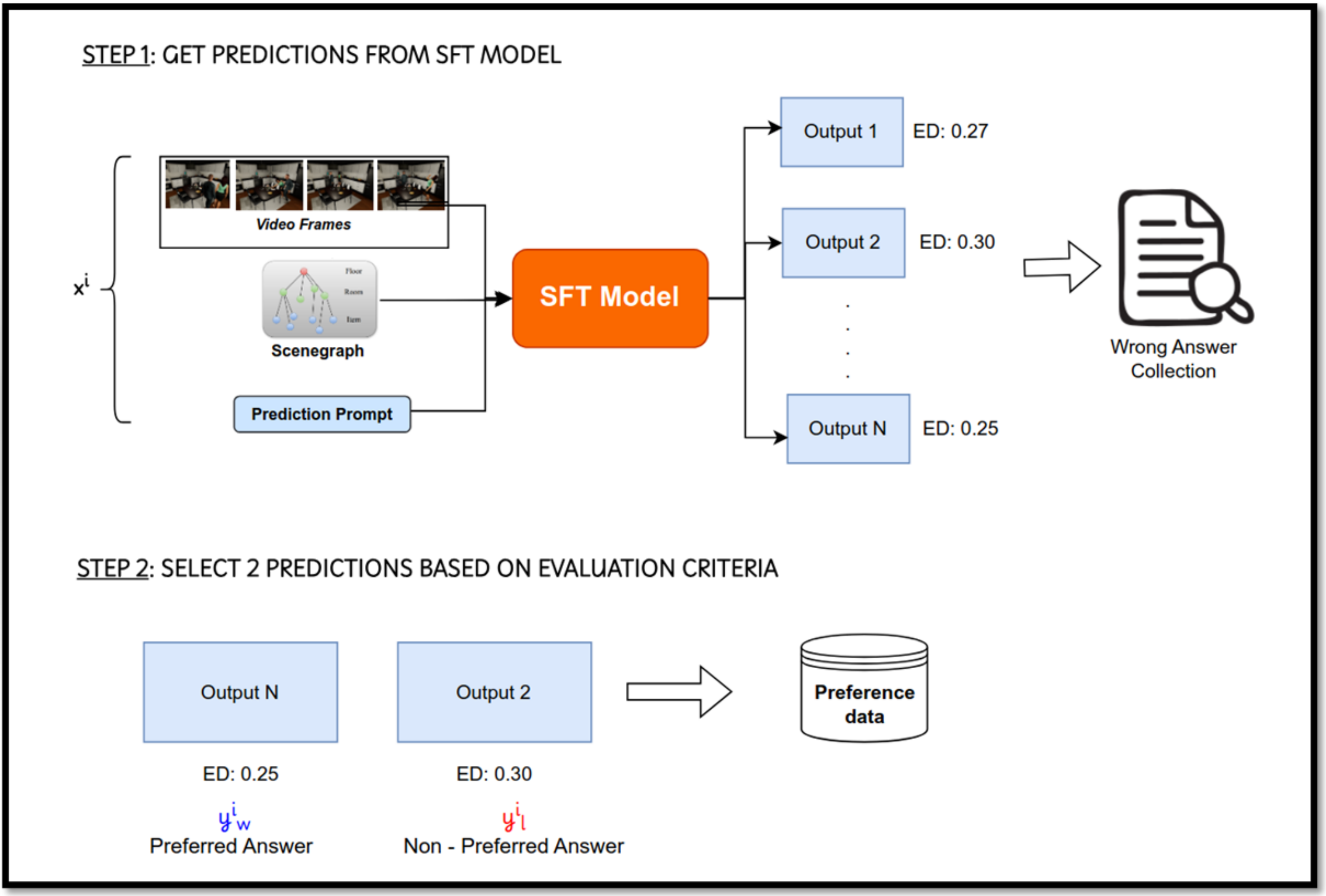


Direct Preference Optimization (DPO)
x: "write me a poem about the history of jazz"
y_w
>
y_l
preference data
maximum likelihood
final LM

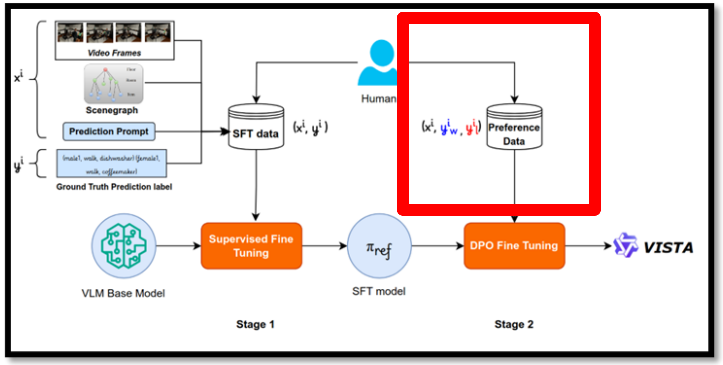
DPO overview

[1] Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." *Advances in Neural Information Processing Systems* 36 (2023): 53728-53741.

DPO Preference Data Building

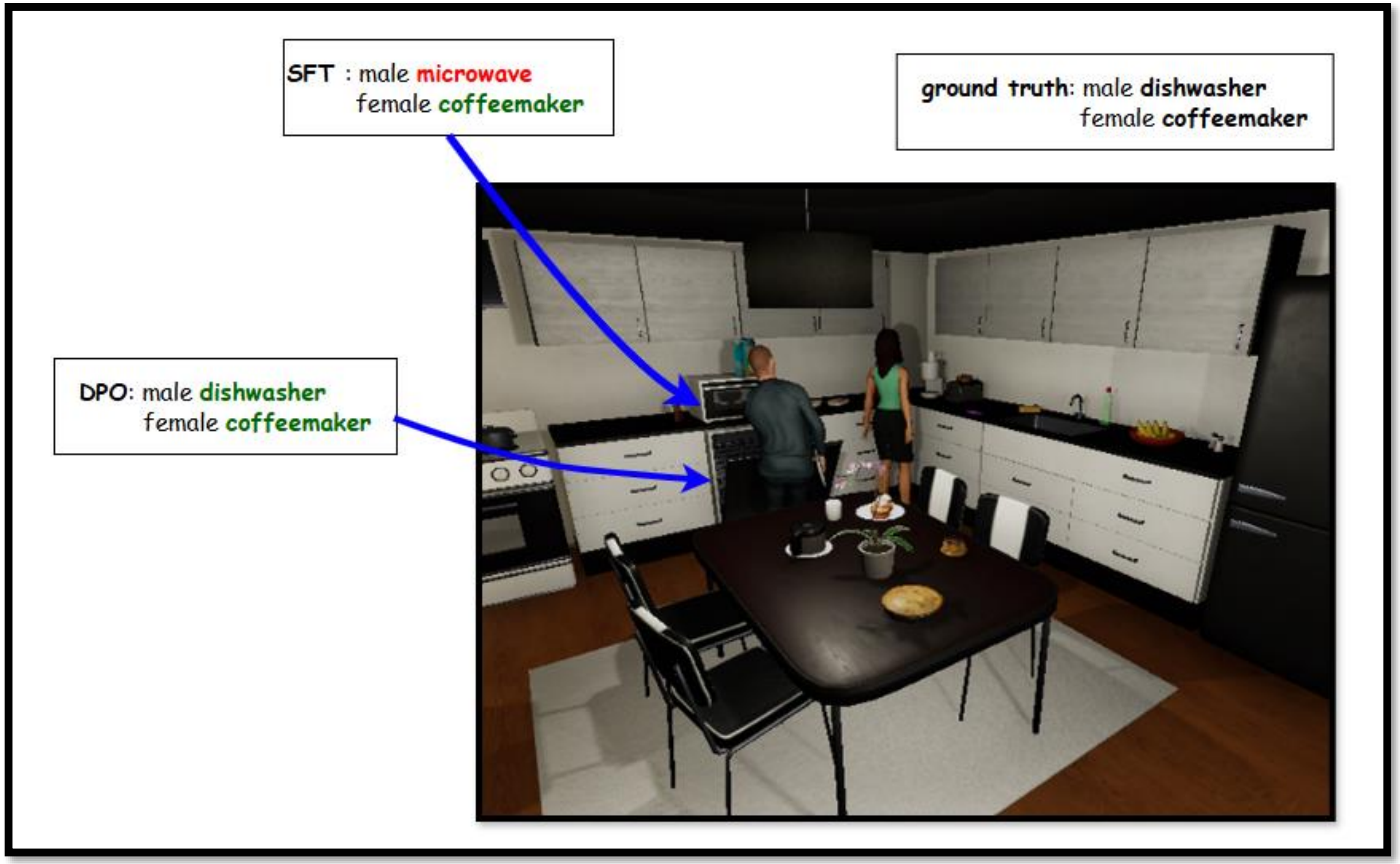


Preference Data Building Process



Why SFT + DPO ?

- Model is refined to generate human preferred answer.
- Target adjustments are needed from SFT model
- To correct inaccurate text regarding Visual Content



Datasets

- ❑ Synthetic Videos: Kitchen, Livingroom & Bedroom Scenario
  - ➢ 30 Videos comprising 1, 2 & 3 agents.

- ❑ Recorded Videos: Kitchen & Communication Zone.
  - ➢ 12 Videos comprising 2 & 3 agents.

Synthetic Video Frames

Recorded Video Frames

Evaluation

Evaluation Overview

Model Selection

Qwen2-VL

GPT - 4

Qwen 2 VL: 2B, 7B, 72B
(for fine tuning)

GPT-4o & GPT-4o-mini
(ablation studies)

Metrics

1. Accuracy : Complete string match
2. Edit Distance : Character match
   - Primary Metric
3. Cosine Similarity : Semantic Similarity

Pred: ”(male, open, microwave) (female, stand, coffemaker)”
GT : “(male, open, dishwasher) (female, stand, coffemaker)”

Baselines

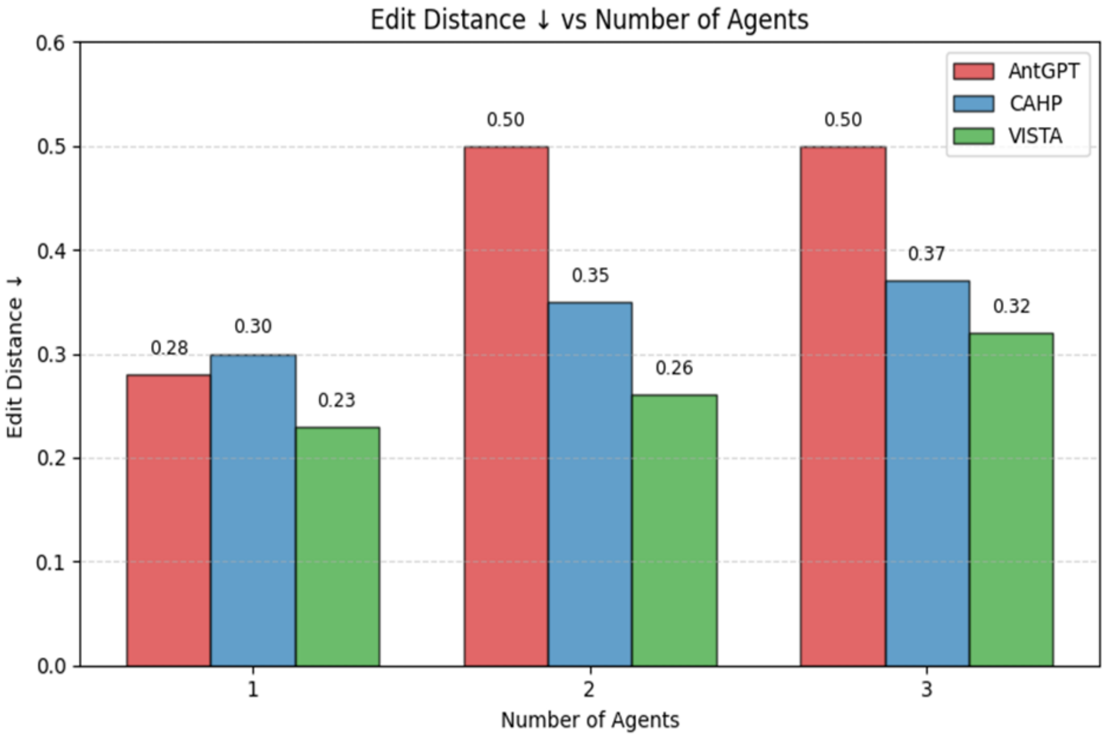
1. Context Aware Human Behavior Prediction – Third Person View
2. AntGPT – First Person View

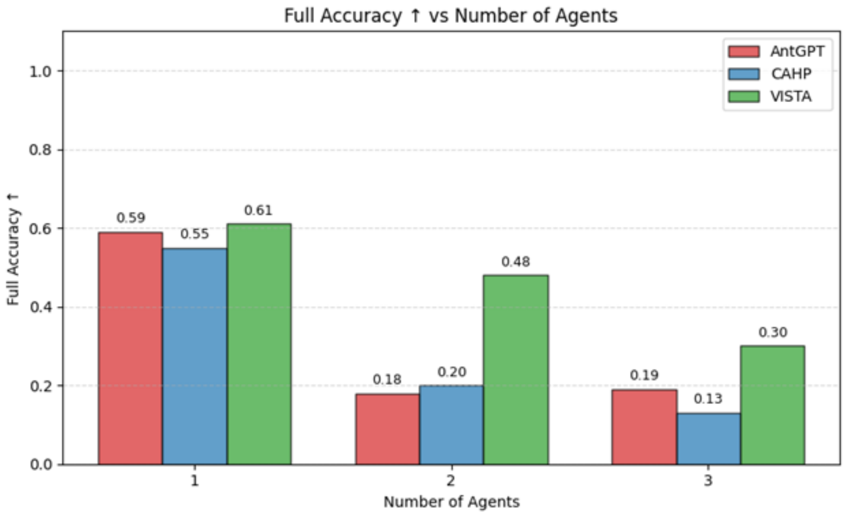
1.Zhao, Qi, et al. "Antgpt: Can large language models help long-term action anticipation from videos?." *arXiv preprint arXiv:2307.16368* (2023).
2.Liu, Yuchen, et al. "Context-Aware Human Behavior Prediction Using Multimodal Large Language Models: Challenges and Insights." *arXiv preprint arXiv:2504.00839* (2025).

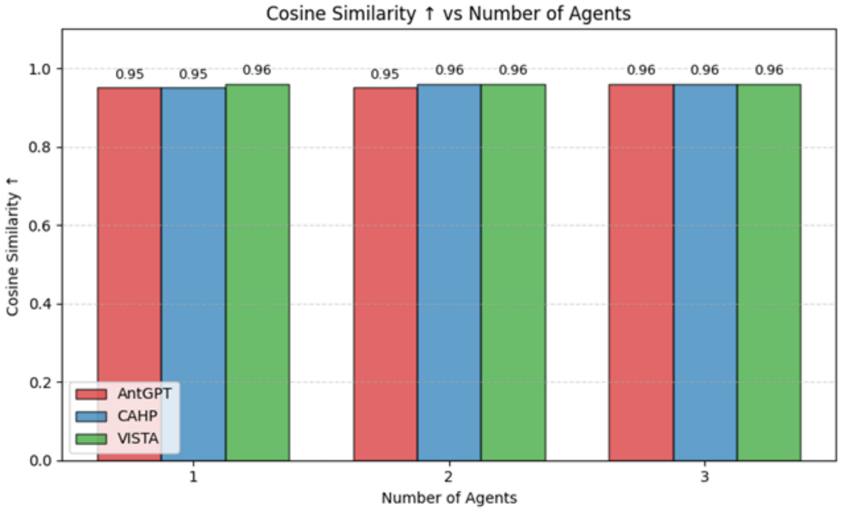
Results - Overall

- VISTA vs Baselines
- With increasing number of humans.

- Edit Distance: lower is better
- **13%** improvement in two humans
- **8%** improvement in three humans



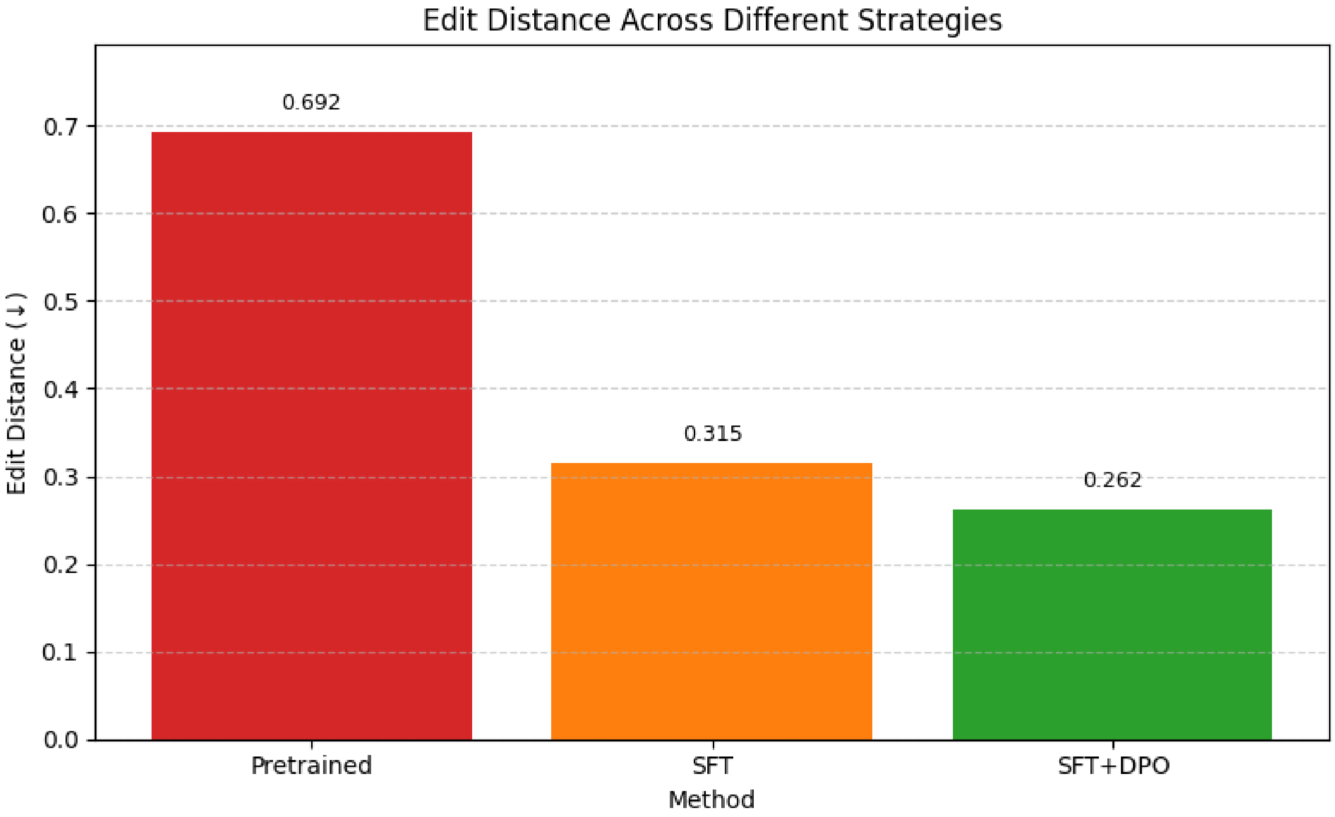




Results – Fine Tuning Strategies

1. Pretrained
2. SFT
3. SFT+DPO

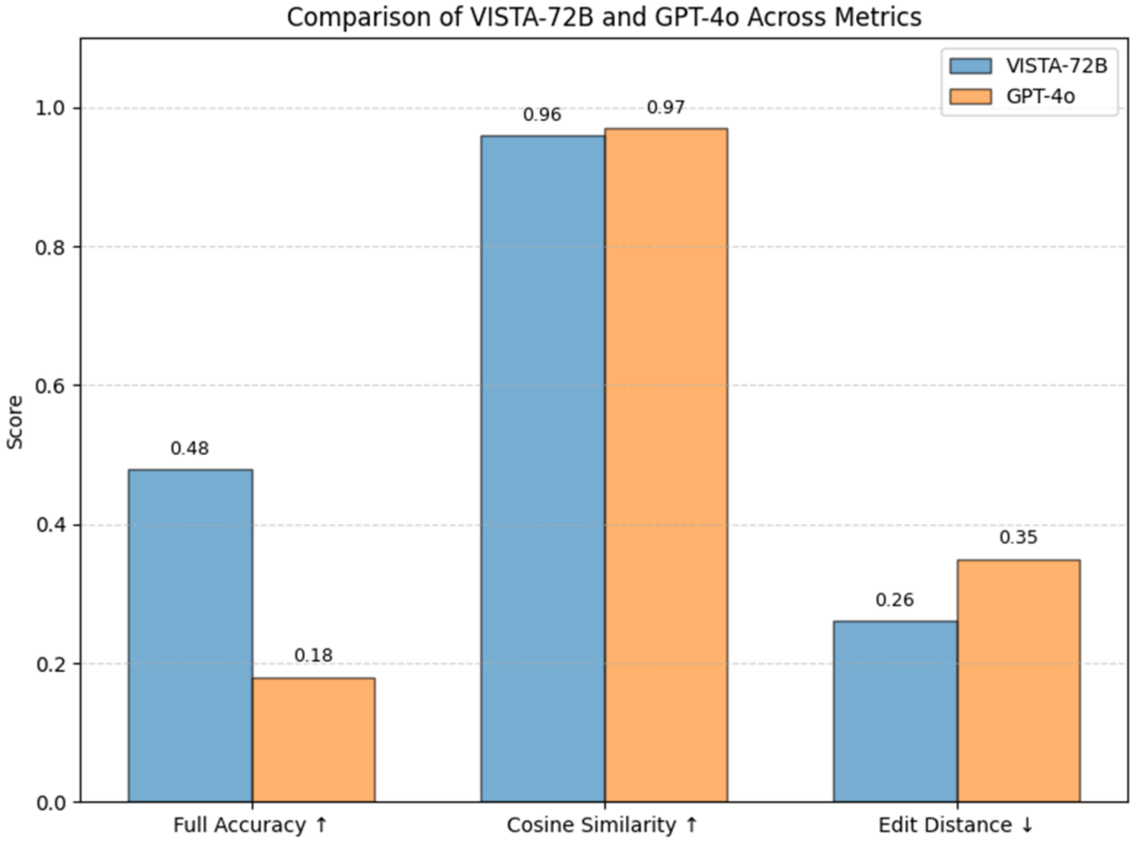
- Edit Distance: lower is better



Improvement → **54%** → **16%**

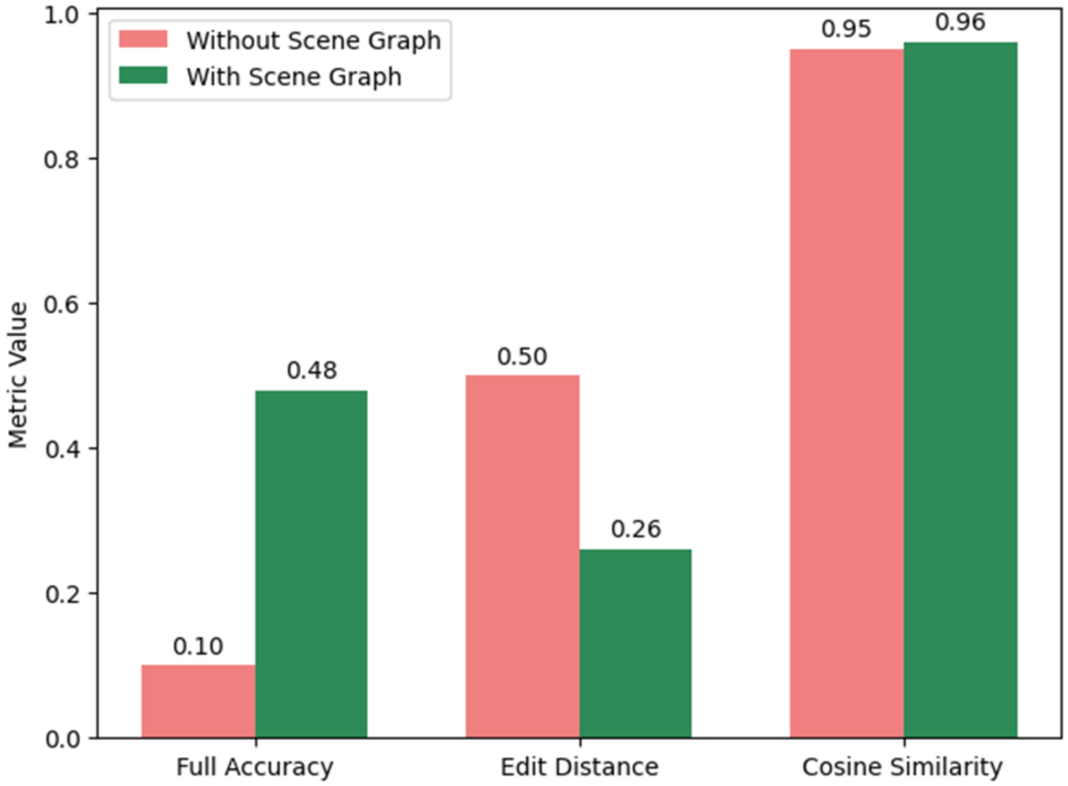
Results – GPT-4o vs VISTA

- Edit Distance: lower is better
- **13.8**% improvement in Edit Distance



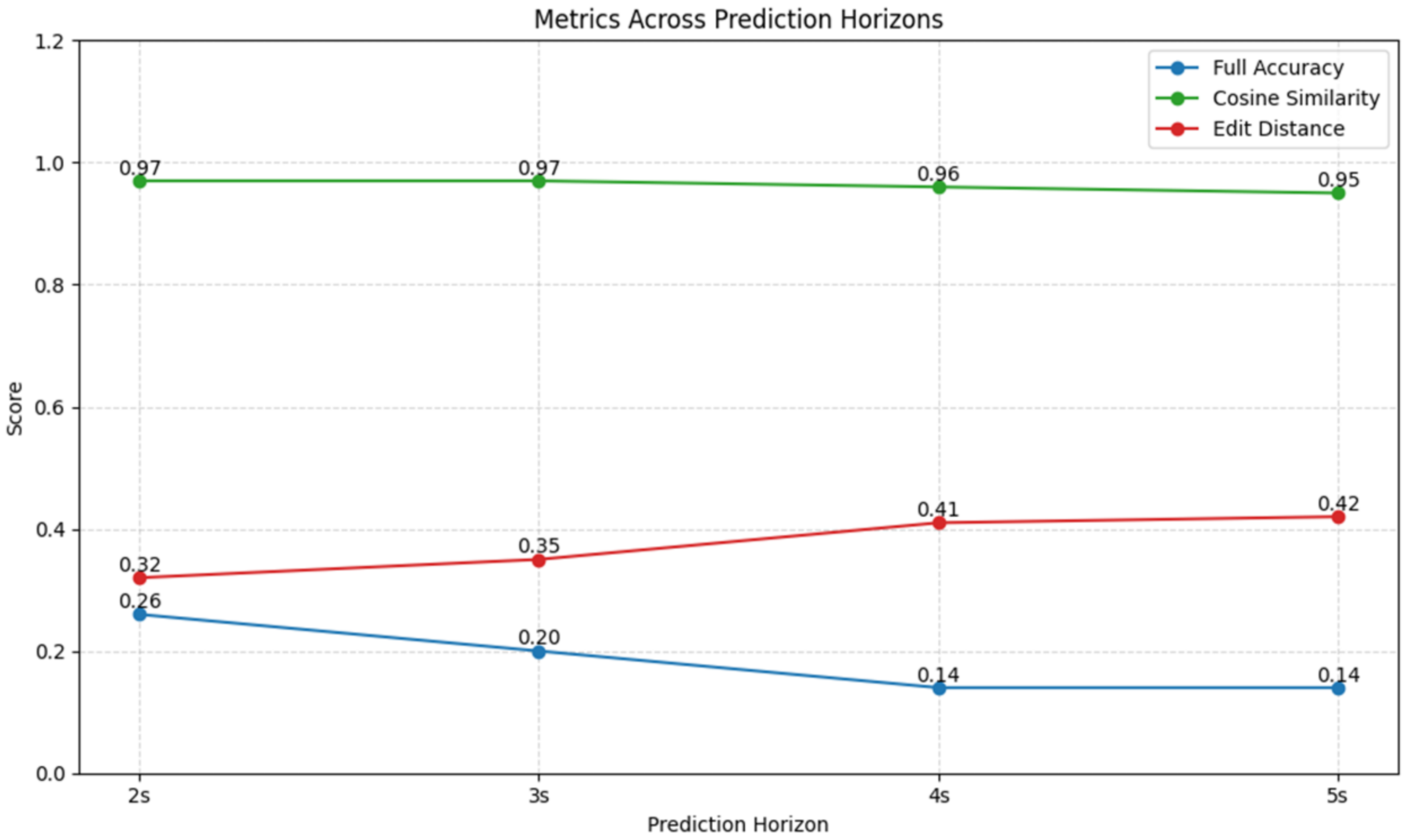
Results – with & w/o Scene graph

- Edit Distance: lower is better
- **52%** improvement in Edit Distance



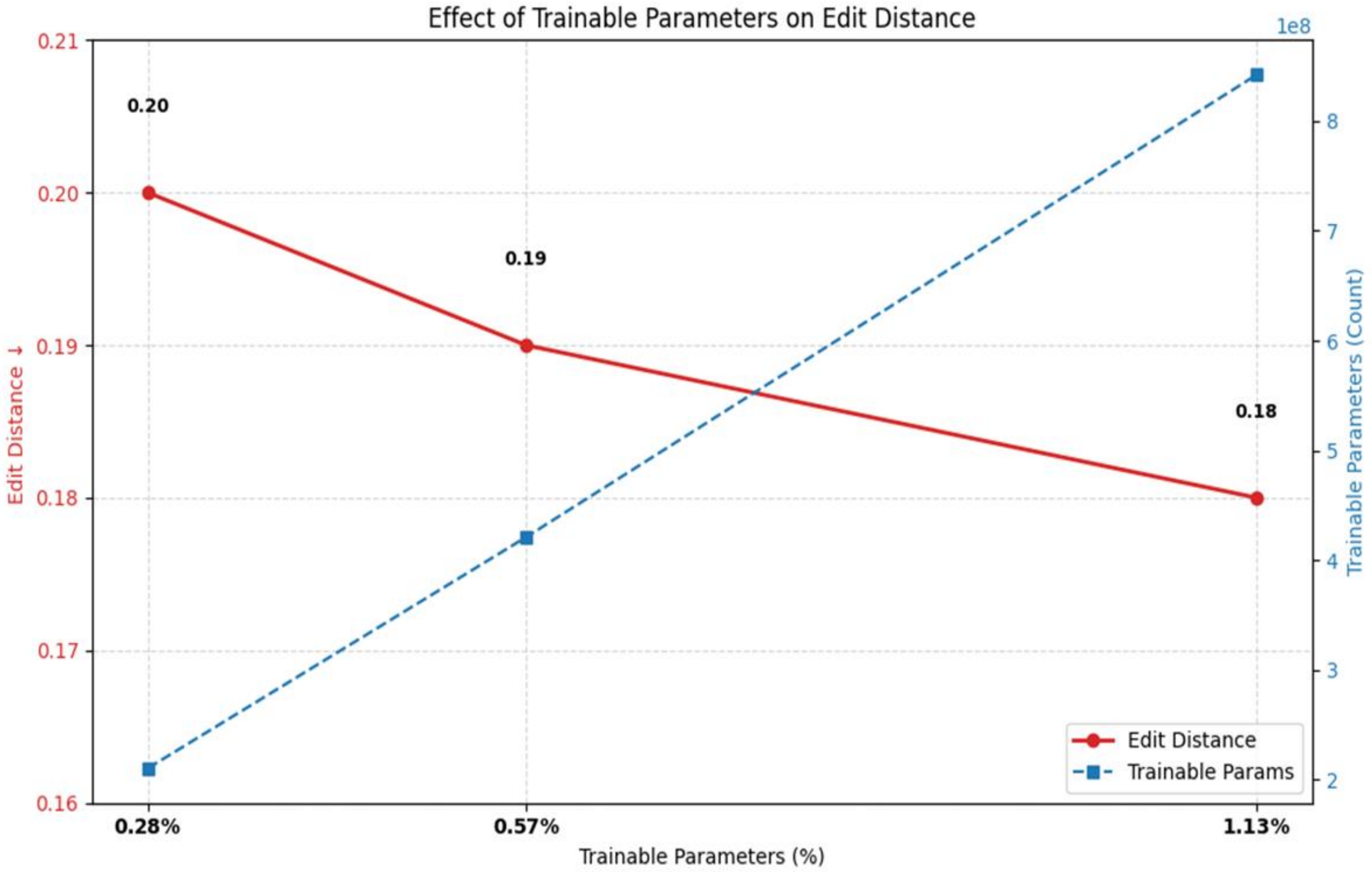
Results – Increasing Prediction Horizon

- Increasing prediction horizon results in lower metrics.



Results – Hyperparameters

- We only train 1.13% of total model parameters (72B).



Conclusion

Conclusion

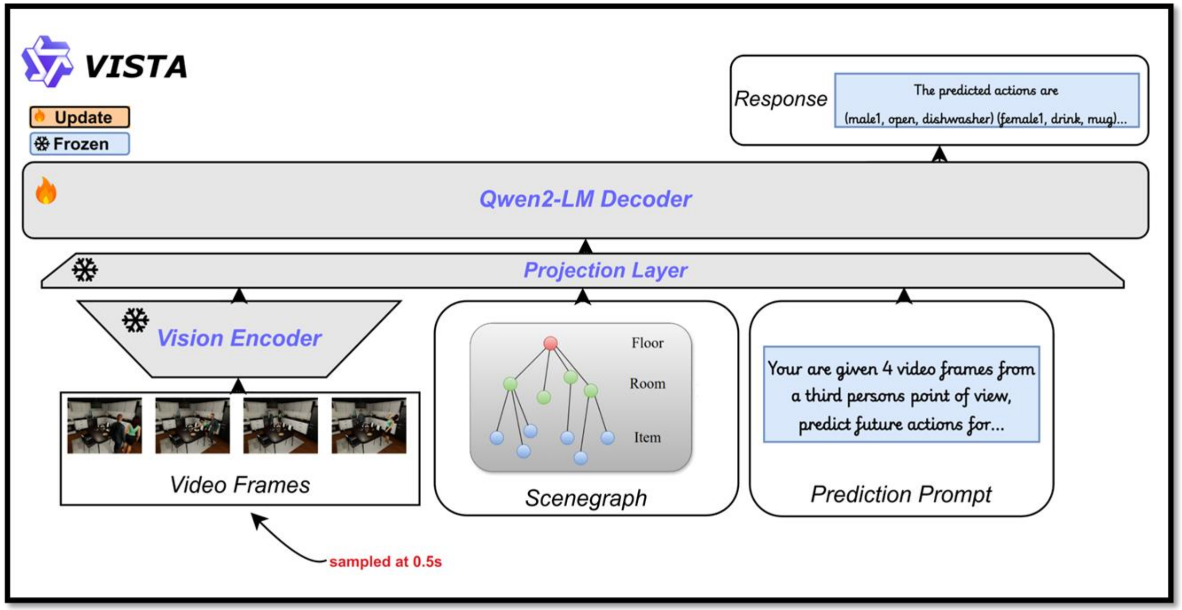
- Introduce VISTA: VLM based framework for multiple human behavior prediction.
- Address gaps in current research of Human Behavior Prediction.
- Outline the fine tuning process.
- Report 13% improvement over existing baselines.

Limitations

- Limited availability of Scene graph.
- Hardware constraints while Fine tuning & inferencing VLM.

Outlook

- Integrate additional modalities.
- Longer Horizons > 5s.



The predicted actions are (male1, open, dishwasher) (female1, drink, mug)...

Your are given 4 video frames from a third persons point of view, predict future actions for...

Thank You !