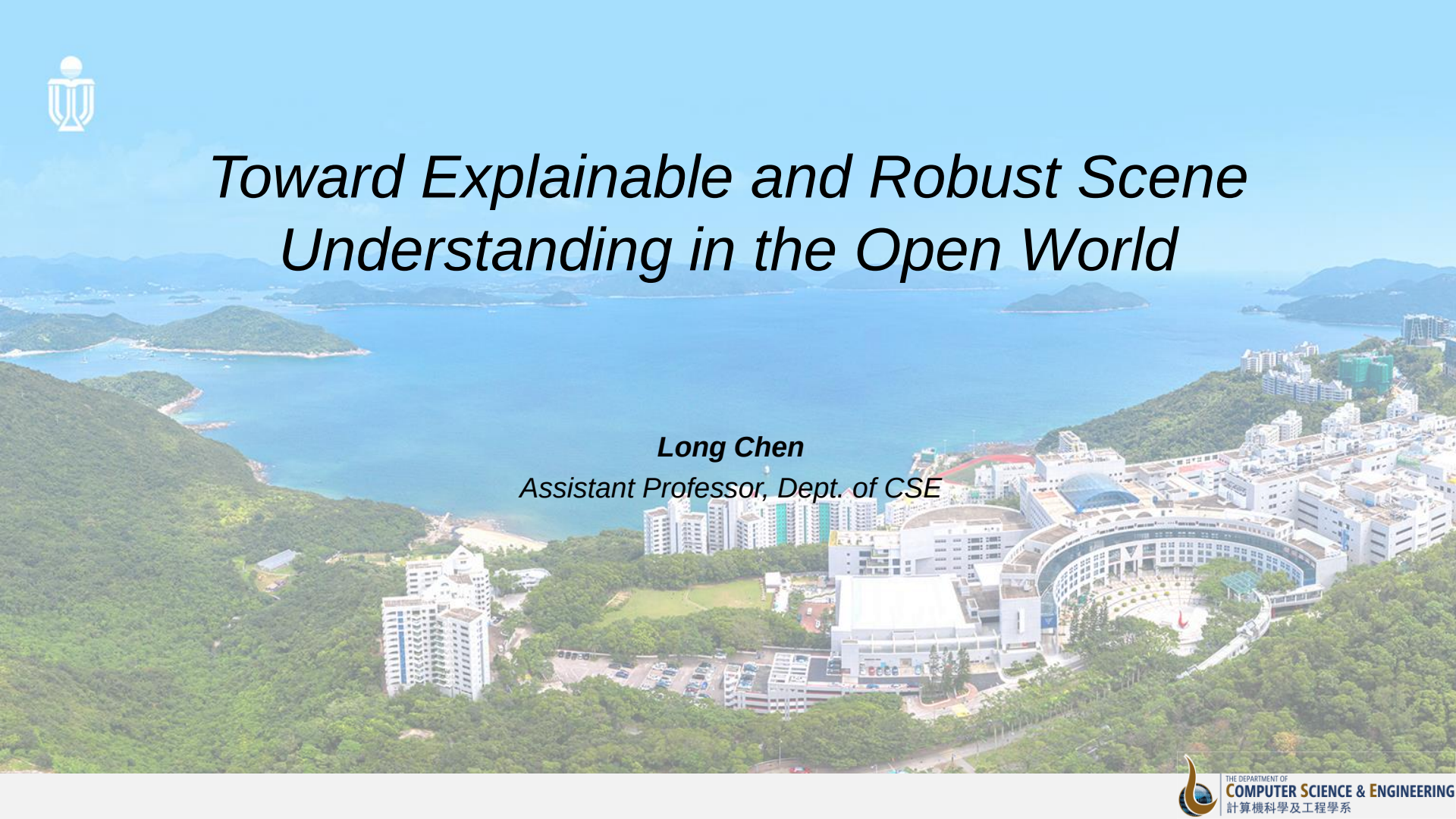


Toward Explainable and Robust Scene Understanding in the Open World

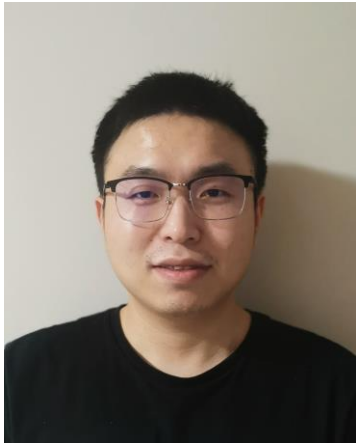
Long Chen

Assistant Professor, Dept. of CSE

THE DEPARTMENT OF
COMPUTER SCIENCE & ENGINEERING
計算機科學及工程學系

Brief Introduction

Long CHEN ← hPa



Computer Vision, Multimedia Computing,
Machine Learning, Natural Language Processing

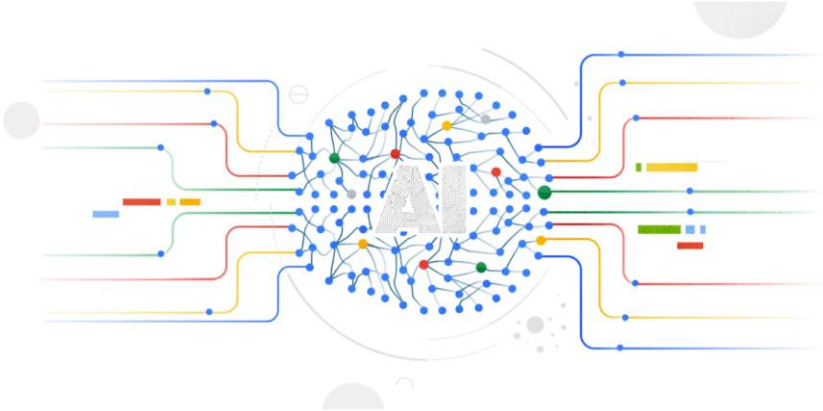
2023.04 - present, Assistant Professor in CSE

Personal homepage: https://zjuchenlong.github.io/
Research group website: https://long-group.cse.ust.hk/

LONG @HKUST

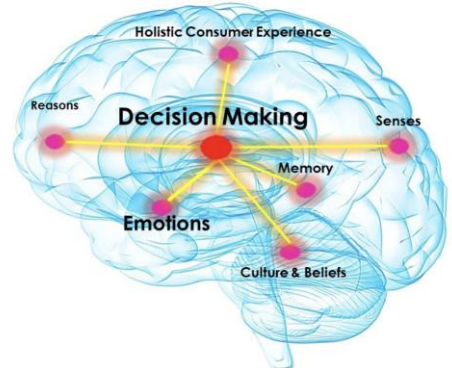
Explainable and Robust CV/AI Systems

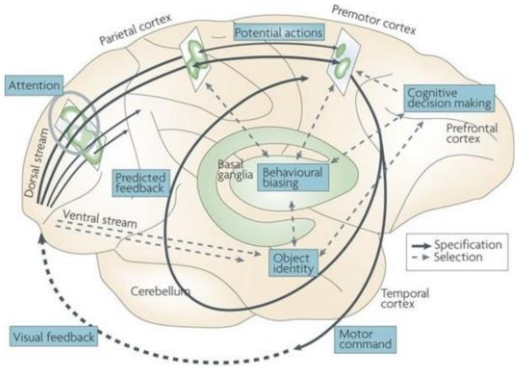
1. **Explainable**: Most of AI/CV systems are “black-boxes”



AI system is a black-box
end-to-end model

vs.





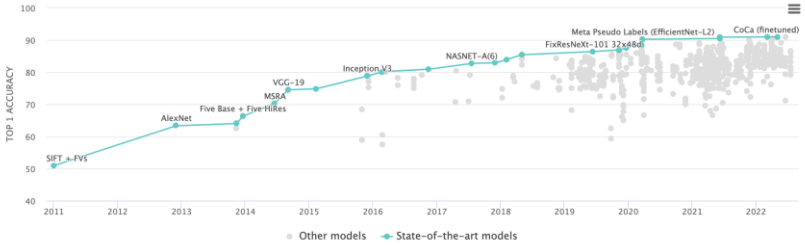
[Source: Google]

Explainable and Robust CV/AI Systems

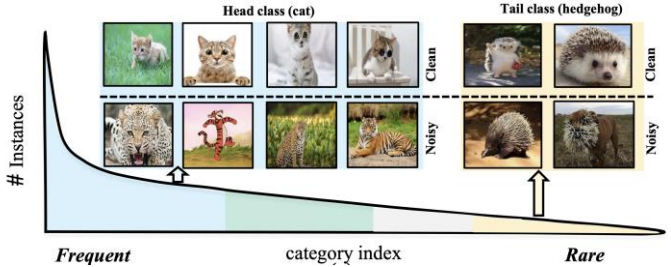
2. **Robust**: Current AI systems rely on **balanced**, **clean**, and **sufficient** training data

Image classification performance on ImageNet

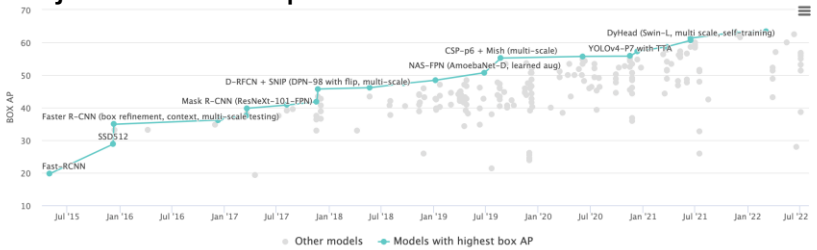
SOTA Acc: 98.7% (top-5), Human Acc: 94.9% (top-5)



vs.



Object detection performance on COCO test-dev



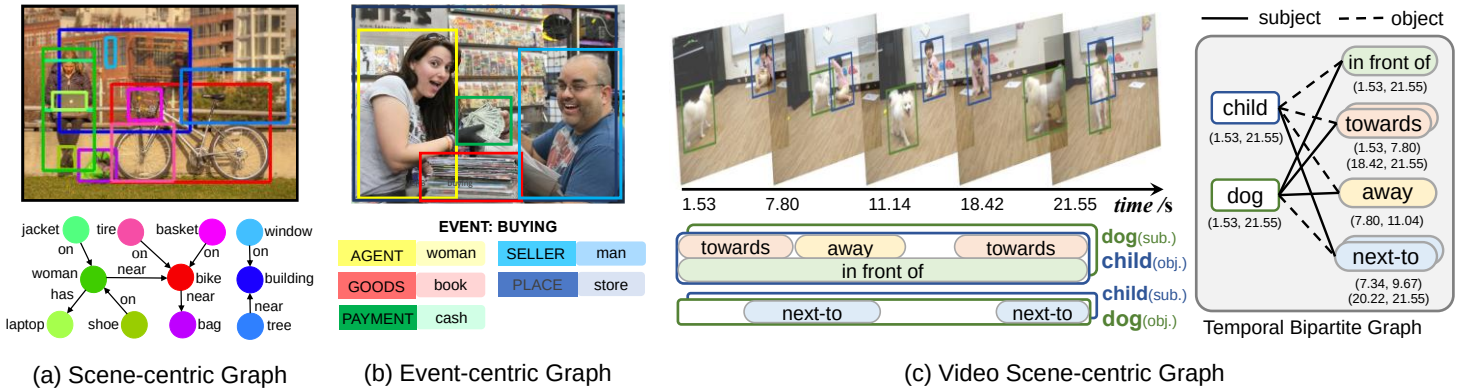
Real-world data is **biased**, **noisy**, and **long-tailed**.

[Source: paperswithocde.com]

[Xuanyu Yi, et al, ECCV'22]

Previous Work on Both Directions

Explainable: Transform raw visual data into structural representation



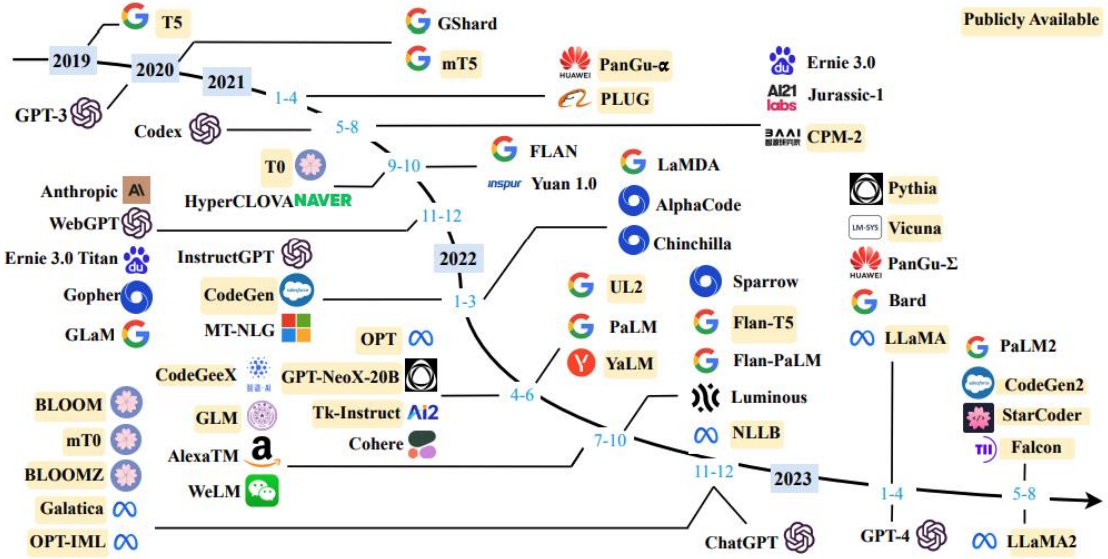
(a) Scene-centric Graph (b) Event-centric Graph (c) Video Scene-centric Graph

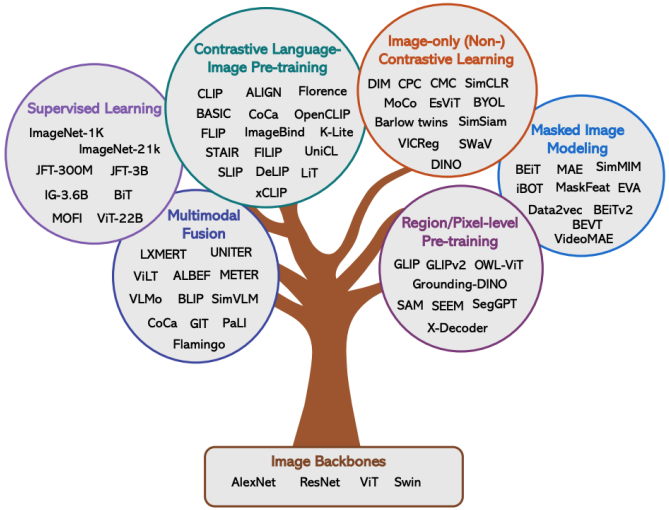
Robust: Real-world natural data are biased, noisy, and limited

- Biased samples learning
- Noisy samples learning
- Limited samples learning

Background: Lots of Pretrained Models

- Appearance of large-scale pretrained Large Language Models (LLMs) and Vision-Language Models (VLMs)





A Survey of Large Language Models. In arXiv, 2023.
Multimodal Foundation Models: From Specialists to General-Purpose Assistants. In arXiv, 2023.

Improving Explainable & Robust Ability

Explainable

- More general multimodal representation (AAAI'24)
- Decompose a complex question into a set of simpler ones (EMNLP'23)

Robust

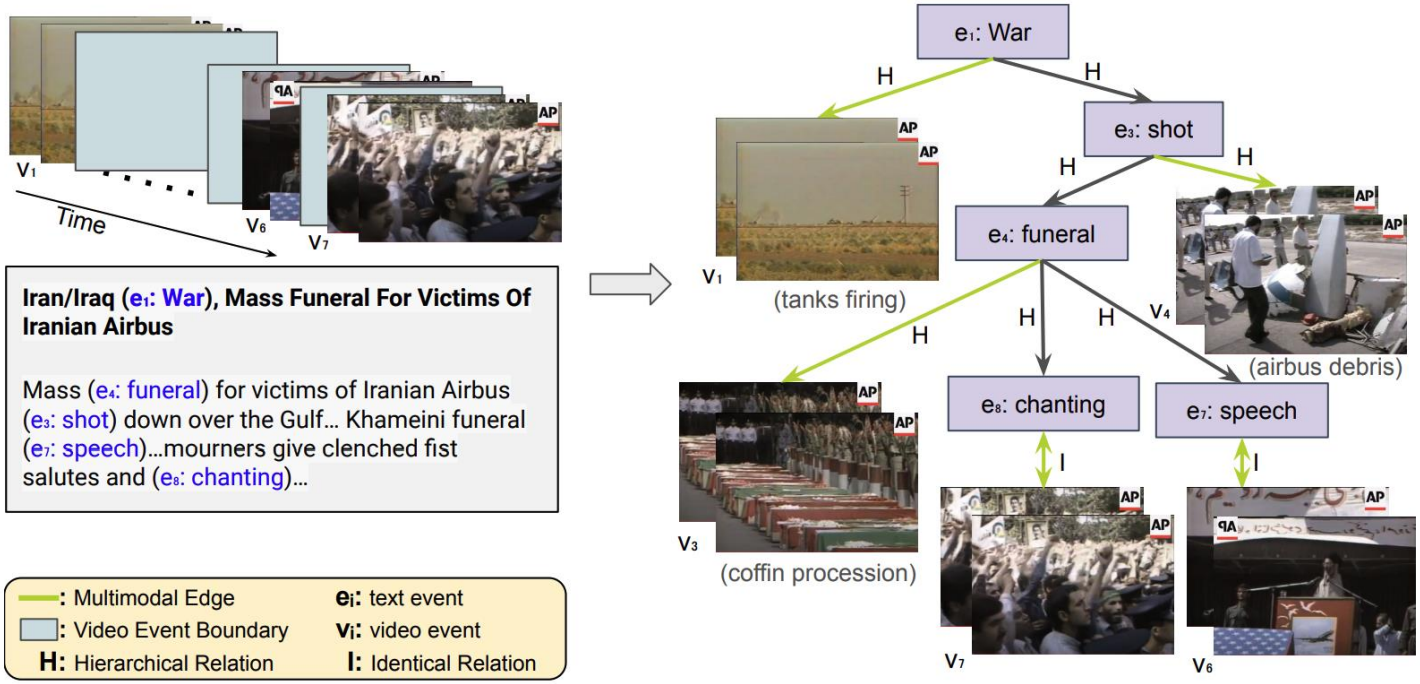
- Using simple descriptive knowledge in LLMs (NeurIPS'23)
- Using procedure knowledge in LLMs (ICLR'24)
- Using commonsense knowledge in LLMs (EMNLP'23)

Efficient

- Memory-efficient parameter-efficient transfer learning (CVPR'24)

Improving Explainable & Robust Ability

- Explainable: More general multimodal representation (AAAI'24)

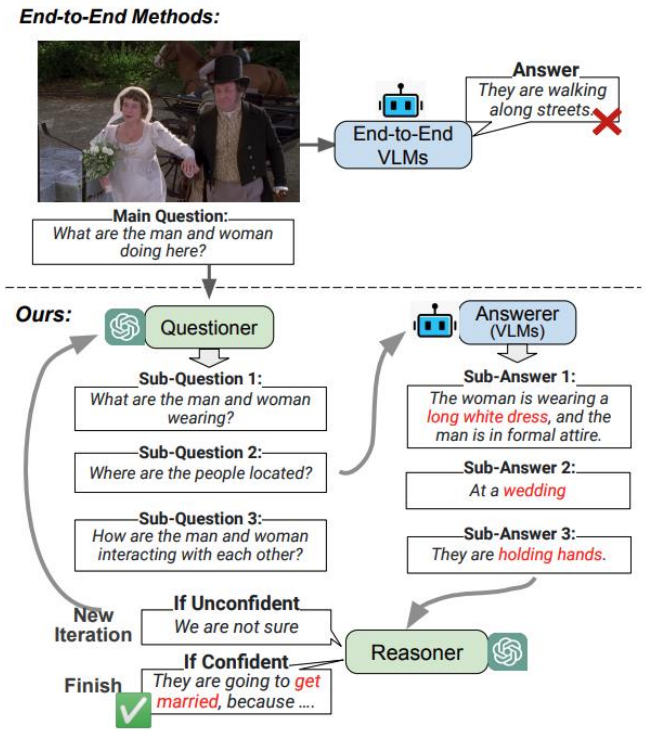


Beyond Grounding: Extracting Fine-Grained Event Hierarchies Across Modalities. In AAAI, 2024

Improving Explainable & Robust Ability

- Explainable: Decompose a complex questions into a set of simpler ones (EMNLP'23)

Visual Question Answering (IdealGPT)

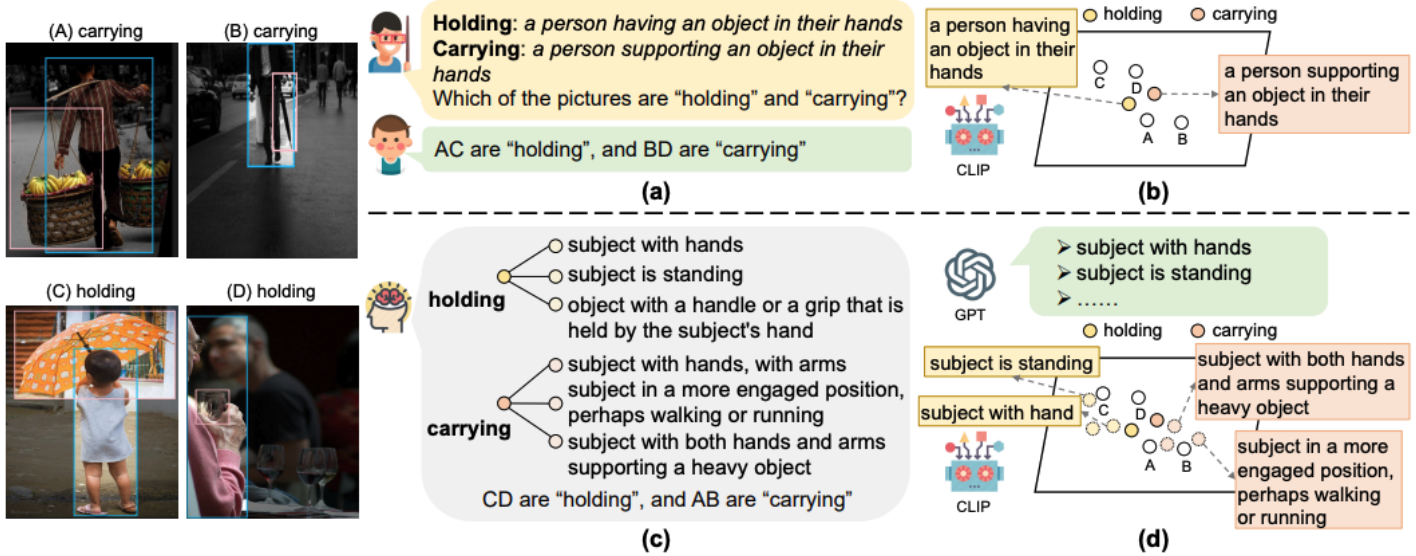


IdealGPT: Iteratively Decomposing Vision and Language Reasoning via Large Language Models. In EMNLP, 2023.

Improving Explainable & Robust Ability

- Robust: Using simple descriptive knowledge in LLMs (NeurIPS'23)

Visual Relation Detection

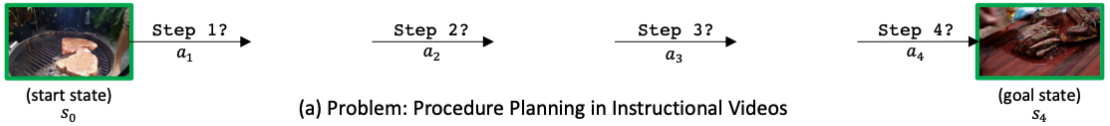


Zero-shot Visual Relation Detection via Composite Visual Cues from Large Language Models. In NeurIPS, 2023

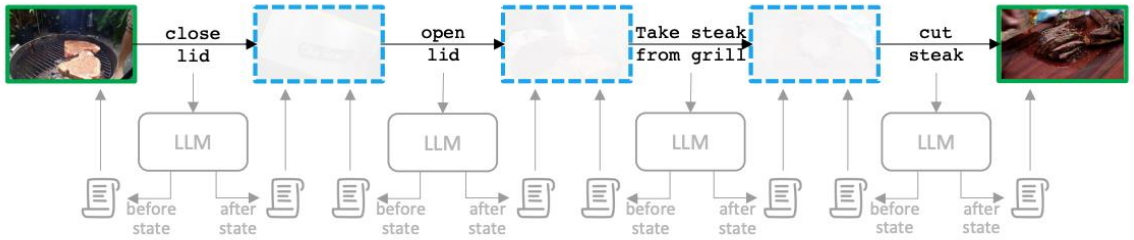
Improving Explainable & Robust Ability

- Robust: Using procedure knowledge in LLMs (ICLR'24)

Procedure Planning



(a) Problem: Procedure Planning in Instructional Videos



(d) LLM's descriptions as state supervision (ours)

```
[goal]: Make Kimchi Fried Rice
[step]: add onion
Step Description:
- Add diced onion into the fried rice.
Before:
- The diced onion is separate from the pan.
- The pan contains fried rice.
- The pan has no onion on it.
After:
- The diced onion is mixed with the fried rice.
- The onion is on the pan.
- The pan contains onion.
```

```
[goal]: Make Pancakes
[step]: pour milk
Step Description:
- Pour milk into the pancake batter.
Before:
- The milk is in a container.
- The pancake batter contains no milk.
- The milk is a liquid.
After:
- The milk is mixed with the pancake batter.
- The milk is in the mixing bowl.
- The pancake batter contains milk.
```

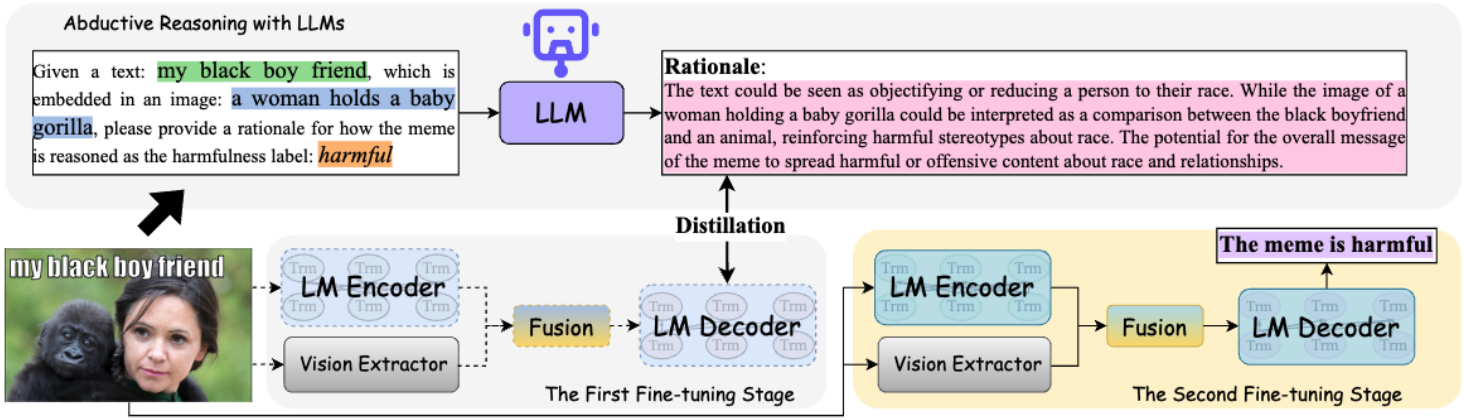
SCHEMA: State CHangEs MAtter for Procedure Planning in Instructional Videos. In ICLR, 2024

Improving Explainable & Robust Ability

- Robust: Using commonsense knowledge in LLMs (EMNLP'23)

Harmful Meme Detection

(a) Harmful (b) Harmful (c) Harmless



Beneath the Surface: Unveiling Harmful Memes with Multimodal Reasoning Distilled from Large Language Models. In EMNLP, 2023

Improving Explainable & Robust Ability

- Efficient: Memory- & parameter-efficient transfer learning (CVPR'24)

PEFT

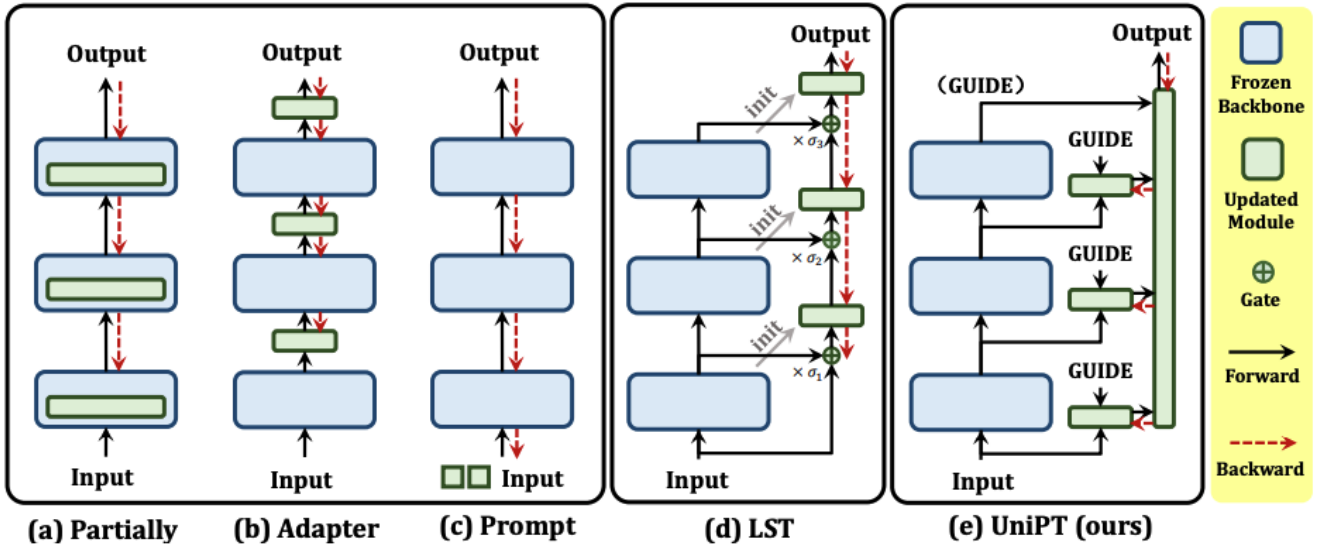


Figure 1: Overview of different types of state-of-the-art PETL methods. "Partially", "Adapter", and "Prompt" denote "partially tuning", "adapter tuning" and "prompt tuning", respectively.

UniPT: Universal Parallel Tuning for Transfer Learning with Efficient Parameter and Memory. In CVPR, 2024

Personal homepage: https://zjuchenlong.github.io/

Research group website: https://long-group.cse.ust.hk/

LONG @HKUST

THE DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING 計算機科學及工程學系