ICCV
OCT 19-23, 2025
HONOLULU
HAWAII

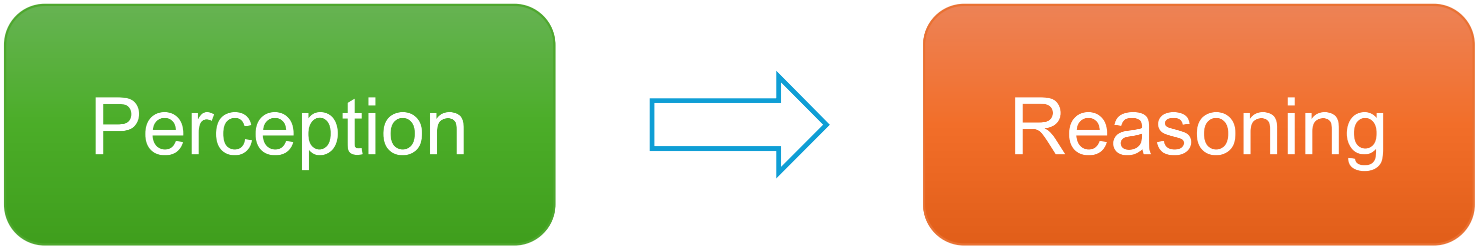
ICCV 2025 Tutorial on

Towards Comprehensive Reasoning in Vision-Language Models

Oct 19 8.30am- 12pm

University at Buffalo

From Perception to Reasoning for VLM



Perception

Reasoning

- Pattern recognition ≠ Reasoning
- Reasoning means compositional structure, causal understanding, and multi-step inference.
- Compared with language-only models, VLMs still lag when visual evidence must be integrated step-by-step.

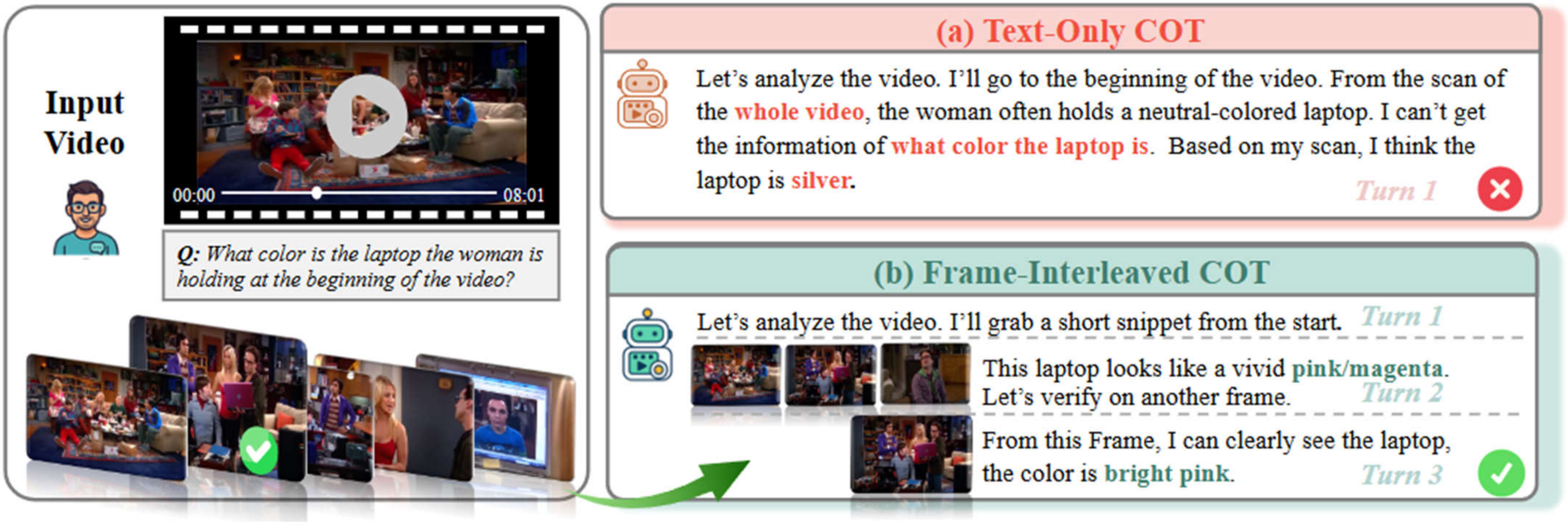
Bring Comprehensive Reasoning Capabilities into VLM

Pattern → Process Reasoning: Modeling reasoning as a process rather than a shortcut.

Single shot → Structured and Reliable Reasoning: Pursuing consistency and robustness for trustworthy inference.

Static → Interactive and Agentic Reasoning: Grounding reasoning in real-world perception and action.

Our Recent Explorations on Process Reasoning



Q: What color is the laptop the woman is holding at the beginning of the video?

(a) Text-Only COT

Let's analyze the video. I'll go to the beginning of the video. From the scan of the whole video, the woman often holds a neutral-colored laptop. I can't get the information of what color the laptop is. Based on my scan, I think the laptop is silver.

(b) Frame-Interleaved COT

Let's analyze the video. I'll grab a short snippet from the start.

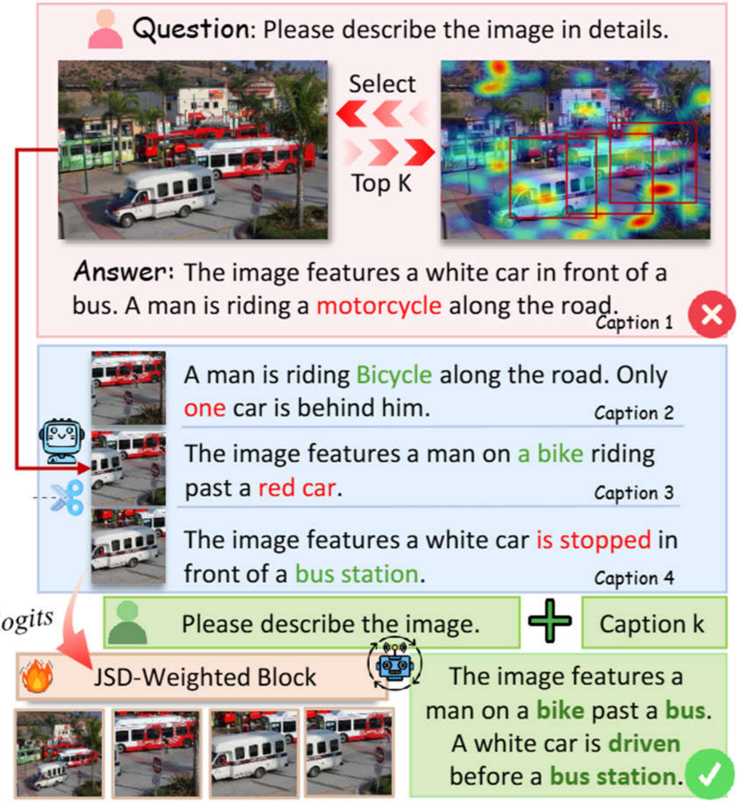
This laptop looks like a vivid pink/magenta. Let's verify on another frame.

From this Frame, I can clearly see the laptop, the color is bright pink.

In **FrameMind**, we studied how interleaved visual-textual chains to enforce multi-step compositional reasoning via Reinforcement Learning for VLM video understanding.

Ge, H., Wang, Y., Chang, K. W., Wu, H., & Cai, Y. (2025). FameMind: Frame-Interleaved Video Reasoning via Reinforcement Learning. arXiv e-prints, arXiv-2509.

Our Recent Explorations on Structured and Reliable Reasoning



Question: Please describe the image in details.

Answer: The image features a white car in front of a bus. A man is riding a motorcycle along the road.

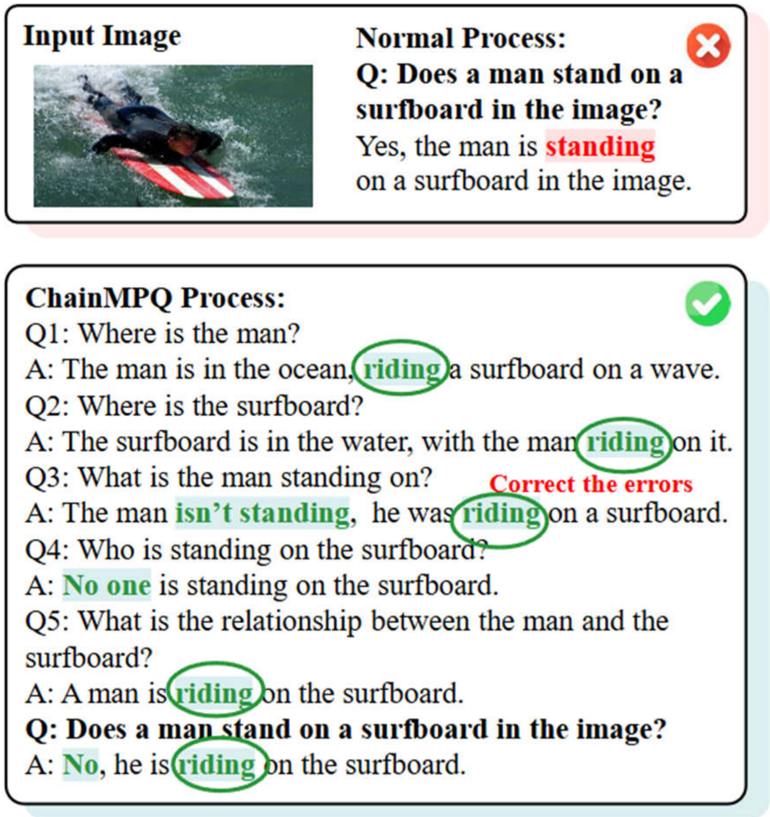
A man is riding Bicycle along the road. Only one car is behind him.

The image features a man on a bike riding past a red car.

The image features a white car is stopped in front of a bus station.

The image features a man on a bike past a bus. A white car is driven before a bus station.

MRFD for VLM hallucinations [EMNLP 2025]



Normal Process:

Q: Does a man stand on a surfboard in the image?

Yes, the man is standing on a surfboard in the image.

ChainMPQ Process:

Q1: Where is the man?

A: The man is in the ocean, riding a surfboard on a wave.

Q2: Where is the surfboard?

A: The surfboard is in the water, with the man riding on it.

Q3: What is the man standing on?

A: The man isn't standing, he was riding on a surfboard.

Q4: Who is standing on the surfboard?

A: No one is standing on the surfboard.

Q5: What is the relationship between the man and the surfboard?

A: A man is riding on the surfboard.

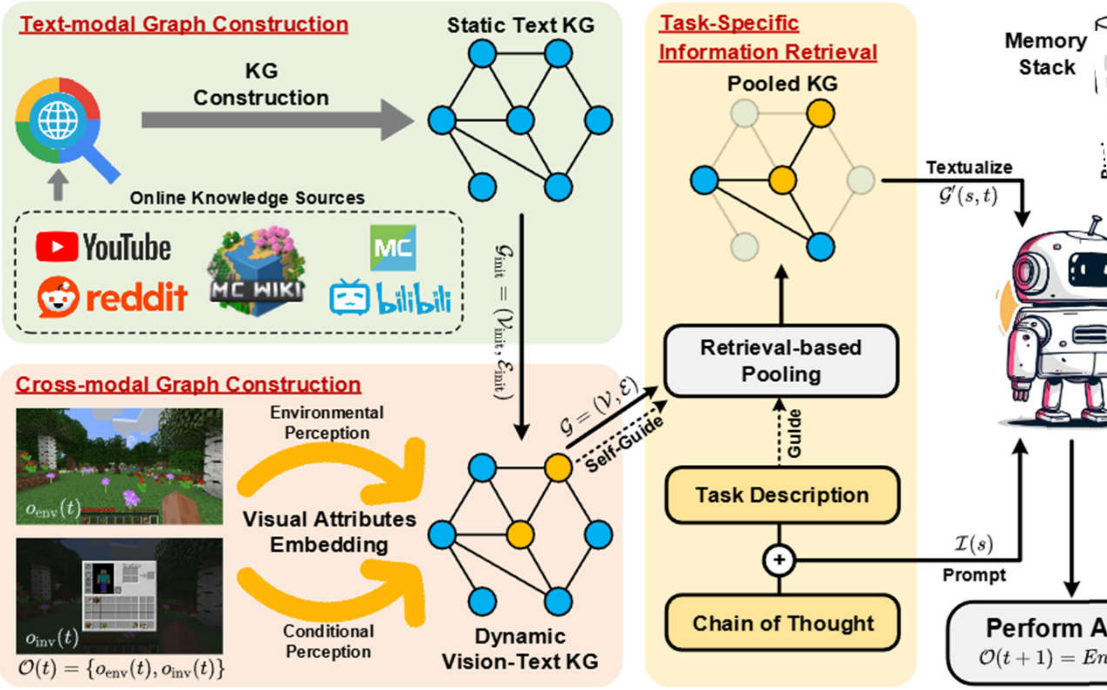
Q: Does a man stand on a surfboard in the image?

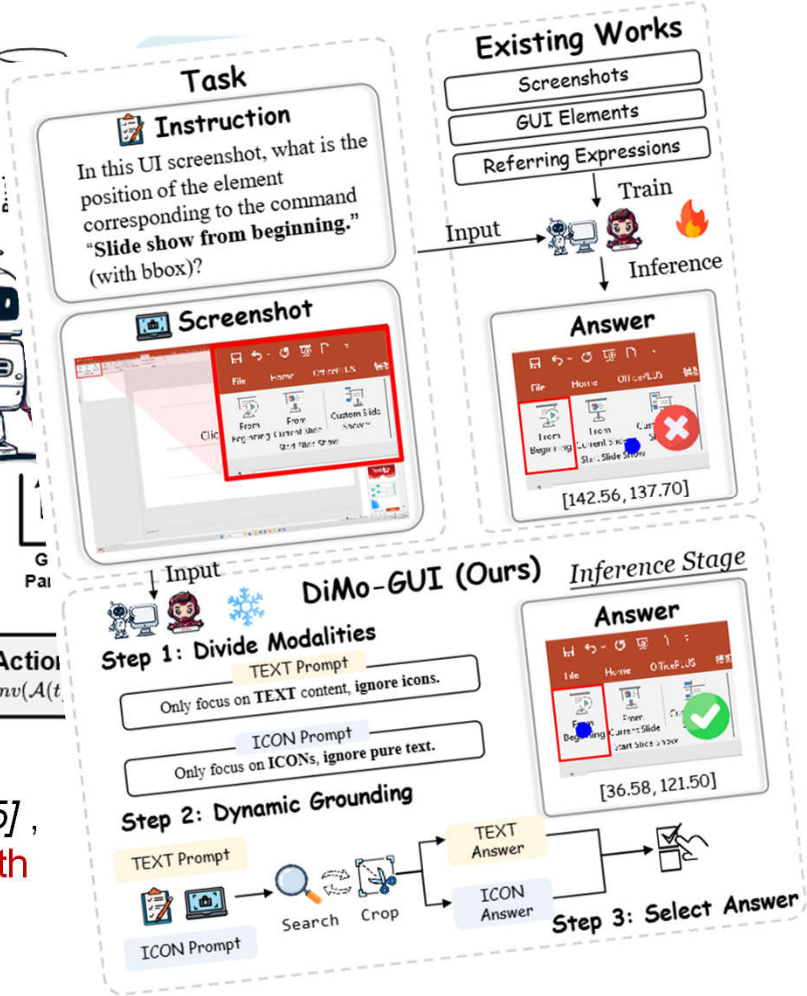
A: No, he is riding on the surfboard.

CHAINMPQ for VLM Relation Understanding [arxiv 2025]

We proposes works like ***MRFD*** and ***CHAINMPQ*** to enhance reliability and self-consistency to make reasoning more trustworthy.

Our Recent Explorations on Interactive and Agentic Reasoning





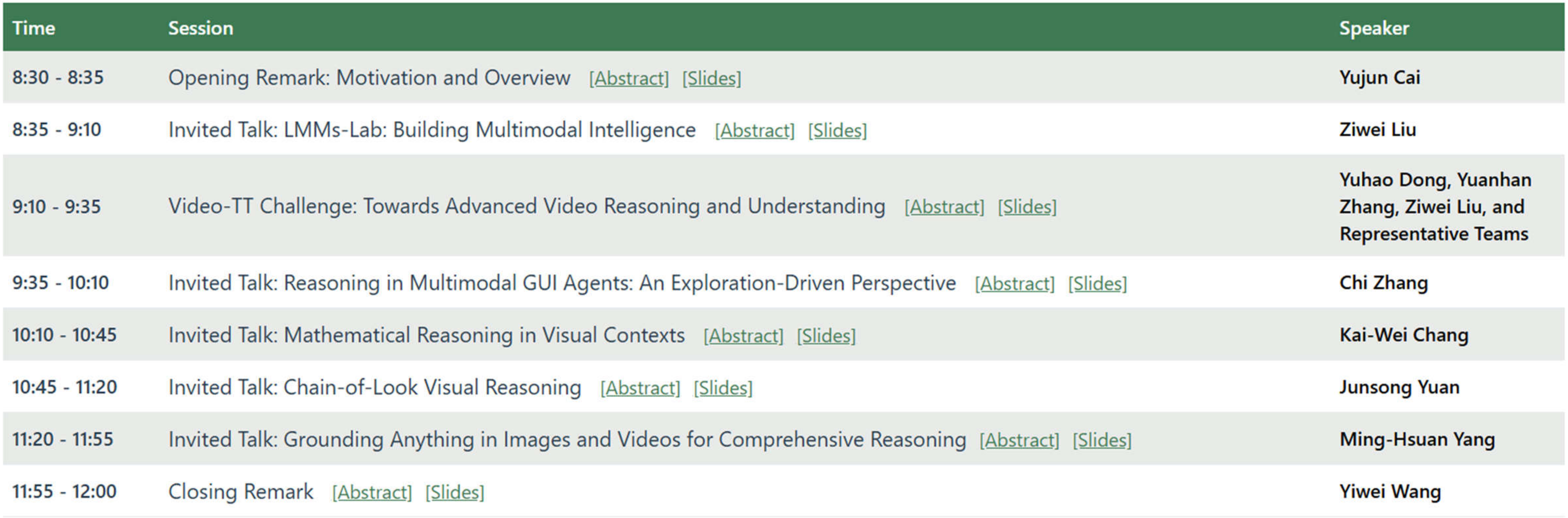
In this UI screenshot, what is the position of the element corresponding to the command **"Slide show from beginning."** (with bbox)?

Only focus on **TEXT** content, **ignore icons.**

Only focus on **ICONs, ignore pure text.**

Through ***Dimo-GUI*** *[EMNLP2025]* and ***Vistawise*** *[EMNLP2025]* , we explore how reasoning emerges when a model interacts with its environment.

Tutorial Schedule

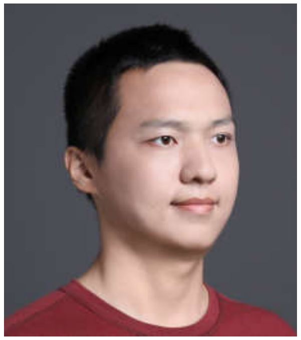


Time	Session	Speaker
8:30 - 8:35	Opening Remark: Motivation and Overview [Abstract] [Slides]	**Yujun Cai**
8:35 - 9:10	Invited Talk: LMMs-Lab: Building Multimodal Intelligence [Abstract] [Slides]	**Ziwei Liu**
9:10 - 9:35	Video-TT Challenge: Towards Advanced Video Reasoning and Understanding [Abstract] [Slides]	**Yuhao Dong, Yuanhan Zhang, Ziwei Liu, and Representative Teams**
9:35 - 10:10	Invited Talk: Reasoning in Multimodal GUI Agents: An Exploration-Driven Perspective [Abstract] [Slides]	**Chi Zhang**
10:10 - 10:45	Invited Talk: Mathematical Reasoning in Visual Contexts [Abstract] [Slides]	**Kai-Wei Chang**
10:45 - 11:20	Invited Talk: Chain-of-Look Visual Reasoning [Abstract] [Slides]	**Junsong Yuan**
11:20 - 11:55	Invited Talk: Grounding Anything in Images and Videos for Comprehensive Reasoning [Abstract] [Slides]	**Ming-Hsuan Yang**
11:55 - 12:00	Closing Remark [Abstract] [Slides]	**Yiwei Wang**

Organizers & Speakers

Yujun Cai
@UQ

Jun Liu
@Lancaster



Yiwei Wang
@UCM

Ming-Hsuan Yang
@UCM

Kai-Wei Chang
@UCLA

Junsong Yuan
@UB

Ziwei Liu
@NTU

Chi Zhang
@Westlake

Yuanhan Zhang
@NTU

Yuhao Dong
@NTU