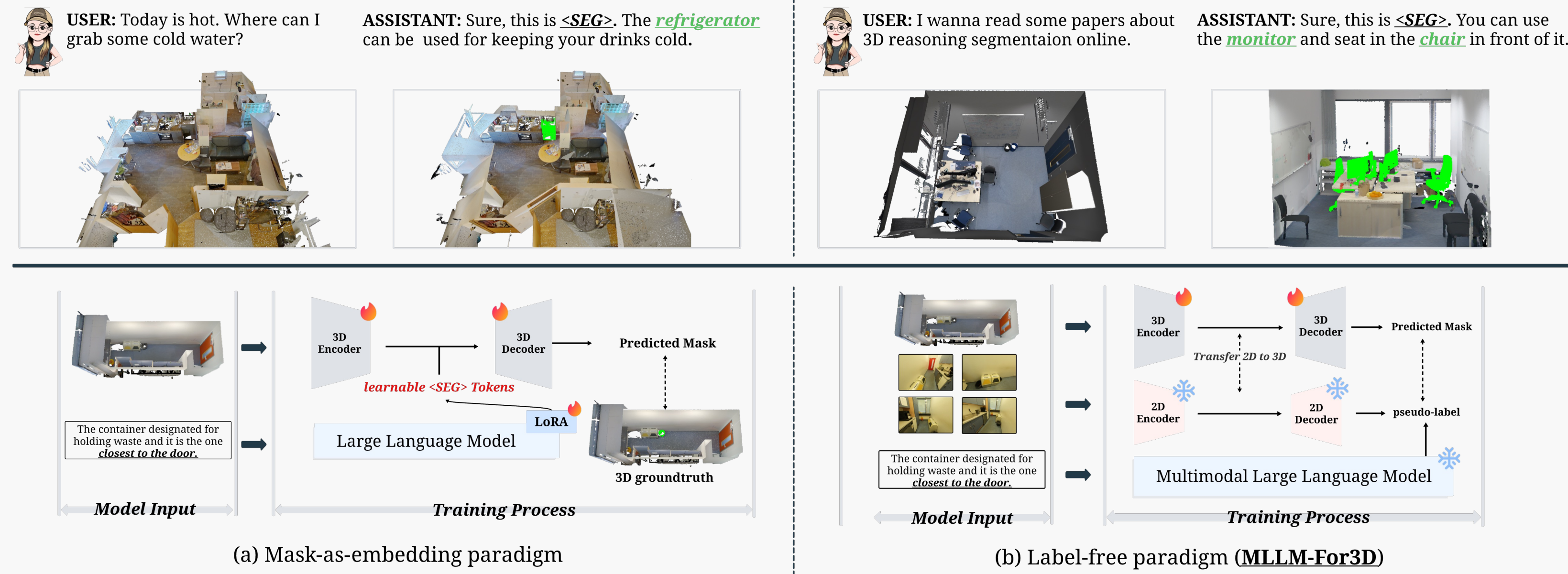


MLLM-For3D: Adapting Multimodal Large Language Model for 3D Reasoning Segmentation

Jiaxin Huang¹, Runnan Chen^{†2}, Ziwen Li¹, Zhengqing Gao¹, Xiao He⁴, Yandong Guo⁴, Mingming Gong^{1,3}, Tongliang Liu^{†1,2}
¹ Mohamed Bin Zayed University of Artificial Intelligence, ² The University of Sydney, ³ The University of Melbourne, ⁴ AI2Robotics

Background & Motivation

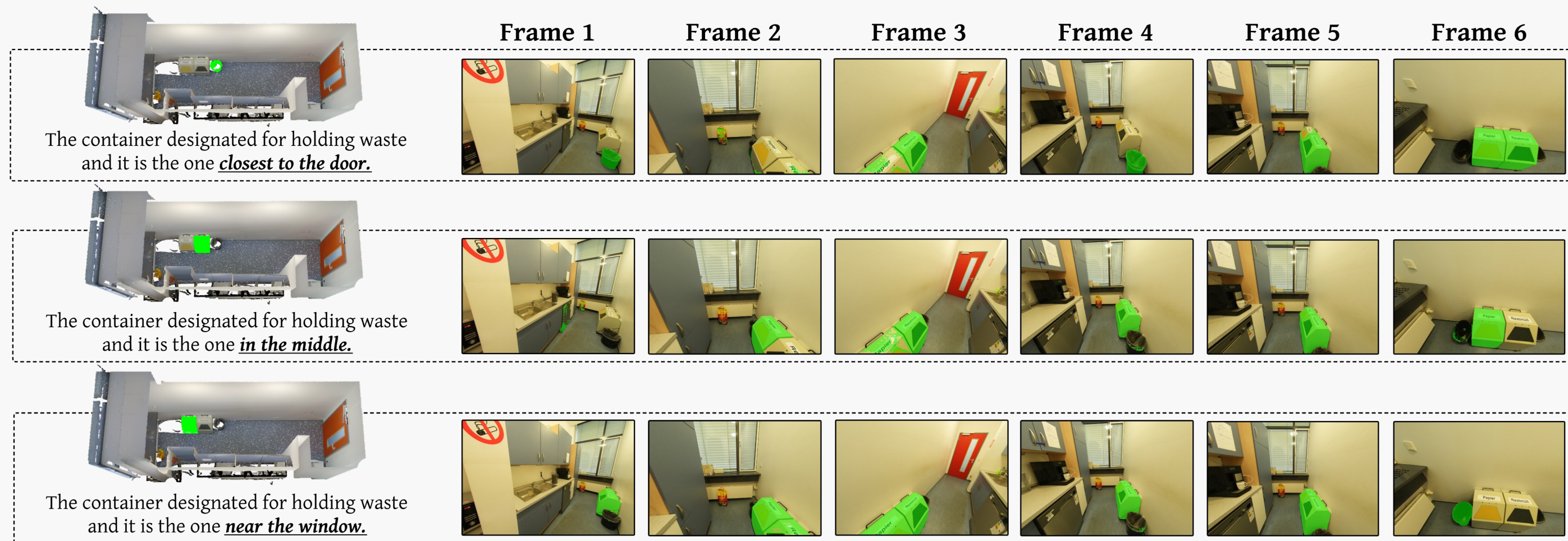


Motivation: Leverage abundant semantic knowledge and reasoning-based inference ability in 2D LLMs (LISA) for **Label-free 3D reasoning segmentation**

Two Problems to be solved: **3D Data Dilemma; High Costs of Generating**

The gap: Lack of 3D LLMs adept at 3D reasoning segmentation. Existing 3D LLMs allow detailed conversations and enhanced capability of referring and grounding but there is no model for 3D reasoning segmentation.

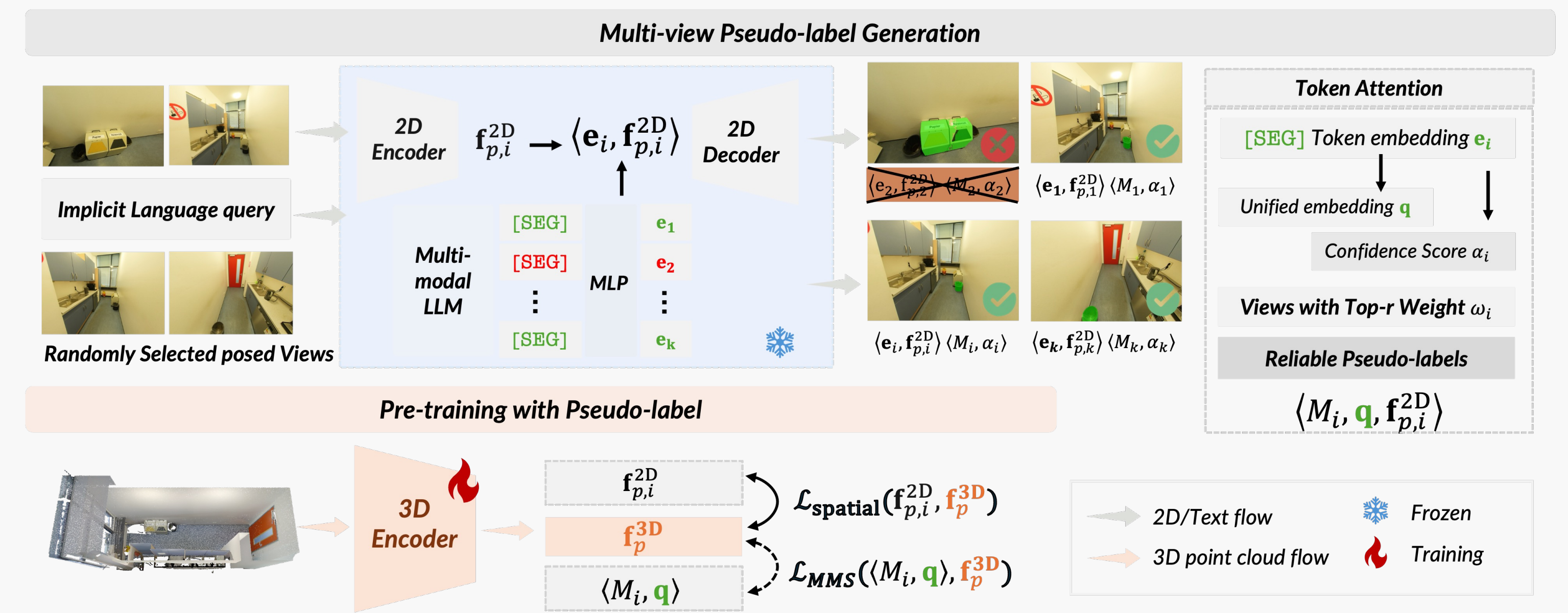
Pilot Study



Two key limitations to the reasoning of 2D models in 3D:

- (i) **False positives in invisible views:** A 2D MLLM processing a single view might hallucinate or segment objects that are described by the instruction but not visible in that particular view.
- (ii) **Multi-View prediction inconsistency:** occurs when the model lacks a mechanism to ensure spatial alignment of predictions across views

Overall Framework



Two Key Components in MLLM-For3D:

- 1. token-for-Query:** multimodal semantic alignment, enabling consistent identification of the same object across different views.
- 2. Spatial consistency:** to enforce that segmentation masks remain coherent in the 3D space, effectively capturing the geometry of the scene.

Experimental Results



Personal
Website



Open to
Discuss!



Github



ArXiv Paper