[Code] [Paper] [Lab Homepage]

ECCV 2026
Malmö

Reasoning Path and Latent State Analysis for Multi-view Visual Spatial Reasoning: A Cognitive Science Perspective

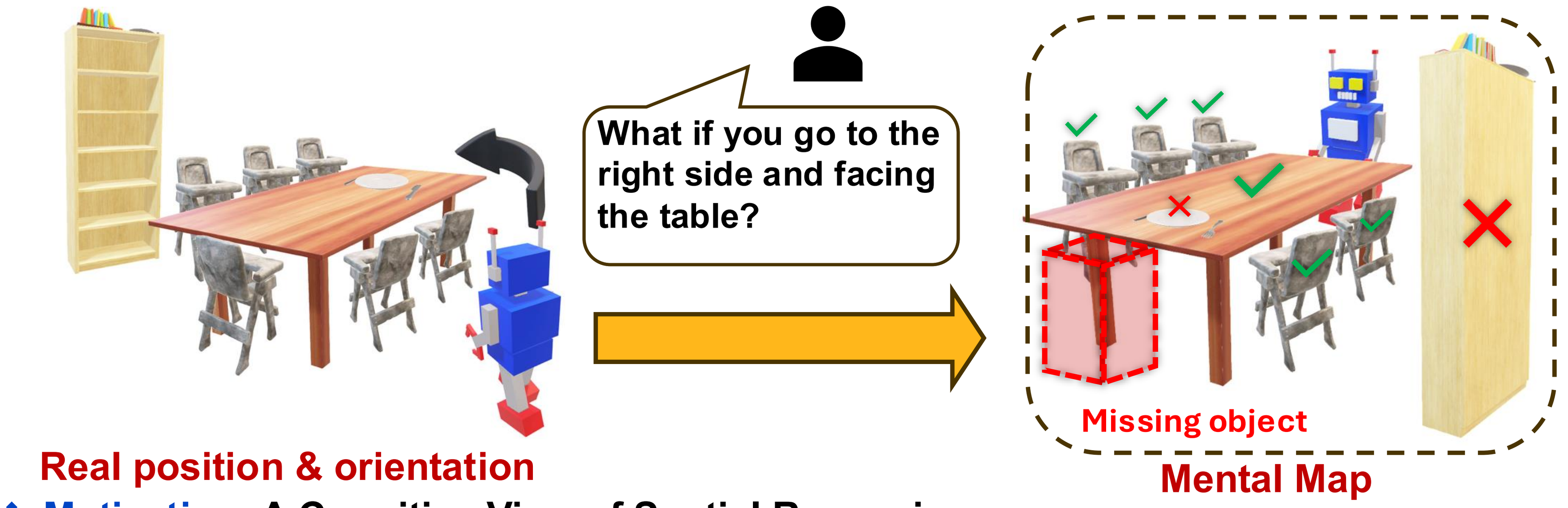
Qiyao Xue*, Haoming Wang*, Weichen Liu*, Shiqi Wang*, Haoming Wang*, Yuyang Wu† , Wei Gao*

University of Pittsburgh*, Carnegie Mellon University†

Background and Motivation

❖ **Background: Multi-view Spatial Reasoning**

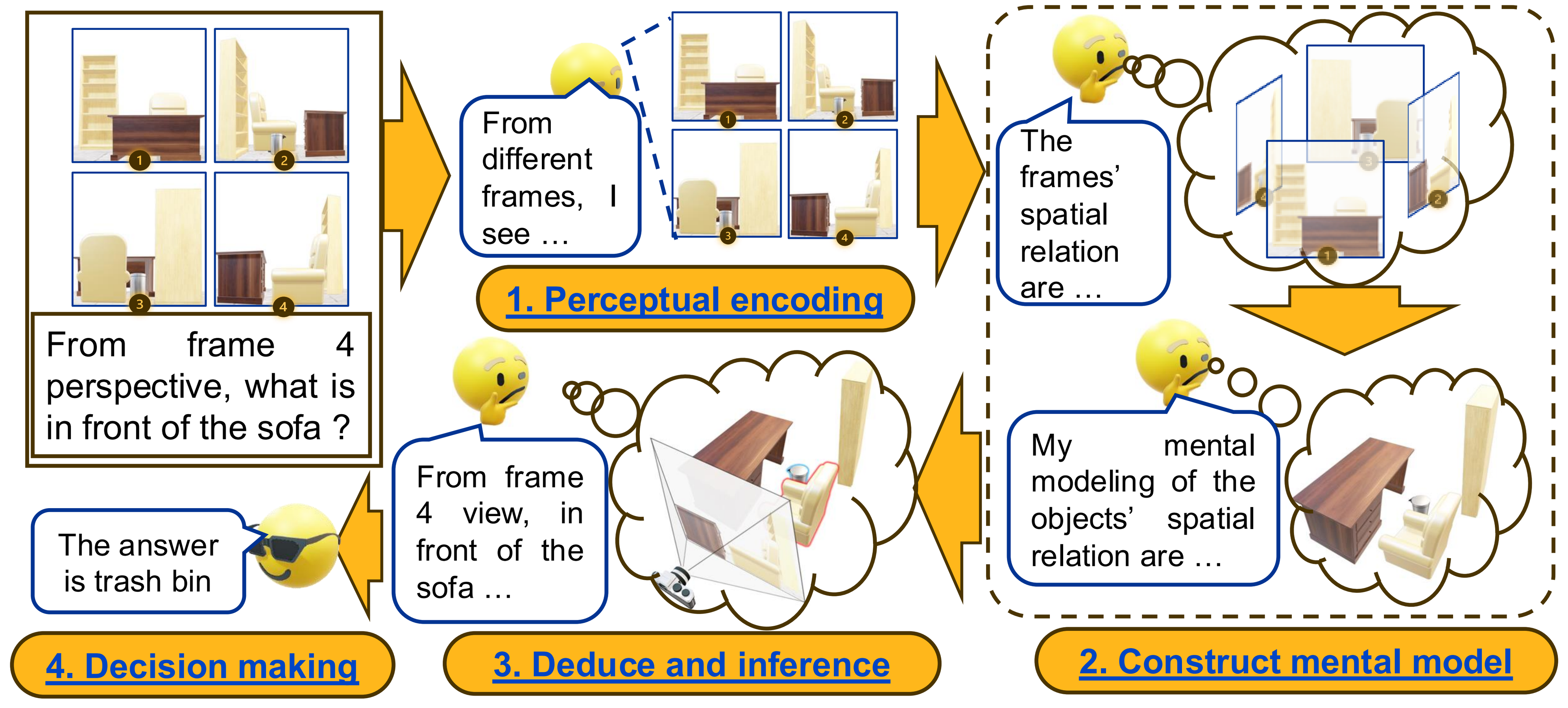
•Multi-view spatial reasoning requires a model to integrate observations from different viewpoints while maintaining a coherent representation of object positions, orientations, and relations.



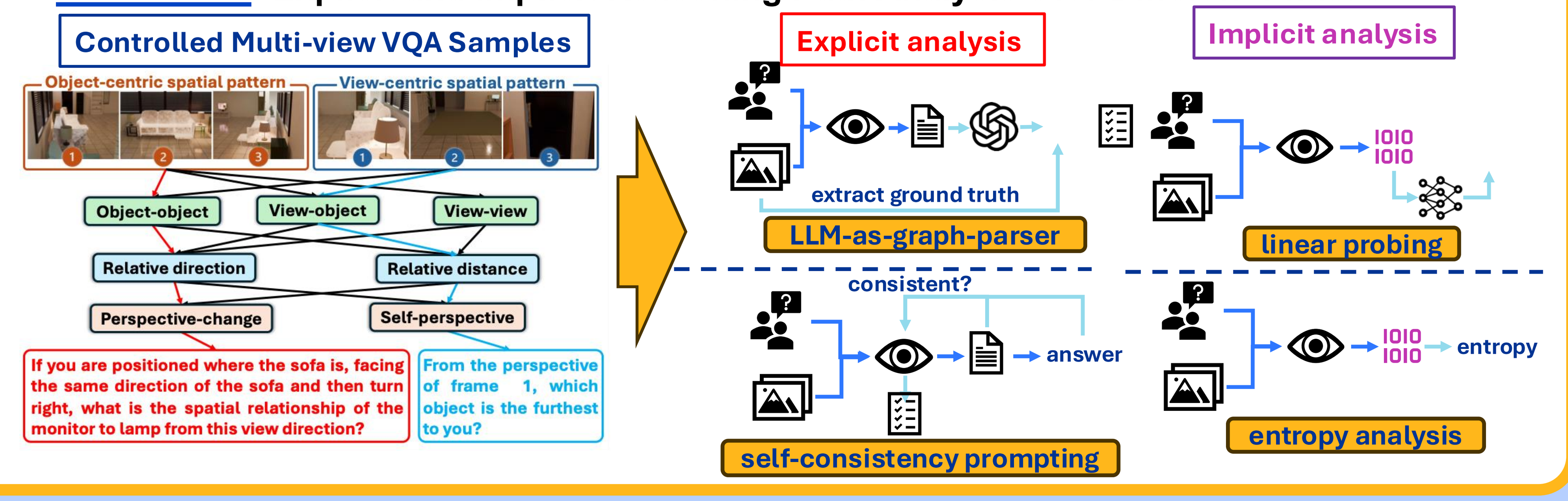
❖ **Motivation: A Cognitive View of Spatial Reasoning**

•Inspired by cognitive science, we analyze spatial reasoning as a sequence of four phases:

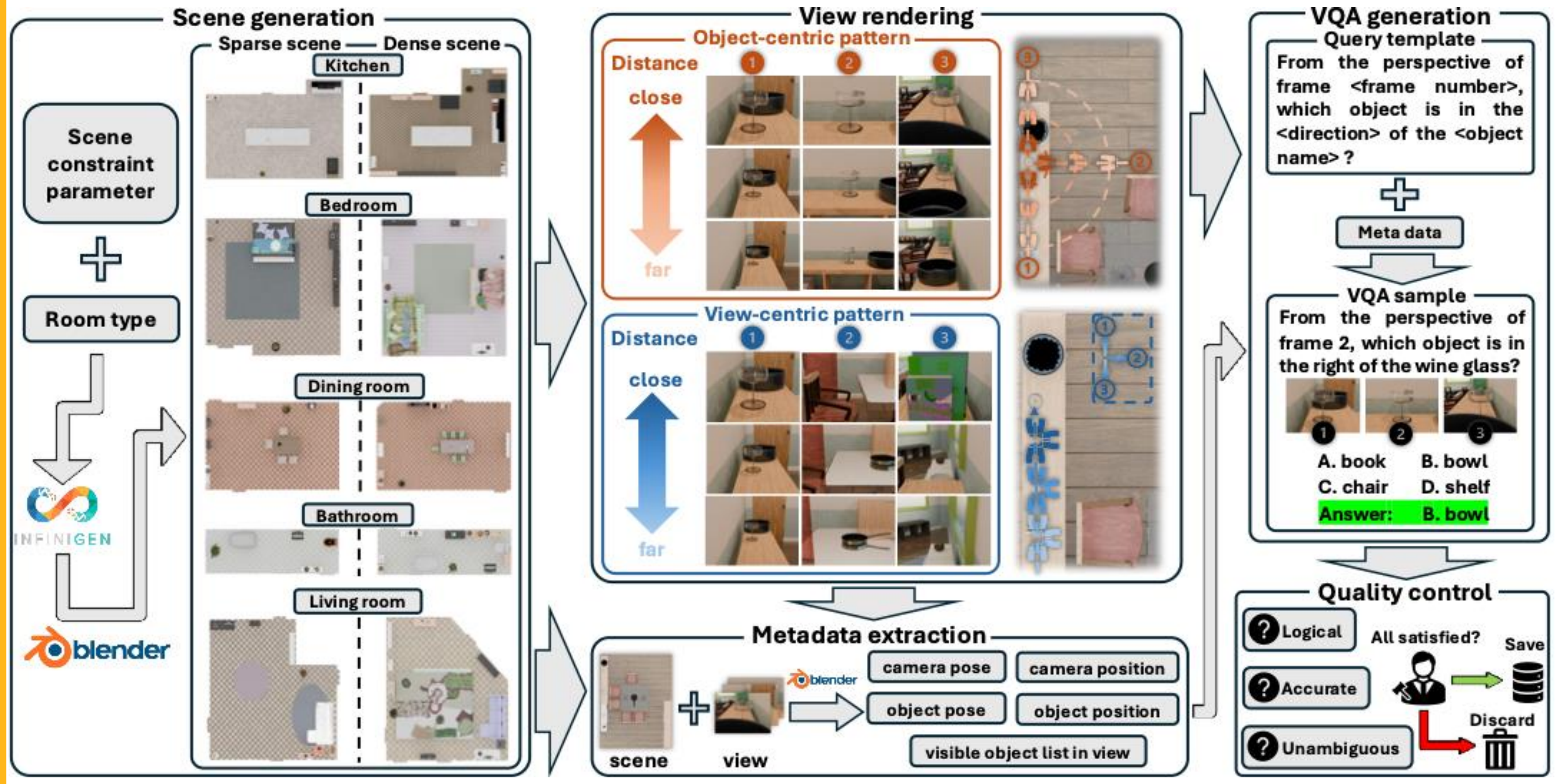
•This phase-based view allows us to identify failures that cannot be explained by final-answer acc.



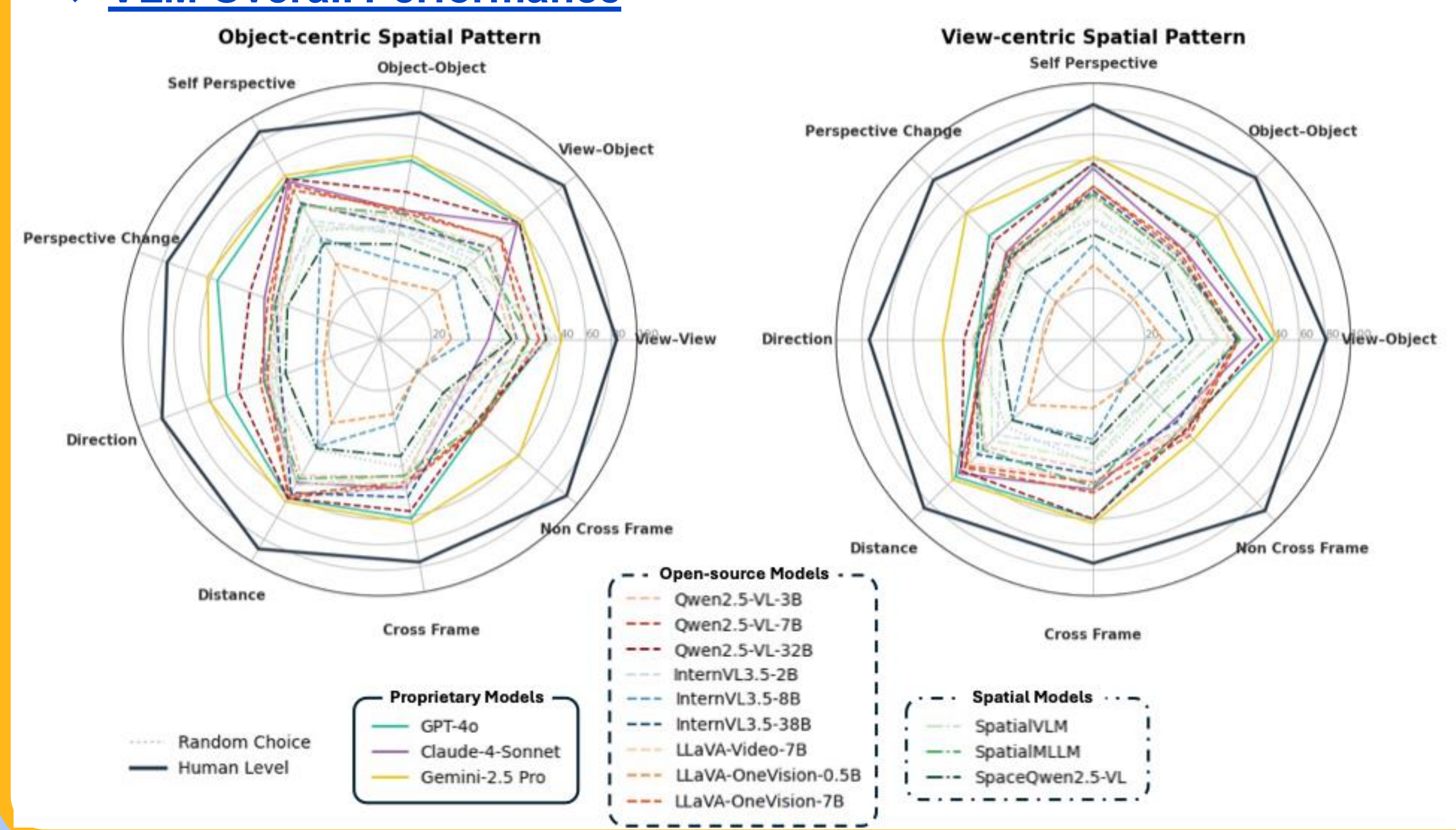
❖ **Our Method: Explicit and Implicit Reasoning Path Analysis over Different Phases**



Data Generation



❖ **VLM Overall Performance**



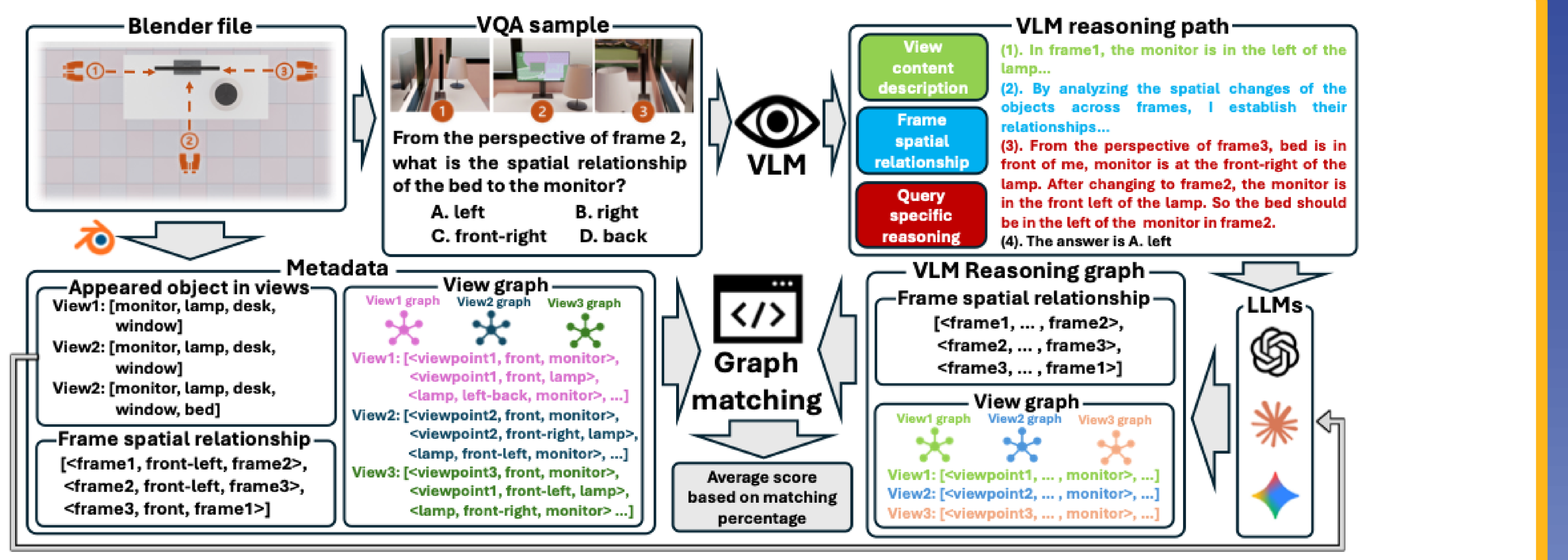
• Object-centric reasoning is consistently easier than view-centric reasoning.
• Perspective-changing queries cause a performance drop of more than four percentage points.
• Cross-frame reasoning is substantially harder than within-frame reasoning.

Explicit VLM Spatial Reasoning Analysis

❖ **Analysis using LLM as Graph Parser**

Method:

• Extract visible objects and spatial relations from Blender metadata to construct ground-truth view and frame graphs.
• Prompt the VLM through four reasoning phases, then use an LLM to parse each phase into spatial triplets: (subject, relation, object).
• Match predicted triplets with ground-truth graphs to measure phase-wise correctness and locate failures in perception, cross-view alignment, or inference.



Findings:

• Models perform well in perceptual encoding but deteriorate during cross-view alignment and inference.
• Larger models degrade more slowly, showing greater phase stability.

Model	Correct (%)			Incorrect (%)		
	1	2	3	1	2	3
Qwen2.5-VL-3B	73.2	43.8	32.6	68.7	31.4	18.5
Qwen2.5-VL-7B	76.1	58.2	36.3	72.5	43.6	21.9
Qwen2.5-VL-32B	**81.9**	**63.5**	**38.8**	**76.8**	**44.2**	**24.6**

Table 1: Phase-wise reasoning correctness of Qwen2.5-VL.

❖ **Self Consistency Prompting**

Method: Re-evaluate the multi-view input, reasoning path, and answer to measure reasoning-answer agreement.

Findings:

• Correct answers exhibit substantially higher self-consistency than incorrect answers.
• Errors are associated with contradictory reasoning and unstable relational grounding.

Model	Correct (%)	Incorrect (%)	Overall (%)
Qwen2.5-VL-3B	64.8	32.5	48.6
Qwen2.5-VL-7B	74.1	41.7	59.3
Qwen2.5-VL-32B	**83.6**	**54.2**	**68.4**

Table 2: Self-consistency prompting of Qwen2.5-VL family.

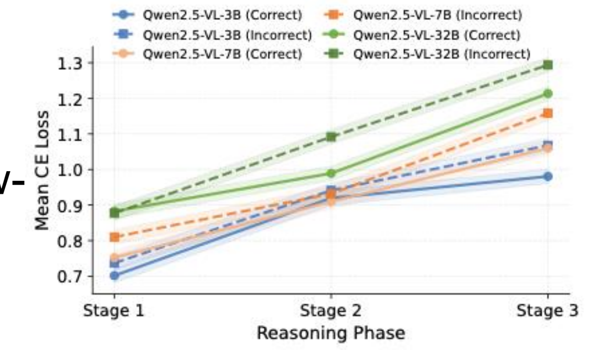
Implicit VLM Spatial Reasoning Analysis

❖ **Linear Probing Analysis**

• Train probes on token representations from each phase to predict the final-answer signal.
• Use probe loss to measure how much task-relevant information remains accessible.

Findings:

• Probe loss increases across phases, indicating progressive information degradation.
• Object-object relations are most stable, while view-view relations degrade the most.



❖ **Entropy Dynamics**

Method:

• Measure the mean and variance of token-level entropy at each reasoning phase.
• Compare uncertainty patterns between correct and incorrect reasoning paths.

Findings:

• Correct reasoning maintains lower entropy, while incorrect reasoning remains more uncertain.
• View-centric, perspective-changing, and cross-frame queries introduce uncertainty at early, middle, and late phases, respectively.