VQA-based with Robot Navigation for Worker Monitoring

Chaewon Yang
M.S. Student
Dept. of Architecture & Architectural Engineering
Seoul National University
chaewony@snu.ac.kr

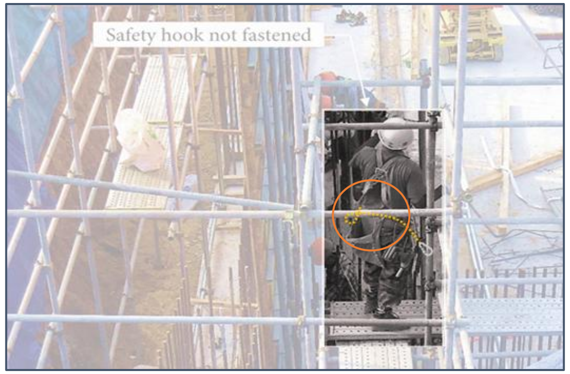
2025. 08. 05.

Research Background

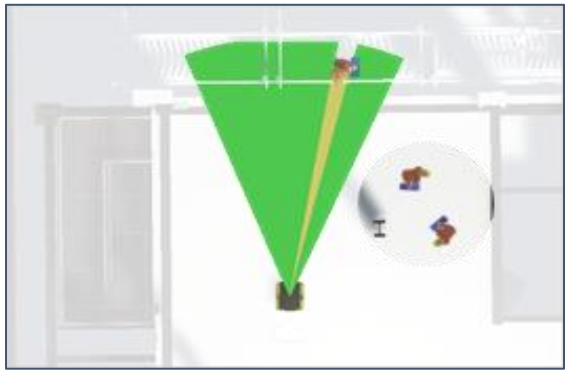
▪ Computer Vision Applications for Safety Monitoring

- Vision sensors are used for **verifying personal protective equipment (PPE) compliance through object detection models** on construction sites.
- Fixed-viewpoint monitoring systems struggle **with occlusion** and **limited Field of View (FoV),** which hinder accurate detection and classification(Paneru et al., 2021).

[Occlusion by Obstacle]

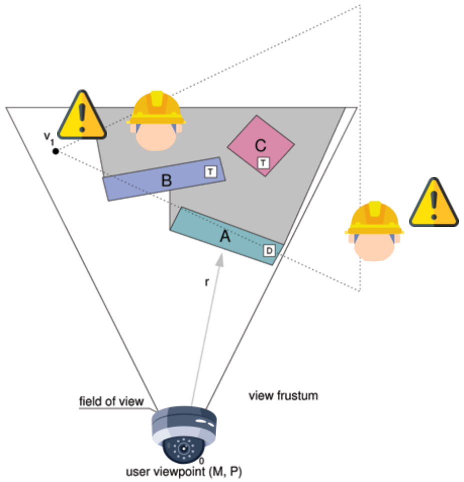


[Self-Occlusion]



[View Occlusion]

✓ Types of Occlusion Problems



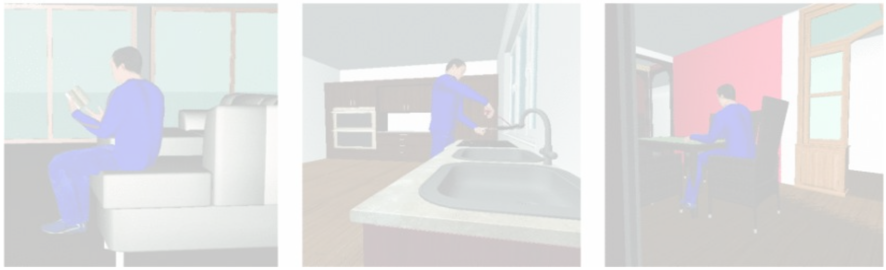
✓ Occlusion problem

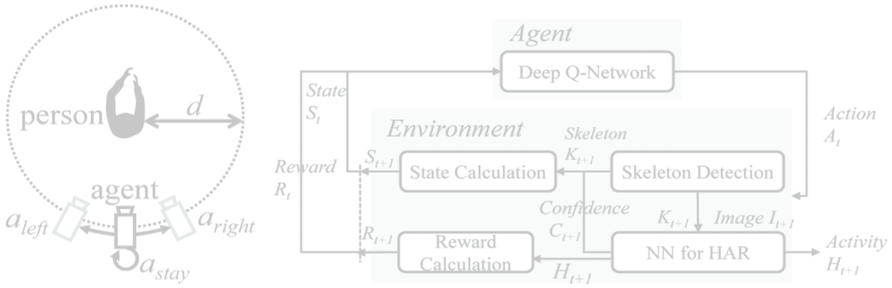
Research Background

▪ Active Vision for Occlusion : Related works

1) Optimizing the robot's view (Kumrai et al. 2020)

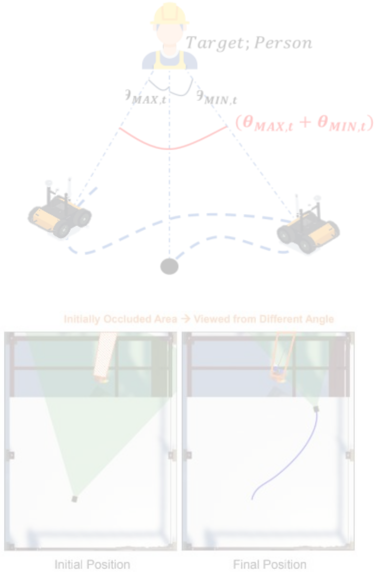
: Strategic camera placement and dynamic view planning using UAVs and UGVs to minimize self-occlusion and enhance perspective diversity.



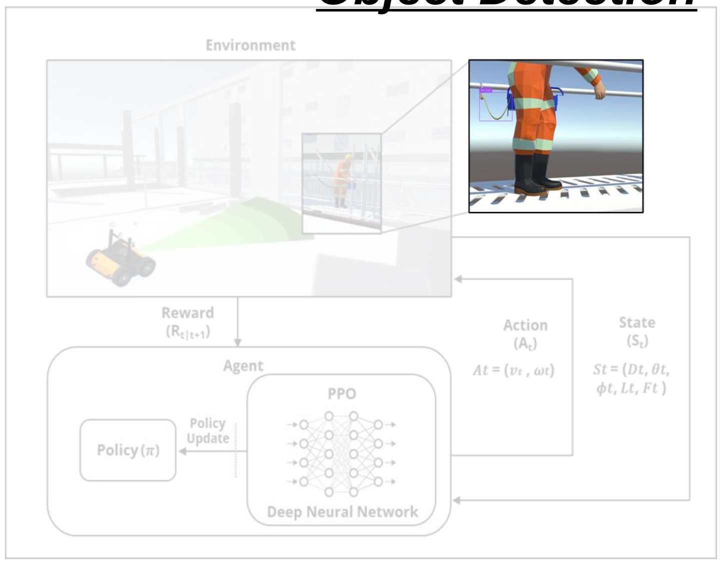


2) RL-based Navigation (Park et al. 2024, Yang et al. 2025)

: Reinforcement learning methods can enhance coverage area of workers and monitoring systems in obstacle-rich environment.

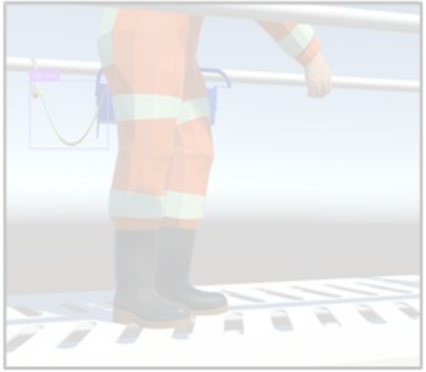


✓ ***Object Detection***



▪ Beyond Detection: Toward Context-Aware Monitoring

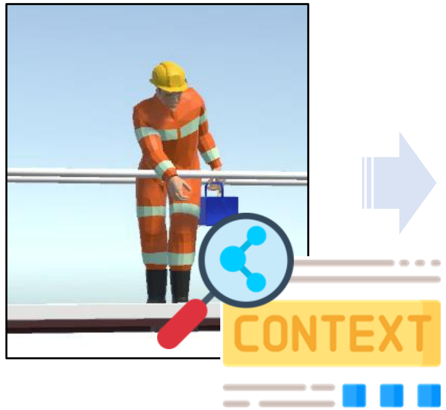
- Limitation of Detection Results based Monitoring



Results

- Detected
- Not Detected

…



Results

- Detected
- Not Detected
- Inconclusive

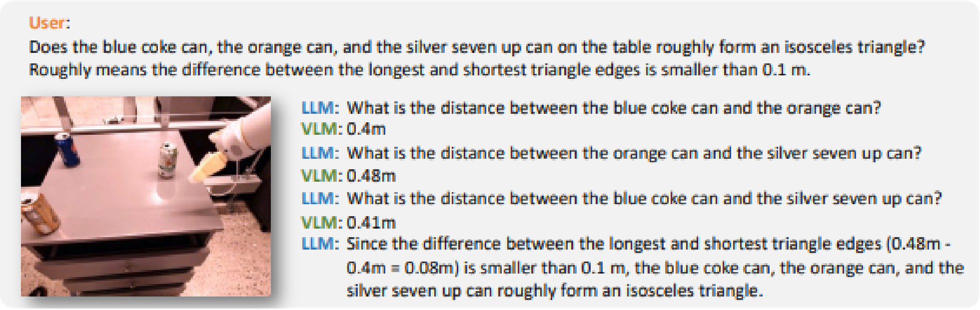
Contextual Difference

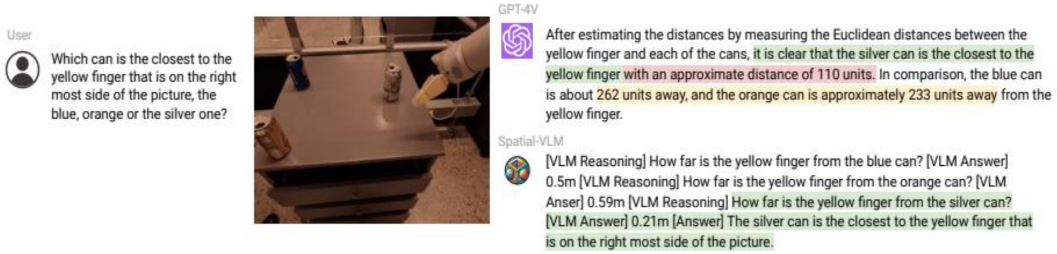
Research Background

- ## Context Inference from Image : Related works

1) Spatial Reasoning_SpatialVLM (Chen et al. 2024)

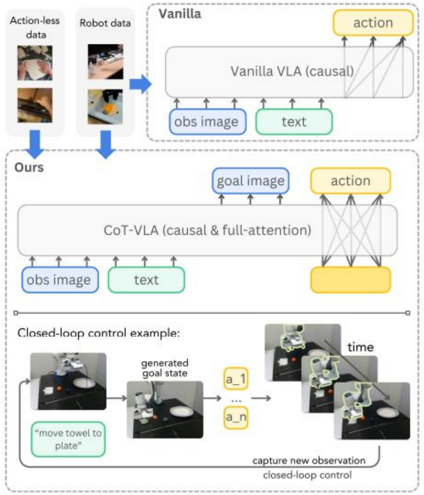
: Quantitative spatial reasoning via large-scale synthetic QA, paving the way for context-aware understanding.

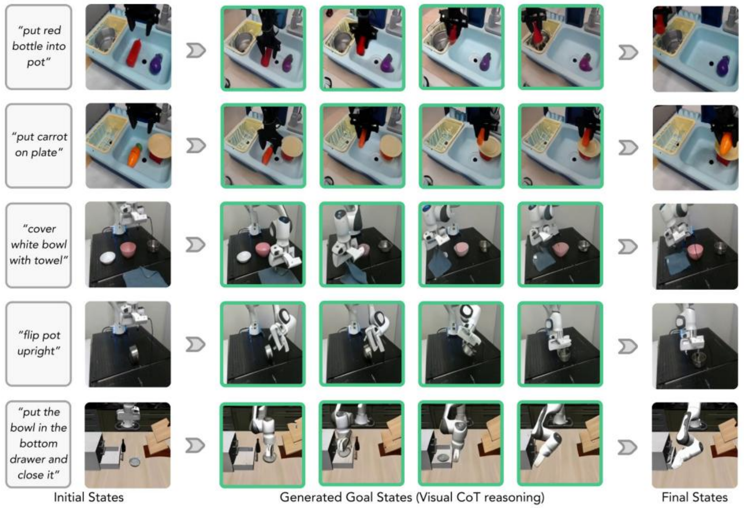




2) Context Understanding_CoT-VLA (Zhao et al. 2025)

: First reasons over what it sees—allowing for deeper situational understanding and context-sensitive decisions.





Research Background

- **Context Inference from Image : Related works**

Context Understanding Pilot Test using GPT-4o

ChatGPT-4o

[GPT-4o] Q: You are monitoring the worker's PPE compliance on a construction site. Based on this image, answer the following question:

Do you see a yellow lifeline attached to the worker?
(Note: We only care about the presence of the lifeline on the worker — not whether it is connected to anything.)

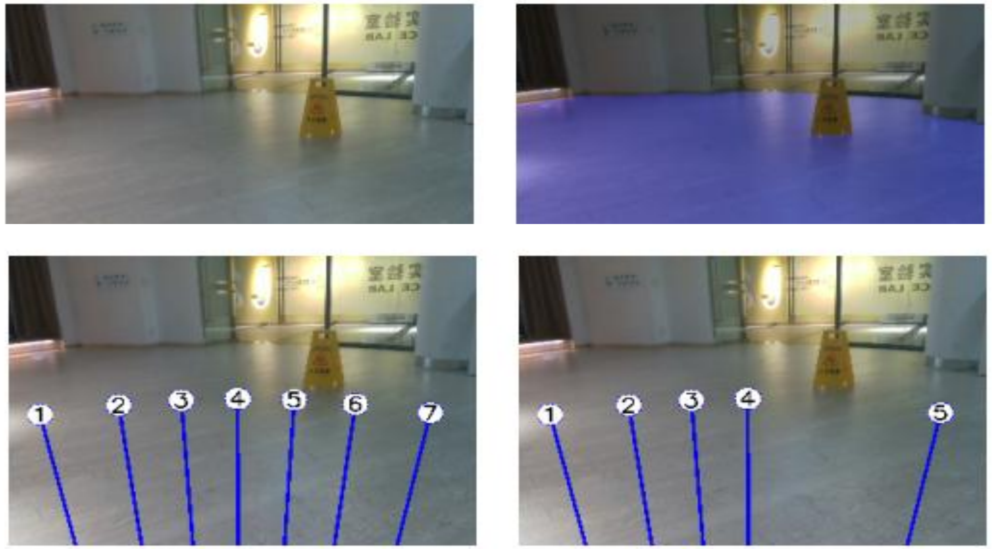
→ A: It is not possible to determine from this image.
Thought: The worker is in the middle distance. The agent should move closer to the worker, ideally to the right side and slightly forward to get a better view of any lifeline attached to the worker.

Research Background

▪ VQA Navigation: Related works

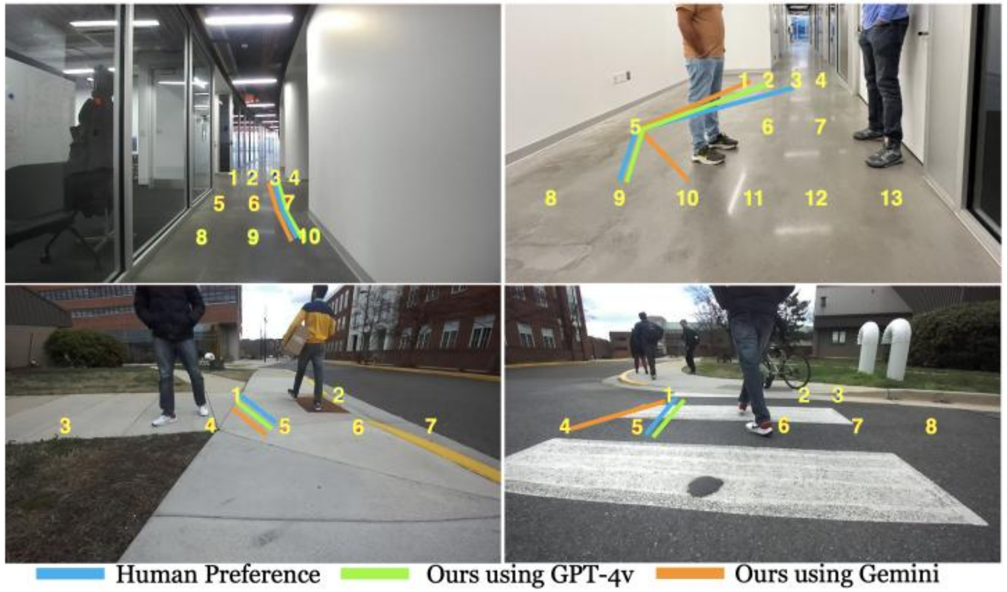
1) Spatial Reasoning_DyNaVLM (Chen et al. 2024)

: The VLM visually reasons about navigation targets and dynamically builds action space, guided by graph memory encoding object-topology relations.



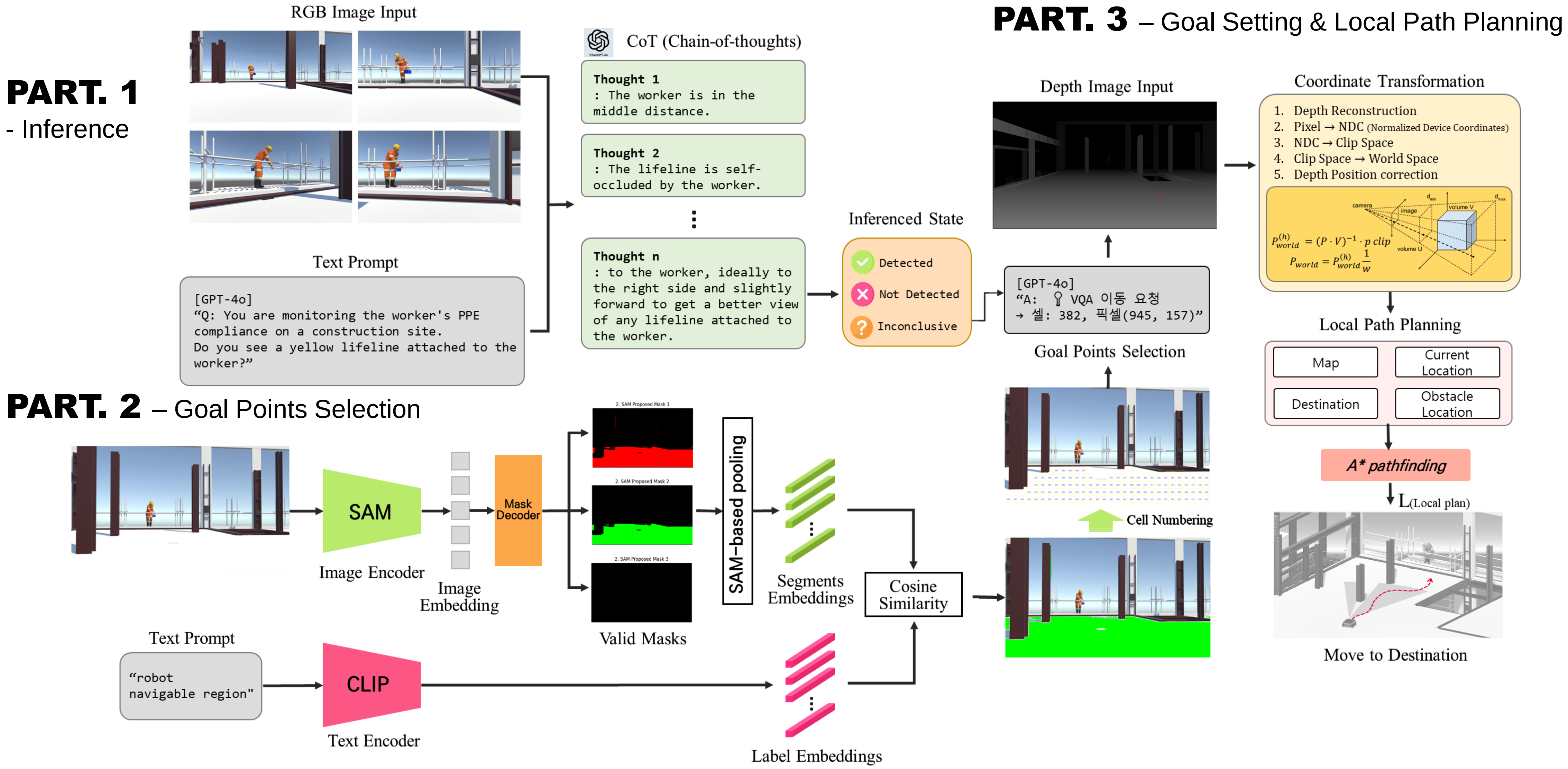
2) Context-aware Navigation (Sathyamoorthy et al. 2024)

: VLM selects a numbered grid cell overlaid on a traversable map image, guiding motion planning via visually grounded navigation.



Research Framework

- **VQA based Navigation Framework**



Research Framework

▪ Research Plan

1. Construct and test a dataset to evaluate the reasoning capabilities of VLMs (Vision-Language Models).
2. Enhance VQA (Visual Question Answering) performance through advanced prompt engineering and CoT (Chain-of-Thought) design.
3. Build a robust data processing system and conducted reality-based testing for reliable evaluation.

▪ Expected Results

- Move beyond binary monitoring based on object detection to enable deeper, inference-driven monitoring.
- Allow robust responses to real-time situations, rather than relying on predefined policies or fixed routes.
- Enable task-oriented navigation without requiring prior training on specific objects.

References

Goetting, D., Singh, H. G., & Loquercio, A. (2024). End-to-end navigation with vision language models: Transforming spatial reasoning into question-answering. arXiv preprint arXiv:2411.05755.

Park, J., Kim, S., Park, M., & Ahn, C. R. A. (2024). Occlusion-Aware Object Detection for Worker Monitoring Using a Reinforcement Learning Approach with a Mobile Robot. Proceedings of the 2024 ASCE International Conference on Computing in Civil Engineering (i3CE).

Ji, Z., Lin, H., & Gao, Y. (2025). DyNaVLM: Zero-Shot Vision-Language Navigation System with Dynamic Viewpoints and Self-Refining Graph Memory. arXiv preprint arXiv:2506.15096.

Mao, Q., Wang, J., Xu, M., & Chen, Z. (2024). NavTr: Object-Goal Navigation With Learnable Transformer Queries. IEEE Robotics and Automation Letters.

Buoso, D., Robinson, L., Averta, G., Torr, P., Franzmeyer, T., & De Martini, D. (2024). Select2plan: Training-free icl-based planning through vqa and memory retrieval. *arXiv preprint arXiv:2411.04006*.

Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., & Xia, F. (2024). Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14455-14465).

Zhao, Q., Lu, Y., Kim, M. J., Fu, Z., Zhang, Z., Wu, Y., ... & Xiang, D. (2025). Cot-vla: Visual chain-of-thought reasoning for vision-language-action models. In Proceedings of the Computer Vision and Pattern Recognition Conference (pp. 1702-1713).

Yang, C., Park, J., Park, M., & Ahn, C. R. A. (2025). Active Robot Vision with Perception-Aware Navigation for Worker Monitoring . *Proceedings of the 2025 ASCE International Conference on Computing in Civil Engineering (i3CE)*.

Thank you!

Chaewon Yang
M.S. Student
Dept. of Architecture & Architectural Engineering
Seoul National University
chaewony@snu.ac.kr