DLCV Final Project : Challenge 1 - Warehouse Spatial Intelligence

(Track 3 of Nvidia AI City Challenge in ICCV 2025)

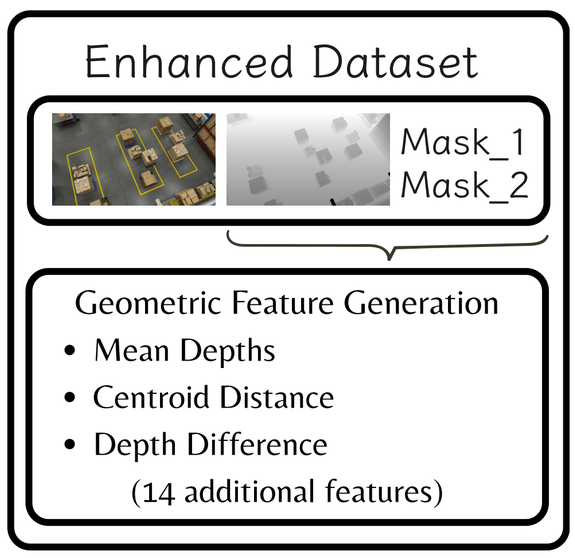
EE4 b11901067 董家愷 EE4 b11901100 王竣平 EE4 b11901176 黃堉翔 EE3 b12901069 段奕鳴

Intro

We present an enhanced Spatial QA Agent based on the "Warehouse Spatial QA" framework. Our system features three key architectural innovations:

- Distance Model: Integrated a Geometric Shortcut with 14 Depth features and optimized via a two-stage training strategy (MSE→Log-MSE), achieving near-perfect precision.
- Inclusion Model: Improved edge-case robustness by fusing 8 explicit geometric features (e.g., IoU, depth diff) into the visual backbone.
- LLM-Driven Parsing: Replaced rule-based rephrasing with LLM extraction for accurate mask identification.

These contributions resulted in a top-tier score of 95.16 on CodaBench. Crucially, this iterative optimization process was guided by our custom automatic error analysis pipeline, which systematically diagnosed baseline failures and validated our geometric refinements.



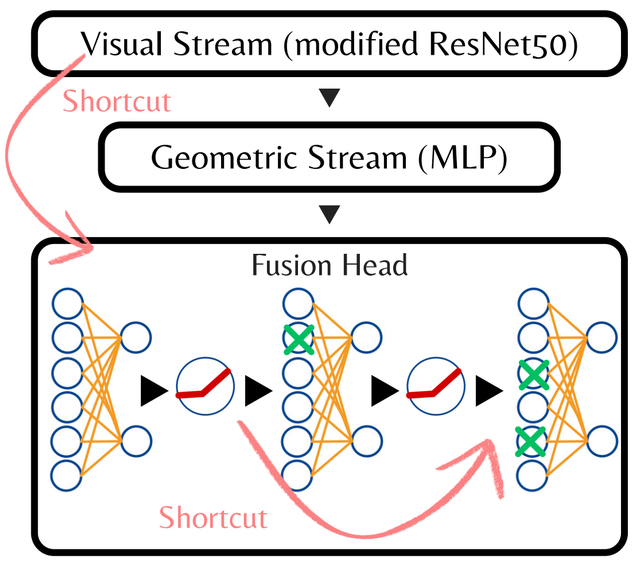


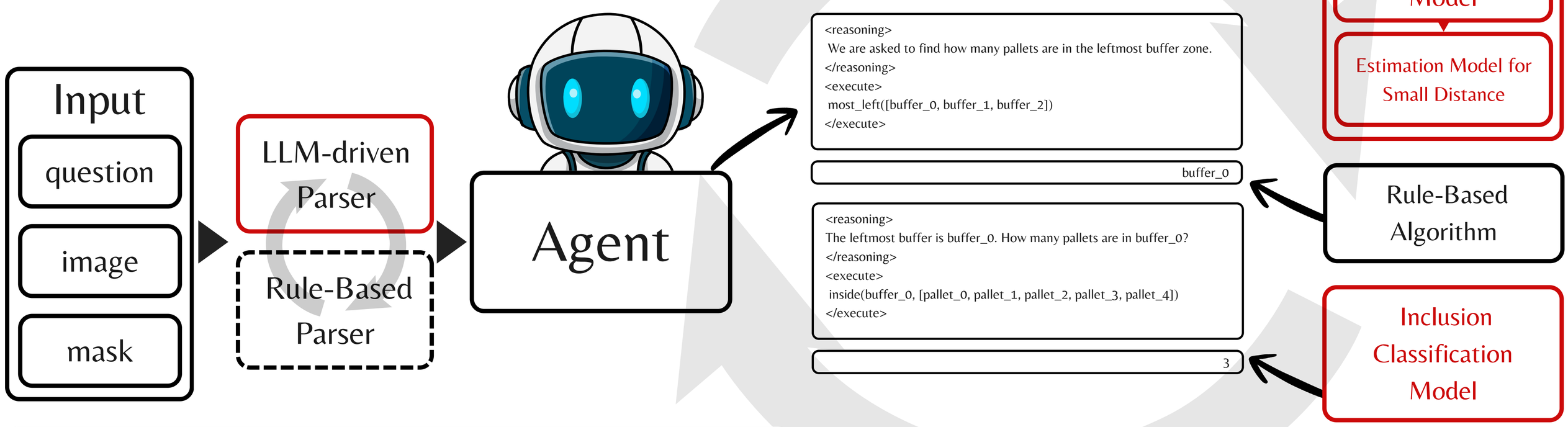
Table 1: Error counts across different distance ranges in the validation set

Range	Total Q.	Baseline	+ Depth + LogMSE	+ Depth + Shortcut + LogMSE
< 1m	9	1	9	3
1 – 3m	20	2	9	1
3 – 10m	68	3	3	1
> 10m	53	5	1	0

Data Curation

During training of the inclusion and the distance estimation model, we created a dataset where each entry is centered on a single image, augmented with region-level annotations (e.g., object masks or region identifiers), and paired with a task-specific target such as a distance label or inclusiveness. Ground truth is extracted from free-form answers or conversations by the LLM.

For the inclusion dataset, negative pairs are additionally generated by selecting non-inclusion pairs with the highest IoU among all masks, along with randomly sampled pairs, with proportions of 30% and 20%, respectively.



Automatic correction pipeline

VLM Diagnosis & Classification

Given the image, the ground-truth answer, and our predicted reasoning path, can you help diagnose which step went wrong?

"from": "gpt",
"value": "The pallet [Region 0] is to the left of the pallet [Region 1]."

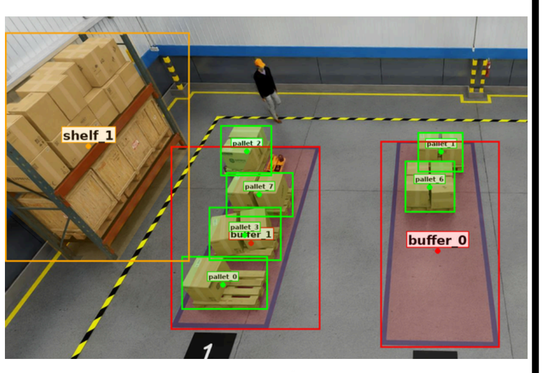
left

GT(freeform answer)

<reasoning>
The tool result from is_left is false. <answer>right</answer>
</reasoning>

right

our prediction



annotated image

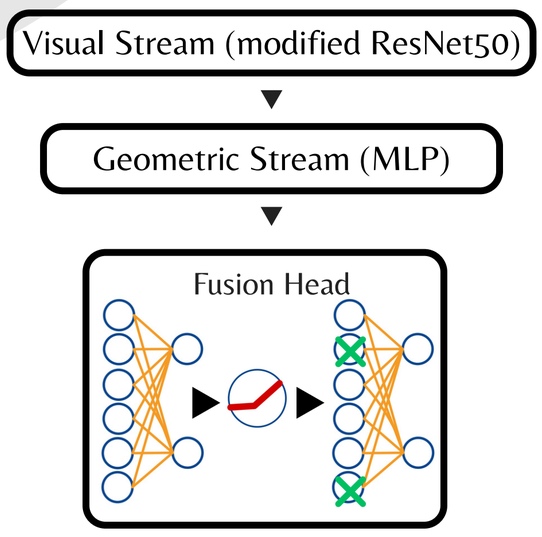
Qwen3-VL-8B-Instruct

"reasoning": The function is left to return a false value since the agent passed wrong argument to the function, check the reasoning process.
"error": reasoning error

distance: 47
inclusion: 25
is_closest: 12
other: 13

By leveraging Qwen3-VL's strong OCR and spatial reasoning capabilities on our annotated images, we automatically categorized over 100 failure cases, significantly accelerating the debugging process and facilitating improvements for the next model version.

The curated dataset for the inclusion model is designed from a relation-centric perspective, explicitly encoding spatial interactions between objects. Features such as IoU and multiple normalized overlap ratios capture containment, intersection, and relative occupancy, which are critical cues for spatial reasoning tasks such as region–object interaction understanding.



Accuracy and Average Error Rate of the Inclusion Model

Method	Baseline	Ours
Accuracy	0.579	0.906
Average Error Rate	0.356	0.054

Overall Perfomance enhance

Method	Count	Distance	Left/Right	MCQ	Overall
Baseline	0.579	0.927	1.000	0.698	0.805
Ours	**0.906**	**0.960**	**1.000**	**0.940**	**0.952**

In conclusion, we introduce an error analysis pipeline and a reproducible cookbook for training and dataset curation of distance estimation and inclusion models, resulting in near-perfect accuracy on a warehouse spatial reasoning benchmark.

Reference

H.-W. Huang, J.-H. Cheng, K.-M. Chen, C.-Y. Yang, B. Alattar, Y.-R. Lin, P. Kim, S. Kim, K. Kim, C.-I. Huang, and J.-N. Hwang, "Warehouse Spatial Question Answering with LLM Agent: 1st Place Solution of the 9th AI City Challenge Track 3," arXiv preprint arXiv:2507.10778v2, Aug. 2025.