

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 黃墳翔 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.

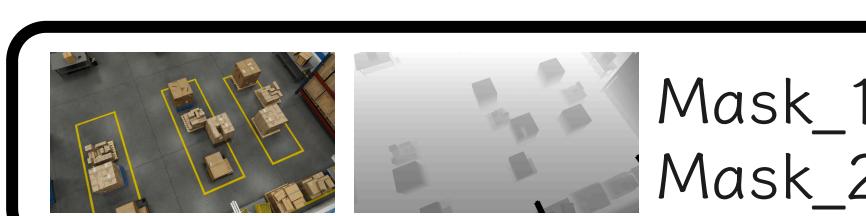
These contributions resulted in a top-tier score of **on CodaBench**. Crucially, this iterative optimization process was guided by our custom automatic error analysis pipeline, which systematically diagnosed baseline failures and validated our geometric refinements.

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.

Enhanced Dataset



- Geometric Feature Generation
- Mean Depths
 - Centroid Distance
 - Depth Difference
(14 additional features)

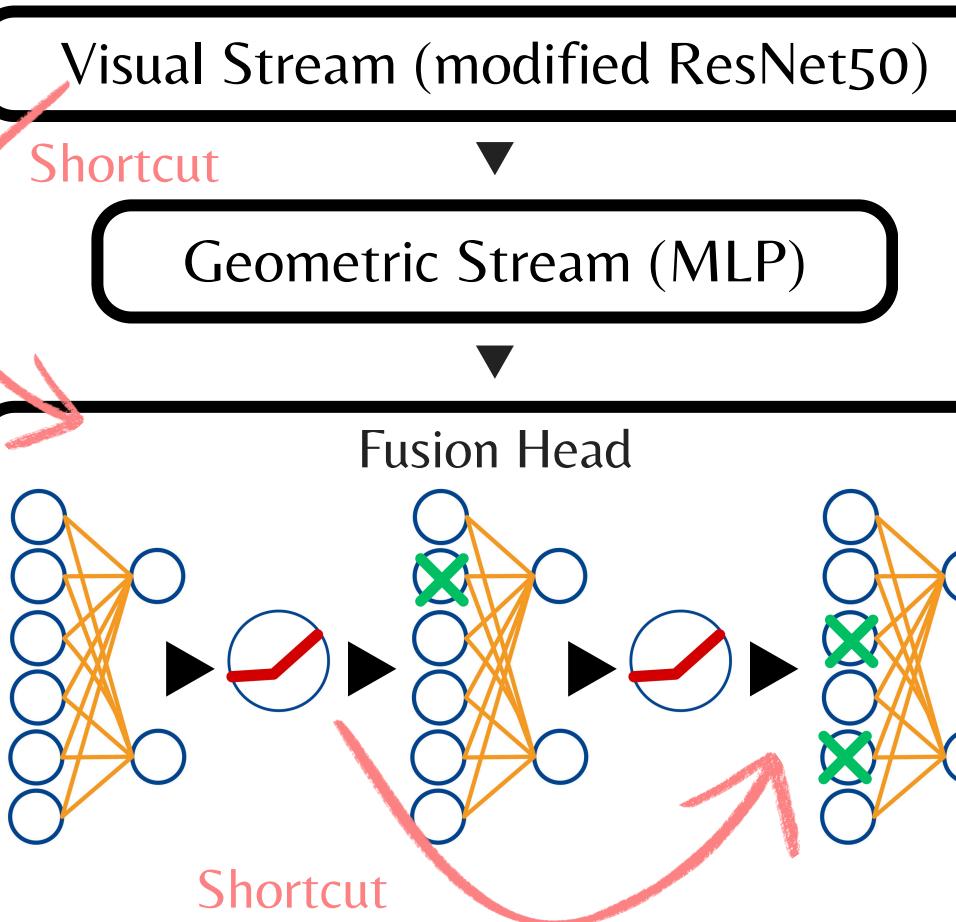


Table 1: Error counts across different distance ranges in the validation set

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

REACT loop

Input

- question
- image
- mask

- LLM-driven Parser
- Rule-Based Parser

Agent

```

<reasoning>
We are asked to find how many pallets are in the leftmost buffer zone.
</reasoning>
<execute>
most_left([buffer_0, buffer_1, buffer_2])
</execute>
```

buffer_0

```

<reasoning>
The leftmost buffer is buffer_0. How many pallets are in buffer_0?
</reasoning>
<execute>
inside(buffer_0, [pallet_0, pallet_1, pallet_2, pallet_3, pallet_4])
</execute>
```

3

Distance Estimation Model

Estimation Model for Small Distance

Rule-Based Algorithm

Inclusion Classification Model

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

