# SpatialVLM → CRISP

**Chen et al.**
**CVPR 2024**

**Zhixing Li and Yinan Yu**
**ECCV 2026**

Spatial / 3D Reasoning and Geometric Structure

**Shaunak Karnik, Rachel Shi, Namhoon Kim, Yung-Ching Sun**

Georgia Tech

# Outline

- Problem Statement
- Related Works
- SpatialVLM
  - Approach
  - Experiments & Results
- CRISP
  - Approach
  - Experiments & Results
- What persisted and what changed?
- Limitations and Implications
- Discussion Questions

Georgia Tech

# Problem Statement

**VLMs have strong 2D semantic perception capabilities, but struggle with 3D spatial perception and reasoning.**

SpatialVLM attributes this to the lack of 3D spatial training data, as most large datasets are made up of simple image-caption pairs.

CRISP argues that traditional QA formats, like that introduced in SpatialVLM, allow models to exploit language priors without actually grounding their reasoning in spatial information.

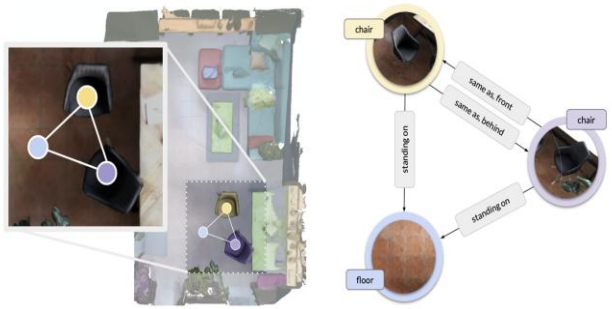
This is an important issue to tackle, as strong 3D spatial perception and reasoning are crucial for embodied tasks in the real world like robotics.

Georgia Tech

# Related Works: Learning Spatial Reasoning

## Explicit Spatial Representations

- Explicit spatial scene memories
- Scene graphs encode structured spatial relationships
- Spatial VQA through explicit pathfinding on scene graphs




Scene graphs allow interpretable, structured, statistical relation learning based on explicit encoded spatial structures

## Implicit Spatial Representations

- VLMs are pretrained on large vision-language datasets
- Spatial knowledge encoded in model weights
- Depth and 3D structure through auxiliary tasks

**SpatialVLM: learns spatial relationships directly in VLMs, without explicit scene graphs**

Georgia Tech

# Related Works: Grounding VLMs for Spatial Reasoning

- LLMs/VLMs gain common sense from internet-scale data
- May lack necessary grounding to perform well at
  - social reasoning
  - physical reasoning
  - embodied tasks
  - spatial reasoning tasks
- Visually Grounded Models
  - Flamingo [1], PaLI [2], PaLM-E [3]
  - Applications: success detection, action prediction, reward prediction
- Research Direction
  - Fine-tune VLMs with generated spatial VQA data
  - Transfer spatial reasoning to embodied applications

[1] Alayrac, J., et al., 2022 [2] Chen, X., et al., 2022 [3] Driess, D., 2023

Georgia Tech

# Related Works: Visual Spatial Intelligence

**Implicit 3D Spatial Cognition:** Understanding 3D world structures from 2D observations.

| Geometric Priors<br>Input / Architecture | Optimization<br>Training | Inference Prompting<br>Inference-time Reasoning |
|---|---|---|
| • Explicitly inject depth maps<br>  ○ SpatialRGPT [1]<br>  ○ SpatialCLIP [2]<br>• Leverage encoders like VGGT to extract spatial features<br>  ○ VG LLM [3]<br>  ○ SpatialMLLM [4] | • Refine pre-training objectives<br>  ○ 3D-VisTA [5]<br>  ○ TIPS [6]<br>• Employ reinforcement learning strategies<br>  ○ SpaceR [7]<br>  ○ SpatialReasoner [8] | • Multimodal Chain-of-Thought (CoT) [9] reasoning<br>  ○ VoT [10]<br>  ○ MVoT [11] |

**Key Idea:** Spatial intelligence can be enhanced through geometric representations, training strategies, and inference-time reasoning.

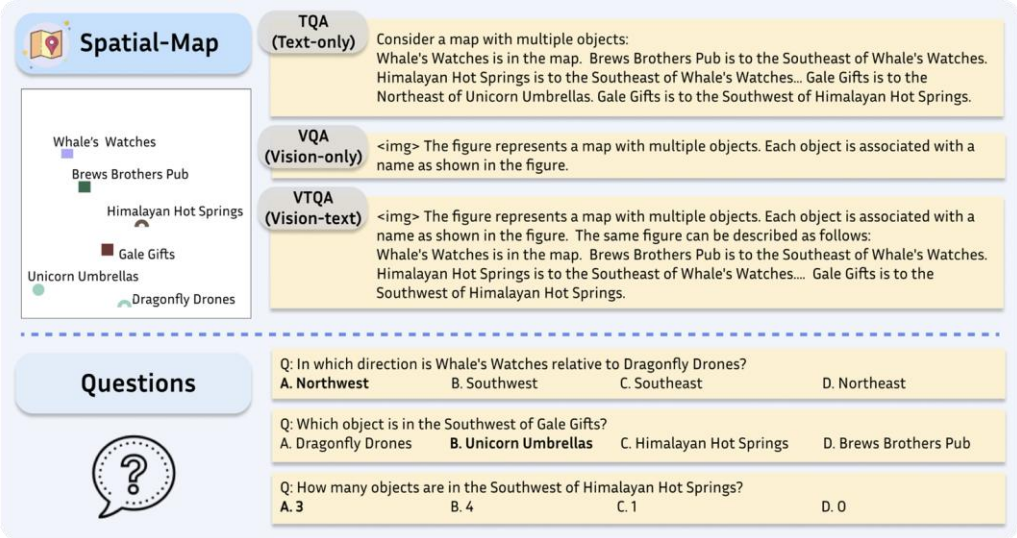
[1] Cheng, A.C., et al., 2024 [2] Wang, X., et al., 2022 [3] Zheng, D., et al., 2025 [4] Wu, G., et al., 2025 [5] Zhu, Z., et al., 2023 [6] Kokitsi Maninis, K., et al., 2025 [7] Ouyang, K., et al., 2025 [8] Ma, W., et al., 2025 [9] Wei, J., et al., 2022 [10] Wu, W., et al., 2025 [11] Li, C., et al., 2025

Georgia Tech

# Related Works: Spatial Reasoning Benchmarks

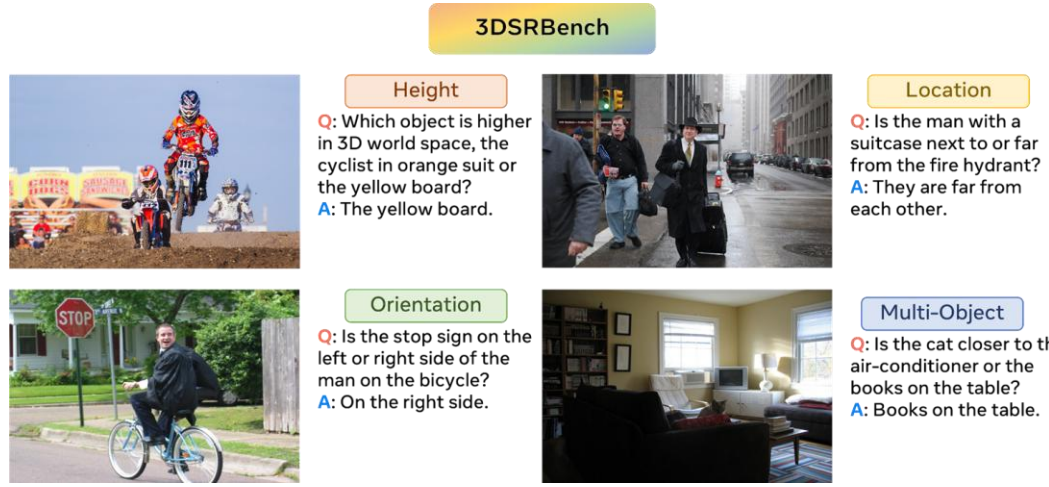
## Abstract Cognitive Probes on Synthetic Images

SPACE [1], BSA [2], SpatialEval [3]




## Embodied Benchmarks in Navigable 3D Spaces

3DSR Bench [4], Open3D-VQA [5], SQA3D [6], Whatsup [7], OmniSpatial [8], MMSI-Bench [9]




**Key Limitations: (1) Rely on discriminative question answering (2) allow VLMs to exploit "semantic shortcuts" to answer questions.**

**CRISP: 3D Scene Graph Construction + Diagnosis of Perception-Reasoning Alignment**


[1] Ramakrishnan, S.K., et al., 2024 [2] Xu, W., et al. 2025 [3] Wang, J., et al. 2024 [4] Ma, W., et al. 2025 [5] Zhang, W., et al. 2025 [6] Ma, X., et al. 2022 [7] Kamath, A., et al. 2023
[8] Johnson, J., et al. 2015 [9] Yang, S., et al. 2025


Georgia Tech

# Approach: SpatialVLM Overview

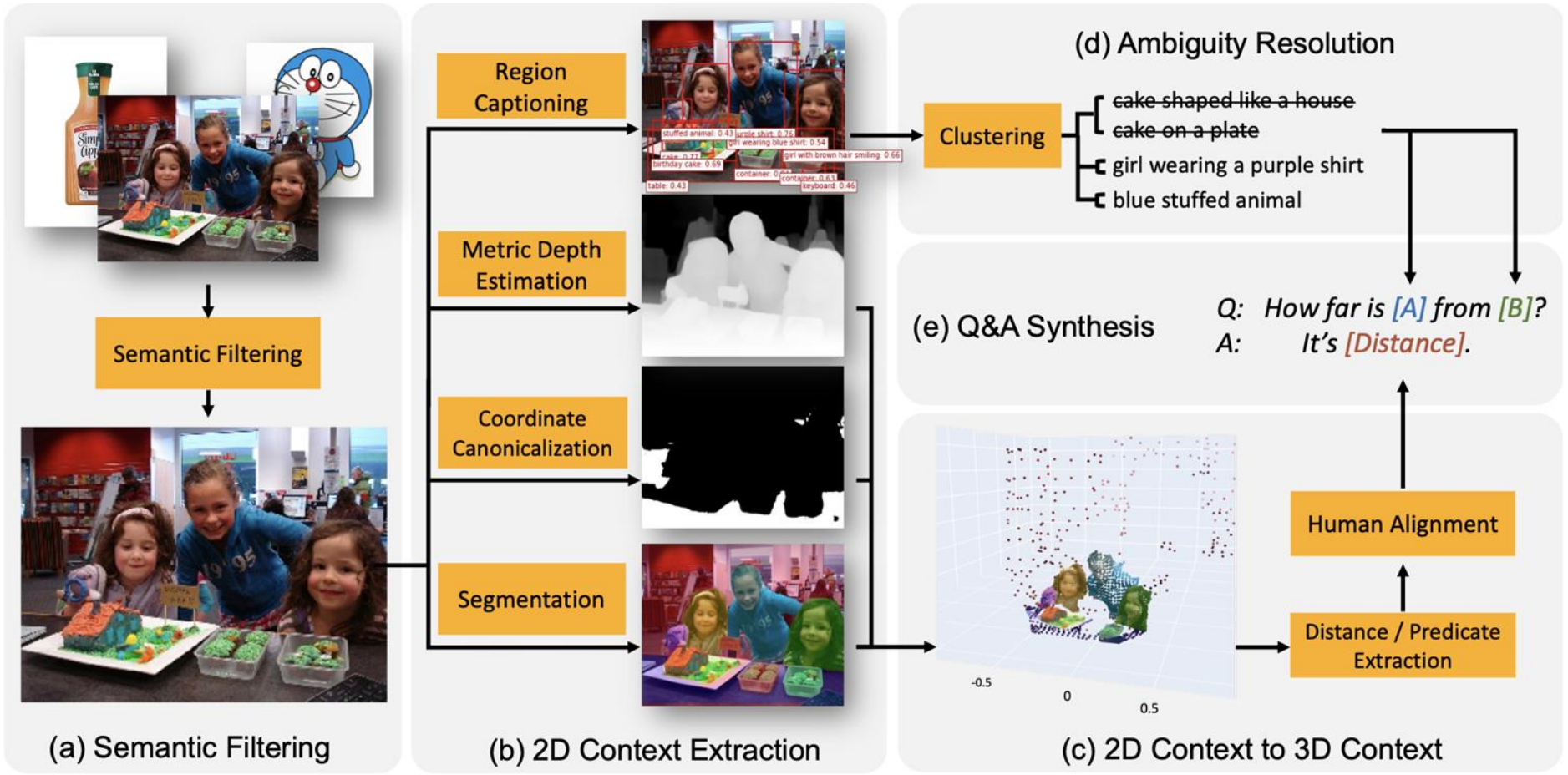
1. Introduce a pipeline to generate internet-scale 3D spatial VQA data

2. Train SpatialVLM on this spatial reasoning data

3. Utilize SpatialVLM for complex reasoning and robotics tasks

Georgia Tech

# Approach: Spatial VQA Data Synthesis Pipeline




*Chen et al., SpatialVLM (CVPR 2024), Figure 2*

Georgia Tech

# Approach: Filtering and 2D Context Extraction

## Semantic Filtering

- Use CLIP to exclude images with a single object without a scene background
- Use the prefix “an iphone photo” in positive image labels to include images with a FOV in the distribution of depth estimation model

## Object-centric Entity Information

- Utilize off-the-shelf models to get object segmentation masks and 1-6 word captions




*Chen et al., SpatialVLM (CVPR 2024), Figure 2*

Georgia Tech

# Approach: 2D to 3D Lifting

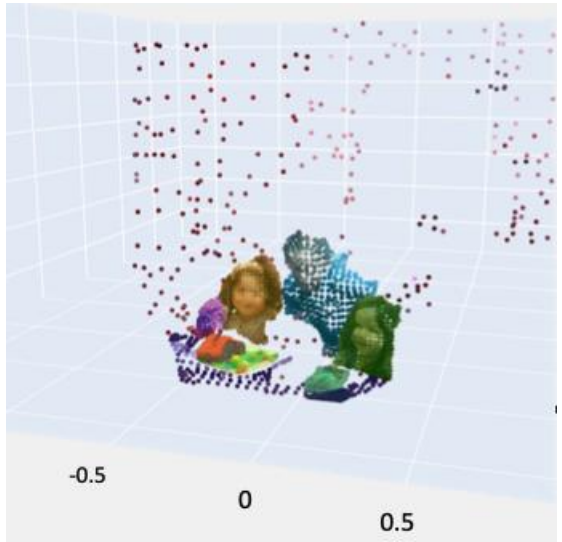
## 2D Image to Point Cloud

- Use ZoeDepth and FOV estimation to lift the images into 3D point clouds
- Clean up the point cloud by removing outliers and removing sparse points

## Coordinate Canonicalization

- Segment horizontal surfaces (floor, tabletop) and fit a plane on its 3D points
- Project the camera frame onto the plane to convert to world frame

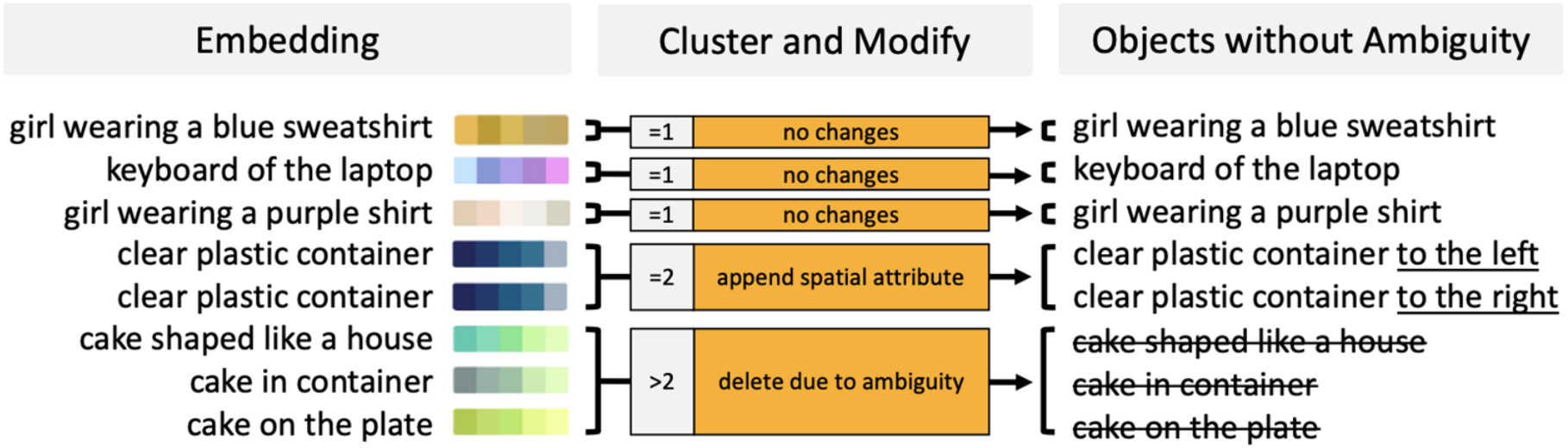




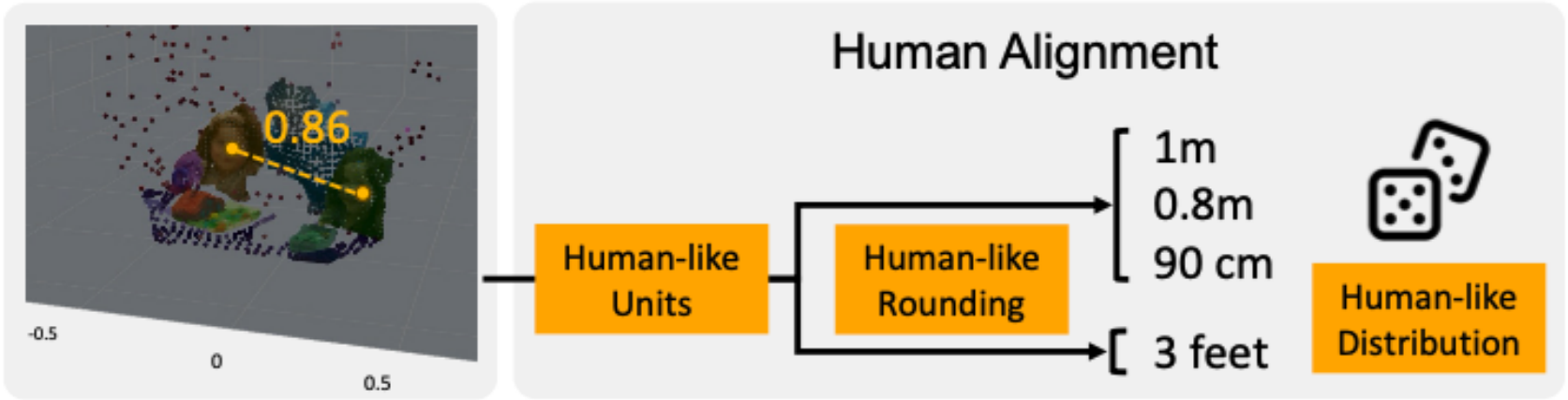

*Chen et al., SpatialVLM (CVPR 2024), Figure 2*

Georgia Tech

# Approach: Ambiguity Resolution and Human Alignment




*Chen et al., SpatialVLM (CVPR 2024), Figure 12*




*Chen et al., SpatialVLM (CVPR 2024), Figure 13*

Georgia Tech

# Approach: Large-Scale Spatial VQA Dataset

## Dataset Format

At most 2 objects (A and B)

Qualitative questions:

- Judges a spatial relation
  - ex. Given objects A and B, which is more to the left/right, bigger/smaller, more elevated, etc.

Quantitative questions:

- Fine-grained answers with number and units
  - ex. How far is object A from B?

## Dataset Scale

**10M** Images

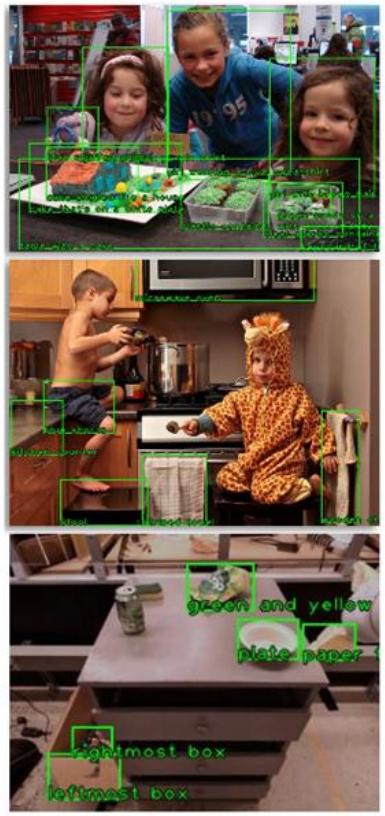
**2B** QA Pairs

**50%** Qualitative

**50%** Quantitative

Georgia Tech

# Approach: Dataset Examples




| Qualitative Q & A | |
|---|---|
| Find out if the girl wearing a purple shirt is situated behind or in front of the girl wearing a blue sweatshirt. | It's the woman wearing white shirt. |
| Does the stool lie in front of the microwave oven? | Yes, the stool is in front of the microwave oven. |
| Is the plate positioned on the left or right side of the green and yellow napkin? | The plate is to the right. |

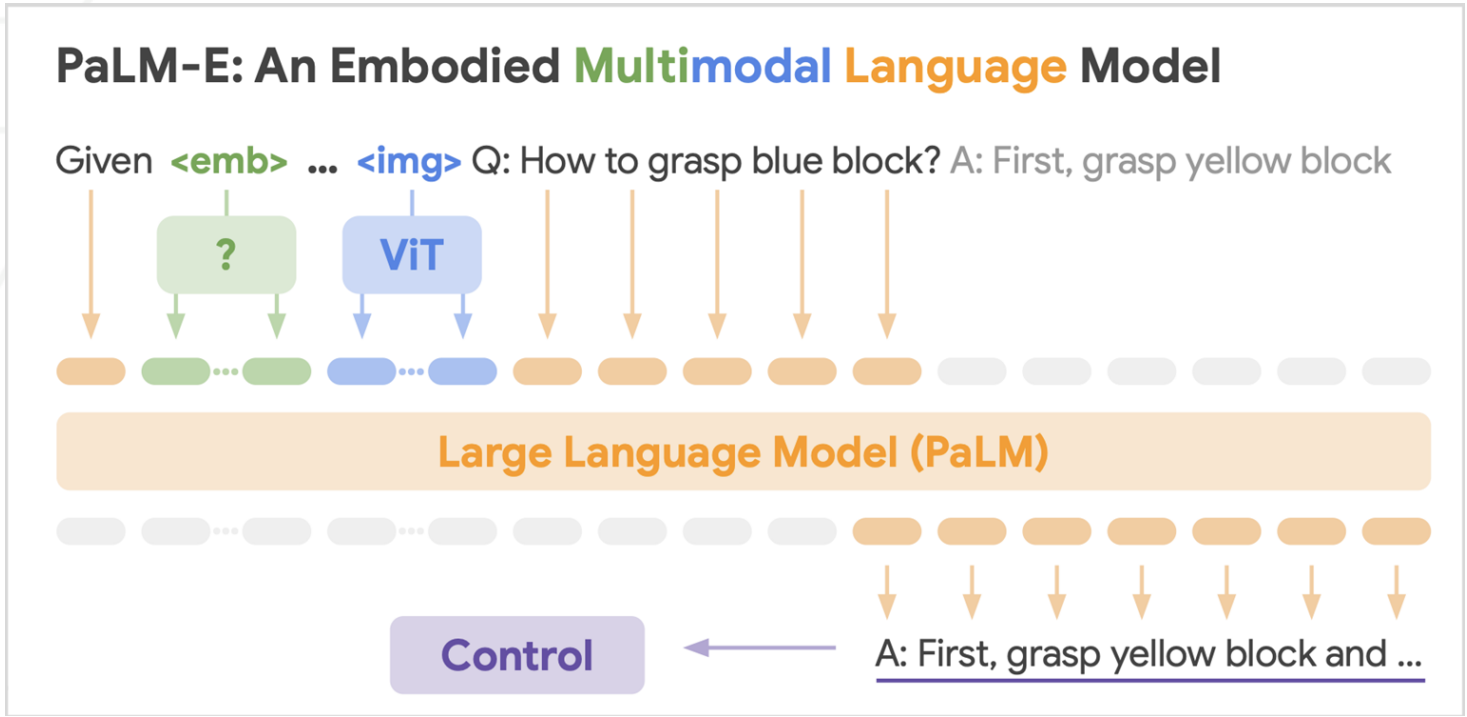
| Quantitative Q & A | |
|---|---|
| Measure the distance from the cake shaped like a house to the girl wearing a purple shirt? | Around half a meter |
| Can you provide the distance measurement between the microwave oven and the kitchen counter? | They are 1 meter apart. |
| How wide is the plate? | 20 centimeters |

*Chen et al., SpatialVLM (CVPR 2024), Figure 3*

Georgia Tech

# Approach: SpatialVLM Architecture and Objective

## PaLM-E architecture with PaLM 2-S language backbone




*Driess et al., PaLM-E (ICML 2023), Figure 1*

### Input
Image and query of a spatial task

### Output
Answer text

### Objective
The same autoregressive text-generation objective as standard VQA:

$$\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p_\theta(a_t \mid I, Q, a_{<t})$$

Georgia Tech

# Approach: SpatialVLM Training Data

## PaLM-E training data w/ 5% of tokens allocated to their spatial reasoning data

PaLM-E data consists of mostly general vision-language tasks with 8.9% robot data

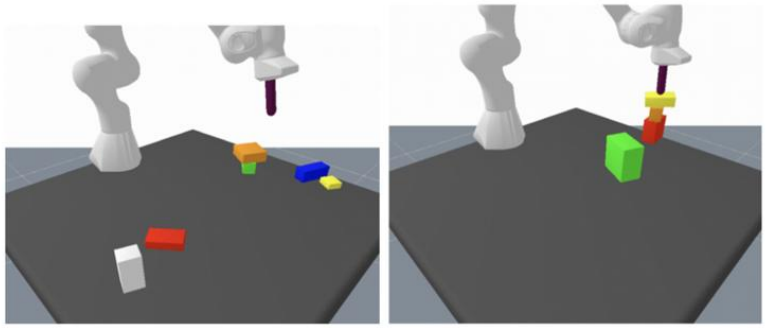


Figure 8: Two TAMP environment test examples. Left with 6 objects (training data contains 3-5 objects), right with 4 objects.

- $q_2$: object-table relation. Example prompt: `Given <img>. Q: Is the red object left, right, or center of the table?`. Target: `A: The red object is in the center of the table.`

*Driess et al., PaLM-E (ICML 2023), Figure 8*

## SpatialVLM scales this, introduces quantitative tasks, and uses real-world images.

Georgia Tech

# Approach: SpatialVLM Training Recipe

| Component | SpatialVLM Configuration |
| --- | --- |
| Initial Training Stage | 110k training steps with a frozen vision encoder |
| Further Training | 70k steps, 2 runs comparing frozen and unfrozen ViT |
| Optimizer | Adam, 2e-4 learning rate, and 512 batch size |

# Approach: Chain-of-Thought Spatial Reasoning

A text-only coordinator LLM uses SpatialVLM for complex, multi-step reasoning by composing simple spatial reasoning questions to solve a more complex problem.

**User**:
Does the blue coke can, the orange can, and the silver seven up can on the table roughly form an isosceles triangle? Roughly means the difference between the longest and shortest triangle edges is smaller than 0.1 m.



**LLM**: What is the distance between the blue coke can and the orange can?
**VLM**: 0.4m
**LLM**: What is the distance between the orange can and the silver seven up can?
**VLM**: 0.48m
**LLM**: What is the distance between the blue coke can and the silver seven up can?
**VLM**: 0.41m
**LLM**: Since the difference between the longest and shortest triangle edges (0.48m - 0.4m = 0.08m) is smaller than 0.1 m, the blue coke can, the orange can, and the silver seven up can roughly form an isosceles triangle.

*Chen et al., SpatialVLM (CVPR 2024), Figure 4*

Georgia Tech

# Experimental Design

## Research Questions

- Does SpatialVLM's VQA data generation and training pipeline improve spatial reasoning capabilities?
- How does the noisy synthetic spatial VQA data and different training strategies affect the learning performance
- Does the VLM equipped with "direct" spatial reasoning capabilities unlock new capabilities such as CoT reasoning and embodied planning?

## Experimental Setup

- Trained SpatialVLM with PaLM-E training set + spatial VQA dataset

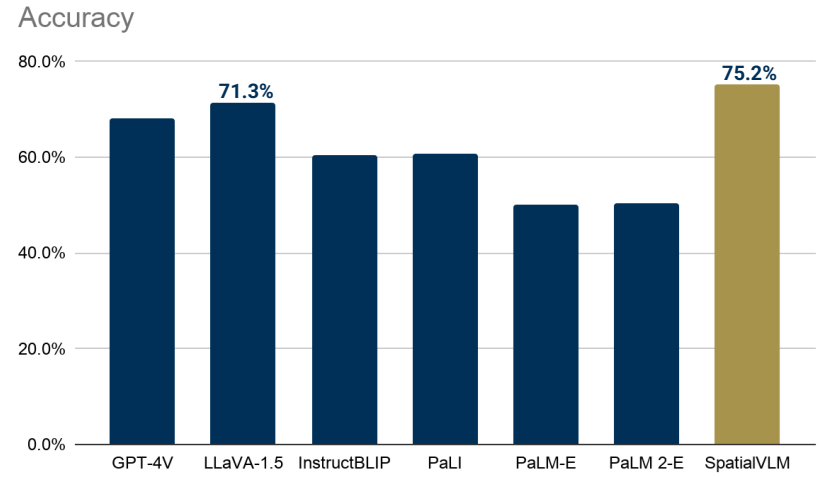
## Baselines

- **GPT-4V** [1]: supports multimodal input, achieves SOTA performance in VL tasks.
- **LLaVA-1.5** [2]
- **InstructBLIP** [3]
- **PaLI** [4]: trained on multi-lingual corpora, shows SOTA performance on captioning and VQA.
- **PaLM-E** [5]: trained on internet-scale vision, language, and robotics data, shows SOTA performance in OKVQA benchmark.
- **PaLM 2-E** [5]: an updated version of PaLM-E, shared network architecture and training procedure with SpatialVLM.

[1] OpenAI, 2023 [2] Liu, H., et al., 2023 [3] Dai, W., et al., 2023 [4] Chen, X., et al., 2022 [5] Driess, D., 2023

# Qualitative Spatial VQA Performance

- Evaluated with human annotated 331 spatial reasoning VQA pairs
- SpatialVLM achieves 75.2% accuracy, outperforms other baselines
- The 2nd best model is LLaVA-1.5 (71.3% accuracy), which leverages bounding boxes and captions in visual instruction tuning
- LLaVA-1.5 performs well in 2D spatial relationship inference, but not that good in 3D reasoning






Q: Is the fireplace screen with red doors smaller than the dog standing on the floor in width?
A: no



Q: Compared to the little boy in a red shirt, which side is the man wearing a blue shirt on?
A: left



Q: Are the windows positioned to the left of the black television?
A: yes



Q: Is the palm tree in distance taller than the parked white car?
A: yes

Georgia Tech

# Quantitative Spatial VQA Performance

- Evaluated with 215 quantitative spatial VQA pairs
- Metrics
  - **Output number rate**
    - The percentages of the model output the valid format numbers
    - Evaluates whether VLM can understand the quantitative spatial reasoning question
  - **In range accuracy**
    - The percentages of the VLM answers that fall into 50% to 200% of the ground truth value
    - Represents how accurate the VLM's quantitative estimates are



Q: Could you provide the distance between the sign and the motorcyclist?
A: about 0.5 meter



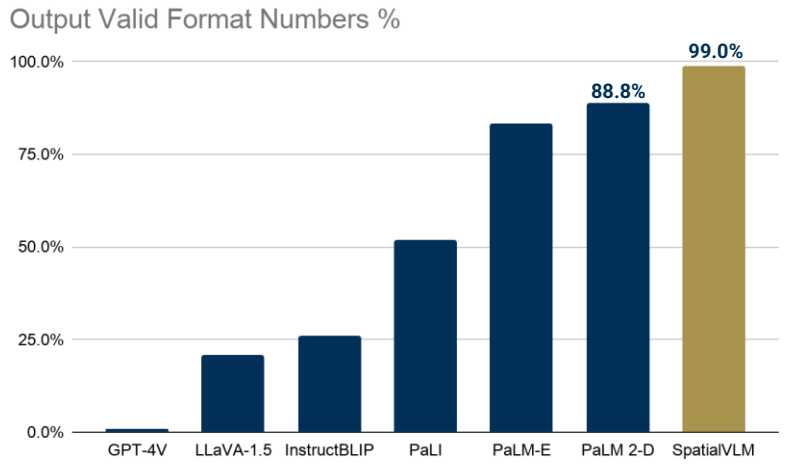
Q: Determine the distance of the fence from the giraffes in a zoo relative to the camera.
A: about 5 meters

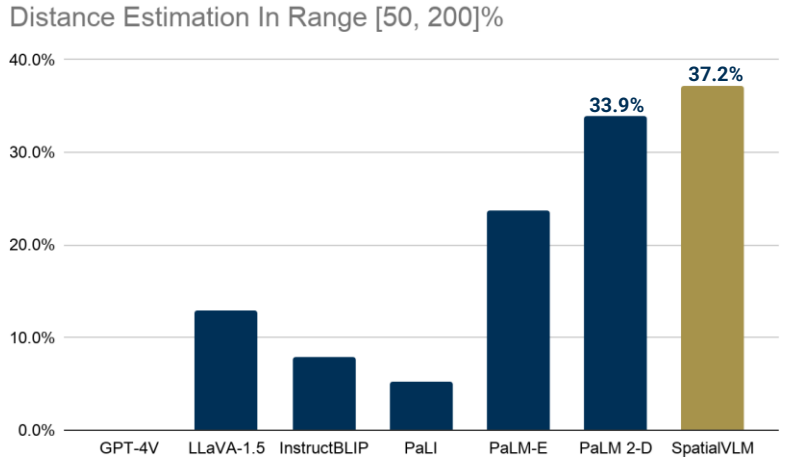


Q: How far is the striped tie towards the left from the black cell phone?
A: about 0.2 meter



Q: How much distance is there between the sand and the people that are standing on the beach?
A: 0, as the people are directly standing on the sand

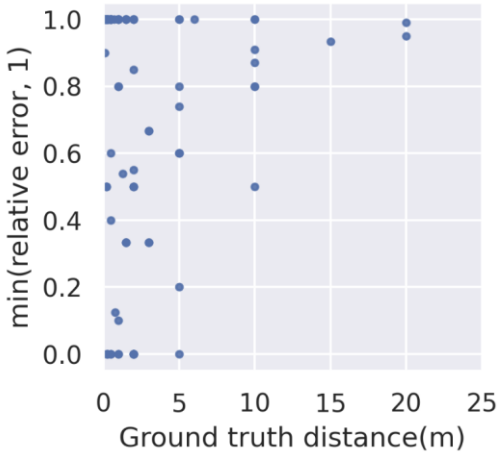


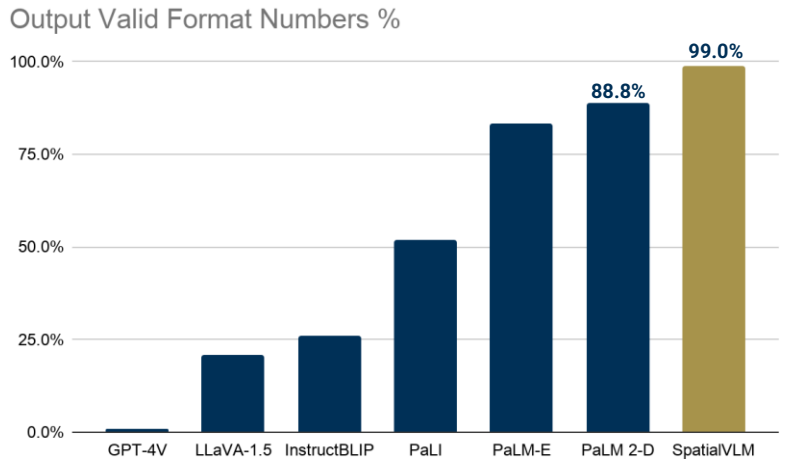

Georgia Tech

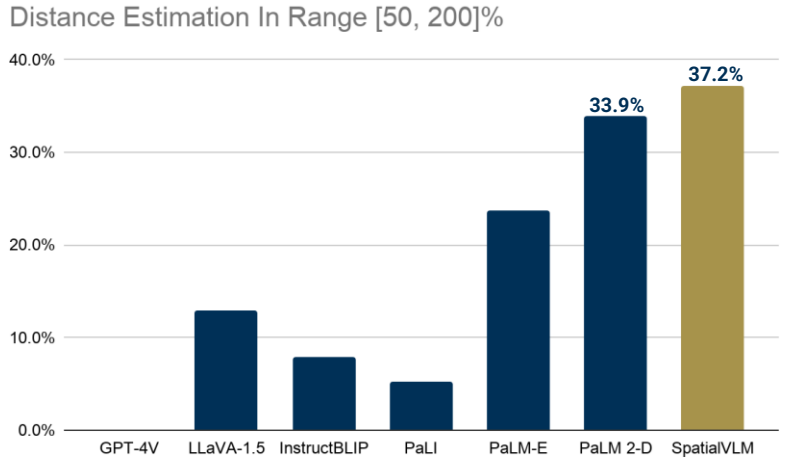
# Quantitative Spatial VQA Performance

- Baseline VLMs are reluctant to give answers with metric values
  - e.g., GPT-4v answered “I cannot provide exact distance as the image does not offer precise references for measurement”
- SpatialVLM is able to output valid format 99% of the time




- The plot of the distance estimation relative error w.r.t. ground truth distance shows that there are generally larger errors for bigger distance
- SpatialVLM achieves well estimation performance with objects 1-10 meters from the camera. (same as the range where ZoeDepth reliably outputs metric depth estimates



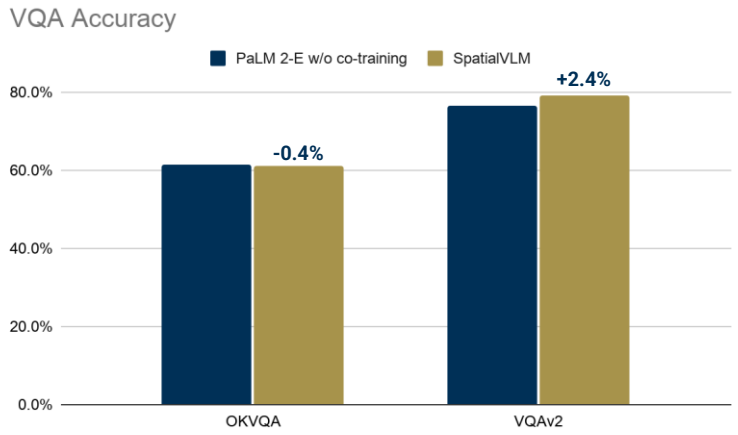




Georgia Tech

# Ablation Study

## Effect of Spatial VQA Data to General VQA

Will the VLM performance degrade since co-train with large spatial VQA data?

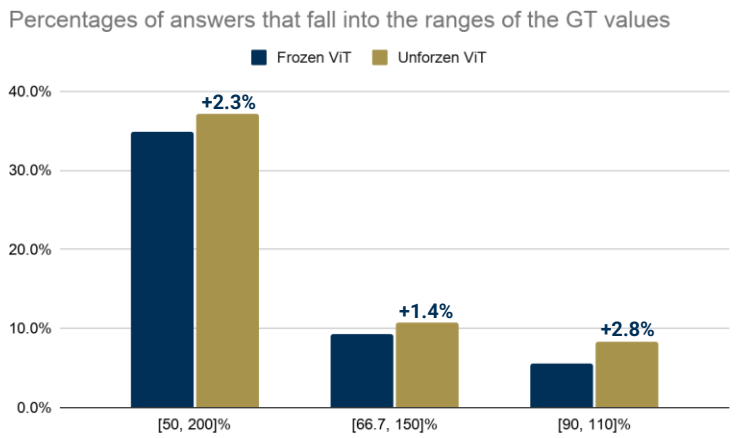



Spatial co-training preserves similar performance on **OKVQA** (limited spatial reasoning questions), while improved the performance on **VQA v2** benchmark.

VLMs are underfitting in spatial reasoning tasks, and can benefit from spatial VQA supervisions w/o hurting general VQA capabilities.

## Effect of ViT Encoder in Spatial Reasoning

Does the frozen ViT encode enough information to perform spatial reasoning?




Setup: Run branch from the 110k training step and train for another 70k steps.

Unfrozen ViT improves metric estimation, but precise predictions remain challenging.

## Effect of Noisy Quantitative Spatial Answers

Can VLMs learn generalizable quantitative estimations from noisy training data?

| Gaussian Noise | 0 | 0.1 | 0.2 | 0.3 |
|---|---|---|---|---|
| MSE [m] | 0.046 | 0.053 | 0.039 | 0.048 |
| In range accuracy [50, 200]% | 59.0% | 55.8% | 61.1% | 61.1% |

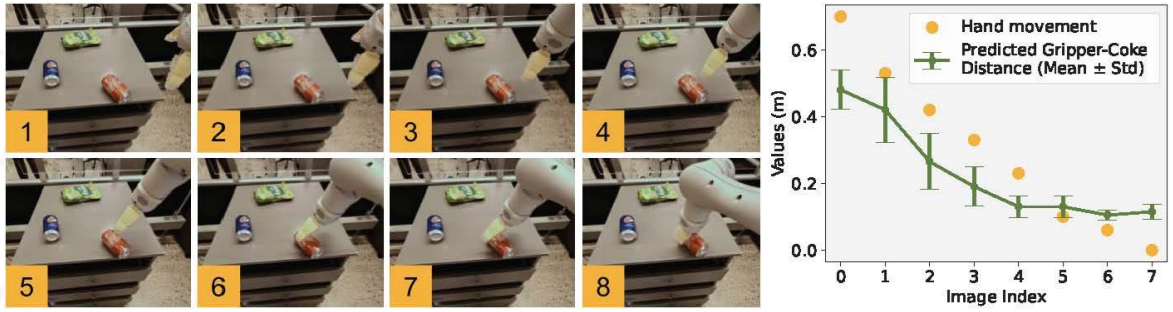
Performance remains similar across the tested random noise levels.

The model can learn useful estimates despite random perturbations in data.

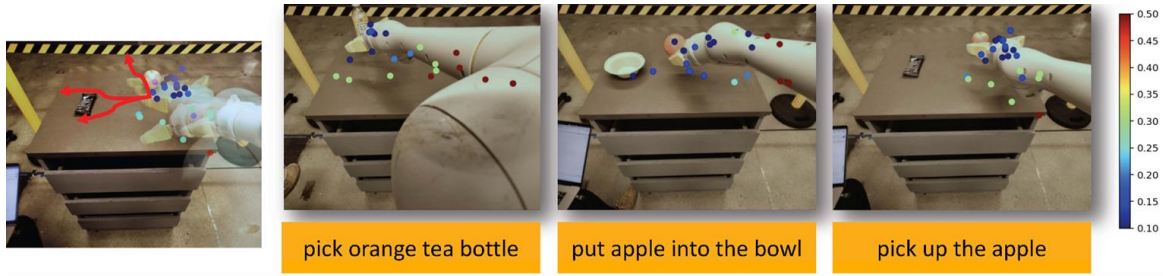
Georgia Tech

# Applications

## Robotics




SpatialVLM successfully predicted distance decreases as the gripper approaches the target. This can be used as a reward signal for reinforcement learning.




SpatialVLM can be prompted to annotate dense rewards for open-vocabulary robotic tasks.

## Multi-Step Spatial Reasoning

**User**:
Does the blue coke can, the orange can, and the silver seven up can on the table roughly form an isosceles triangle? Roughly means the difference between the longest and shortest triangle edges is smaller than 0.1 m.



**LLM**: What is the distance between the blue coke can and the orange can?
**VLM**: 0.4m
**LLM**: What is the distance between the orange can and the silver seven up can?
**VLM**: 0.48m
**LLM**: What is the distance between the blue coke can and the silver seven up can?
**VLM**: 0.41m
**LLM**: Since the difference between the longest and shortest triangle edges (0.48m - 0.4m = 0.08m) is smaller than 0.1 m, the blue coke can, the orange can, and the silver seven up can roughly form an isosceles triangle.

Use an external LLM (e.g., GPT-4) to decompose a complex question into spatial queries and combines SpatialVLM's spatial reasoning ability allow multi-step spatial reasoning.

Georgia Tech

# SpatialVLM Discussion

## Strengths

Data synthesis pipeline enables internet-scale data for spatial reasoning grounded in real-world images

## Limitations

Uncertainty/Error Propagation

- Data synthesis pipeline chains together multiple off-the-shelf vision models, so the pipeline inherits each of their limitations, and errors propagate through the chain

Depth Scale Ambiguity

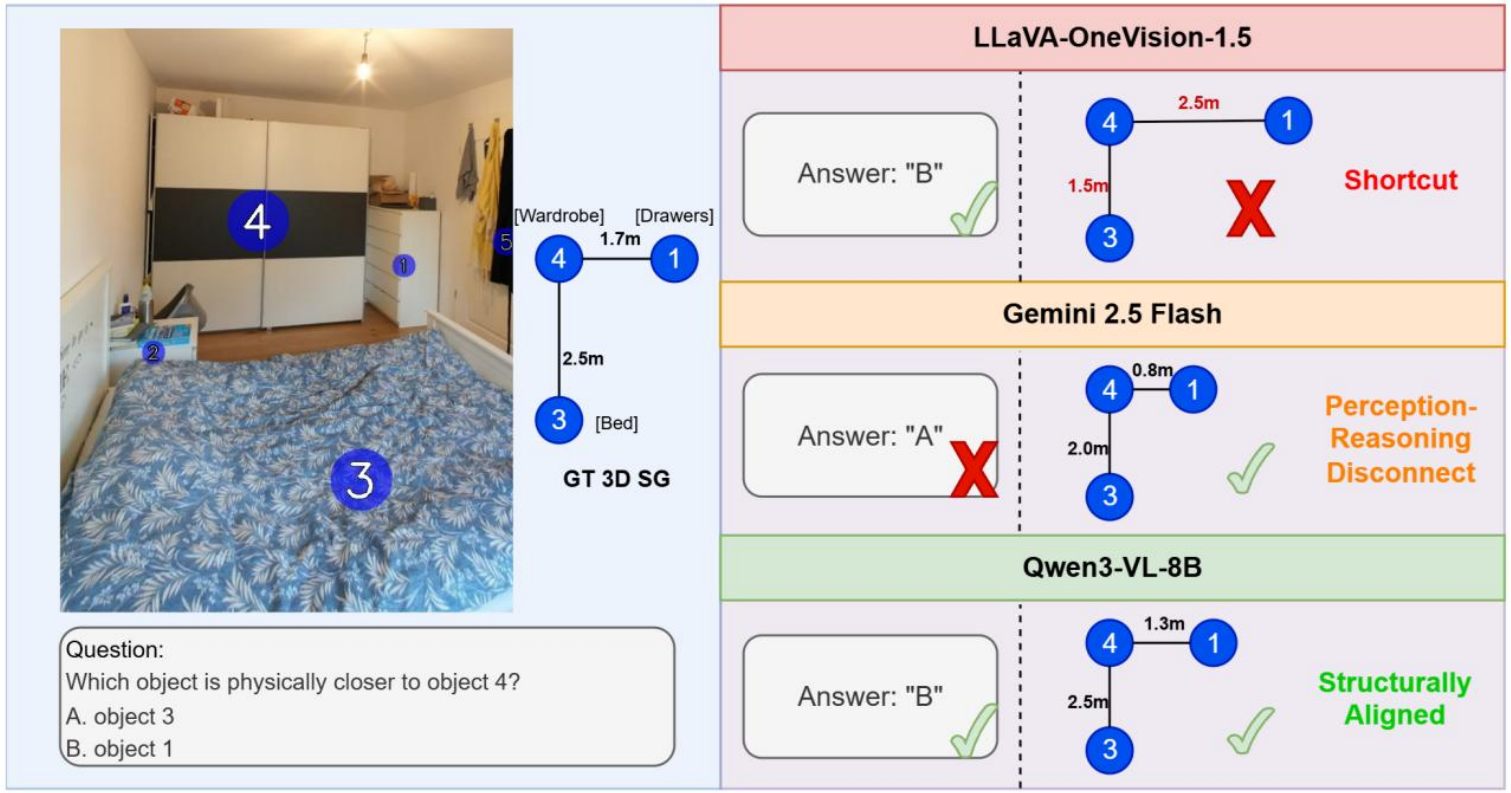
- A 2D image can correspond to multiple 3D scenes at different scales, so depth estimation models rely on learned assumptions. If those assumptions are wrong, the spatial reasoning dataset inherits those biases
- Accuracy depends on the estimator's reliable range

Can't do complex reasoning on its own

Georgia Tech

# Introduction to CRISP:
## Discussion Question

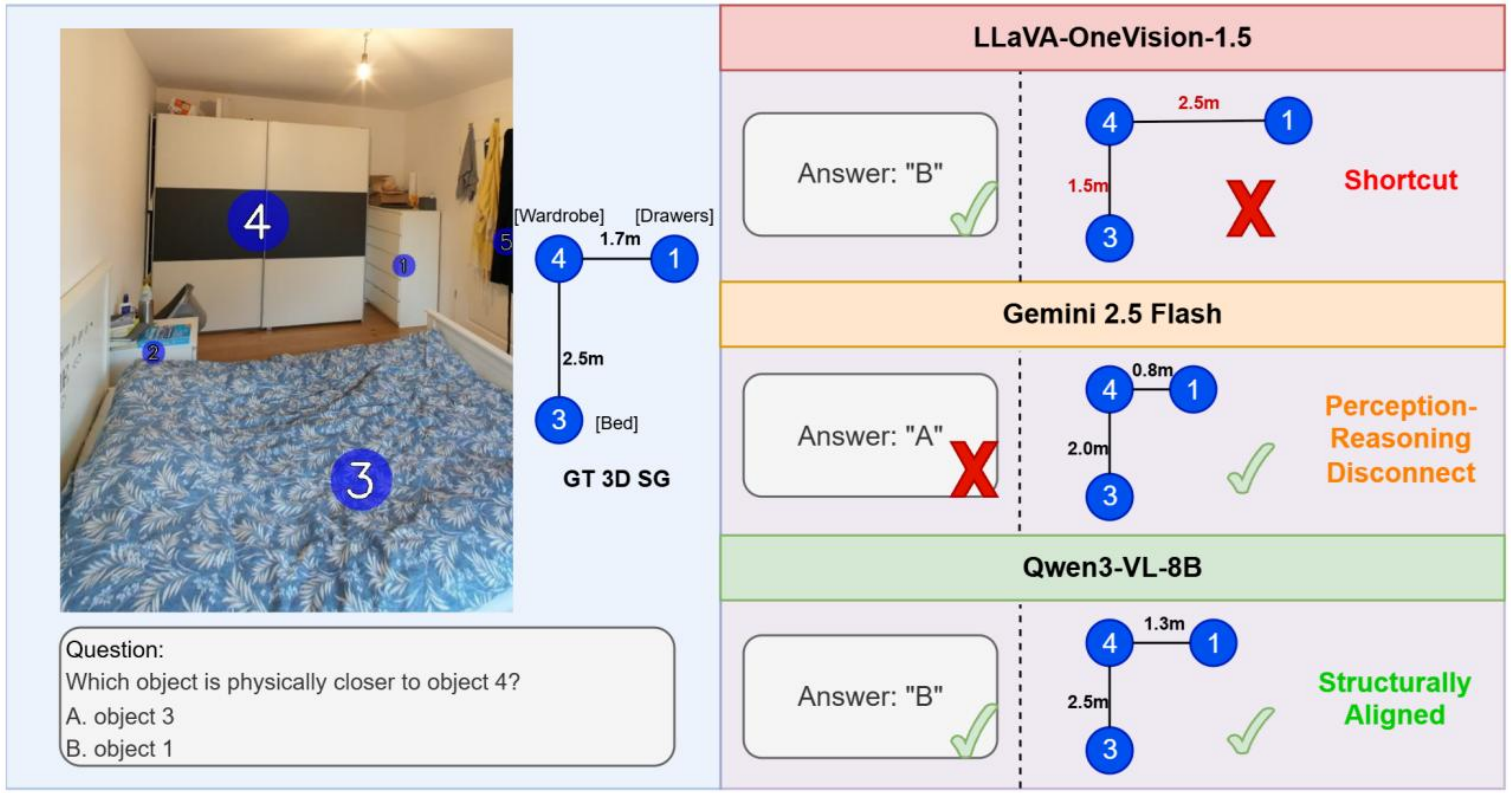
**What are the major failure modes associated with 3D spatial reasoning?**



# Introduction to CRISP:
## Discussion Question

**What are the major failure modes associated with 3D spatial reasoning?**

1. **Perception (atomic perception):** Failure to perceive objects and other scene information correctly
2. **Reasoning (structural spatial perception):** Failure to correctly reason about spatial relationships/measurements
3. **Translation (perception-reasoning disconnect):** Failure to translate reasoning into an answer (that is logically consistent with reasoning)




**Failure Examples:**

1. **Perception:** *"Object #3 is a carpet instead of a bed."* - For the dataset, objects are labeled with a numerical ID in the input image and referred to by this ID in questions. This is to bypass any object detection failures in order to isolate spatial reasoning failures.
2. **Reasoning:** *"The wardrobe is 2.5 meters from the drawers."* - They are visibly closer than 2.5 meters. LLaVA-OneVision-1.5 perceives spatial relationships incorrectly.
3. **Translation:** *"Object 3, the bed, is closer to object 4, the wardrobe."* - Despite having correct internal 3D reasoning, Gemini 2.5 Flash answers the question incorrectly.

Georgia Tech

# CRISP Overview

**CRISP** = **C**onsistency of **R**easoning **I**n **S**patial **P**erception

- new evaluation benchmark
- 3D reasoning from single view image QA

**Traditional QA Benchmarks**

*result-oriented* correctness



**CRISP Benchmark**

*process-oriented* grounding

Georgia Tech

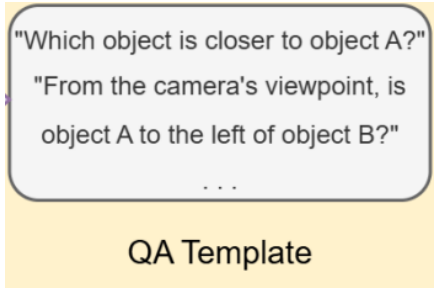
# CRISP Overview

**Problem:** Grounding in geometric reasoning is a **fundamental prerequisite** to answering spatial questions but models do not always ground their answers in their reasoning.

| Problems addressed by CRISP | CRISP Solutions |
|---|---|
| Internal spatial reasoning capabilities are not made explicit in previous 3D QA benchmarks | Evaluate on explicit representations of spatial reasoning: structured 3D Scene Graph **(3D GS) outputs** |
| Implicit spatial reasoning and language reasoning may not be consistent, obfuscating failure modes | **Consistency** between 3D spatial reasoning generation and direct QA answer evaluated |
| Confounding factors confused other QA benchmarks: the number of problems solved through text-only reasoning, object detection failures | Evaluate on text-only input in addition to image and text input. Use numerical IDs. |

Georgia Tech

# CRISP Approach: Tasks



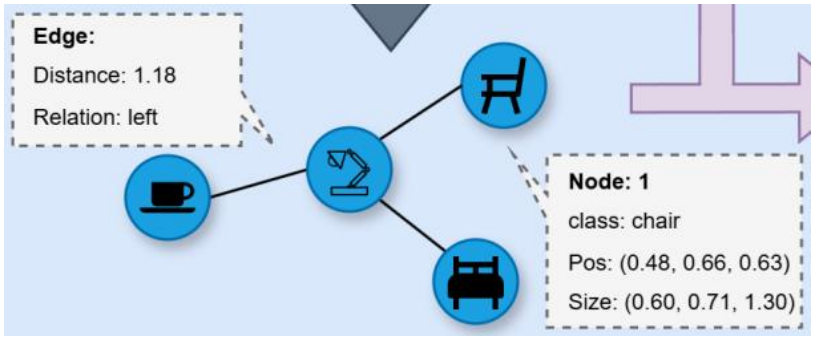

**Spatial QA Task:**

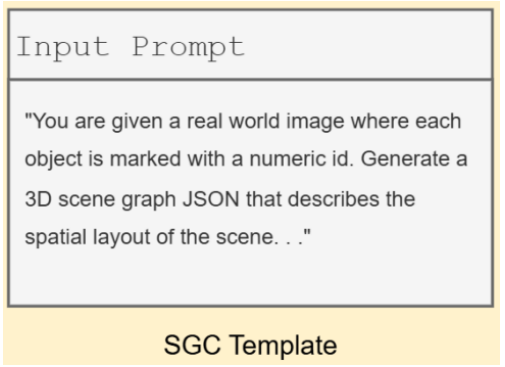
**Goal:** Evaluate spatial perception and reasoning through image-grounded questions.

- **Input:** Image + predefined objects of interest + spatial question
  - spatial questions are in the following forms:
    - MCQs = Multiple-Choice Questions
    - NAQs = Numerical Answer Questions
- **Output:** Multiple-choice or numerical answer.

**Evaluation:**
- MCQs: Accuracy (exact matches)
- NAQs: Mean Relative Accuracy (MRA) across numerical error thresholds.

$$MRA = \frac{1}{10}\sum_{\theta \in \mathcal{C}} \mathbb{1}\left(\frac{|\hat{y} - y|}{y} < 1 - \theta\right)$$




**SGC Task:** (SGC = 3D Scene Graph Construction)

**Goal:** Evaluate geometric precision and semantic correctness of structured 3D scene understanding outputs.

- **Input:** Image + predefined objects of interest
- **Output:** Object-centric, star-topology scene graph:
  - Nodes: 3D dimensions and camera-to-object distances.
  - Edges: Inter-object distances and semantic directions.

**Evaluation:**
- Metric Estimation Score = object size score + distance accuracy score
- Relation Score = directional relationship correctness

$$S_{size} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{3}\sum_{k\in\{w,l,h\}}\frac{\min(k_p, k_g)}{\max(k_p, k_g)}$$ ıatio $$S_{rel} = \frac{1}{N_{edge}}\sum_{i=1}^{N_{edge}}\frac{N_{\text{Correct Pairs}}}{N_{\text{Total Pairs}}}$$ ıre

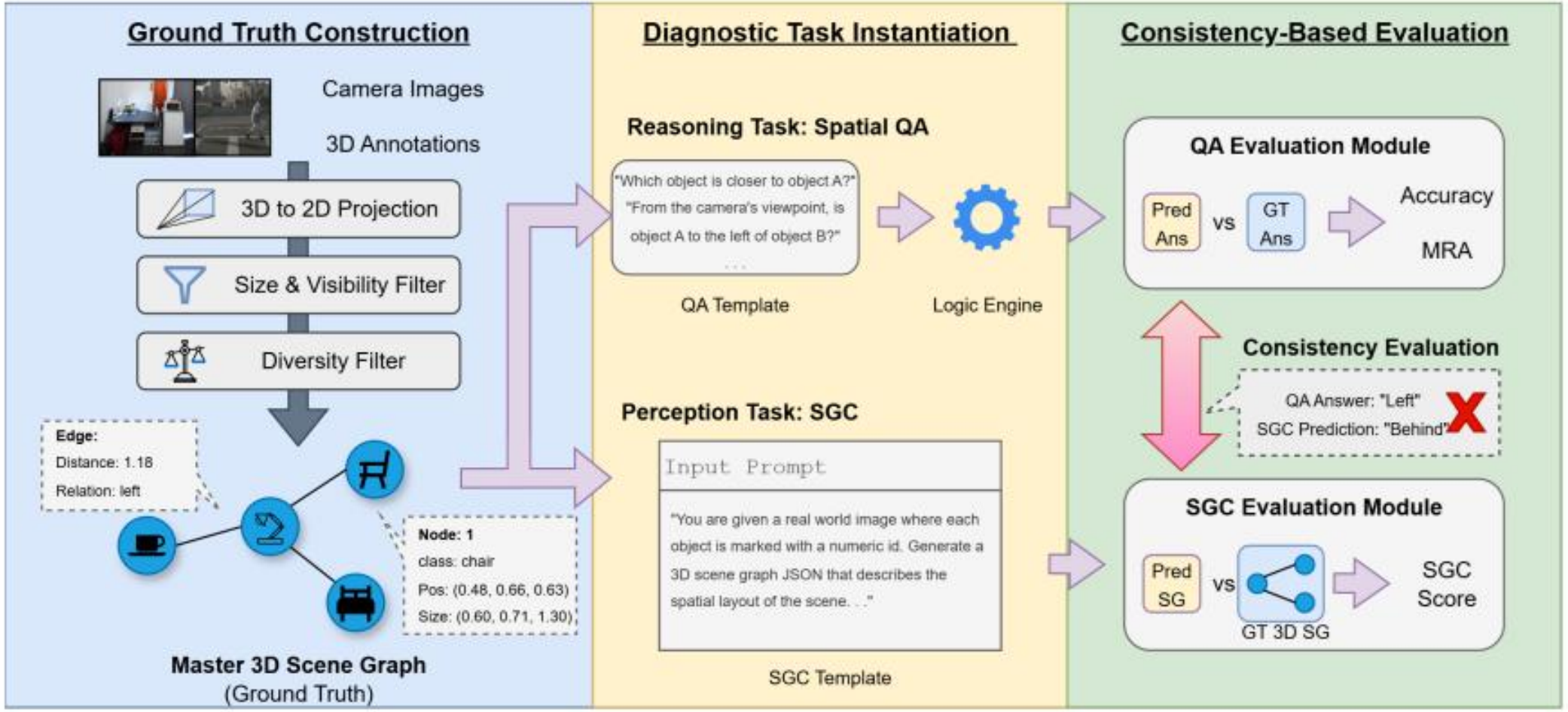
$$S_{dist_to_cam}, S_{dist} = \frac{1}{N}\sum_{i=1}^{N}\max\left(0, 1 - \frac{|d_p - d_g|}{d_g + \epsilon}\right)$$

Georgia Tech

# CRISP Approach: Spatial QA and SGC Tasks

**Task Construction and Examples:**

- both tasks are derived from the same set of **Master 3D Scene Graphs**



# CRISP Approach: Spatial QA Task

**Spatial QA Task:**

Goal: Evaluate spatial perception and reasoning through image-grounded questions.

**7 Competencies for embodied agents:** counting, logical deduction, direction, distance, ranking, size, and view transformation

- **Input:** Image + spatial question
- **Output:** Multiple-choice or numerical answer.

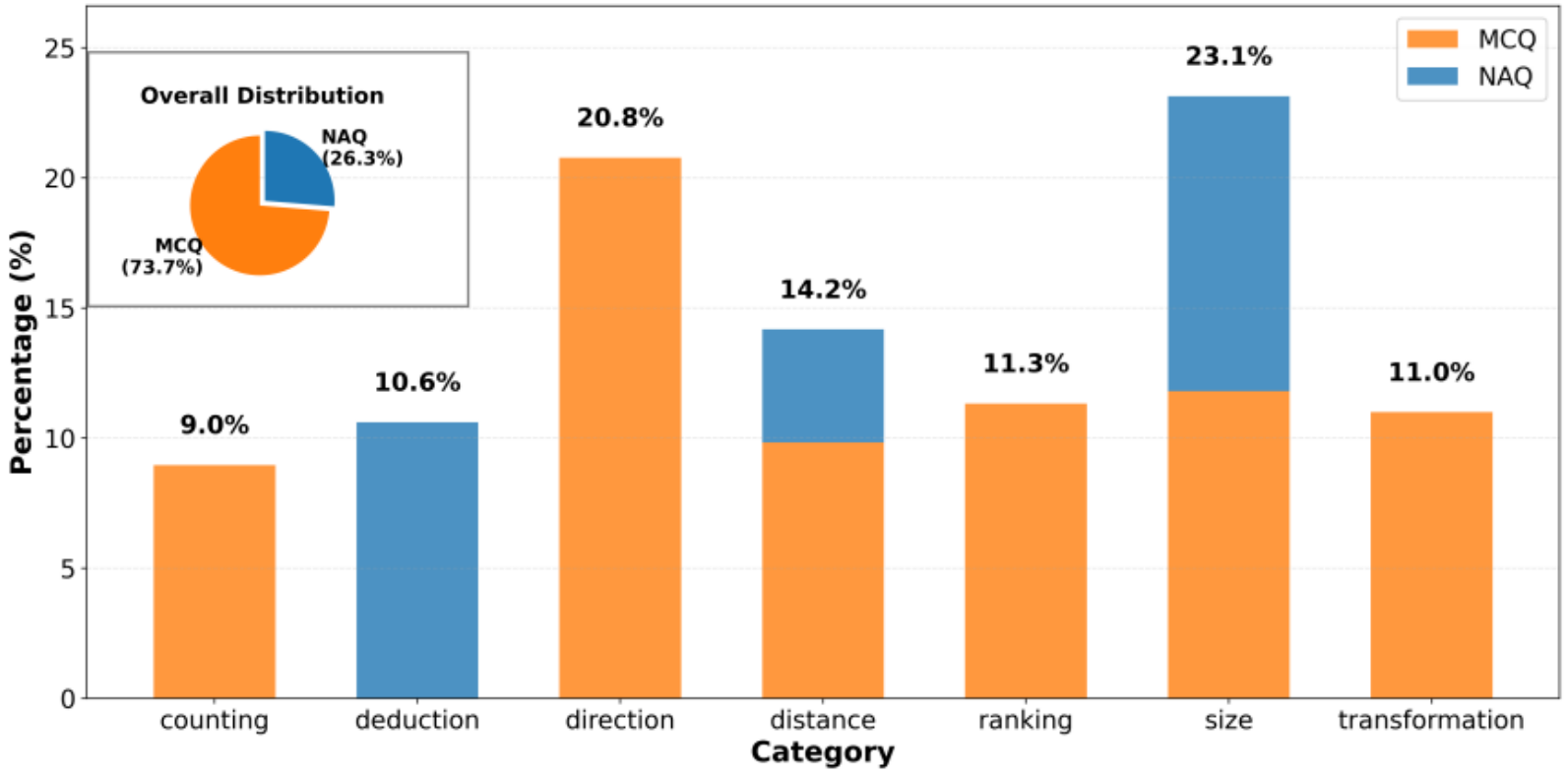



**Table 1: Spatial QA Prompt Templates.**

| Category | Type | Example template |
|---|---|---|
| Directional Perception | MCQ | "From the camera's viewpoint, is <OBJ_A> in front of <OBJ_B>?"<br>"From the camera's viewpoint, which best describes the relative position of <OBJ_A> to <OBJ_B>?" |
| Size Estimation | MCQ | "Considering real-world physical dimensions, which object is larger in volume: <OBJ_A> or <OBJ_B>?"<br>"Which object is taller: <OBJ_A> or <OBJ_B>?" |
| | NAQ | "Approximately how long (in meters) is <OBJ_A>?" |
| Distance Estimation | MCQ | "Which object is physically closer to the camera: <OBJ_A> or <OBJ_B>?"<br>"Which object is physically closer to <OBJ_A>: <OBJ_B> or <OBJ_C>?" |
| | NAQ | "What is the distance (in meters) between <OBJ_A> and <OBJ_B>?" |
| Counting | MCQ | "From the camera's viewpoint and only considering the marked objects, how many objects are <REL_PHRASE> <OBJ_A>?"<br>"Considering the marked objects, how many objects are within <DIST> meters of <OBJ_A>?" |
| Spatial Ranking | MCQ | "Rank the following objects by distance from the camera (nearest to farthest): <OBJ_A>, <OBJ_B>, <OBJ_C>."<br>"Rank the following objects by height (tallest to shortest): <OBJ_A>, <OBJ_B>, <OBJ_C>." |
| Logical Deduction | NAQ | "Suppose we place a new object A in the <DIR> direction from <OBJ_A> at a distance of approximately <DIST> meters. What is the horizontal distance (in meters) between A and <OBJ_B>?" |
| View Transformation | MCQ | "Imagine you are standing at the location of <OBJ_A> and facing directly towards <OBJ_B>. From this perspective, in which direction is <OBJ_C> located relative to you?" |

# CRISP Approach: Spatial QA Task

**Spatial QA Task:**

Goal: Evaluate spatial perception and reasoning through image-grounded questions.

**7 Competencies for embodied agents:** counting, logical deduction, direction, distance, ranking, size, and view transformation

- **Input:** Image + spatial question
- **Output:** Multiple-choice or numerical answer.

## Spatial Question Examples:

| | | |
|---|---|---|
| View Transformation | MCQ | "Imagine you are standing at the location of <OBJ_A> and facing directly towards <OBJ_B>. From this perspective, in which direction is <OBJ_C> located relative to you?" |
| Logical Deduction | NAQ | "Suppose we place a new object A in the <DIR> direction from <OBJ_A> at a distance of approximately <DIST> meters. What is the horizontal distance (in meters) between A and <OBJ_B>?" |

Georgia Tech

# CRISP Approach: SGC Task

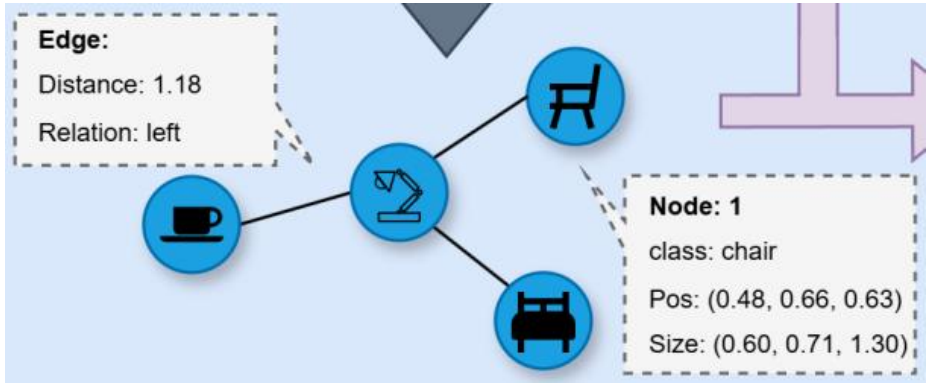
**SGC Task:** (SCG = 3D Scene Graph Construction)

**Goal:** Evaluate geometric precision and semantic correctnes structured 3D scene understanding outputs.

- **Input:** Image + predefined objects of interest

> You are given a real world image where each object is marked with a numeric id (Object 1, Object 2, ...).
> **Task:**
> Generate a 3D scene graph JSON that describes the spatial layout of the scene.

- **Output:** Object-centric, star-topology scene graph:
  - Nodes: 3D dimensions and camera-to-object distances.
  - Edges: Inter-object distances and semantic directions.




**Prompt for 3D Scene Graph Construction**

You are given a real world image where each object is marked with a numeric id (Object 1, Object 2, ...).
**Task:**
Generate a 3D scene graph JSON that describes the spatial layout of the scene.
**Requirements:**
1. Create exactly one entry in "objects" for **every** id listed in "Scene Objects".
2. Create edges from the **Center Object** to **all other** objects.
3. Analyze the spatial relationship from the center object to each target object.
4. For "relation", list **ALL** applicable terms (e.g., ["left", "in front of"]).
5. For "relation", use ONLY these words: ["left", "right", "in front of", "behind", "up", "down"].
**Output Format:**

```json
{
  "objects": [
    {
      "id": int,
      "dist_to_cam": float,
      "size": {
        "w": float,
        "l": float,
        "h": float
      }
    }
  ],
  "edges": [
    {
      "from": int,
      "to": int,
      "distance": float,
      "relation": ["string"]
    }
  ]
}
```

**Input Data:**
Center Object id: `{center_id}`
Scene Objects (ids shown in the image): `{object_list_str}`
**Response:**
Output the valid JSON only, without any extra text.

Georgia Tech

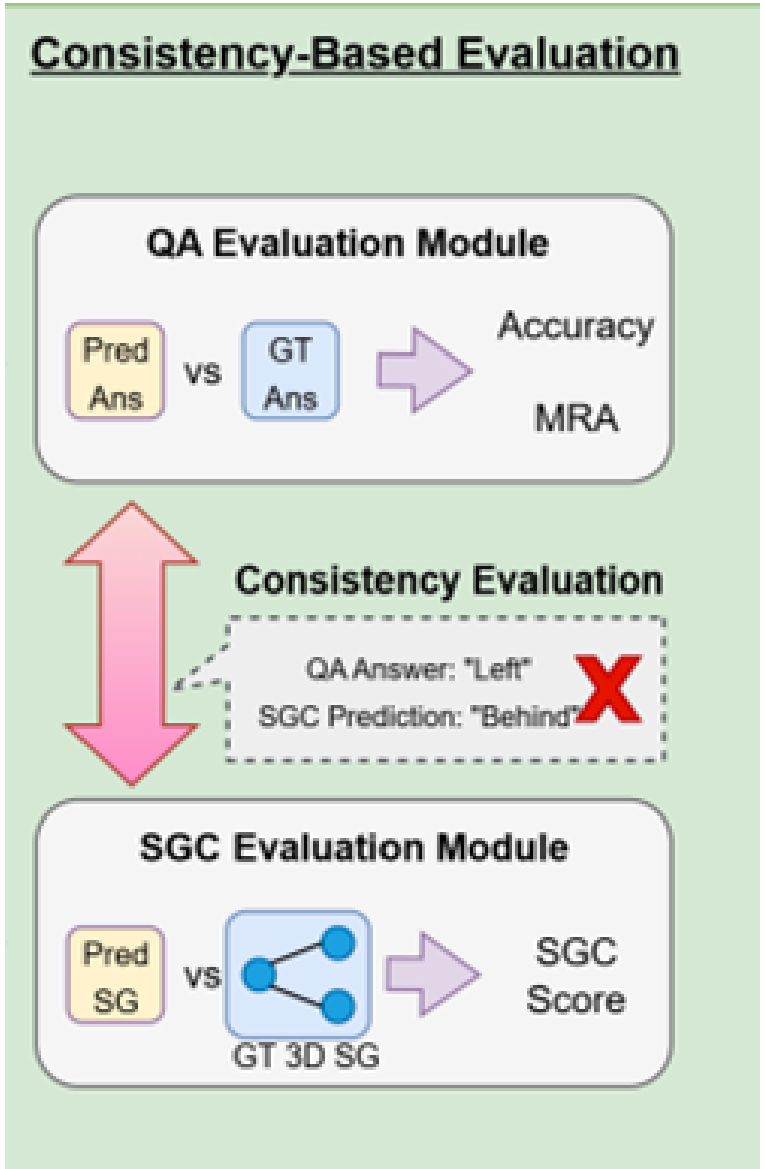
# CRISP Approach: Consistency Evaluation

**Evaluating Self-Consistency between QA Task and SGC Task:**

- only measures logical consistency, not correctness in answering questions

**Steps:** (for each spatial question)

1. Parse 3D SG to infer QA answer deterministically
   - graph can fail to support inference → [FAILED] token
   - **output:** derived answer, $A_{derived}$
2. Grab direct QA response from model
   - **output:** direct QA response, $A_{QA}$
3. Compare SGC derived answer with direct QA response



# CRISP Approach

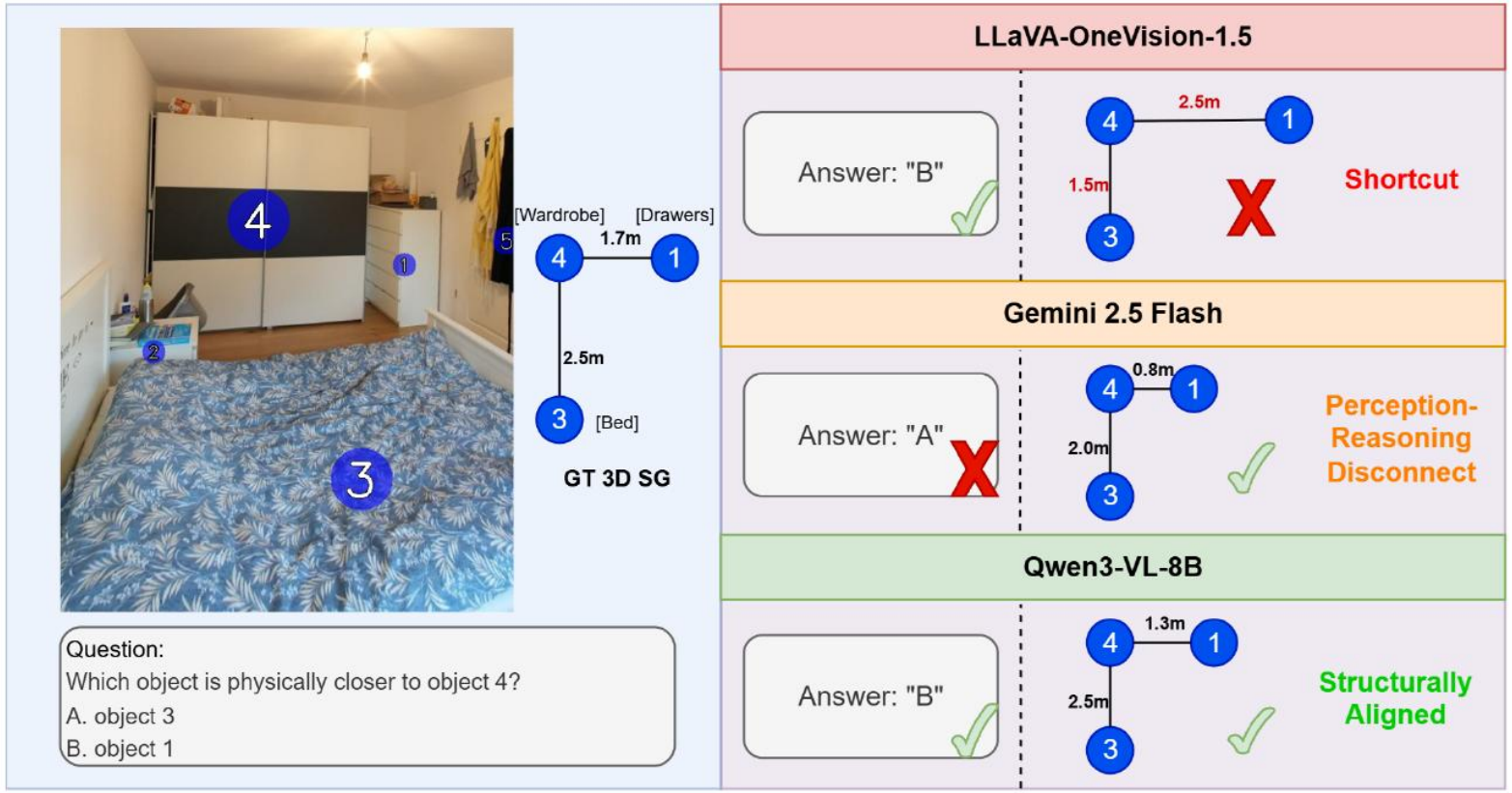
**CRISP Diagnosis:** Using a combination of these QA/SGC/consistency scores, CRISP claims to be able to identify several key problems.

## Problems:

- Semantic Shortcut
  - high QA score
  - low SGC score
  - low consistency score
- Perception-Reasoning Disconnect
  - low QA score
  - high SGC score
  - low consistency score

## Goal: Generalizable Models with

- high QA score
- high SGC score
- **high consistency score**




Georgia Tech

# CRISP Experiments

| Model | QA | $S_{est}$ | $S_{rel}$ | SGC | Con. |
|---|---|---|---|---|---|
| ***Proprietary Models*** | | | | | |
| Gemini-2.5-Flash | 47.68 | 61.24 | 68.61 | 64.92 | 49.98 |
| Gemini-2.5-Pro† | **58.93** | 57.79 | 71.36 | 64.58 | **57.28** |
| Gemini-3-Flash | 53.41 | 65.25 | **72.33** | **68.79** | 55.20 |
| GPT-5-Mini† | 54.15 | 61.39 | 72.01 | 66.70 | 56.75 |
| GPT-5.2 | 48.75 | **66.67** | 64.61 | 65.64 | 48.87 |
| ***Open-source Models*** | | | | | |
| LLaVA-OV-1.5-8B | 47.85 | 44.53 | 56.12 | 50.33 | 38.98 |
| InternVL3.5-8B | 48.37 | 51.32 | 61.10 | 56.21 | 43.59 |
| InternVL3.5-38B | 53.98 | 53.12 | 64.68 | 58.90 | 48.70 |
| Qwen2.5-VL-7B | 44.78 | 40.62 | 44.79 | 42.70 | 35.62 |
| Qwen3-VL-8B | **55.34** | 57.09 | **67.62** | 62.35 | **52.16** |
| Qwen3-VL-32B | 53.98 | **60.21** | 64.69 | **62.45** | 51.56 |
| ***Specialized Baselines*** | | | | | |
| Cambrian-S | **45.95** | **46.76** | 47.50 | 47.13 | 34.22 |
| VG LLM | 38.34 | 40.61 | **53.85** | **47.23** | **35.29** |

†: Thinking mode restricted.

CRISP evaluates 13 VLMs on 9,839 questions from 1,162 scenes. All models use zero-shot prompts and greedy decoding.

QA (Question Answering) scores whether the model answers spatial questions correctly.

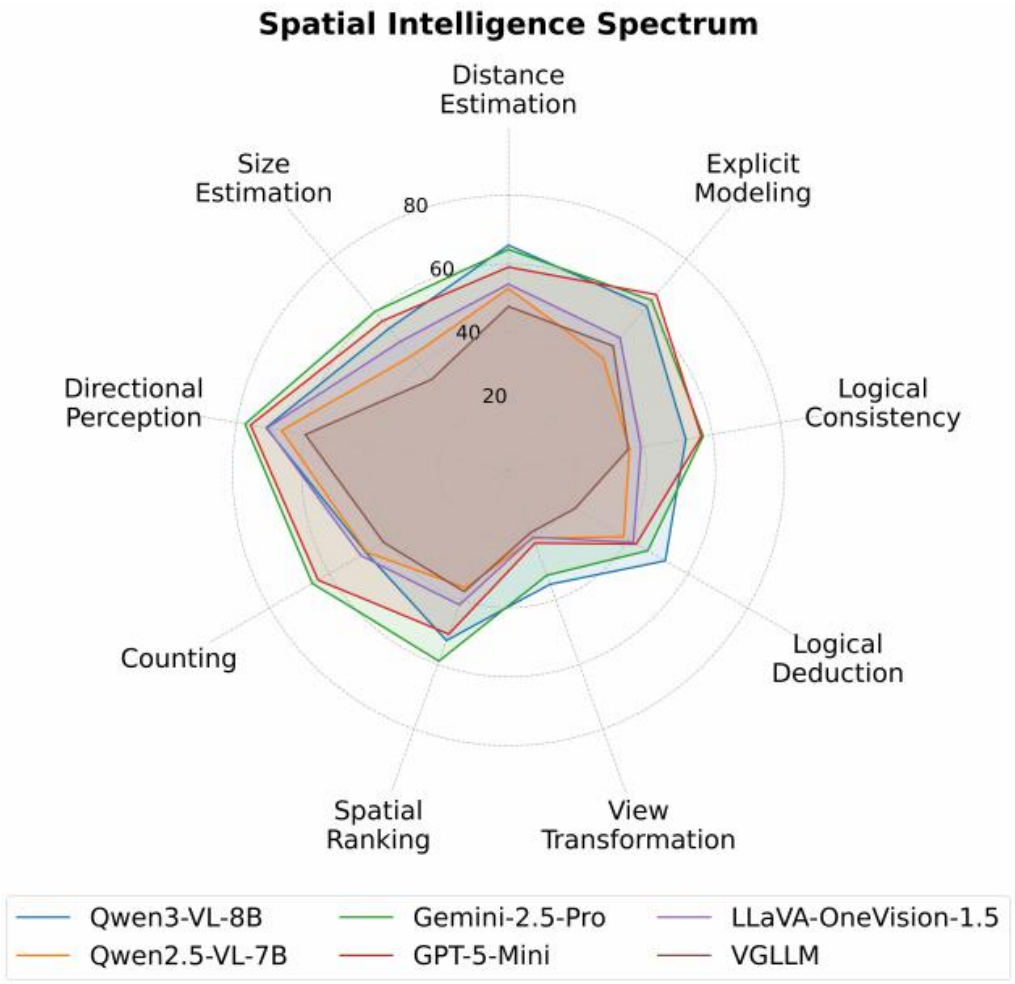
SGC (Scene Graph Construction) scores how accurately its generated graph describes object sizes, distances and spatial relations.

Consistency (Con.) checks whether the direct answer matches what the graph implies. Both can still be wrong.

Cambrian-S (34.22) and VGLLM (35.29) have the lowest consistency scores.

Georgia Tech

# Results by Spatial Task



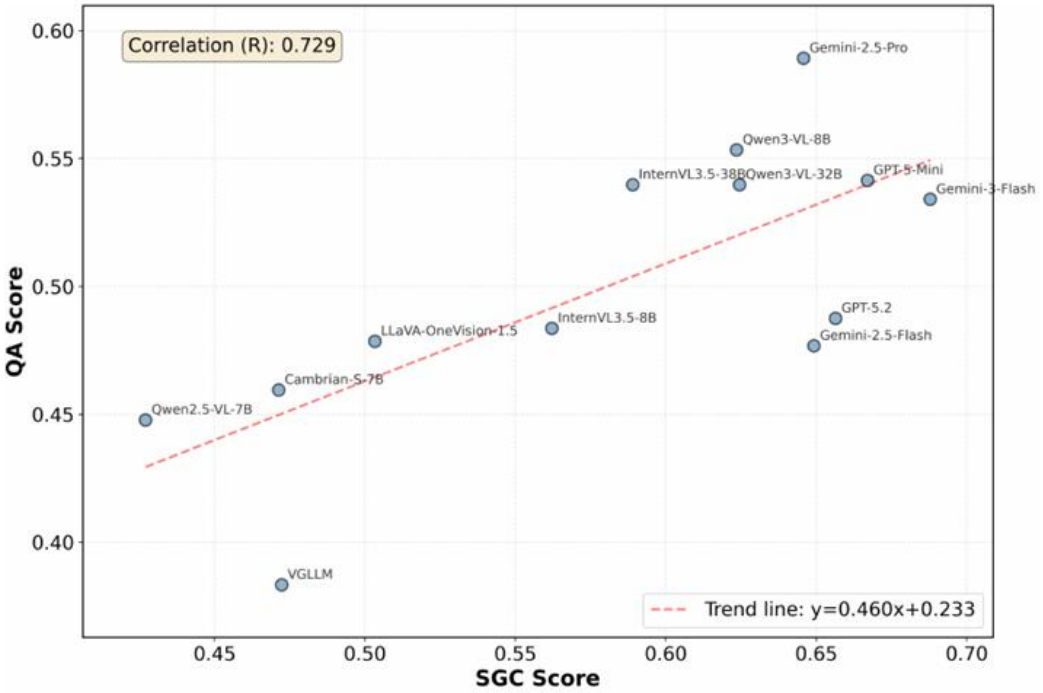

The radar compares six models across nine dimensions of spatial ability, including geometry, question answering and consistency.

Across the six models shown, view transformation receives lower scores than counting and direction.

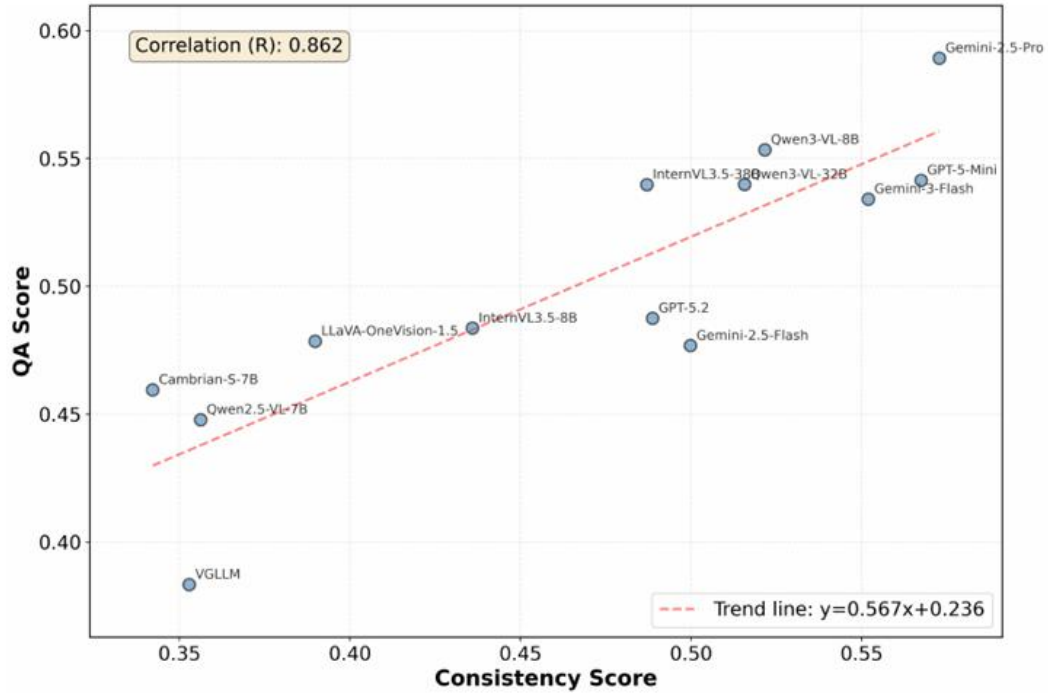
These questions require expressing spatial relationships from a different viewpoint.

Georgia Tech

# QA and Scene Graphs




(a) Perception-Reasoning Alignment.

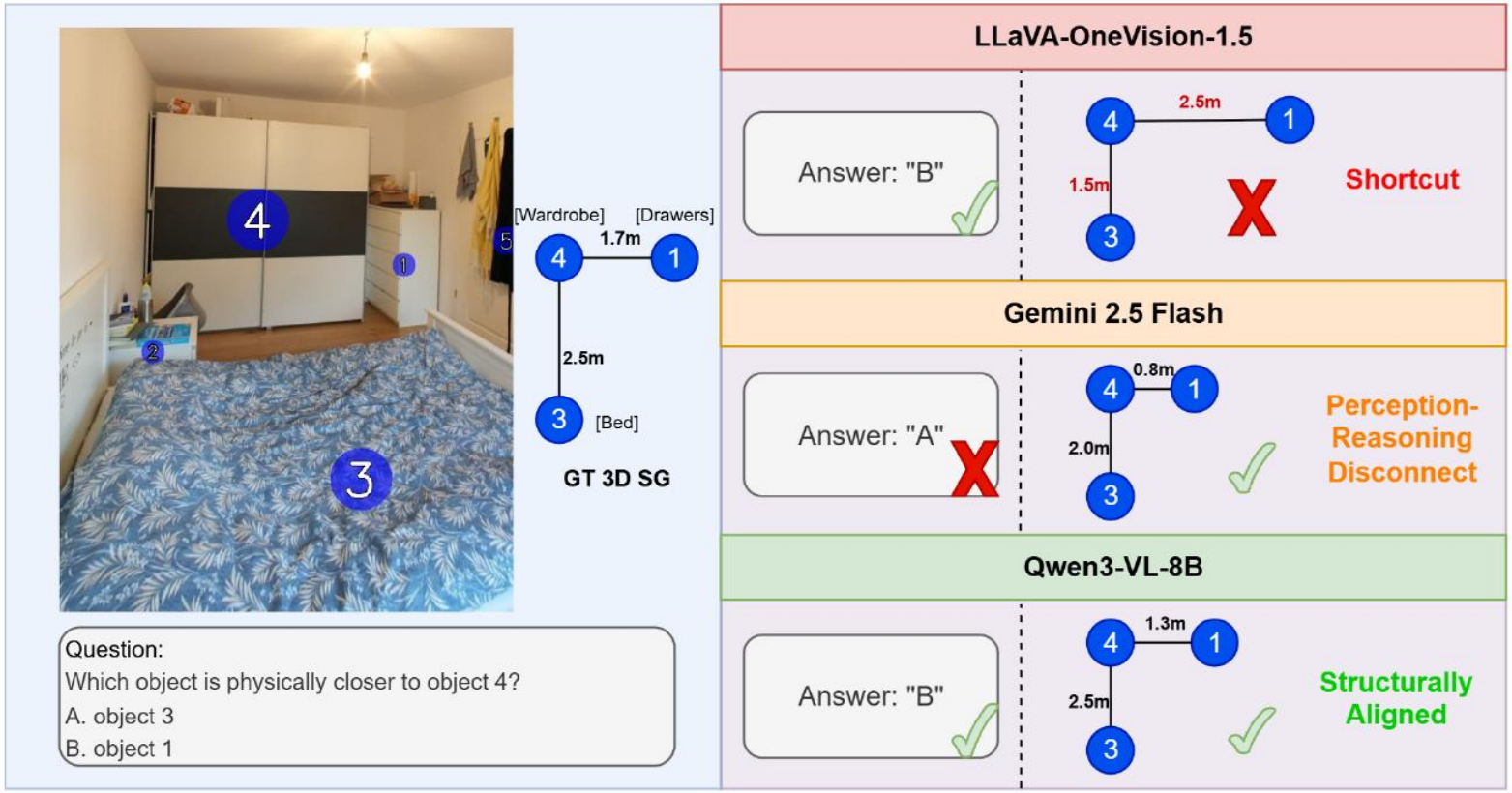



(b) Reliability Verification via Consistency.

Higher QA scores generally accompany better scene graphs and greater consistency, but similar QA scores can hide different structural accuracy. Gemini 2.5 Flash and LLaVA score almost the same on QA (47.68 vs. 47.85), yet their SGC scores differ substantially (64.92 vs. 50.33).

Georgia Tech

# An Example of Disagreement




The drawers are closer to the wardrobe than the bed is, so B is correct. LLaVA answers B but generates a graph with the opposite distance ordering. Gemini generates the correct ordering but answers A. This example shows why CRISP scores answers, graphs and their agreement separately.

Georgia Tech

# Blind Test

| Model | QA | SGC |
|---|---|---|
| ***Proprietary Models**** | | |
| Gemini-2.5-Flash | 28.02 (-19.66) | 45.33 (-19.59) |
| Gemini-2.5-Pro | 26.90 (-32.03) | 37.05 (-27.53) |
| Gemini-3-Flash | 7.94 (-45.47) | 39.47 (-29.32) |
| GPT-5-Mini | 18.31 (-35.84) | 39.32 (-27.38) |
| GPT-5.2 | 30.34 (-18.41) | 38.39 (-27.25) |
| ***Open-source Models*** | | |
| LLaVA-OV-1.5-8B | 22.39 (-25.46) | 43.20 (-7.13) |
| InternVL3.5-8B | 32.31 (-16.06) | 41.45 (-14.76) |
| InternVL3.5-38B | 33.78 (-20.20) | 41.66 (-17.24) |
| Qwen2.5-VL-7B | 28.47 (-16.31) | 39.46 (-3.24) |
| Qwen3-VL-8B | 31.29 (-24.05) | 34.75 (-27.60) |
| Qwen3-VL-32B | 32.78 (-21.20) | 35.56 (-26.89) |
| Random baseline | 25.85 | 28.13 |

*: Affected by text-only refusal behaviors. See Sec. 4.2 for details.

The authors remove the image but keep the prompt to test how much visual input contributes. Parentheses show the score change after removing the image.

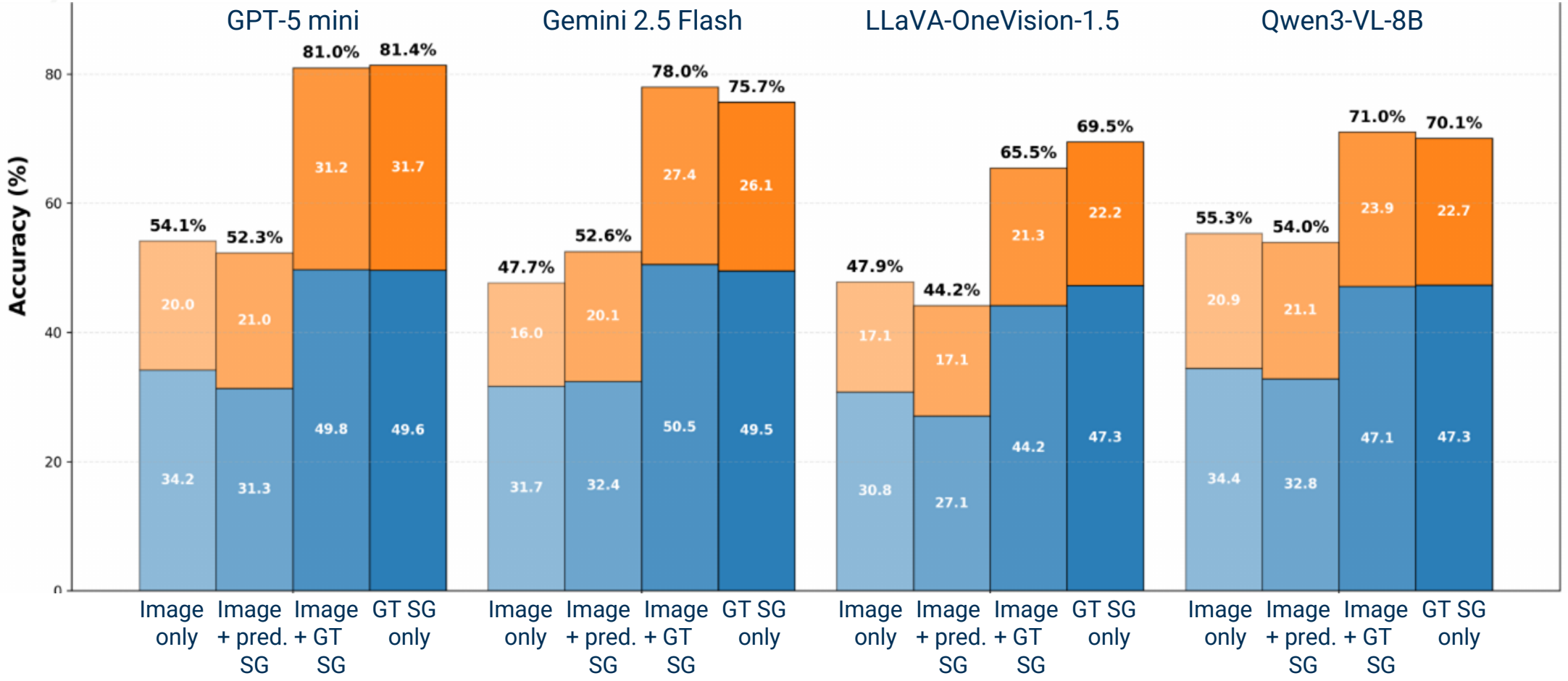
Adding the image increases SGC by 3.24 points for Qwen2.5-VL-7B and 27.60 points for Qwen3-VL-8B. Both models also improve on QA.

The benefit to generated geometry differs substantially between the two models.

Georgia Tech

# Oracle Scene Graphs

Blue: Atomic access (direction, distance, size)  Orange: Compositional logic (ranking, counting, transformation, deduction)



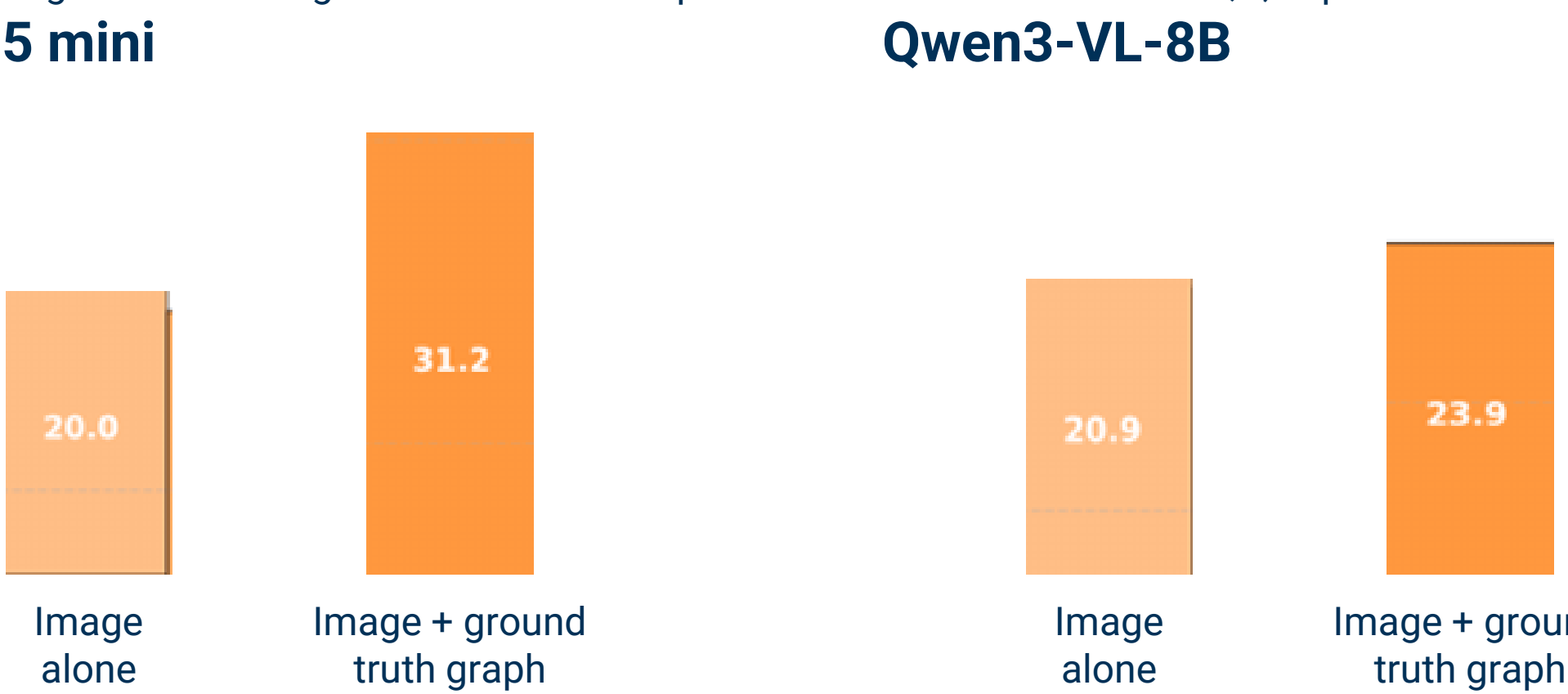

SG = scene graph. pred. = model prediction. GT = ground truth. Every input also includes the question.

Blue and orange show each question group's contribution to total QA. The oracle supplies spatial facts; compositional gains show whether models can combine them, for example to rank objects by distance.

Georgia Tech

# How to Interpret the Oracle Gain

Orange segments from Figure 7 show the compositional contribution to total QA, in points.



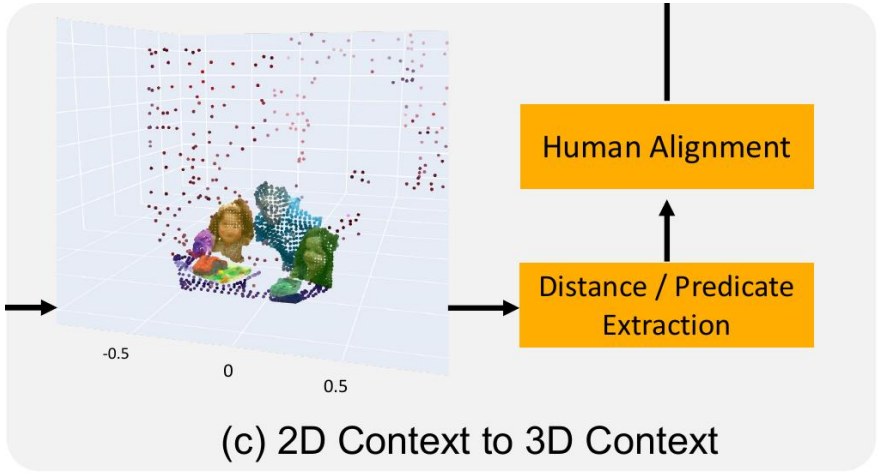

The compositional gain is 11.2 points for GPT-5 mini and 3.0 for Qwen3-VL-8B. The oracle adds correct geometry and explicit structure, so this experiment cannot isolate perception as the sole source of error.

Georgia Tech

# What Persisted

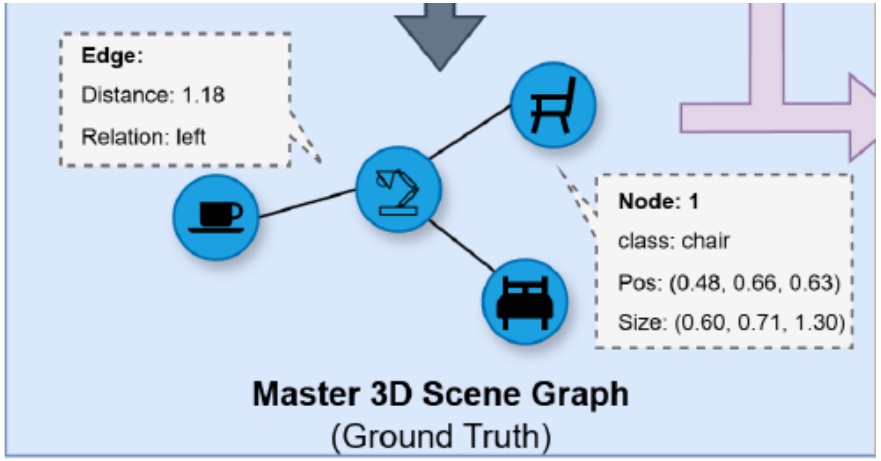
Both papers connect the objects named in a question to explicit measurements of their size, distance and relative position in the scene.

## SpatialVLM




SpatialVLM estimates 3D object properties from images and uses them to generate spatial questions and answers.

## CRISP




CRISP records these geometric quantities in reference scene graphs and compares them with the model's outputs.

Georgia Tech

# What Changed

## SpatialVLM

SpatialVLM uses 3D estimates to create training examples. Its experiments ask whether spatial supervision improves the model's answers.

The experiments add spatial training data and compare a frozen vision encoder with one that is updated.

## CRISP

CRISP uses 3D graphs to evaluate spatial knowledge. It checks whether a model's answers agree with its generated geometry and the reference scene.

The experiments change the information available at inference by removing the image or supplying a predicted or ground truth graph.

SpatialVLM addresses a shortage of spatial supervision. CRISP addresses a different gap: correct answers can coexist with inaccurate geometry, so answer accuracy alone cannot verify spatial grounding.

Georgia Tech

# Limitations

- CRISP while identifying a disconnect between existing internal 3D reasoning (explicit 3D Scene Graph outputs) and answers to QA questions, **does not provide a solution** to bridge this disconnect
- VLM Benchmarks (including CRISP) rely on **many assumptions**
  - these numerous assumptions limit the scope of any results we observe
  - should be a reason for skepticism
- CRISP assumes that explicit 3D Scene Graph outputs are reflective of internal 3D reasoning
  - All evaluations on reasoning are done on observed model behaviors rather than proven causal mechanisms
  - We **do not have a clear understanding of the internal 3D reasoning mechanisms** behind the CRISP results
  - Consistency scores only offer an indication that internal 3D reasoning mechanisms are fragmented, but no measure of:
    - how often this disconnect in reasoning happens between tasks
    - what training mechanisms, data mixtures, etc. cause the model to diverge in 3D reasoning
- Single-view
  - multi-view reasoning adds an additional layer of complexity to spatial reasoning
- Singular snapshot in time
  - evaluations do not extend to video models or temporal reasoning in general
- Simple Reasoning Problems
  - more complex reasoning tasks might need to be chained together

Georgia Tech

# Implications

- Model and data scaling are **not enough** to scale spatial reasoning in models
  - logical consistency between **implicit 3D models and explicit textual reasoning** must be encouraged
  - logical consistency between **tasks** sharing the same ground truth source must be encouraged
- Specialized spatial models (like Cambrian-S and VG LLM) do not perform well compared to generalized models on the CRISP benchmark
  - The big question is: **Why?**
  - We may need to rethink how we train/finetune these models
  - **spatial post-training** might need to be done differently
    - maybe with explicit grounding mechanisms during training

Georgia Tech

# Discussion Questions

1. Will we be able to achieve high performance on spatial reasoning tasks through scale alone? Do we need spatial grounding?
    - At what scales of data would you expect human-level spatial reasoning?
1. Are output-based measures of the latent reasoning space (such as 3D Scene Graphs in CRISP) reflective enough of internal reasoning pathways to diagnose errors in final answers?
    - Could such outputs actually be taking completely different pathways through the model?
1. In what ways should we seek to ground reasoning in models?

# Additional Material for Discussion

Georgia Tech

# CRISP Approach: Consistency Evaluation

**Evaluating Self-Consistency between QA Task and SGC Task:**

- only measures logical consistency, not correctness in answering questions

**Steps:** (for each QA answer)

1. Parse SGC to infer QA answer deterministically
   - graph can fail to support inference → [FAILED] token
   - output: derived answer
2. Grab direct QA response from model
3. Compare SGC derived answer with direct QA response

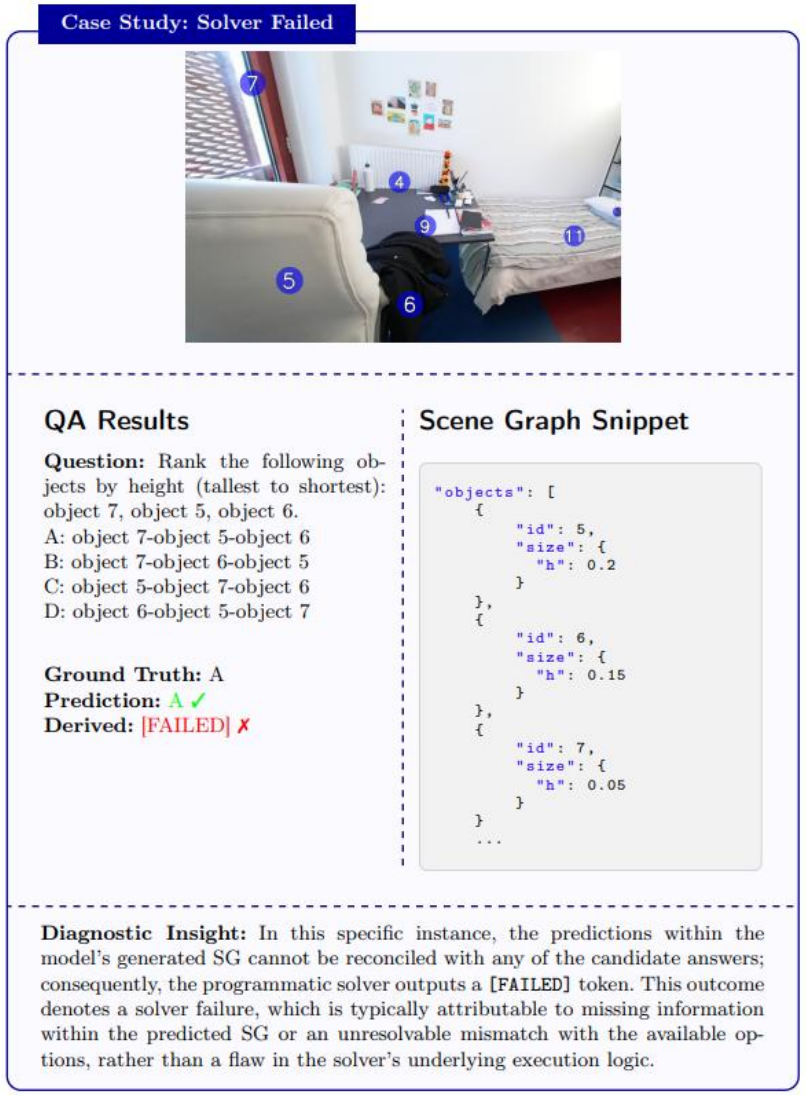



**Fig. 4:** A qualitative example of LLaVA-OneVision1.5-8B demonstrating the solver failed.

Georgia Tech

# High 3D SG Format Compliance Rate

We can see that many of the models had a near **perfect format compliance rate for generating 3D Scene Graphs**

**Table 3:** Detailed evaluation results of SGC. The definitions for all evaluated metrics are detailed in Sec. 3.3. Specifically, Overall denotes the SGC Score, while Comp Rate refers to the 3D SG format compliance rate. A generation is classified as "compliant" if and only if the model outputs a syntactically correct JSON structure without omitting any nodes or edges. Bold denotes the best performance.

| Model | Size | Cam Dist | Dist | Est | Rel | Overall | Comp Rate |
|---|---|---|---|---|---|---|---|
| ***Proprietary Models*** | | | | | | | |
| Gemini-2.5-Flash | 69.17 | 58.59 | 55.97 | 61.24 | 68.61 | 64.92 | 100.00 |
| Gemini-2.5-Pro† | **69.85** | 44.14 | 59.38 | 57.79 | 71.36 | 64.58 | 100.00 |
| Gemini-3-Flash | 69.69 | 57.80 | **68.26** | 65.25 | **72.33** | **68.79** | 100.00 |
| GPT-5-Mini† | 68.25 | 59.73 | 56.20 | 61.39 | 72.01 | 66.70 | 100.00 |
| GPT-5.2 | 68.60 | **64.29** | 67.12 | **66.67** | 64.61 | 65.64 | 100.00 |
| ***Open-source Models*** | | | | | | | |
| LLaVA-OV-1.5-8B | 45.59 | 44.67 | 43.33 | 44.53 | 56.12 | 50.33 | 99.91 |
| InternVL3.5-8B | 60.65 | 46.34 | 46.98 | 51.32 | 61.10 | 56.21 | 99.91 |
| InternVL3.5-38B | **65.43** | 49.01 | 44.93 | 53.12 | 64.68 | 58.90 | 100.00 |
| Qwen2.5-VL-7B | 52.06 | 35.78 | 34.02 | 40.62 | 44.79 | 42.70 | 87.87 |
| Qwen3-VL-8B | 59.97 | 59.57 | 51.72 | 57.09 | **67.62** | 62.35 | 99.48 |
| Qwen3-VL-32B | 63.64 | **60.90** | **56.09** | **60.21** | 64.69 | **62.45** | 100.00 |
| ***Specialized Baselines*** | | | | | | | |
| Cambrian-S | **54.58** | **44.62** | **41.09** | **46.76** | 47.50 | 47.13 | 91.48 |
| VG LLM | 46.69 | 38.57 | 36.56 | 40.61 | **53.85** | **47.23** | 77.54 |

†: Thinking mode restricted.

Georgia Tech

# Derived QA Score vs. Base QA Score

## Derived QA Score (from SGC)

**Table 4: Comparison of Base QA and Derived QA Scores.**

| Model | Base QA | Derived QA |
|---|---|---|
| ***Proprietary Models*** | | |
| Gemini-2.5-Flash | 47.68 | 55.03 (+7.35) |
| Gemini-2.5-Pro | 58.93 | 55.61 (-3.32) |
| Gemini-3-Flash | 53.41 | 59.36 (+5.95) |
| GPT-5-Mini | 54.15 | 56.43 (+2.28) |
| GPT-5.2 | 48.75 | 58.77 (+10.02) |
| ***Open-source Models*** | | |
| LLaVA-OV-1.5-8B | 47.85 | 41.95 (-5.90) |
| InternVL3.5-8B | 48.37 | 46.40 (-1.97) |
| InternVL3.5-38B | 53.98 | 48.66 (-5.32) |
| Qwen2.5-VL-7B | 44.78 | 37.75 (-7.03) |
| Qwen3-VL-8B | 55.34 | 52.83 (-2.51) |
| Qwen3-VL-32B | 53.98 | 54.70 (+0.72) |
| ***Specialized Baselines*** | | |
| Cambrian-S | 45.95 | 37.79 (-8.16) |
| VG LLM | 38.34 | 34.90 (-3.44) |

Georgia Tech