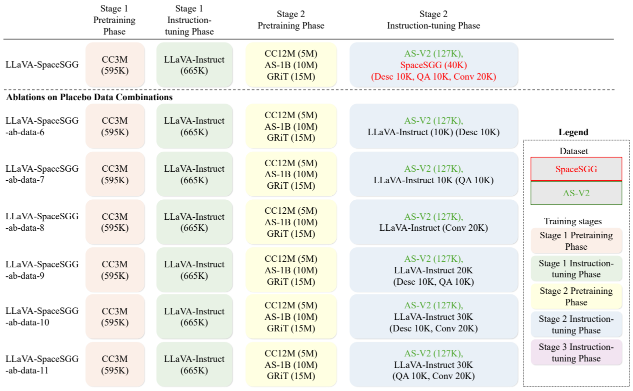
LLaVA-SpaceSGG: Visual Instruct Tuning for Open-vocabulary Scene Graph Generation with Enhanced Spatial Relations
WACV2025 Supplementary Material

Mingjie Xu[1*], Mengyang Wu[2*], Yuzhi Zhao[3†], Jason Chun Lok Li[4], Weifeng Ou[5]
parasolohalo@gmail.com, yzzhao2-c@my.cityu.edu.hk

[1]Independent Researcher [2]The Chinese University of Hong Kong [3]City University of Hong Kong
[4]The University of Hong Kong [5]Dongguan University of Technology



	Stage 1 Pretraining Phase	Stage 1 Instruction-tuning Phase	Stage 2 Pretraining Phase	Stage 2 Instruction-tuning Phase
LLaVA-SpaceSGG	CC3M (595K)	LLaVA-Instruct (665K)	CC12M (5M) AS-1B (10M) GRiT (15M)	AS-V2 (127K), SpaceSGG (40K) (Desc 10K, QA 10K, Conv 20K)
Ablations on Placebo Data Combinations				
LLaVA-SpaceSGG -ab-data-6	CC3M (595K)	LLaVA-Instruct (665K)	CC12M (5M) AS-1B (10M) GRiT (15M)	AS-V2 (127K), LLaVA-Instruct (10K) (Desc 10K)
LLaVA-SpaceSGG -ab-data-7	CC3M (595K)	LLaVA-Instruct (665K)	CC12M (5M) AS-1B (10M) GRiT (15M)	AS-V2 (127K), LLaVA-Instruct 10K (QA 10K)
LLaVA-SpaceSGG -ab-data-8	CC3M (595K)	LLaVA-Instruct (665K)	CC12M (5M) AS-1B (10M) GRiT (15M)	AS-V2 (127K), LLaVA-Instruct (Conv 20K)
LLaVA-SpaceSGG -ab-data-9	CC3M (595K)	LLaVA-Instruct (665K)	CC12M (5M) AS-1B (10M) GRiT (15M)	AS-V2 (127K), LLaVA-Instruct 20K (Desc 10K, QA 10K)
LLaVA-SpaceSGG -ab-data-10	CC3M (595K)	LLaVA-Instruct (665K)	CC12M (5M) AS-1B (10M) GRiT (15M)	AS-V2 (127K), LLaVA-Instruct 30K (Desc 10K, Conv 20K)
LLaVA-SpaceSGG -ab-data-11	CC3M (595K)	LLaVA-Instruct (665K)	CC12M (5M) AS-1B (10M) GRiT (15M)	AS-V2 (127K), LLaVA-Instruct 30K (QA 10K, Conv 20K)

Legend
Dataset
SpaceSGG
AS-V2
Training stages
Stage 1 Pretraining Phase
Stage 1 Instruction-tuning Phase
Stage 2 Pretraining Phase
Stage 2 Instruction-tuning Phase
Stage 3 Instruction-tuning Phase

Figure 1. We conduct placebo ablation studies by testing the same data combination, replacing components in SpaceSGG with equivalent ones from the LLaVA-Instruct dataset.

The supplementary material contains (1) more ablation studies testing effectiveness of the proposed dataset; (2) more examples about the SpaceSGG dataset including 3 components (SpaceSGG-Desc, SpaceSGG-QA and SpaceSGG-Conv); (3) more visual examples about our proposed LLaVA-SpaceSGG prediction compare with other models (TextPSG, ASMv2).

1. A. More Ablation Studies

To further validate the effectiveness of the proposed dataset, we replaced each element with equivalent components from the LLaVA-Instruct dataset, ensuring the same number of entries were sampled. The experimental settings are illustrated in Figure 1. As shown in Table 1, these replacements did not improve the model's SGG performance or spatial understanding, further highlighting the significance of our dataset.

Ablation Setting	**Recall**	**mRecall**	**Accuracy (%)**
LLaVA-SpaceSGG -ab-data-6	9.49	8.18	30.415
LLaVA-SpaceSGG -ab-data-7	14.95	11.74	51.775
LLaVA-SpaceSGG -ab-data-8	13.2	8.44	51.8
LLaVA-SpaceSGG -ab-data-9	0	0	0
LLaVA-SpaceSGG -ab-data-10	0	0	26.895
LLaVA-SpaceSGG -ab-data-11	13.07	8.66	39.075
LLaVA-SpaceSGG	15.43	13.23	52.48

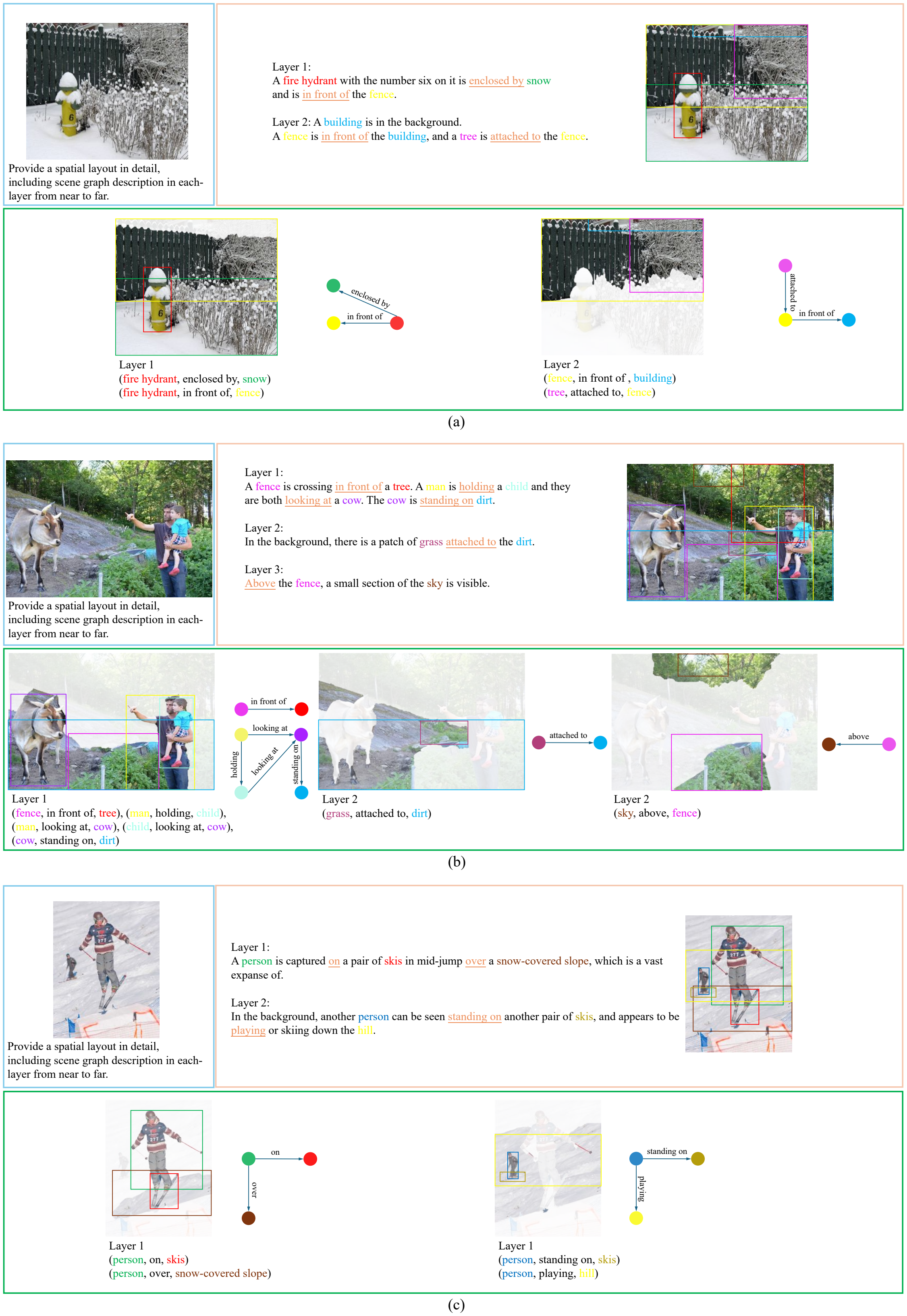
Table 1. We experimented with different mixing ratios of replaced placebo data, using refabricated data combinations for the experimental settings. The red, blue, and green colors denote the best, the second highest and the third highest results, respectively. For detailed experimental settings, please refer to Figure 1.

2. B. SpaceSGG dataset examples

We provide three types examples of SpaceSGG data, see in Figure 2, Figure 3 and Figure 4.

3. C. PSG Dataset Evaluation Comparison

We report more visual evaluation results of LLaVA-SpaceSGG compared with ASMv2 and TextPSG, see in Figure 5.



Provide a spatial layout in detail, including scene graph description in each-layer from near to far.

Layer 1:
A fire hydrant with the number six on it is enclosed by snow and is in front of the fence.

Layer 2: A building is in the background.
A fence is in front of the building, and a tree is attached to the fence.

Layer 1
(fire hydrant, enclosed by, snow)
(fire hydrant, in front of, fence)

Layer 2
(fence, in front of , building)
(tree, attached to, fence)

(a)

Provide a spatial layout in detail, including scene graph description in each-layer from near to far.

Layer 1:
A fence is crossing in front of a tree. A man is holding a child and they are both looking at a cow. The cow is standing on dirt.

Layer 2:
In the background, there is a patch of grass attached to the dirt.

Layer 3:
Above the fence, a small section of the sky is visible.

Layer 1
(fence, in front of, tree), (man, holding, child), (man, looking at, cow), (child, looking at, cow), (cow, standing on, dirt)

Layer 2
(grass, attached to, dirt)

Layer 2
(sky, above, fence)

(b)

Provide a spatial layout in detail, including scene graph description in each-layer from near to far.

Layer 1:
A person is captured on a pair of skis in mid-jump over a snow-covered slope, which is a vast expanse of.

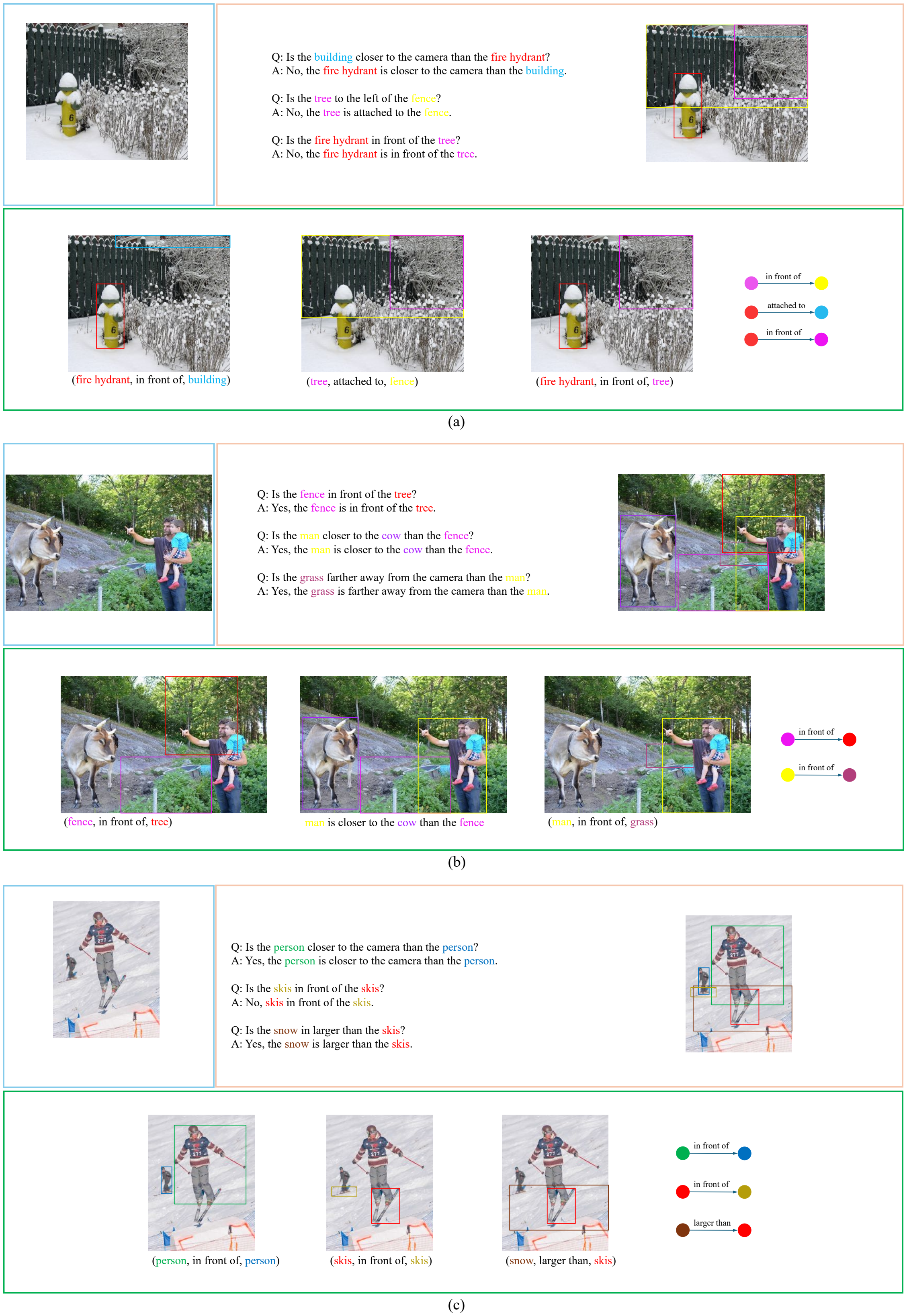
Layer 2:
In the background, another person can be seen standing on another pair of skis, and appears to be playing or skiing down the hill.

Layer 1
(person, on, skis)
(person, over, snow-covered slope)

Layer 1
(person, standing on, skis)
(person, playing, hill)

(c)

Figure 2. Data Examples of SpaceSGG-Desc in SpaceSGG.



Q: Is the building closer to the camera than the fire hydrant?
A: No, the fire hydrant is closer to the camera than the building.

Q: Is the tree to the left of the fence?
A: No, the tree is attached to the fence.

Q: Is the fire hydrant in front of the tree?
A: No, the fire hydrant is in front of the tree.

(fire hydrant, in front of, building)
(tree, attached to, fence)
(fire hydrant, in front of, tree)
in front of
attached to
in front of

(a)

Q: Is the fence in front of the tree?
A: Yes, the fence is in front of the tree.

Q: Is the man closer to the cow than the fence?
A: Yes, the man is closer to the cow than the fence.

Q: Is the grass farther away from the camera than the man?
A: Yes, the grass is farther away from the camera than the man.

(fence, in front of, tree)
man is closer to the cow than the fence
(man, in front of, grass)
in front of
in front of

(b)

Q: Is the person closer to the camera than the person?
A: Yes, the person is closer to the camera than the person.

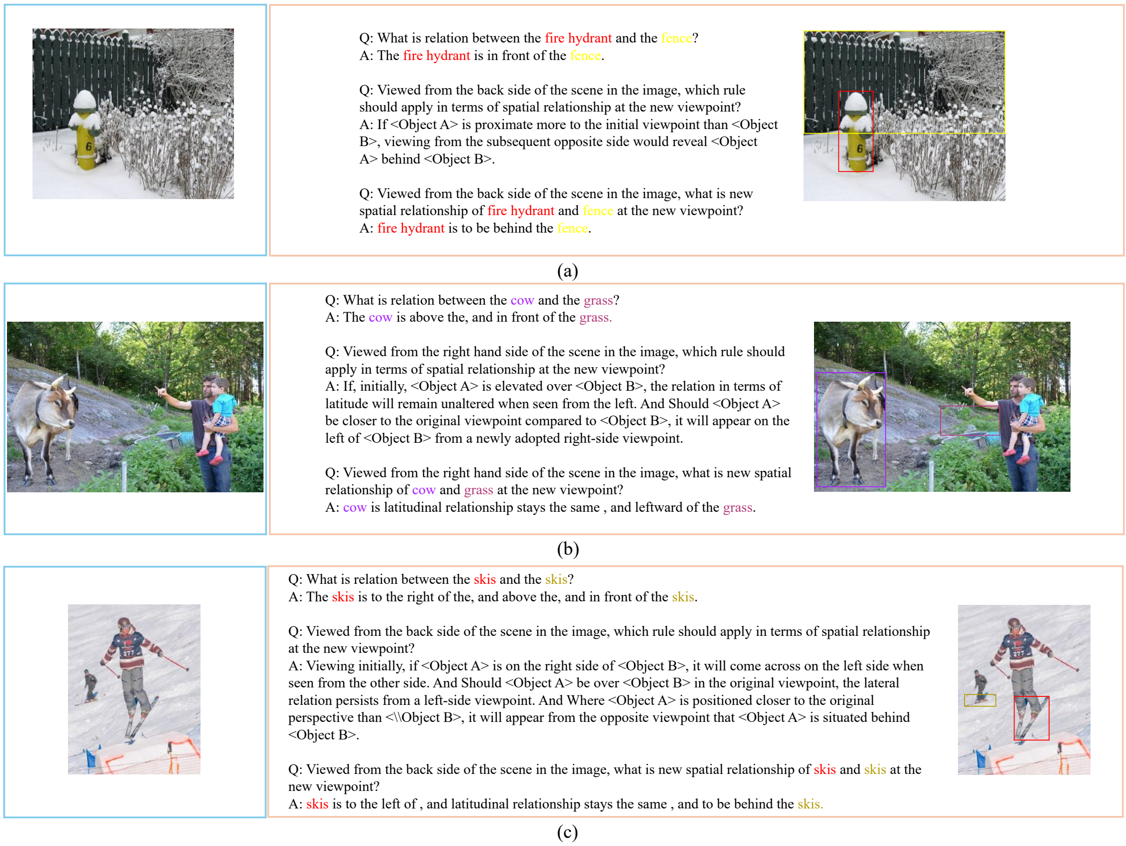
Q: Is the skis in front of the skis?
A: No, skis in front of the skis.

Q: Is the snow in larger than the skis?
A: Yes, the snow is larger than the skis.

(person, in front of, person)
(skis, in front of, skis)
(snow, larger than, skis)
in front of
in front of
larger than

(c)

Figure 3. Data Examples of SpaceSGG-QA in SpaceSGG.



Q: What is relation between the fire hydrant and the fence?
A: The fire hydrant is in front of the fence.

Q: Viewed from the back side of the scene in the image, which rule should apply in terms of spatial relationship at the new viewpoint?
A: If <Object A> is proximate more to the initial viewpoint than <Object B>, viewing from the subsequent opposite side would reveal <Object A> behind <Object B>.

Q: Viewed from the back side of the scene in the image, what is new spatial relationship of fire hydrant and fence at the new viewpoint?
A: fire hydrant is to be behind the fence.

(a)

Q: What is relation between the cow and the grass?
A: The cow is above the, and in front of the grass.

Q: Viewed from the right hand side of the scene in the image, which rule should apply in terms of spatial relationship at the new viewpoint?
A: If, initially, <Object A> is elevated over <Object B>, the relation in terms of latitude will remain unaltered when seen from the left. And Should <Object A> be closer to the original viewpoint compared to <Object B>, it will appear on the left of <Object B> from a newly adopted right-side viewpoint.

Q: Viewed from the right hand side of the scene in the image, what is new spatial relationship of cow and grass at the new viewpoint?
A: cow is latitudinal relationship stays the same , and leftward of the grass.

(b)

Q: What is relation between the skis and the skis?
A: The skis is to the right of the, and above the, and in front of the skis.

Q: Viewed from the back side of the scene in the image, which rule should apply in terms of spatial relationship at the new viewpoint?
A: Viewing initially, if <Object A> is on the right side of <Object B>, it will come across on the left side when seen from the other side. And Should <Object A> be over <Object B> in the original viewpoint, the lateral relation persists from a left-side viewpoint. And Where <Object A> is positioned closer to the original perspective than <\\Object B>, it will appear from the opposite viewpoint that <Object A> is situated behind <Object B>.

Q: Viewed from the back side of the scene in the image, what is new spatial relationship of skis and skis at the new viewpoint?
A: skis is to the left of , and latitudinal relationship stays the same , and to be behind the skis.

(c)

Figure 4. Data Examples of SpaceSGG-Conv in SpaceSGG.

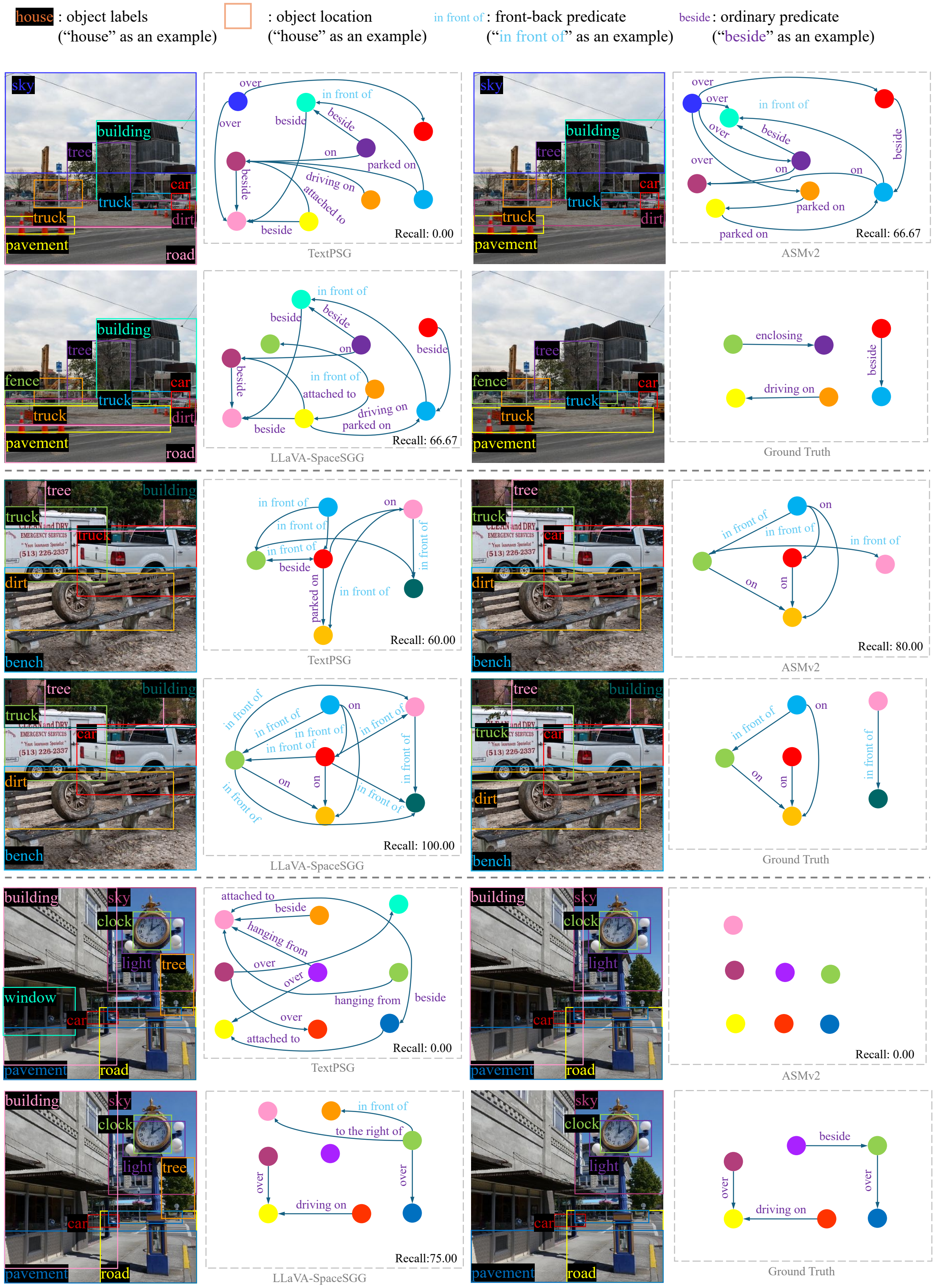


Figure 5. Additional examples of LLaVA-SpaceSGG Open-Vocabulary SGG prediction compared with others on PSG validation set.