



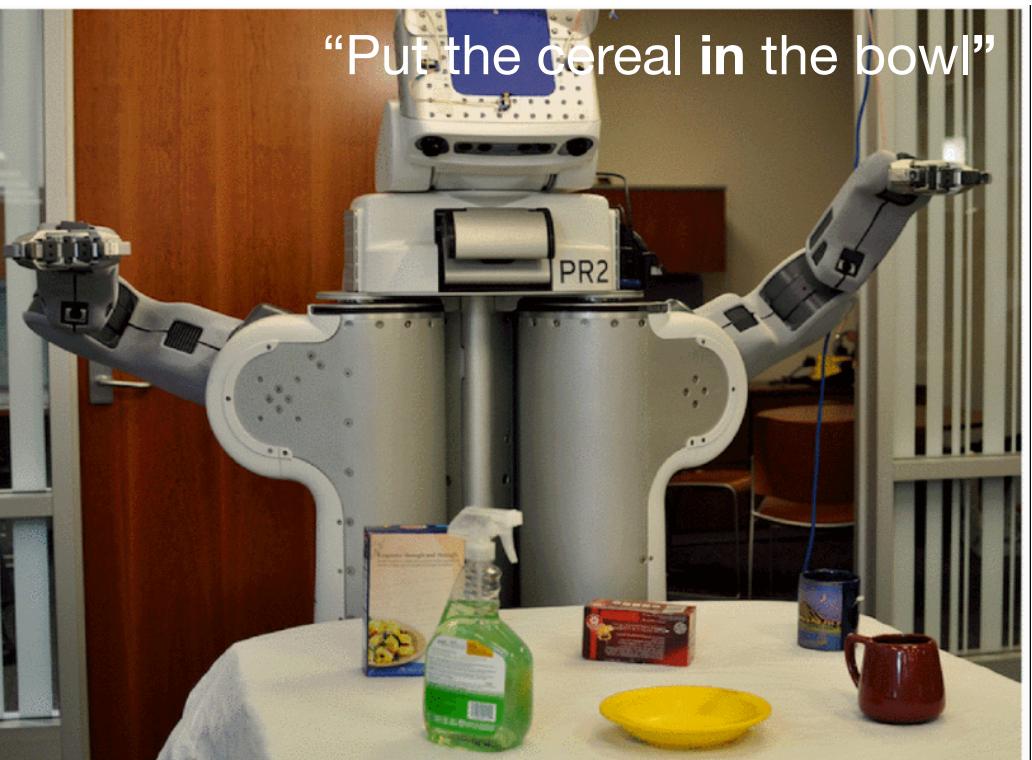
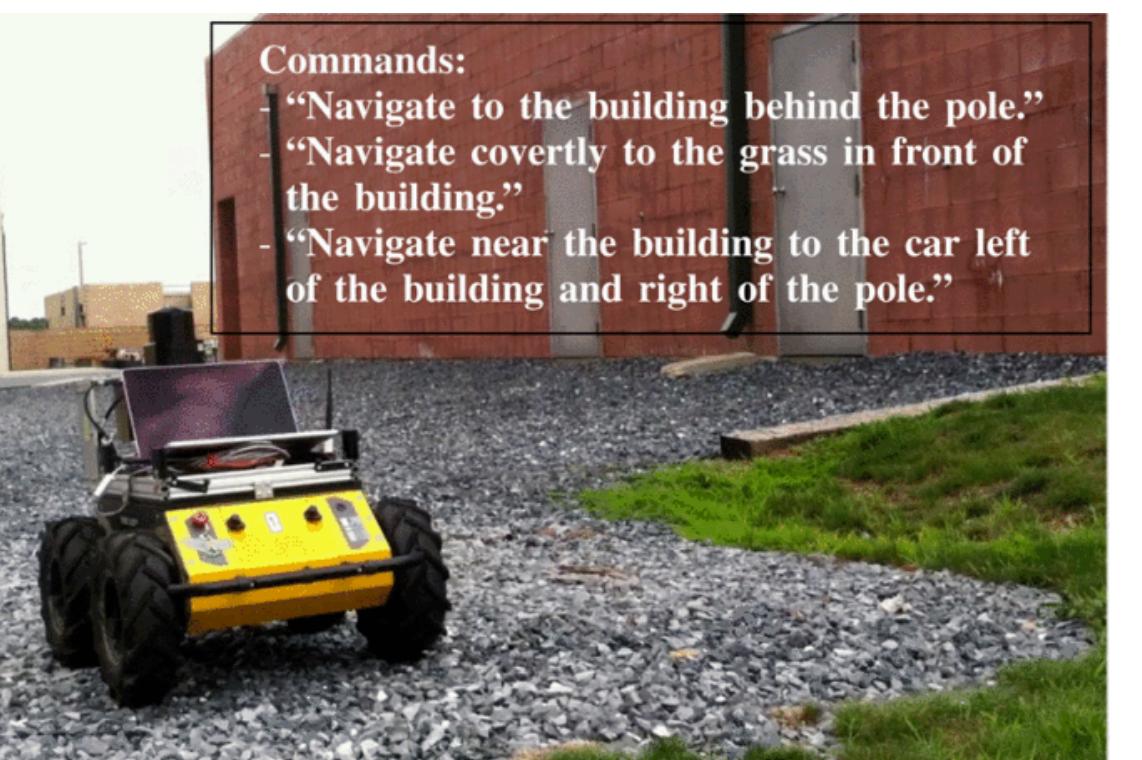
Rel3D: A Minimally Contrastive Benchmark for Grounding Spatial Relations in 3D

Ankit Goyal* Kaiyu Yang* Dawei Yang*[▲] Jia Deng*

University of Michigan[▲] Princeton University*



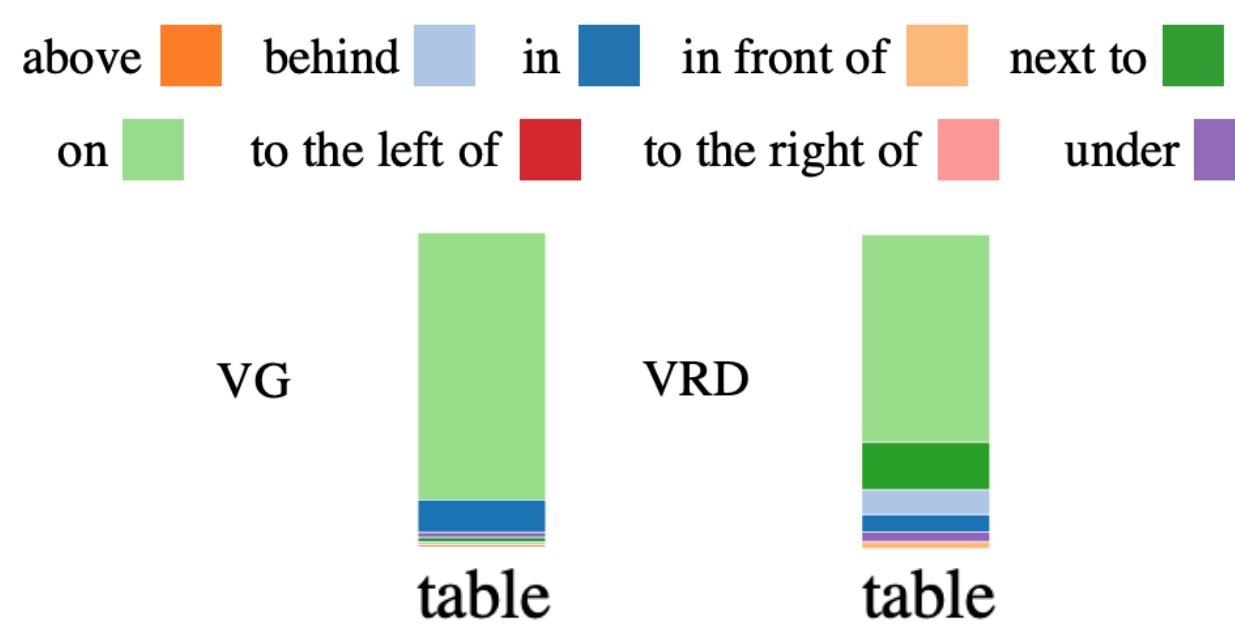
Motivation



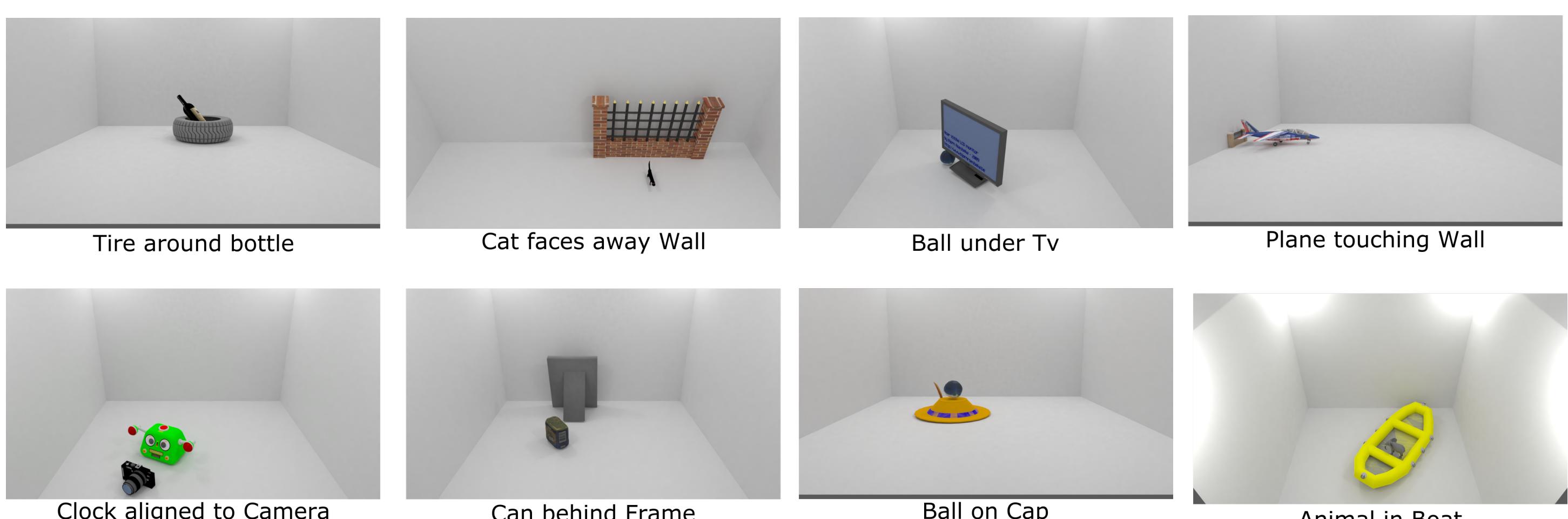
Grounding spatial relations is crucial in for ai agents like robots for navigation [1] and object manipulation [2]

Two critical issues in prior datasets:

- 1. Suffer from language and 2D bias [3].
- 2. Limited to only 2D images — 3D cues like depth, pose are critical for spatial grounding relations [4, 5]



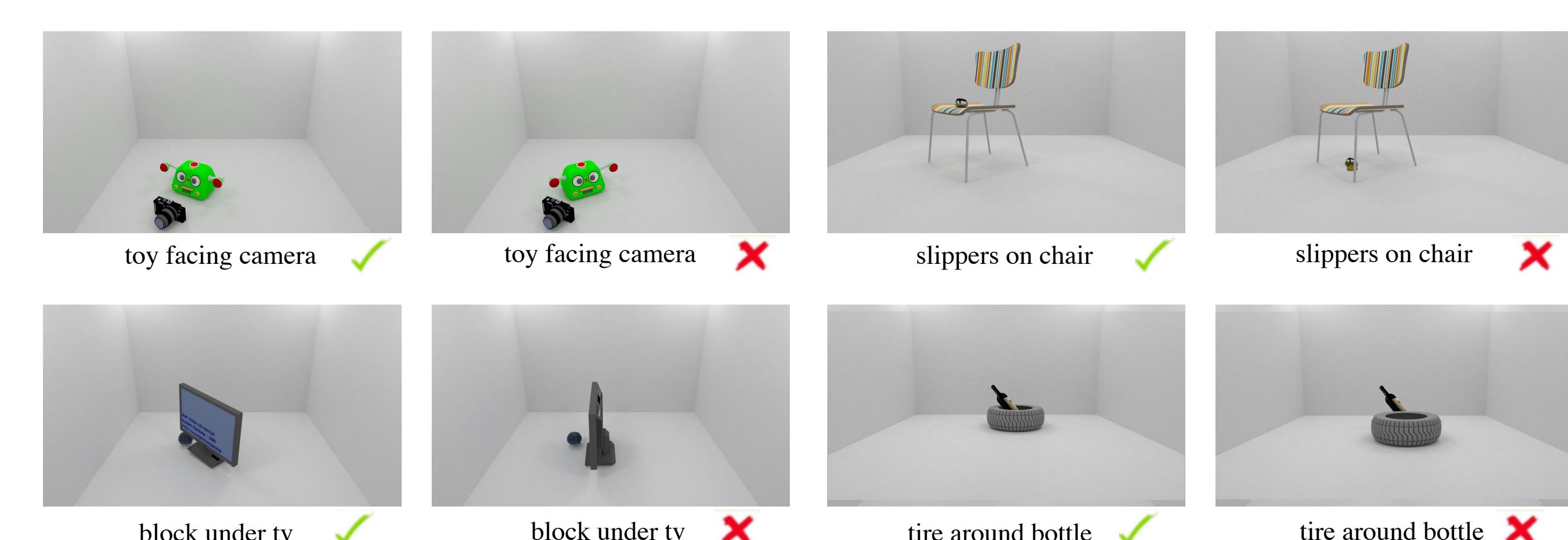
Rel3D



10K synthetic 3D scenes with human annotated spatial relations

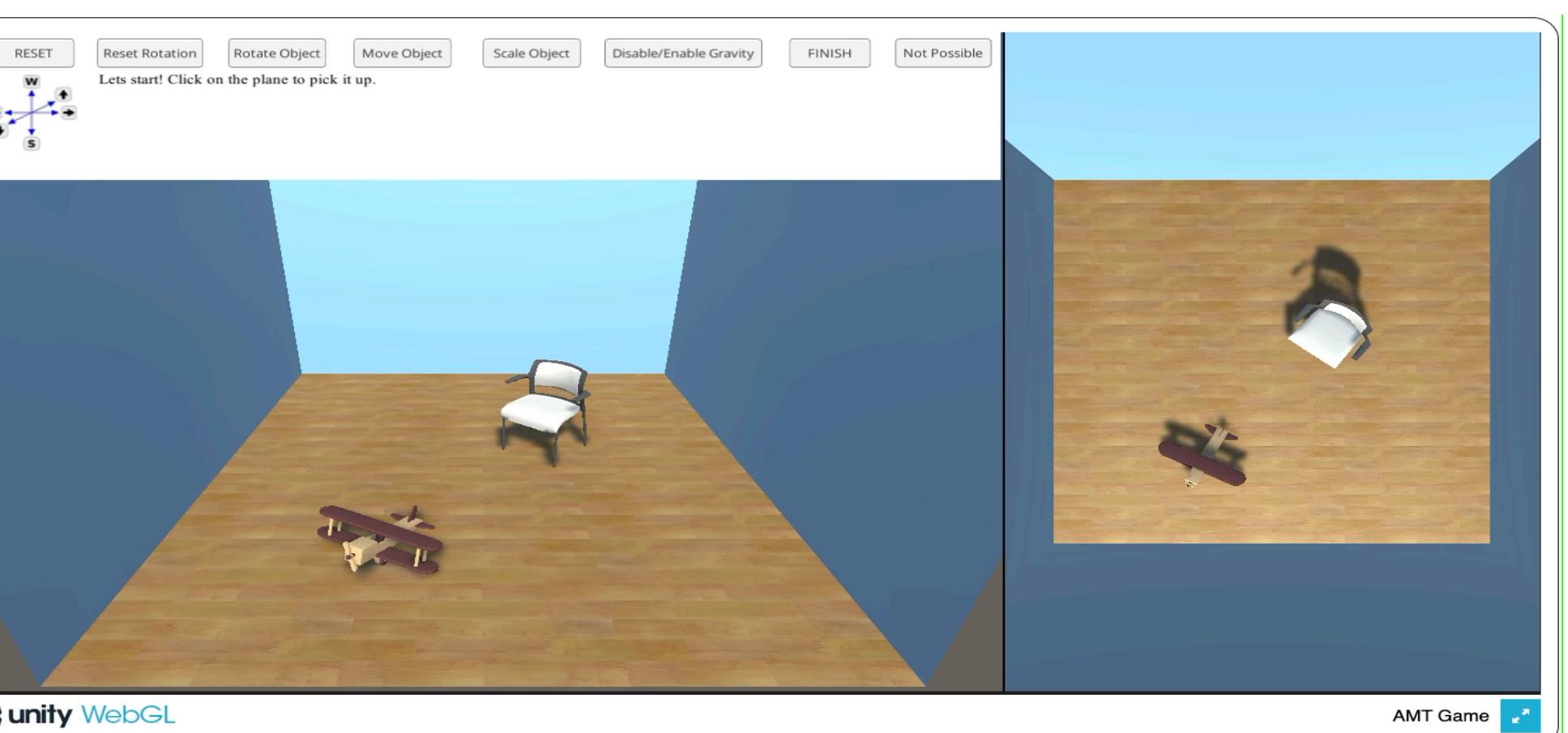
Rel3D is **minimally contrastive**:

- Scenes occur in pairs; relation holds in one and does not hold in other.
- Same objects, same room, same camera, same lighting.

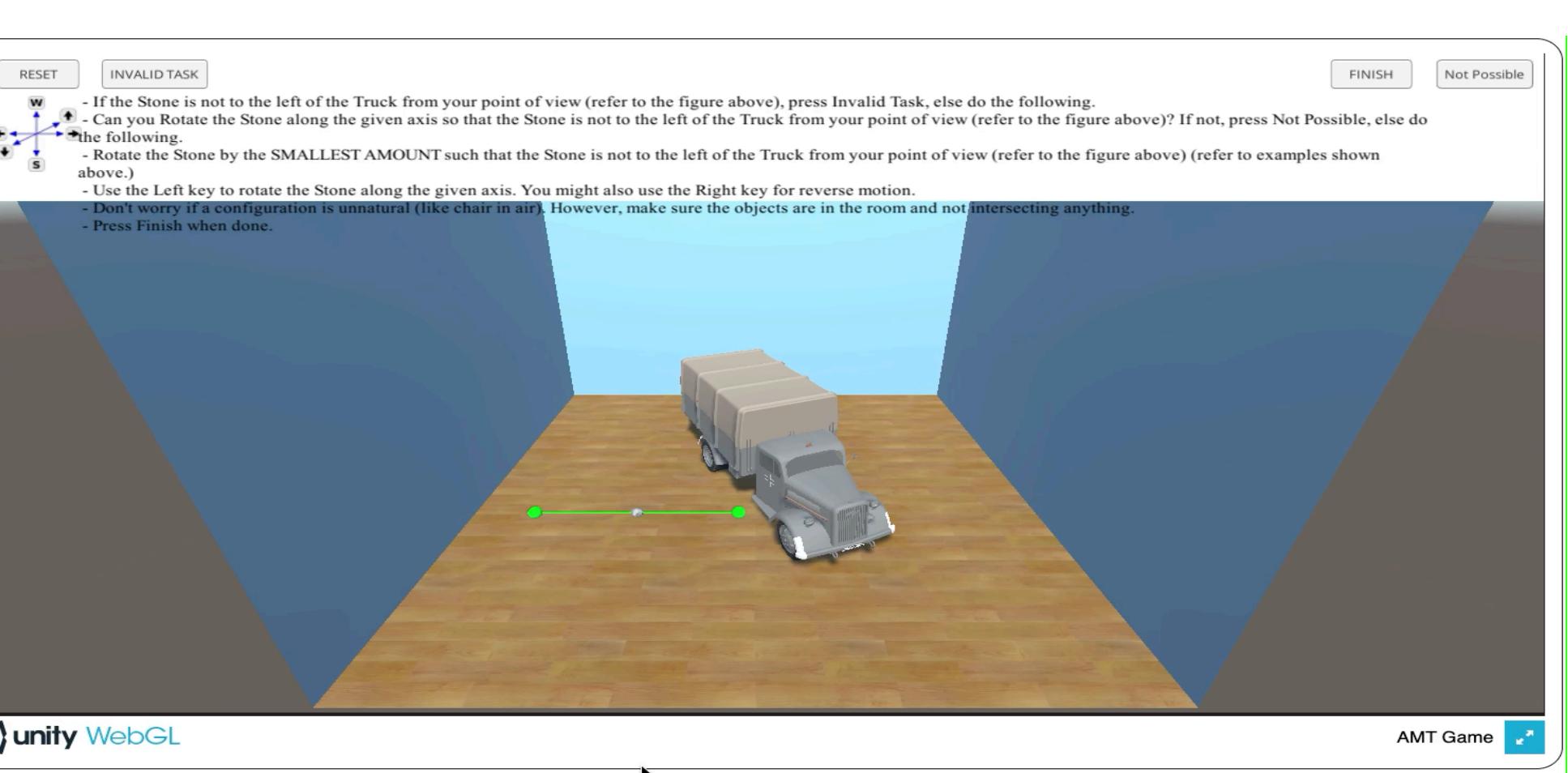


Data Collection

Stage 1:
AMT workers manipulate and resize objects so that they satisfy a relation in unity webgl interface



Stage 2:
Independent AMT workers minimally modify the scene so that the relation does not hold; object position or rotation changed along a given axis

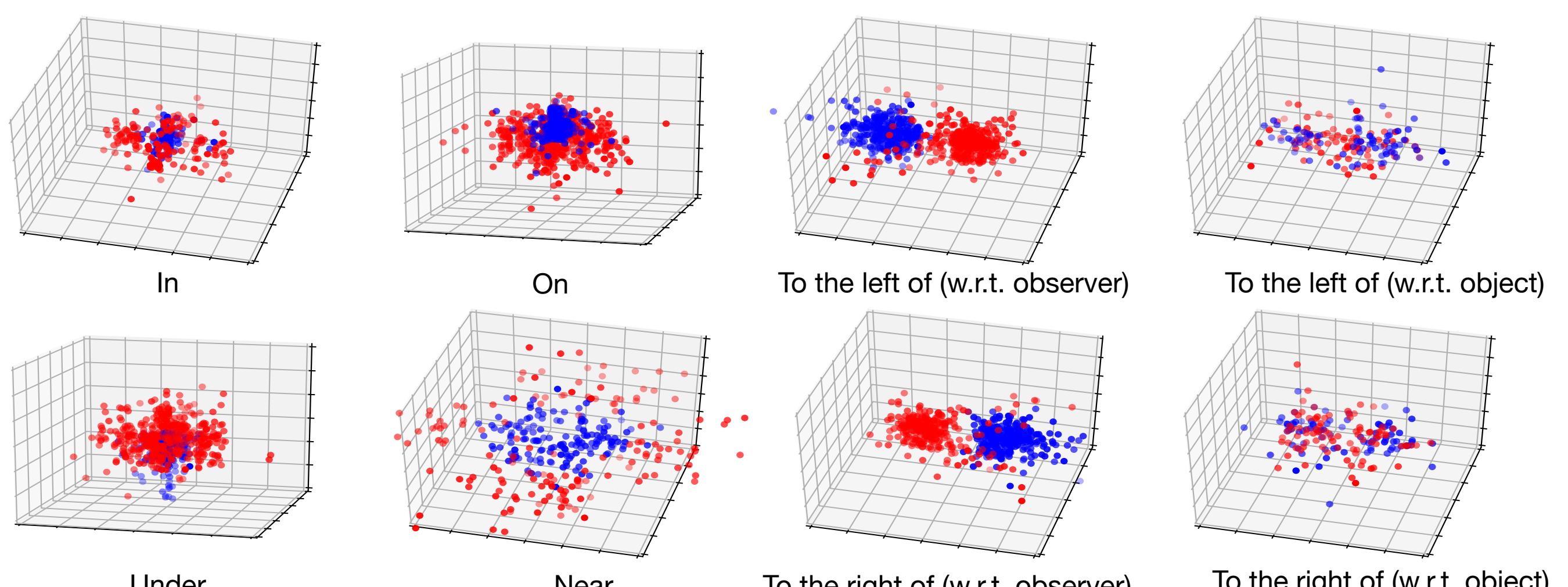


Stage 3:
Independent workers verify images rendered from different viewpoints; Images added to dataset only when both in a pair are correct.

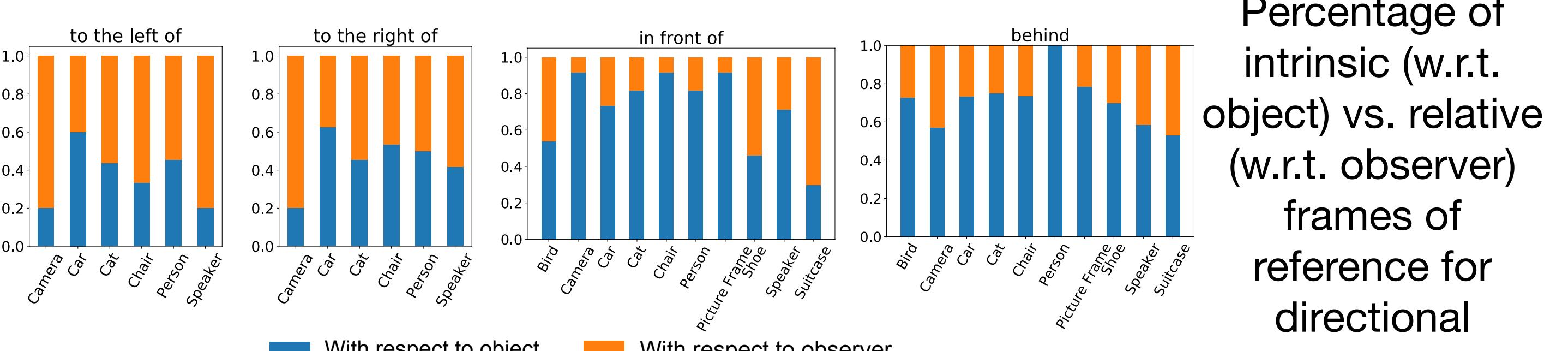


In the image above, would you say that the ball is below the TV? Yes No

Analysis



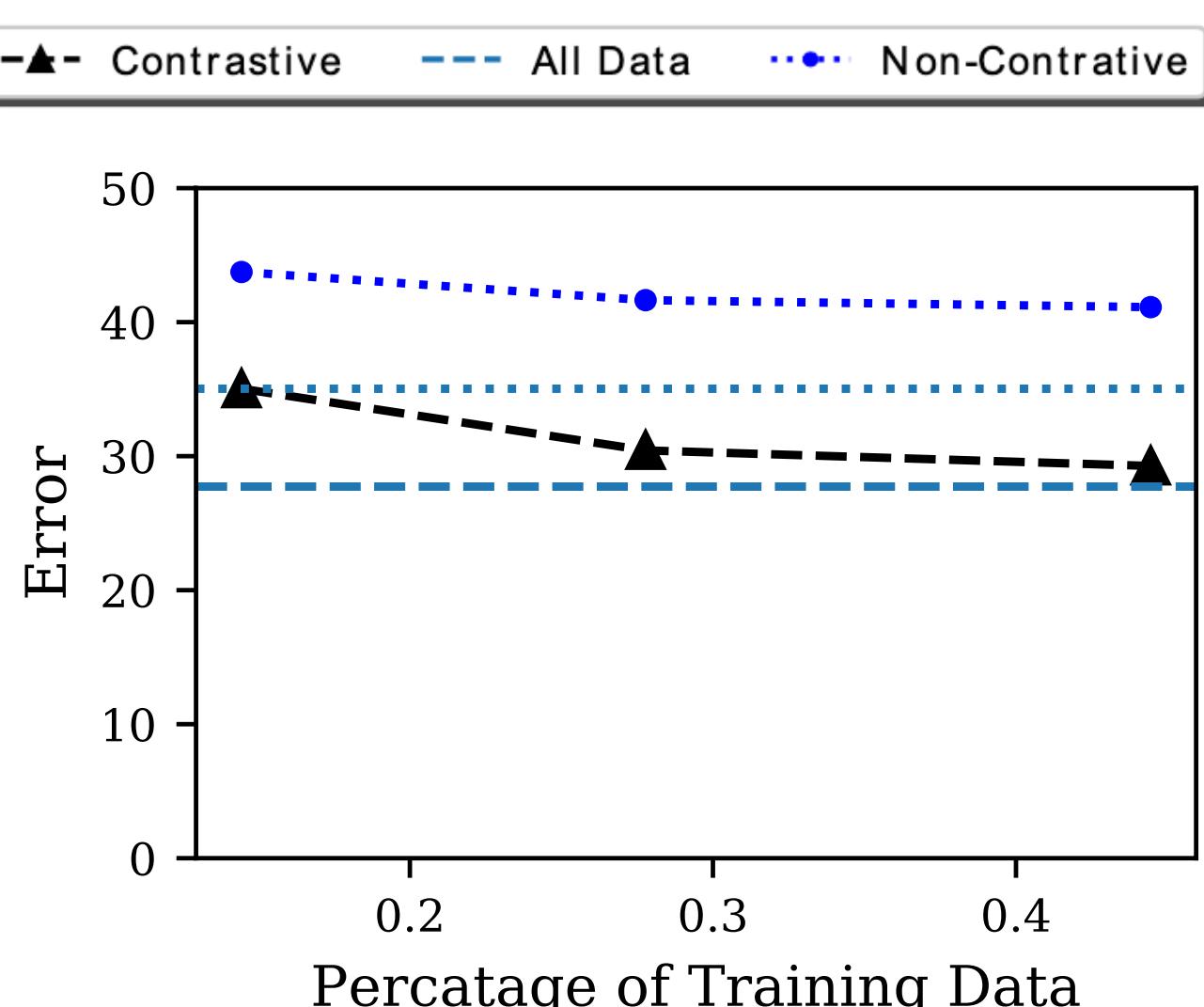
The relative position of the object w.r.t. to the subject (subj.) in the observer's (obs.) reference frame.



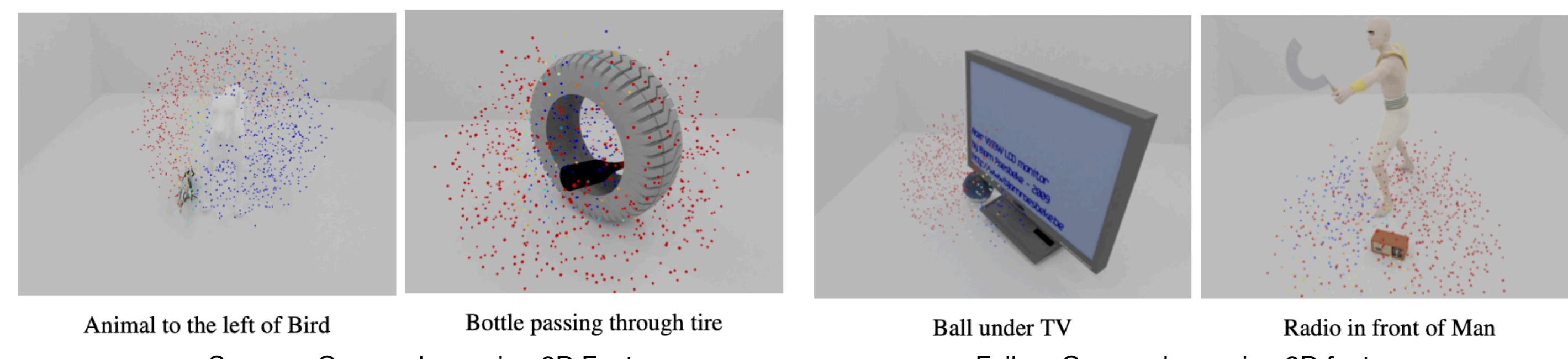
Experiments

Model	Accuracy
Random	50.00%
Language Only	50.00%
Bounding Box Only	74.14%
DRNet [6]	73.25%
Vip-CNN [7]	72.32%
VTransE [8]	72.27%
PPR-FCN [9]	73.30%
MLP - Raw Features	81.24%
MLP - Aligned Features	85.03%
Human	94.25%

- Language only model perform randomly
- SOTA models do not outperforms bounding box only model
- 3D features help, scope for improvement



Minimally contrastive samples lead to sample efficient training; models trained on contrastive subsets outperform those trained on non-contrastive subsets using less than 1/3rd samples.



- 3D Features: object scale, rotation, position and up and front direction
- **Ball under TV:** Approximates the TV as a cuboid and predicts some regions underneath the screen as not under the TV.
- **Radio in front of Man:** Ambiguous case where whether the front of a person is defined w.r.t to their face or torso.

References

- [1] Bouali, Abdeslam, et al. "Grounding spatial relations for outdoor robot navigation." ICRA 2015
- [2] Guadarrama, Sergio, et al. "Grounding spatial relations for human-robot interaction." IROS 2013
- [3] Yang, Kaiyu, et al. "SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition." ICCV 2019.
- [4] Ye, Jun et al. "Exploiting depth camera for 3d spatial relationship interpretation" ACM Multimedia 2013
- [5] Guadarrama, Sergio, et al. "Grounding spatial relations for human-robot interaction." IROS 2013
- [6] Dai, Bo, Yuqi Zhang, and Dahua Lin. "Detecting visual relationships with deep relational networks." CVPR 2017.
- [7] Li, Yikang, et al. "Vip-cnn: Visual phrase guided convolutional neural network." CVPR 2017.
- [8] Zhang, Hanwang, et al. "Visual translation embedding network for visual relation detection." CVPR 2017.
- [9] Zhuang, Bohan, et al. "Towards context-aware interaction recognition for visual relationship detection." ICCV 2017.