Supplementary – Embodied Language Grounding with 3D Visual Feature Representations

Mihir Prabhudesai[*], Hsiao-Yu Fish Tung[*], Syed Ashar Javed[*]
Maximilian Sieb[†], Adam W. Harley, Katerina Fragkiadaki
{mprabhud, htung,sajaved,msieb,aharley,katef}@cs.cmu.edu
Carnegie Mellon University
Project website: https://mihirp1998.github.io/project_pages/emblang/

1. Model details: Language-conditioned 3D visual imagination

We train our stochastic generative networks using conditional variational autoencoders. For the *what* generative module, our inference network conditions on the word embeddings of the adjectives and the noun in the noun phrase, as well as the 3D feature tensor obtained by cropping the 3D feature map $\mathbf{M} = \text{GRNN}(I)$ using the ground-truth 3D bounding box of the object the noun phrase concerns. For the *where* generative module, the corresponding inference network conditions on one-hot encoding of the spatial expression, as well as the 3D relative spatial offset, available from 3D object box annotations. Inference networks are used only at training time. Our *what* and *where* decoders take the posterior noise and predict 3D object appearance feature tensors, and cross-object 3D spatial offsets, respectively, for each object. We add predicted object feature tensors at predicted 3D locations in a 3D feature canvas. Our reconstruction losses ask the language-generated and image-inferred 3D feature maps from GRNNs to be close in feature distance, both in 3D and after 2D neural projection using the GRNN 3D-to-2D neural decoder, and the predicted cross-object 3D relative spatial offsets to be close to the ground-truth cross-object 3D relative offsets.

2. Experimental details: Language conditioned scene generation

We visualize our model's predictions in two ways: i) **neurally rendered** are obtained by feeding the generated 3D assembled canvas to the 3D-to-2D neural projection module of GRNNs, ii) **Blender rendered** are renderings of Blender scenes that contain object 3D meshes selected by small feature distance to the language generated object 3D feature tensors, and arranged based on the predicted 3D spatial offsets.

We consider a database of 300 object 3D meshes to choose from. To get the object feature tensor for a candidate 3D object model, we render multi-view RGB-D data of this object in Blender, and input them to the GRNN to obtain the corresponding feature map, which we crop using the groundtruth bounding box. Blender renders better convey object appearance because the neurally rendered images are blurry. Despite pixel images being blurry, our model retrieves correct object meshes that match the language descriptions.

3. Additional experiments

Scene generation conditioned on natural language In Figure 1, we compare our model with the model of Deng et al (5) on language to scene generation with utterances longer than those used during training time. We show both neural and Blender rendering of scenes predicted from our model. We remind the reader that a Blender rendering is computed by using the cross-object relative 3D offsets predicted by our model, and using the generated object 3D feature tensors to retrieve the closest matching meshes from a training set. Our training set is comprised of 100 objects with known 3D bounding boxes, and for each we compute a 3D feature tensor by using the 2D-to-3D unprojection module described above, and cropping the corresponding subtensor based on the 3D bounding box coordinates of the object. Despite our neural rendering being blurry, we show the features of our generative networks achieve correct nearest neighbor retrieval. The generation results show our model can generalize to utterances that are much longer than those in the training data. In Figure 2, we show rendering results from our model on our real world dataset.

[*]Equal contribution
[†]Work done while in Carnegie Mellon University.

One key feature of our model is that it generates a scene as opposed to an independent static image. In Figure 3, we show rendering images from the 3D feature tensor across different viewpoints. The rendering images are consistent across viewpoints. For a 2D baseline (5), it is unclear how we can obtain a set of images that not only match with input sentence but also are consistent with each others.

We show in Figures 4-5 more neural and Blender rendering of scenes predicted from our model, conditioning on parse trees of natural language utterances. In 6, we show rendering results learned from our real world dataset.

Scene generation conditional on natural language and visual context In Figures 7-9 we show examples of scene generation from our model when conditioned on both natural language and the visual context of the agent. In this case, some objects mentioned in the natural language utterance are present in the agent's environment, and some are not. Our model uses a 3D object detector to localize objects in the scene, and the learnt 2D-to-3D unprojection neural module to compute a 3D feature tensor for each, by cropping the scene tensor around each object. Then, it compares the object tensors generated from natural language to those generated from the image, and if a feature distance is below a threshold, it grounds the object reference in the parse tree of the utterance to object present in the environment of the agent. If such binding occurs, as is the case for the "green cube" in the top left example, then our model uses the image-generated tensors of the binded objects, instead of the natural language generated ones, to complete the imagination. In this way, our model grounds natural language to both perception and imagination.

Affordability inference based on 3D non-intersection Objects do not intersect in 3D. Our model has a 3D feature generation space and can detect when this basic principle is violated. The baseline model of (5) directly generates 2D images described in the utterances (conditioned on their parse tree) without an intermediate 3D feature space. Thus, it performs such affordability checks in 2D. However, in 2D, objects frequently occlude one another, while they still correspond to an affordable scene. In Figure 10, we show intersection over union scores computed in 3D by our model and in 2D by the baseline. While for our model such scores correlate with affordabilty of the scene (e.g., the scenes in 1st, third, and forth columns in the first row are clearly non-affordable as objects inter-penetrate) the same score from the baseline is not an indicator of affordability, e.g., the last column in the last row of the figure can in fact be a perfectly valid scene, despite the large IoU score.

Language-guided placement policy learning In Figure 11, we show the initial and final configurations of the learned policy using different referential expression. The robot can successfully place the object to the target location given the referential expression. We also show in the supplementary a video of a real robot executing the task.

4. Additional related work

The inspiring experiments of Glenberg and Robertson (10) in 1989 demonstrated that humans can easily judge the plausibility —they called it *affordability*—of natural language utterances, such as *"he used a newspaper to protect his face from the wind"*, and the implausibility of others, such as *"he used a matchbox to protect his face from the wind"*. They suggest that humans associate words with actual objects in the environment or prototypes in their imagination that retain perceptual properties—how the objects look—and affordance information (8)—how the objects can be used. A natural language utterance is then understood through perceptual and motor *simulations* of explicitly and implicitly mentioned nouns and verbs, in some level of abstraction, that encode such affordances. For example, the matchbox is too small to protect a human face from the wind, while a newspaper is both liftable by a human and can effectively cover a face when held appropriately. This hypothesis is currently better known as simulation semantics (6; 7; 2; 4) and has extensive empirical support: reaction times for visual or motor operations are shorter when human subjects are shown a related sentence (9; 3), and MRI activity is increased in the brain's vision system or motor areas when human subjects are shown vision- or motor-related linguistic concepts, respectively (1; 11; 12). This paper proposes an initial computational model for the simulation semantics hypothesis for the language domain of object spatial arrangements.

References

[1] L. Aziz-Zadeh, C. Fiebach, S. Narayanan, J. Feldman, E. Dodge, and R. Ivry. Modulation of the ffa and ppa by language related to faces and places. *Social neuroscience*, 3:229–38, 02 2008. 2

[2] B. Bergen. Mental simulation in spatial language processing. 01 2005. 2

[3] B. Bergen. Experimental methods for simulation semantics. *Methods in cognitive linguistics*, pages 277–301, 2007. 2

[4] B. Bergen. Embodiment, simulation and meaning. *The Routledge Handbook of Semantics*, pages 142–157, 01 2015. 2

[5] Z. Deng, J. Chen, Y. FU, and G. Mori. Probabilistic neural programmed networks for scene generation. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 4028–4038. Curran Associates, Inc., 2018. 1, 2, 3, 11

[6] J. Feldman and S. Narayanan. Embodied meaning in a neural theory of language. *Brain and language*, 89:385–92, 06 2004. 2

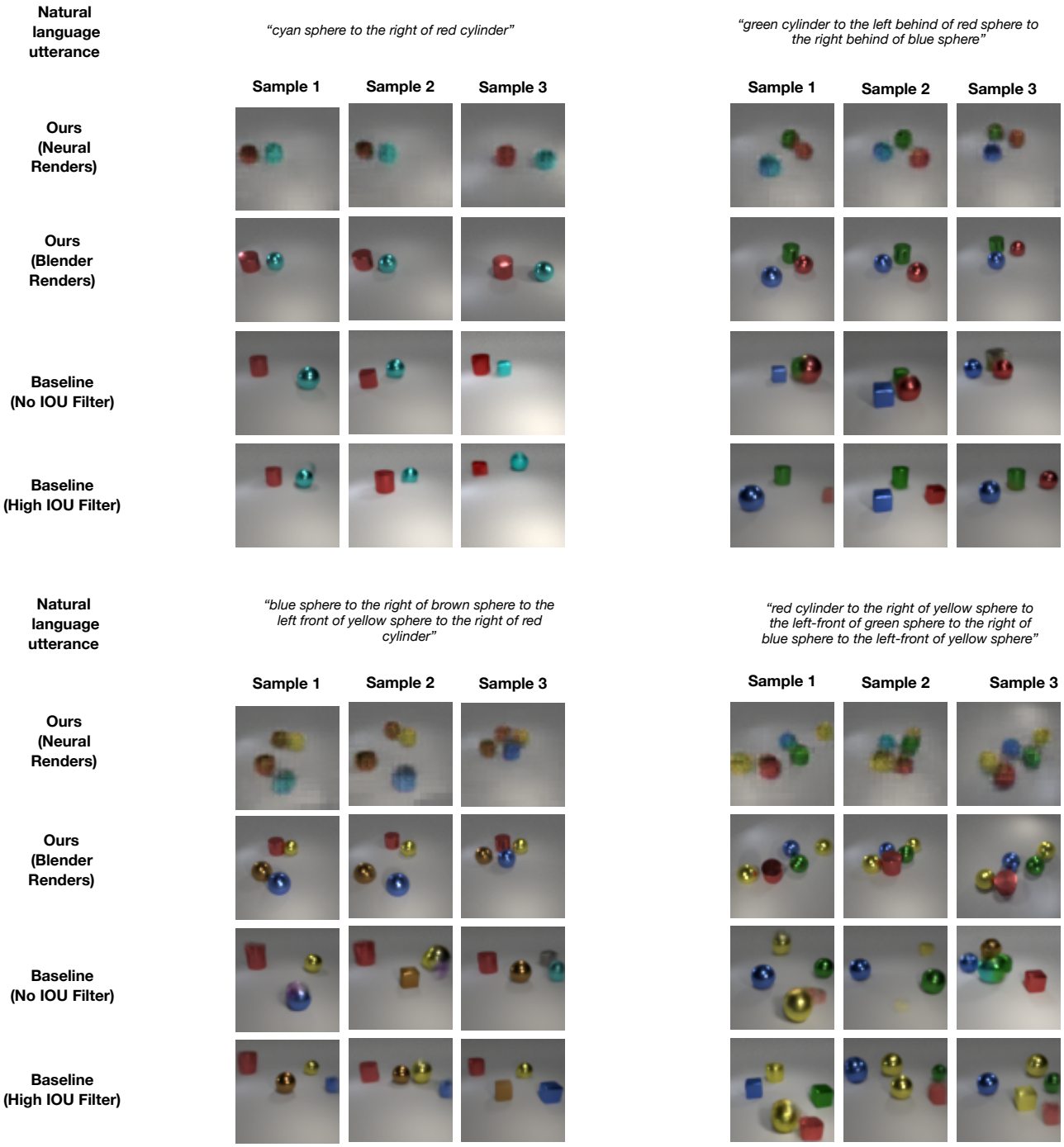

Figure 1: **Language to scene generation** from our model (Row1, Row2) and the model of Deng et al (5) (Row 3, 4) for utterances longer than those encountered at training time. Both our model and the baseline are stochastic, and we sample three generated scenes/images per utterance. (Row 1 and Row 2) shows neural and Blender rendering results from our model. For the Blender rendering, we retrieve the closet 3D object mesh using the features of our generative networks, and place the retrieved objects in the corresponding locations in Blender to render an image. (Row 3) shows image generation results with no IoU constraint during sampling for the baseline. This means objects might go out of the field of view. (Row 4) shows result with high IoU constraint.

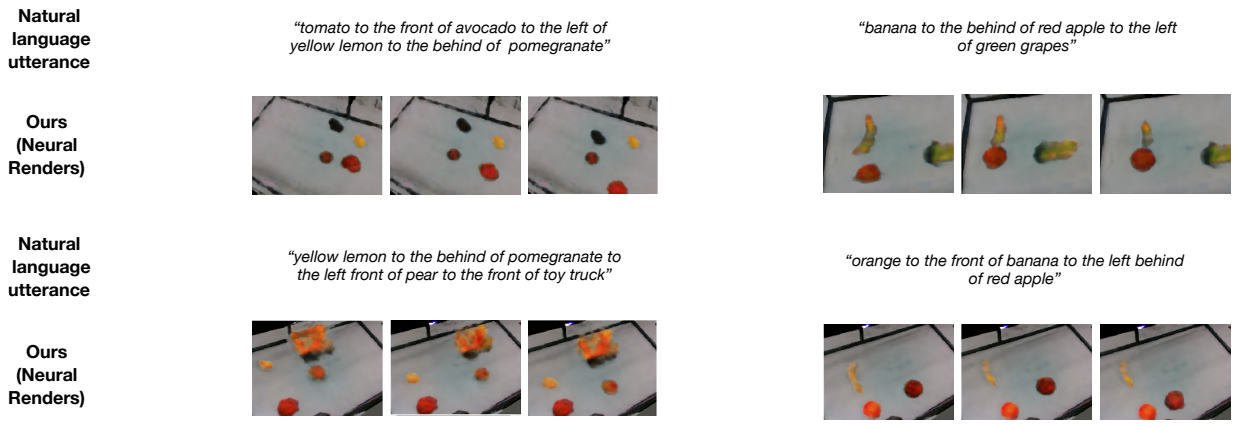

Figure 2: **Language to image generation** on our real world data. We sample three different scenes for each natural language utterances.

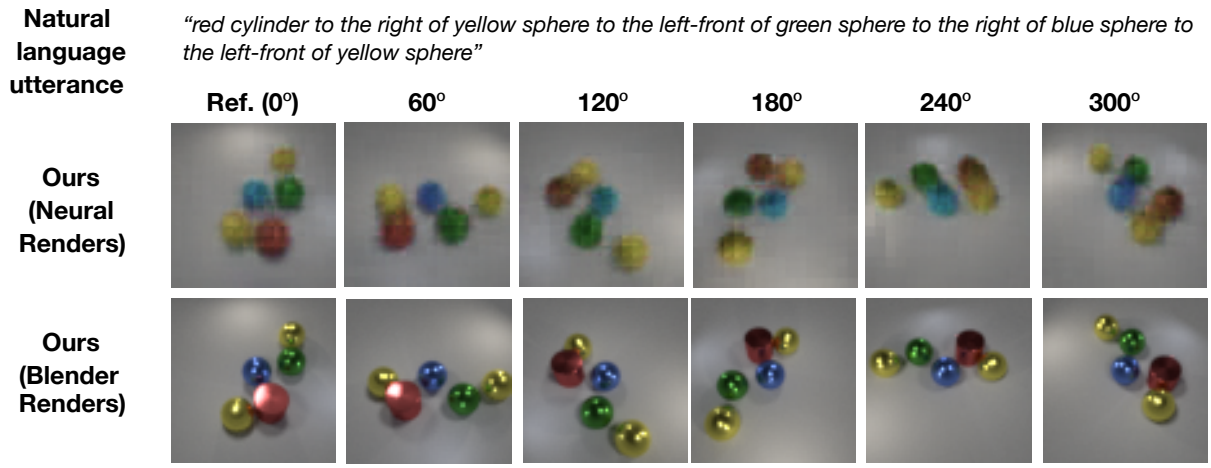

Figure 3: **Consistent scene generation** . We render the generated 3D feature canvas from various viewpoints in the first row using the neural GRNN decoder, and compare against the different viewpoint projected Blender rendered scenes. Indeed, our model correctly predicts occlusions and visibilities of objects from various viewpoints, and can generalize across different number of objects. 2D baselines do not have such imagination capability.

[7] J. A. Feldman. *From Molecule to Metaphor: A Neural Theory of Language*. MIT Press, Cambridge, MA, 2006. 2

[8] J. J. Gibson. *The Ecological Approach to Visual Perception*. Houghton Mifflin, 1979. 2

[9] A. Glenberg and M. Kaschak. Grounding language in action. *Psychonomic Bulletin and Review*, 9(3):558–565, 9 2002. 2

[10] A. Glenberg and D. Robertson. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language*, 2000. 2

[11] R. M Willems, I. Toni, P. Hagoort, and D. Casasanto. Neural dissociations between action verb understanding and motor imagery. *Journal of cognitive neuroscience*, 22:2387–400, 11 2009. 2

[12] A. P. Saygin, S. Mccullough, M. Alac, and K. Emmorey. Modulation of bold response in motion-sensitive lateral temporal cortex by real and fictive motion sentences. *Journal of cognitive neuroscience*, 22:2480–90, 11 2009. 2

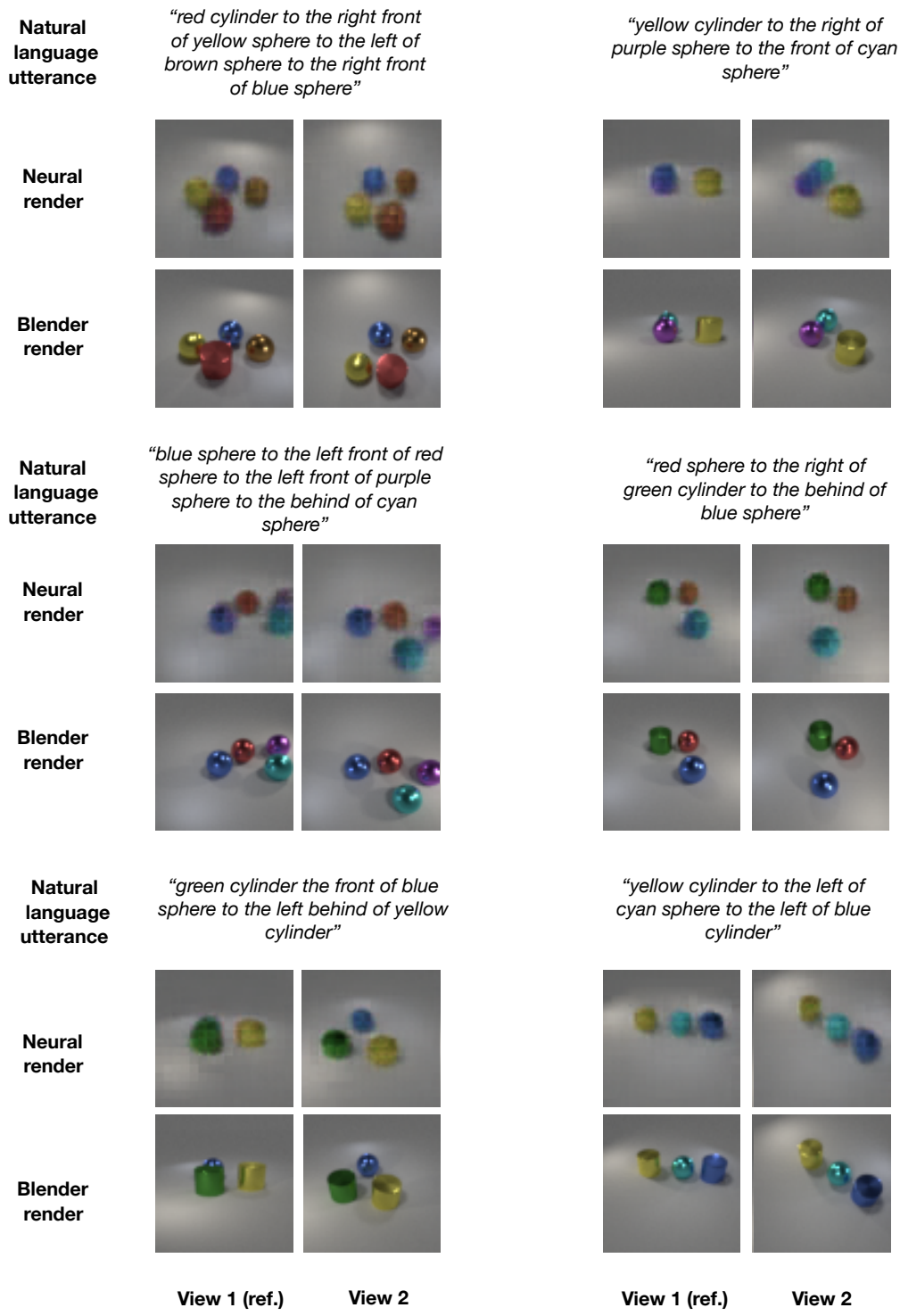

Figure 4: **Natural language conditioned neural and blender scene renderings generated by the proposed model.** We visualize each scene from two nearby views, a unique ability of our model, due to its 3-dimensional generation space.

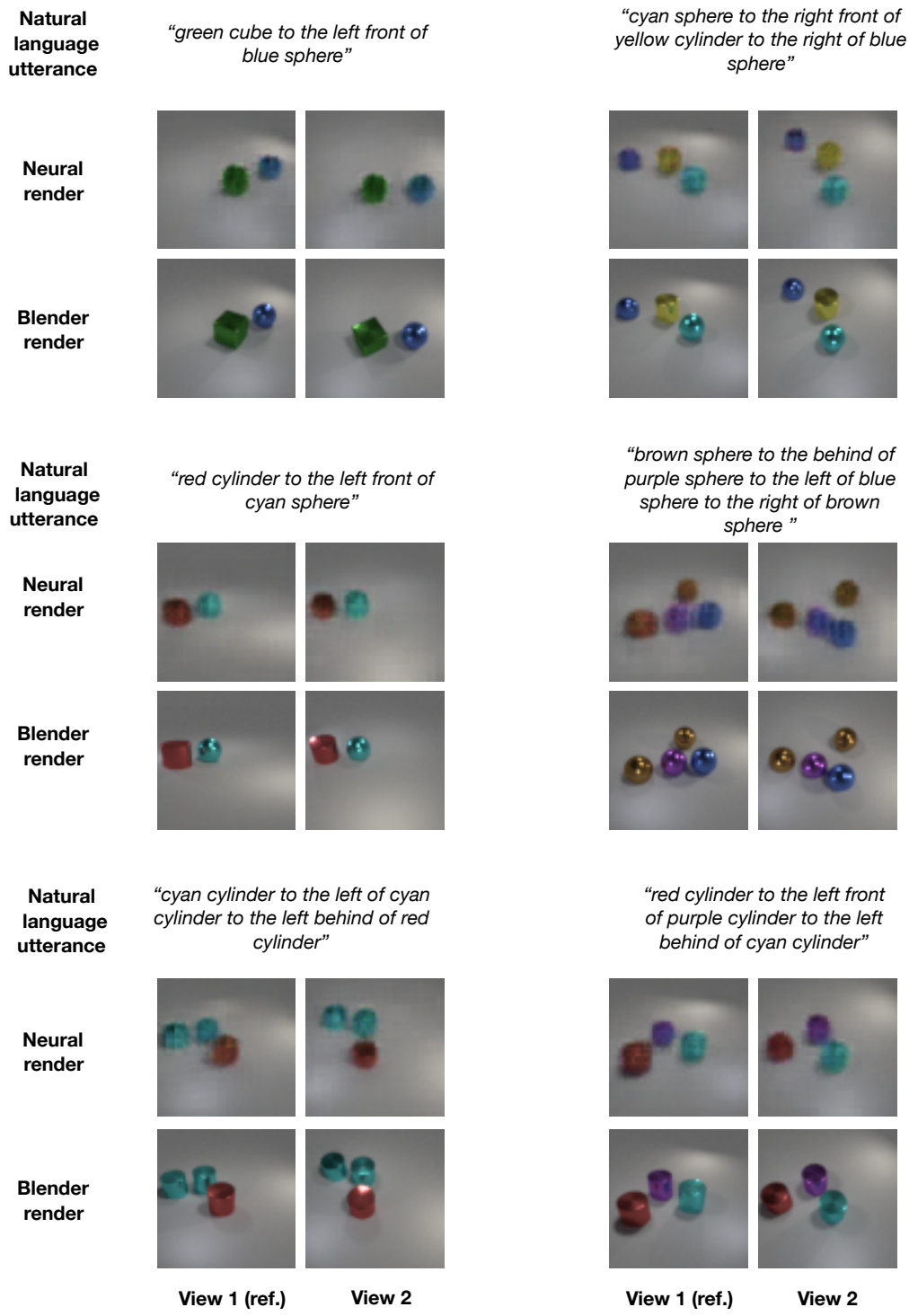

Figure 5: **(Additional) Natural language conditioned neural and blender scene renderings generated by the proposed model.**

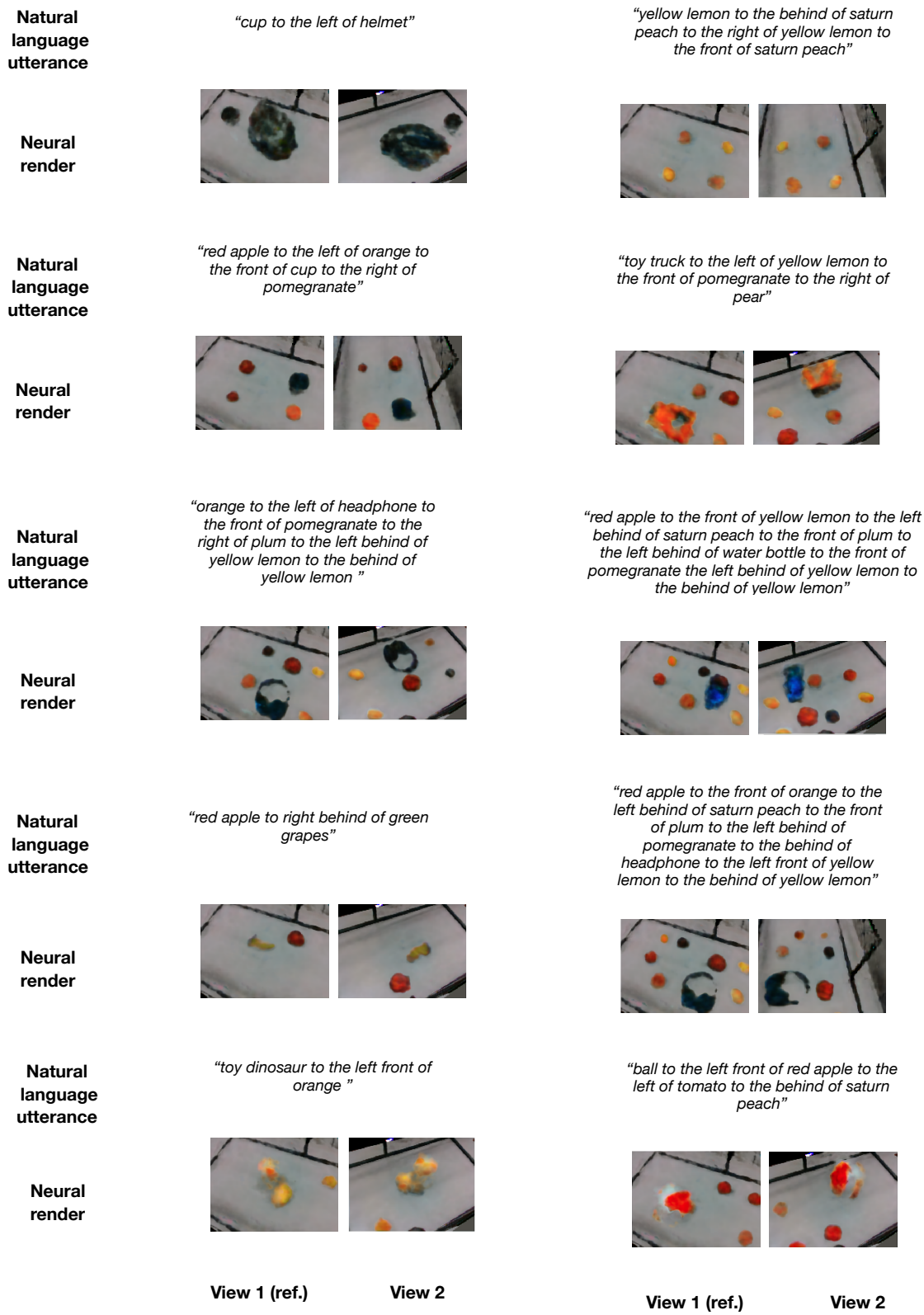

Figure 6: **(Additional) Natural language conditioned neural scene renderings generated by the proposed model over our real world dataset.**

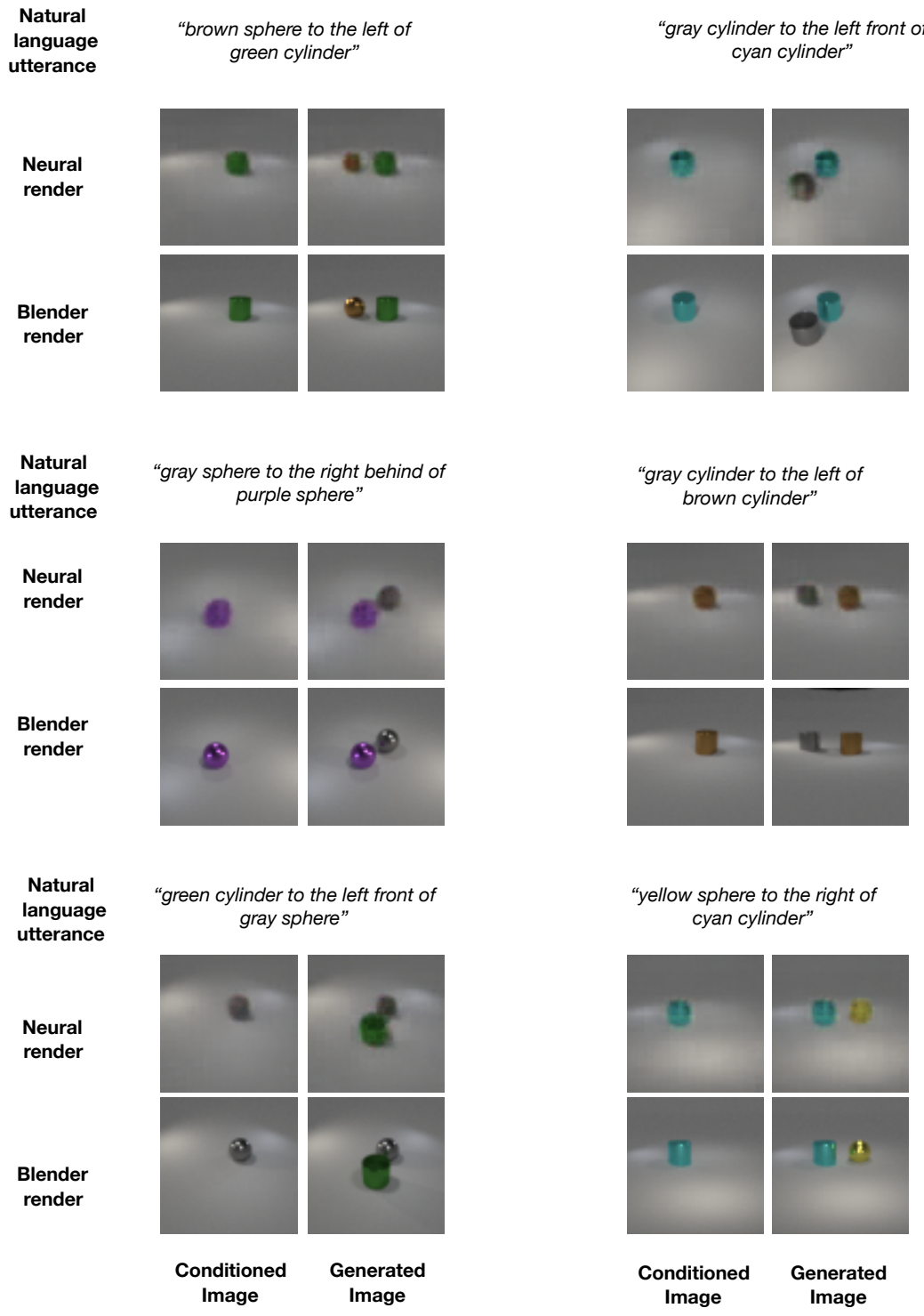

Figure 7: **Neural and blender scene renderings generated by the proposed model, conditioned on natural language *and* the visual scene.** Our model uses a 3D object detector to localize objects in the scene, and the learnt 2D-to-3D unprojection neural module to compute a 3D feature tensor for each, by cropping accordingly the scene tensor. Then, it compares the natural language conditioned generated object tensors to those obtained from the image, and grounds objects references in the parse tree of the utterance to objects presents in the environment of the agent, if the feature distance is below a threshold. If such binding occurs, as is the case for the "green cube" in top left, then, our model used the image-generated tensors of the binded objects, instead of the natural language generated ones, to complete the imagination. In this way, our model grounds natural language to both perception and imagination.

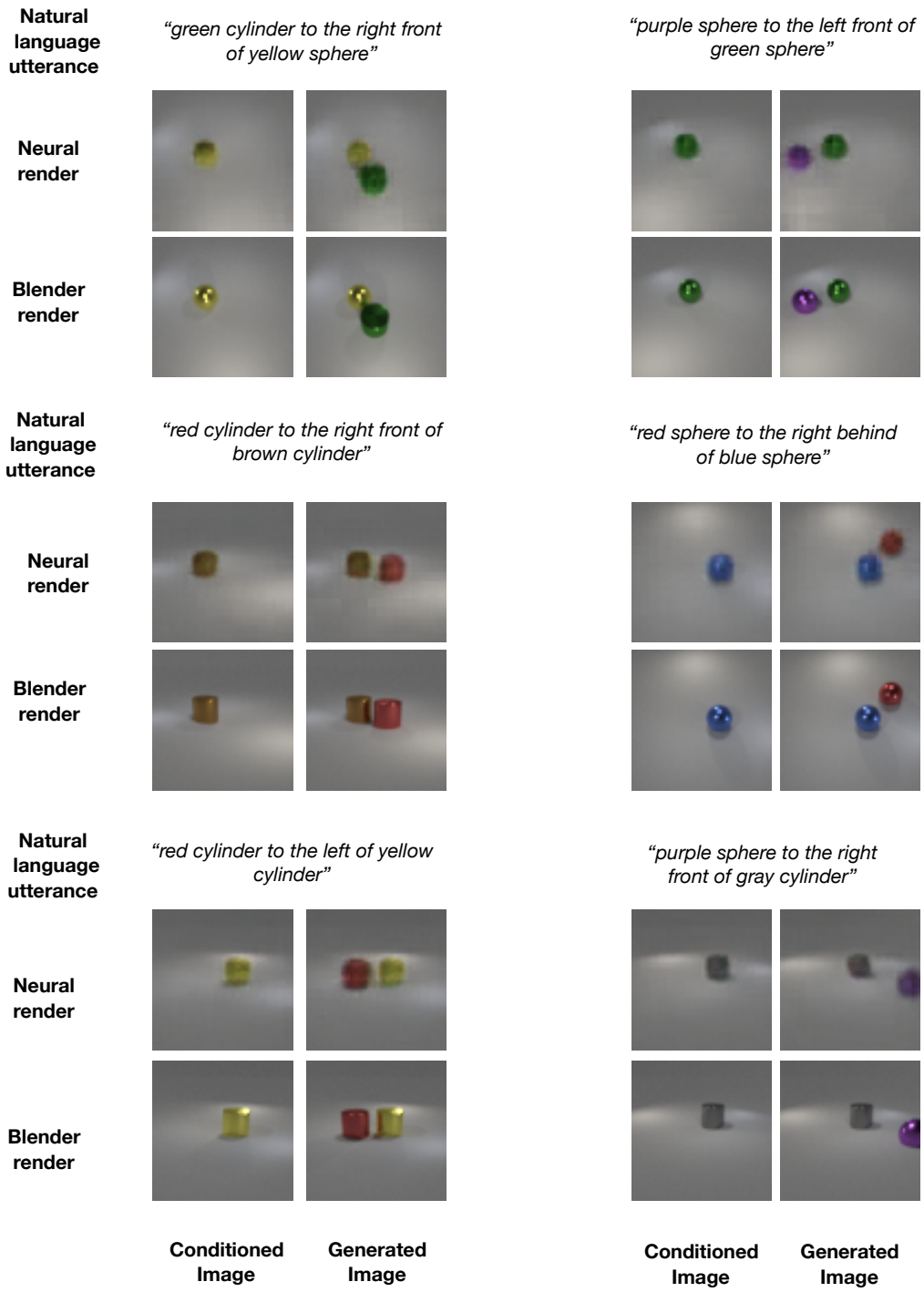

Figure 8: **(Additional) Neural and blender scene renderings generated by the proposed model, conditioned on natural language *and* the visual scene.**

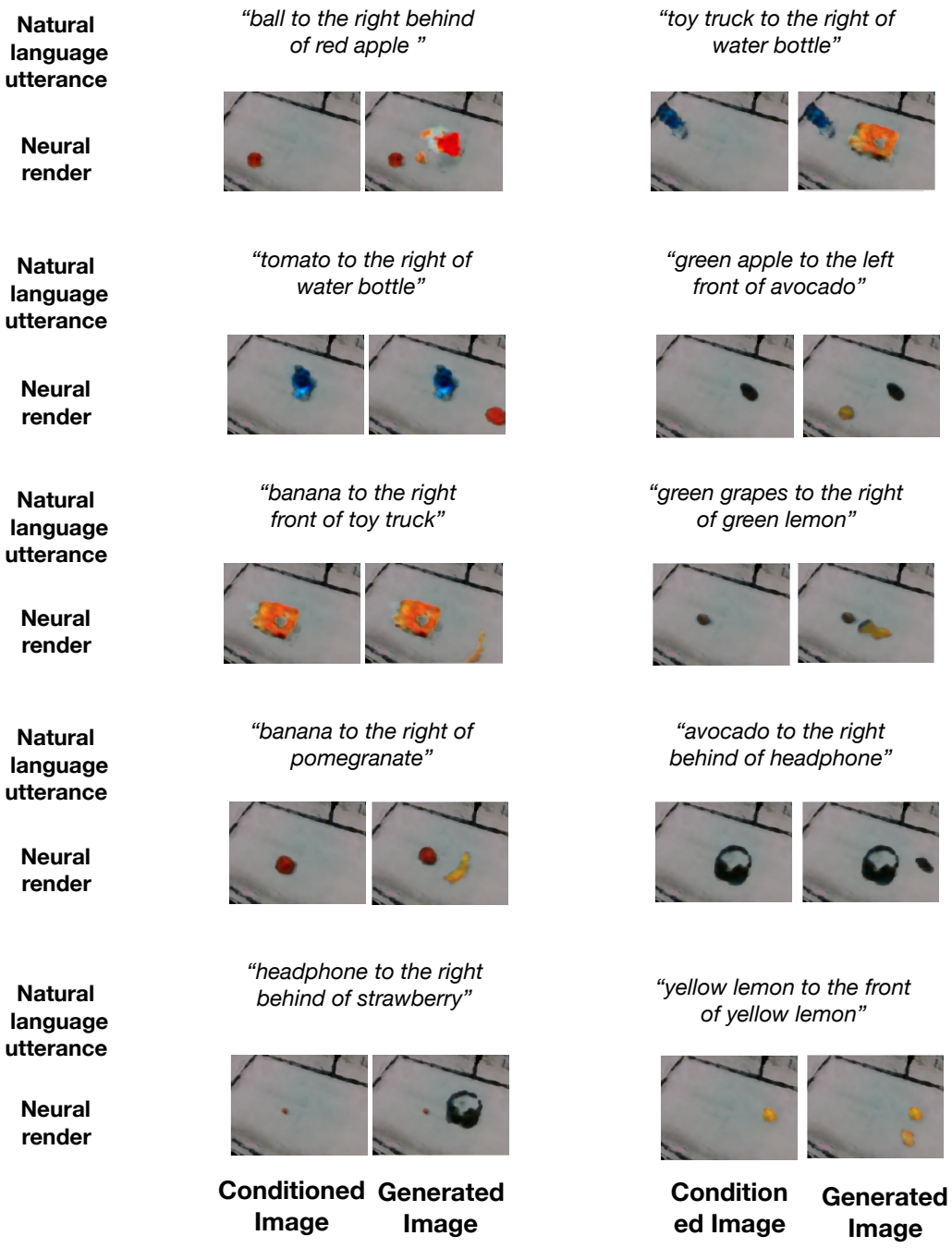

Figure 9: **(Additional) Neural scene renderings generated by the proposed model, conditioned on natural language *and* the visual scene from our real world dataset.**

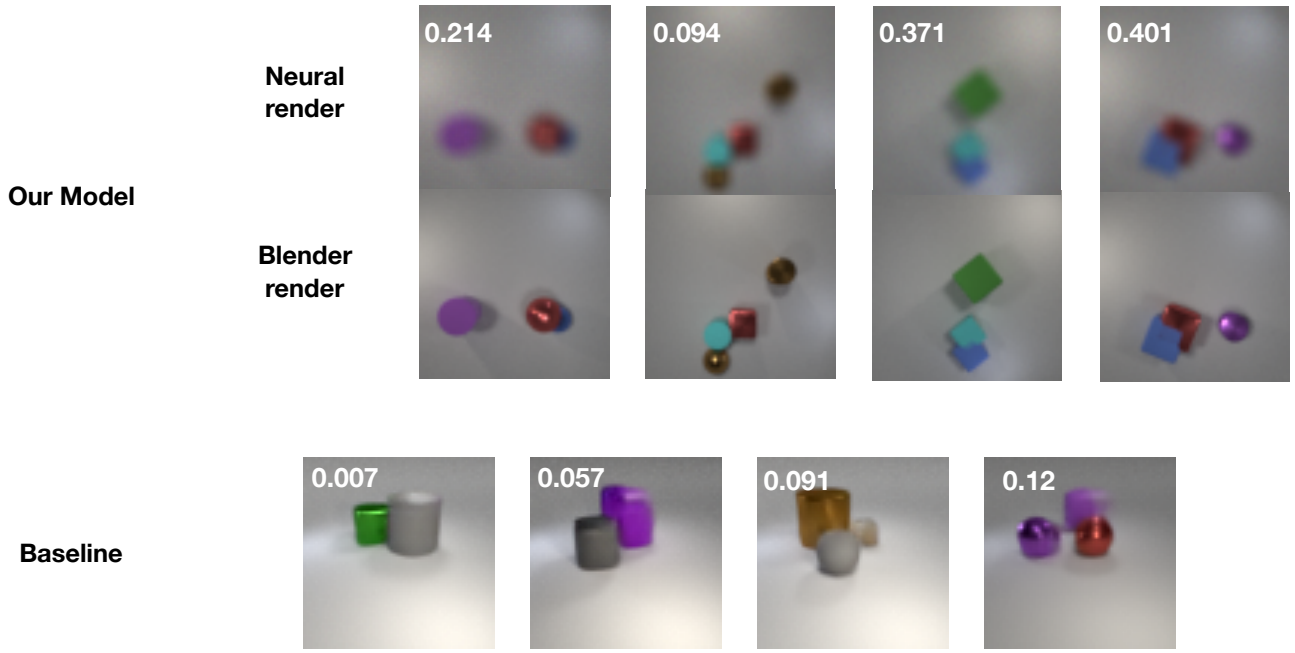

Figure 10: **Affordability prediction comparison of our model with the baseline work of (5).** In the top 2 rows, we show the Neural and Blender renderings of our model. Since we reason about the scene in 3D, our model allows checks for expression affordability by computing the 3D intersection-over-union (IoU) scores. In contrast, the bottom row shows the baseline model which operates in 2D latent space and hence cannot differentiate between 2D occlusions and overlapping objects in 3D.

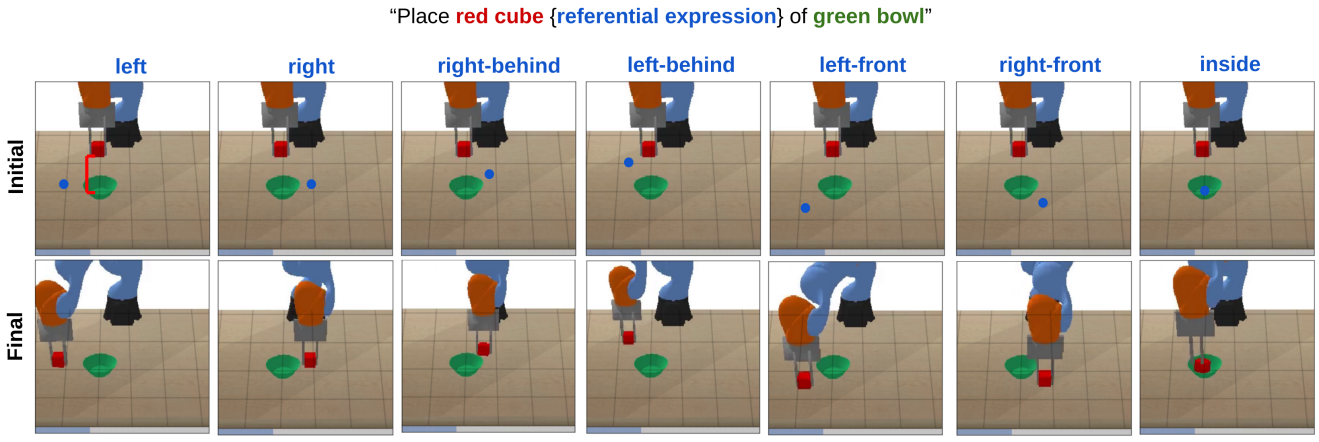

Figure 11: **Language-guided placement policy learning.** We show the final configurations of the learned policy using different referential expressions for the utterance "Place red cube {referential expression} of green bowl." *Top:* Initial robot configuration with the goal position generated by our method indicated as a blue dot. *Bottom:* Final robot configuration. We can see that the robot successfully places the cube with respect to the bowl according to the given referential expression.