# On the Robustness of Language Guidance for Low-Level Vision Tasks: Findings from Depth Estimation

## Supplementary Material

This supplementary contains details of the activity-level semantic transformations, modified train-test split, distribution of scenes in Sun RGB-D we leveraged for our experiments, along with additional illustrations.

## A. Activity-Level Transformations

Table 8 presents the transformations we leveraged to evaluate the semantic, activity level understanding of language-guided depth estimators, as mentioned in Section 4. We curate these descriptions from ChatGPT.

## B. Train-Test Split : Out-of-Distribution Supervised Setting

In Table 9, we present the new train-test split as described in Section 5, for the supervised setting. Out of 24,231 images, 17,841 were used in training , while the remaining 6,390 were used in testing.

## C. Scene-Level Sentence Type Distribution

We present in Figure 9, an overall distribution of objects and the corresponding sentences created by our framework, across scenes on the NYUv2 Test Split. We use the NLTK tokenizer for calculating the average number of words in a caption. For the pairwise relationships, we present results on all objects in a scene, irrespective of their uniqueness. Hence, we notice a direct correlation between the # of objects and the # of relationships across scenes. Also, interestingly we find that the # of vertical relationships are lesser in comparison to depth and horizontal; this can be attributed to NYUv2 being an indoor dataset having lesser variation in height.

## D. Distribution of Scenes in SUN RGB-D

In Figure 10, we illustrate the # of images for a given scene type in the Sun-RGBD dataset. As shown in Section 5, despite having 50% overlap in the scene-type between its training and testing distribution, VPD has the largest drop in performance.

## E. Additional Illustrations

Figure 11 presents additional depth map illustrations as generated by models trained with varying types of natural language guidance. Lastly, in Figure 12, we present illustrative results when VPD is evaluated in a zero-shot setting across multiple language modalities.

| Original Scene Name | Activity Level Description |
|---|---|
| printer room | room to access and operate printing equipment |
| bathroom | room to attend to personal hygiene and grooming |
| living room | place to relax, socialize, and entertain guests in a house |
| study | room to focus on reading, learning, and researching |
| conference room | room to hold meetings and discussions |
| study room | room to concentrate on academic or professional tasks |
| kitchen | room to prepare and cook meals |
| home office | place to work on professional tasks from home |
| bedroom | room to sleep and rest in a home |
| dinette | place to have informal meals |
| playroom | place to engage in recreational activities and games for kids |
| indoor balcony | place to enjoy views and relax indoors |
| laundry room | room to clean and maintain clothing and fabrics |
| basement | place for storage, recreation, or utilities usually below ground level |
| exercise room | room to workout and engage in physical activities |
| foyer | area of the house to welcome guests and as an entryway |
| home storage | storage area in a house to store items and belongings |
| cafe | place to enjoy beverages and light meals in a social setting |
| furniture store | place to browse and purchase furniture items |
| office kitchen | place to prepare refreshments and snacks in an office |
| student lounge | place to relax and interact in a university or school setting for students |
| dining room | room to have formal meals with family or guests |
| reception room | room to welcome and accommodate visitors |
| computer lab | lab to use computers for learning or work purposes |
| classroom | room to attend educational lectures and lessons |
| office | place to carry out professional tasks and responsibilities |
| bookstore | place to browse and purchase books and literary materials |

Table 8. Original scene names in the NYUv2 dataset and their corresponding activity level descriptions.

| Train Scenes | Test Scenes |
|---|---|
| printer room, bathroom, living room, study, conference room | student lounge, dining room, reception room |
| study room, kitchen, home office, bedroom, dinette, playroom | computer lab, classroom, office, bookstore |
| indoor balcony, laundry room, basement, exercise room | foyer, home storage, cafe, furniture store, office kitchen |

Table 9. Modified Train-Test Split of the NYUv2 dataset, as described in the scene distribution supervised setting.

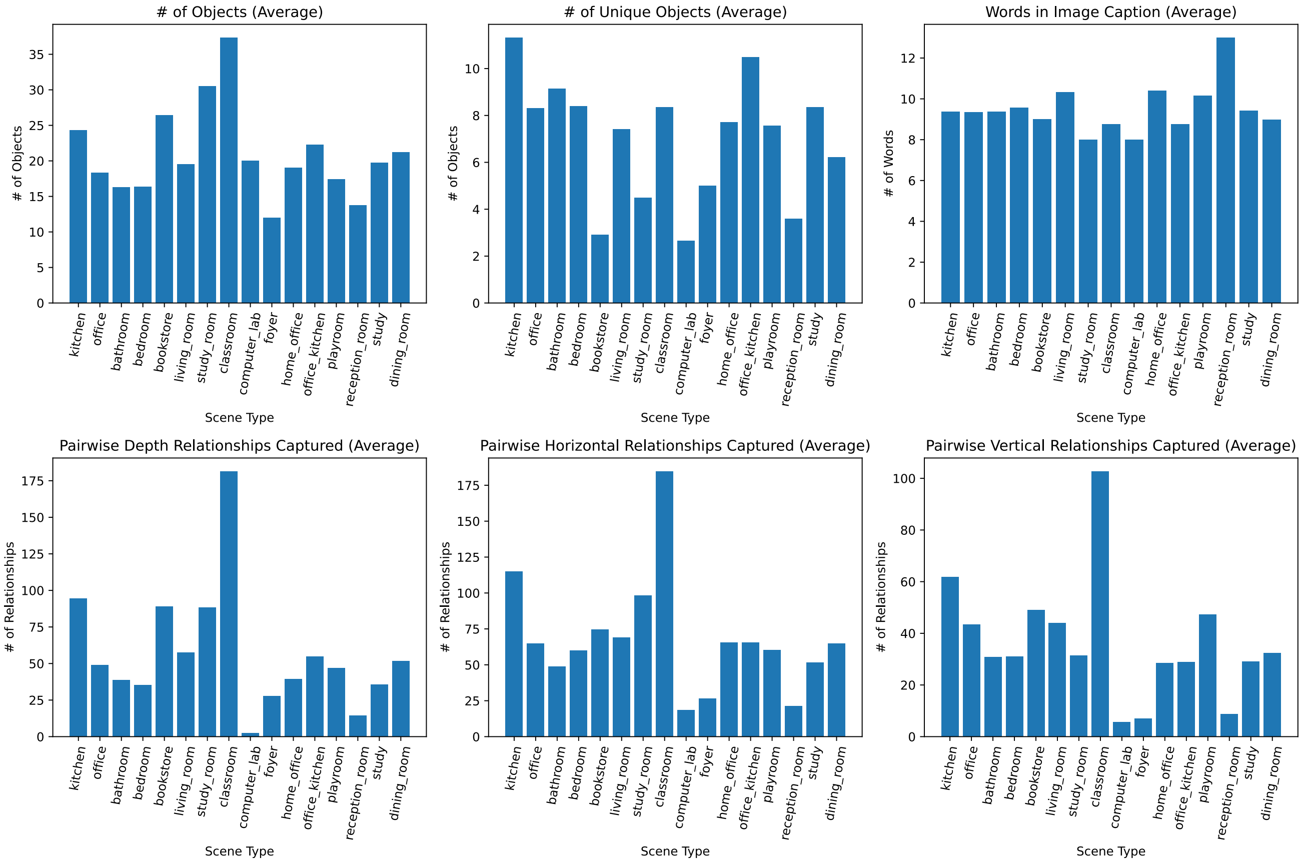

Figure 9. Graphical illustration of the average number of objects (complete and unique), average number of words in an image caption and the number of spatial relationships captured, across scene types. Results shown on NYUv2 test split.

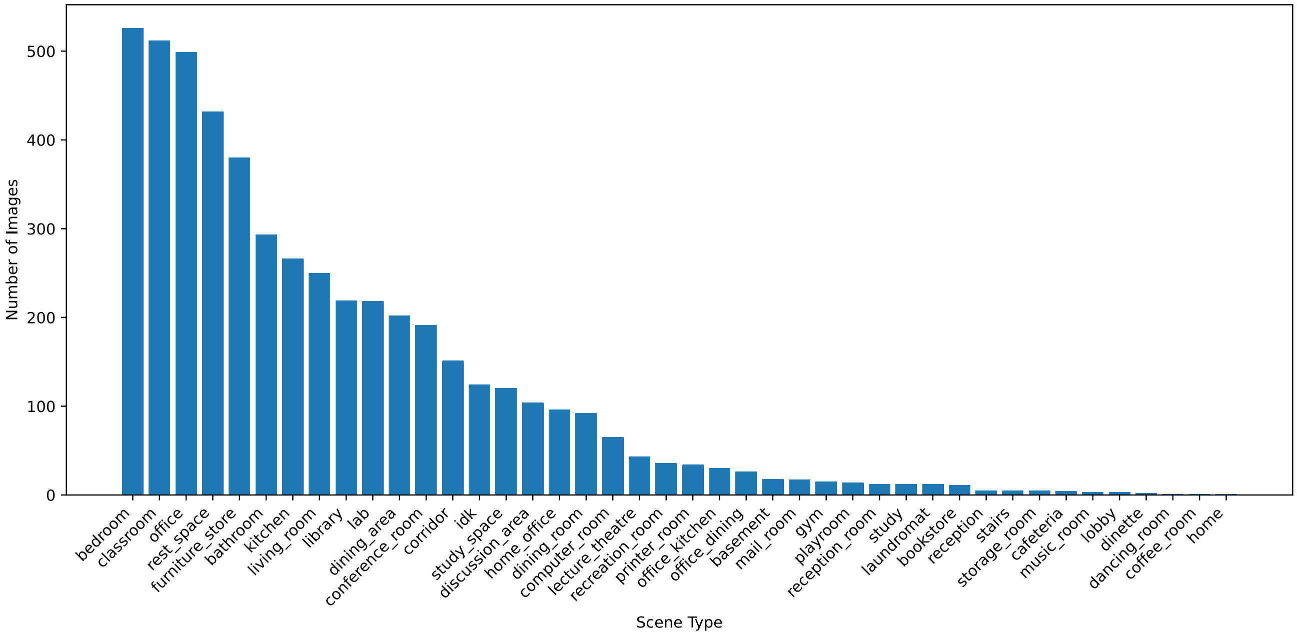

Figure 10. Number of instances per scene types in the SUN RGB-D dataset.

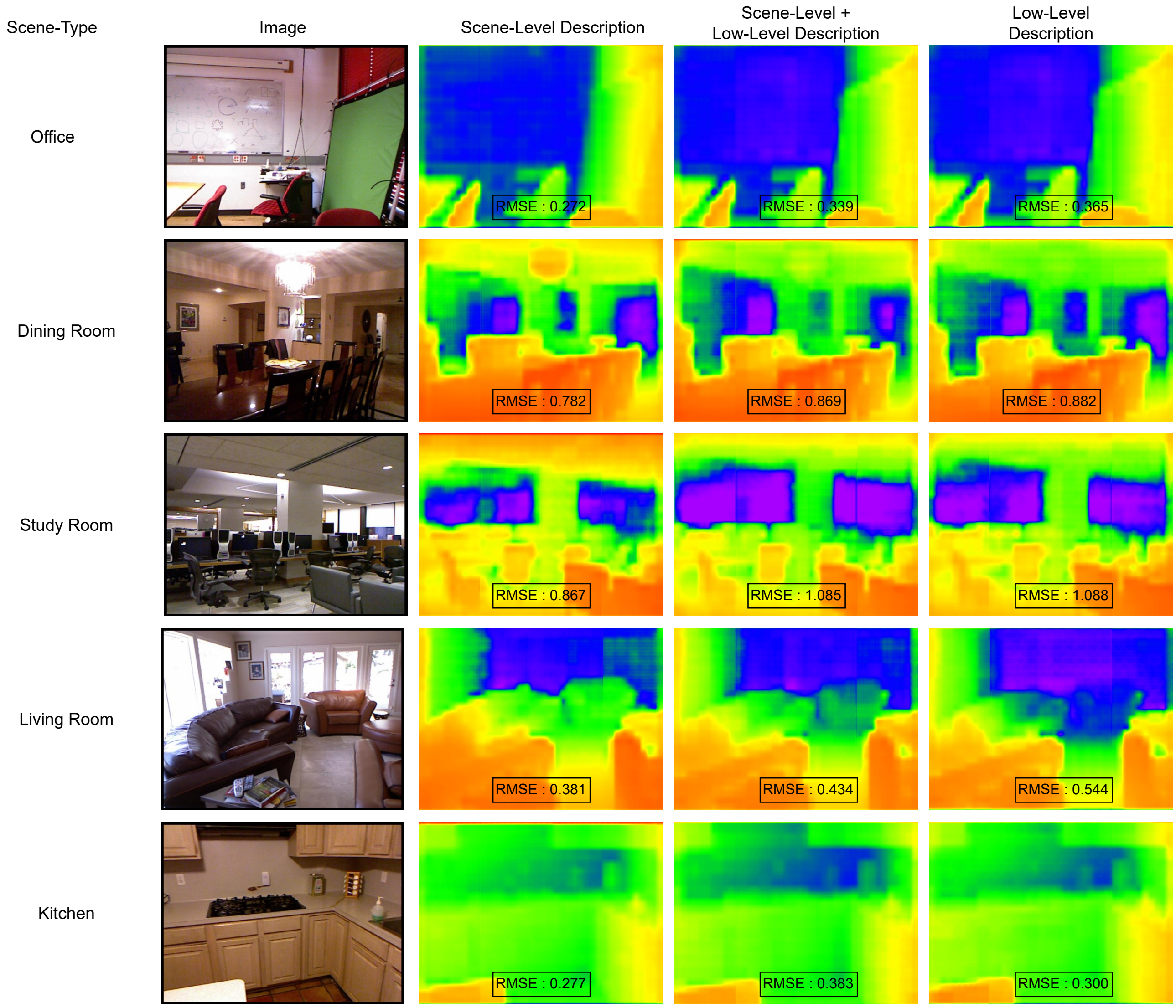

Figure 11. Comparison of Generated Depth Maps, trained in a **supervised** setting, across 5 different scene-types and 3 kinds of natural language guidance.

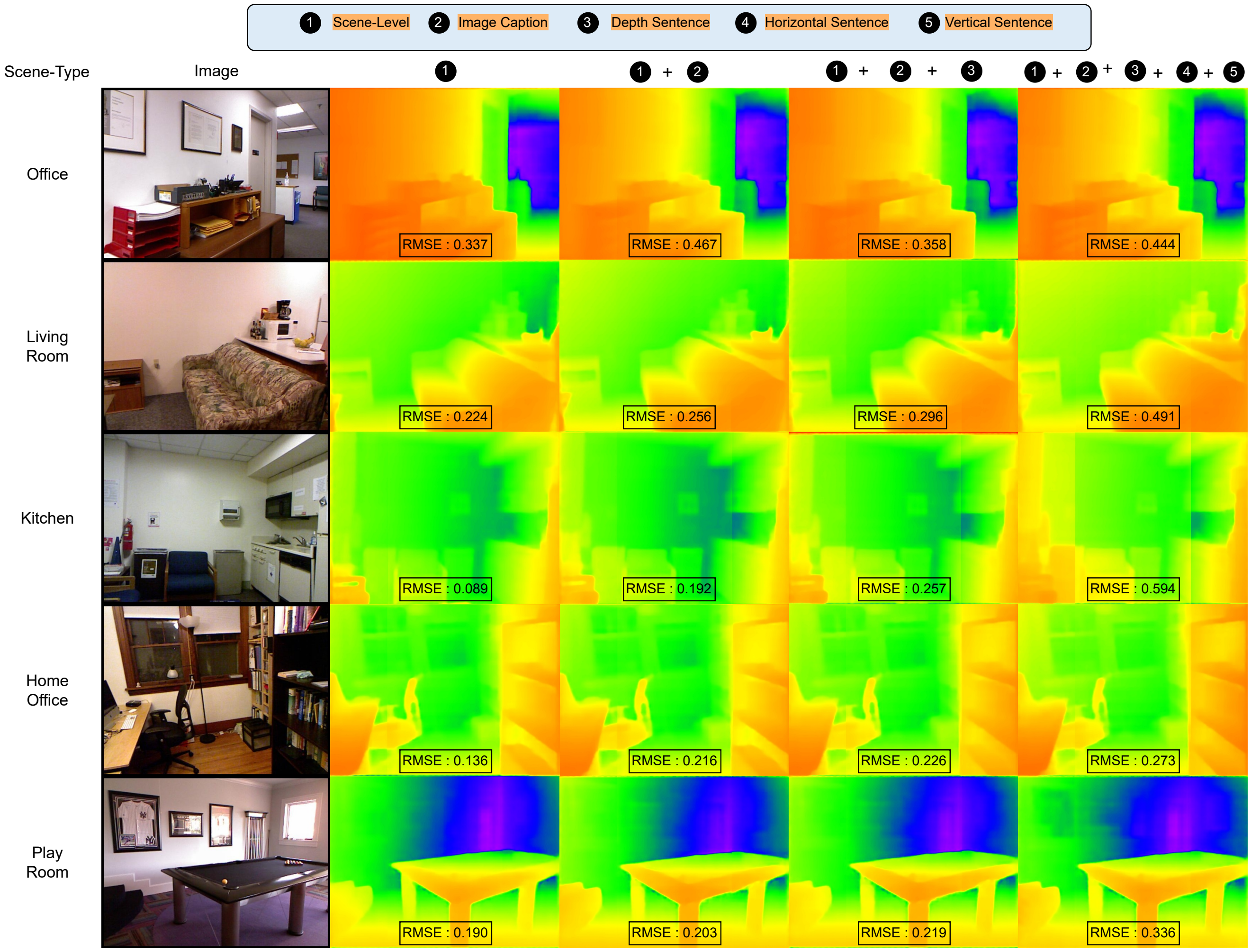

Figure 12. Comparison of Generated Depth Maps, when evaluated in a **zero-shot** setting, across 5 different scene-types and 4 kinds of natural language guidance.