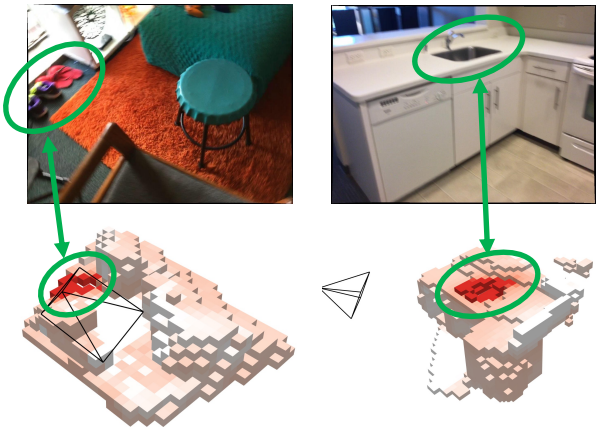
Monocular Open Vocabulary Occupancy Prediction for Indoor Scenes

Supplementary Material

6. Experimental Results

Text-conditioned 3D grounding examples. The two examples in Fig. 6 show that our open-vocabulary occupancy model can localizes fine-grained indoor categories directly in 3D.



Shoes Sink

Figure 6. **Text-conditioned 3D grounding.** Given a reference image and a free-form text query (left: *shoes*; right: *sink*), our language-embedded occupancy highlights corresponding voxels in red; darker red indicates a higher likelihood for the queried category. Green circles and arrows denote image-to-3D correspondence, and the camera frustum denotes the viewing pose.

Analysis on class ambiguity. In our experiments, two error modes emerged: (i) broad catch-all labels (*furniture*/*objects*) lack distinctive geometry and are inconsistently segmented in 2D, leading to mixed features; (ii) visually similar classes (e.g., *sofa*/*chair*) share shape cues under partial views. These findings highlight the importance of reliable semantic supervision and point to practical remedies (e.g., using a hierarchical label space and attribute-augmented prompts for broad categories, or applying pairwise contrastive calibration for look-alike classes.)

6.1. Ablation Studies

Table 5. **Ablation on the number of rendered views for feature alignment.**

number render frames	mIoU	IoU
3	20.58	59.11
5	21.05	59.50
7	20.92	59.19

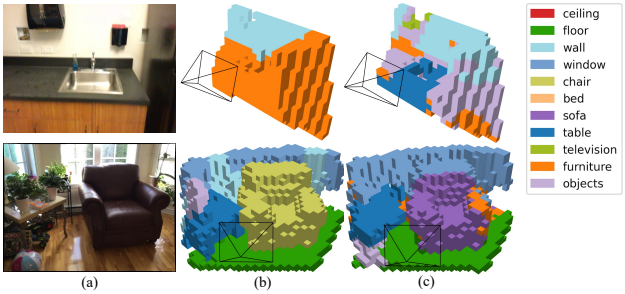


Figure 7. **Analysis of class ambiguity under OV training, closed-set evaluation.** (a) Input image. (b) Ground-truth semantic occupancy. (c) Our open-vocabulary prediction. Top row: frequent confusion between broad categories *furniture* vs. *objects*. Bottom row: confusion between *sofa* and *chair*.

Ablation on the number of rendered views for feature alignment. We ablate the number of rendered views used in L_{feat} in Eq. (11): Gaussians from the current frame are rendered into K nearest views and the rendered features are aligned to frozen open-vocabulary features. As shown in Tab. 5, K=5 gives the highest mIoU/IoU. The small performance variance across K indicates the robustness of our alignment scheme.

Ablation on image size. As shown in Tab. 6, our method attains competitive accuracy even at a small input size of 280, while running in real time. Larger resolutions further improve mIoU and IoU but reduce FPS, suggesting that our approach is well-suited for practical real-time deployment under constrained compute budgets.

Table 6. **Ablation on image size.**

image size	mIoU	IoU	FPS
280	18.99	55.83	54.89
420	20.00	58.36	41.18
518	21.05	59.50	22.47

6.2. Qualitative Visualization

We present qualitative results under both closed in Fig. 8 and open vocabulary Fig. 9 settings.

7. Limitations

While effective under geometry-only supervision, our approach still has two limitations. *(i) Broad/heterogeneous categories.* Classes such as *furniture* and *objects* show high intra-class variability and weak geometric salience; despite

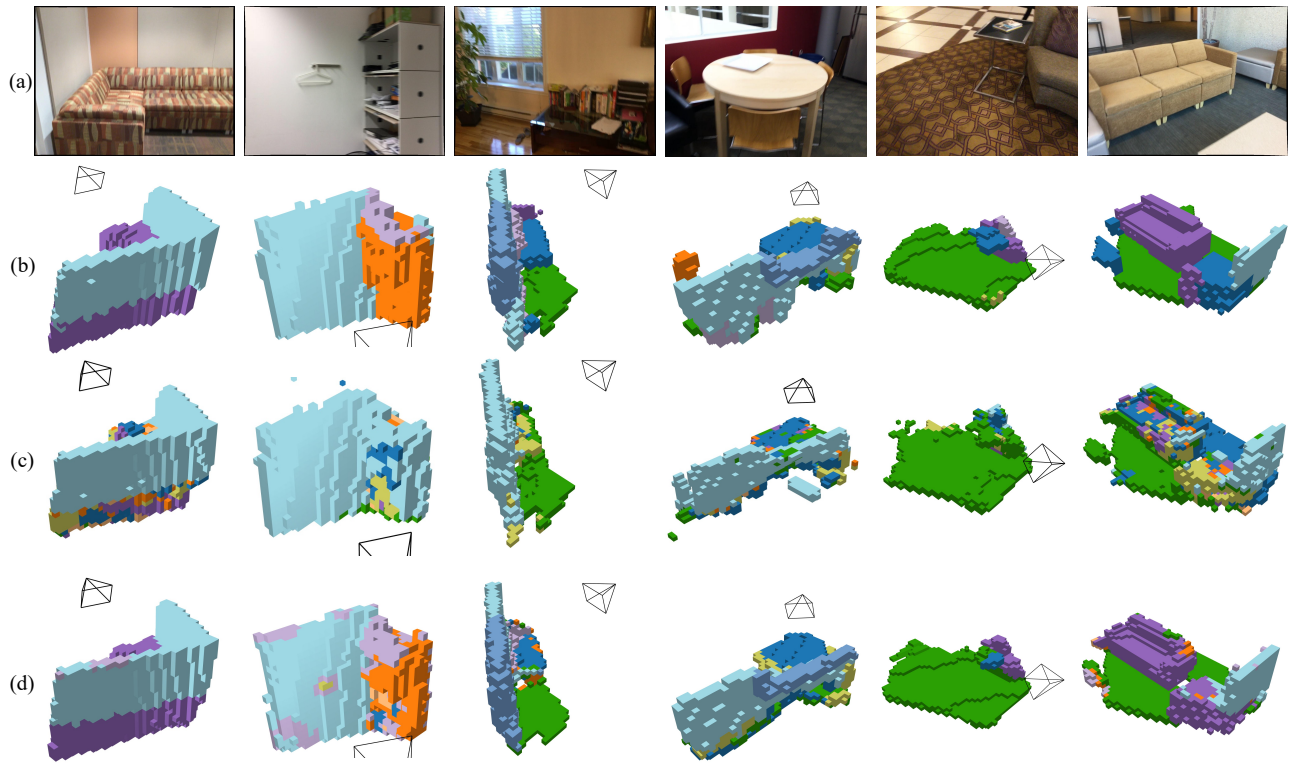


Figure 8. **Qualitative results on Occ-ScanNet.** From top to bottom: (a) input images; (b) ground-truth semantic occupancy; (c) results from the re-implemented LOcc [52] for monocular open-vocabulary occupancy prediction; (d) results of our method. Each column shows a distinct scene and camera viewpoint.

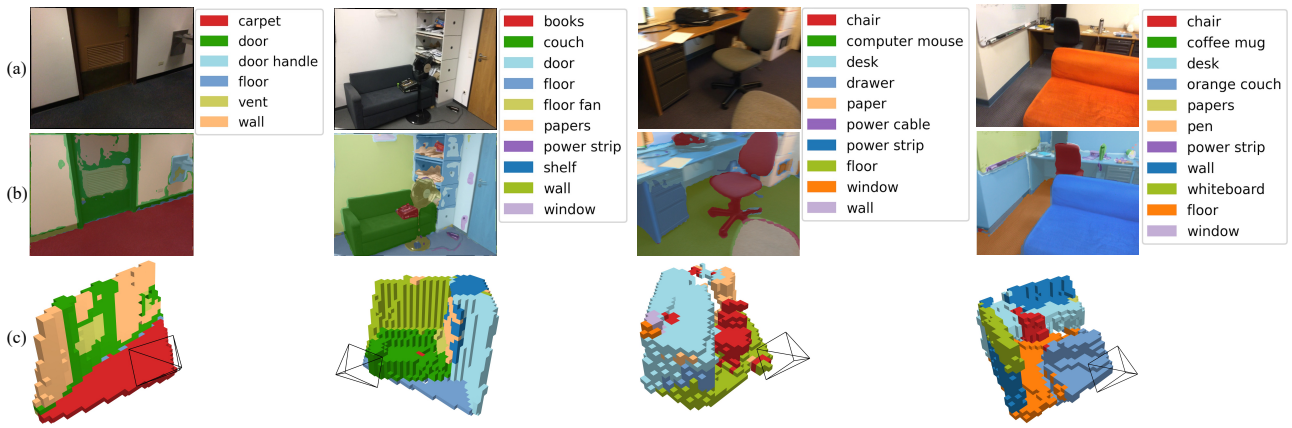


Figure 9. **Open-vocabulary qualitative results.** Legends list the *VLM-extracted object nouns* used as text queries. (a) Input image. (b) Open-vocabulary 2D segmentation for the queried nouns. (c) Our 3D open-vocabulary occupancy, colored by the same category names.

language alignment, their calibration remains challenging and can dilute discrimination. *(ii) Reliance on 3D geometry labels.* Training requires binary occupancy supervision. We find purely self-supervised optimization unstable in highly cluttered indoor scenes; relaxing dependence on explicit 3D occupancy labels (e.g., via stronger multi-view priors or self-distillation) is an important next step.