Supplementary Material for Open-NeRF: Towards Open Vocabulary NeRF Decomposition

Hao Zhang, Fang Li, Narendra Ahuja
University of Illinois Urbana-Champaign
{haoz19, fangli3, n-ahuja}@illinois.edu

In this section, we introduce further details that, due to space constraints, we were unable to incorporate into the main body of the paper. The expanded discussions are categorized into four parts. **(1)** We introduce the novel scenes, which are captured from a diverse range of scenarios, along with examples of our annotations. The purpose of these is to quantitatively evaluate the capability of our model in managing the NeRF decomposition with open vocabularies. **(2)** In the main manuscript, we provide the quantitative comparisons between our proposed approach, Open-NeRF, and LERF [3] within the *Desktop* scene context, and here we deliver more qualitative comparisons. The comparison underlines the superior performance of our method over its counterpart. **(3)** As highlighted in the main manuscript, the open-vocabulary NeRF decomposition capacity of Open-NeRF opens up possibilities for practical applications, particularly in the realm of NeRF editing. To substantiate this assertion, we showcase the results of NeRF decomposition achieved through our approach. Furthermore, we present the outcomes of texture modifications resulting from Open-NeRF outputs, thereby providing tangible evidence of the practical usefulness of our method. **(4)** We reveal the results generated through the use of LSeg [5] and show the shortcomings that LSeg is challenged with scenarios that extend beyond its conventional realm as mentioned in the main manuscript.

1. Proposed Novel Scenes

The evaluation dataset we introduce is compiled from a diverse range of 8K video frames sourced from mobile phones. Every scene within the dataset contains between 150 to 300 image observations, seized from varying viewpoints with distinct scene coverage angle ranges. To further aid the evaluation of decomposition accuracy, we deliver ground truth 2D segmentation for all new scenes collected by us and also the data from Mip-nerf 360 [1]. And we provide an annotation every 10 frames. As shown in Figure 3, we labeled multiple objects in each scene. The integration of these unconstrained in-the-wild scenes allows for a more comprehensive and pragmatic appraisal of Open-NeRF's efficacy in addressing diverse and challenging real-world scenarios.

As demonstrated in Figure 4 of the main manuscript, the scene, denoted as *'Desktop'*, includes a plethora of common items such as 'phones', 'computers', and 'boxes', in addition to less common objects like 'VR glasses', all housed in a moderately dense space. The setup provides a platform to assess the model's proficiency in the accurate decomposition of an open-vocabulary query within a reasonably congested context.

Figure 5, as presented in the main manuscript, illustrates examples of the *'Toy-1,2,3'* and *'Car'* scenes. The *'Toy'* scene categories encompass five divergent scenarios with a variety of toys, including 'Plush Toys', 'Legos', and 'PVC Figuarts', each situated in distinct environments. These provide long-tail samples where each object may correlate with different textual queries, enabling an assessment of the model's capacity to interact with novel objects and process open-vocabulary queries.

The *'Car'* category is made up of three separate collections of image observations, each featuring one or more vehicles within the scenes. One such collection was sourced in a garage setting, while the remaining scenes originate from an outdoor environment. These diverse scenarios further facilitate the examination of Open-NeRF's performance under various conditions.

2. Experiments

2.1. Quantitative Results

In the main manuscript, we show the quantitative results of our proposed method compared with LERF [3] in the scene: *Desktop* and here we provide the visual results as shown in Figure 1. LERF struggles to provide accurate segmentation results for the objects while only providing a rough localization. However, Open-NeRF is able to provide decent 3D segmentation results

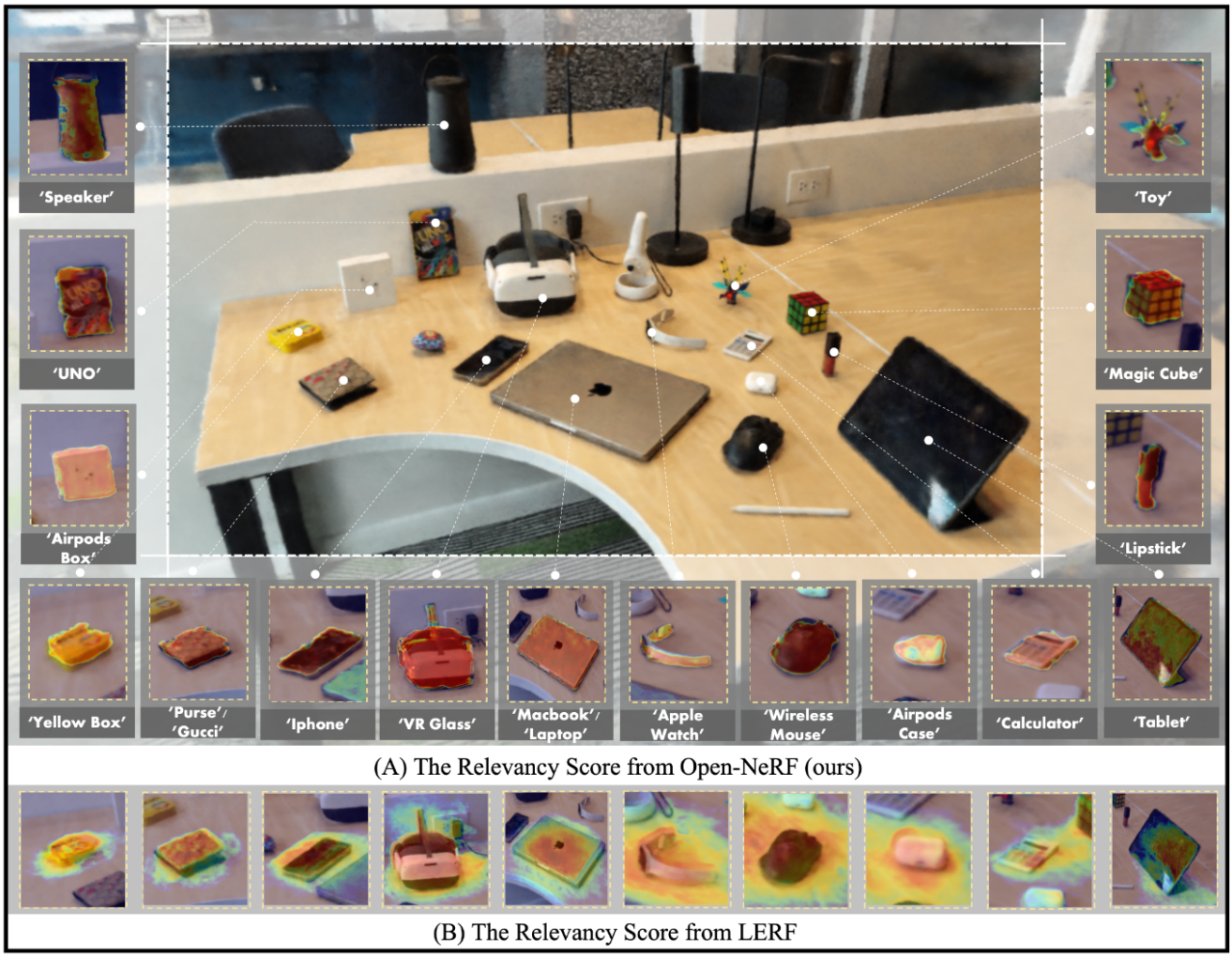

Figure 1. Relevance scores obtained by Open-NeRF and LERF in the *'Desktop'* scene.

2.2. NeRF Decomposition Results

As previously indicated in our main manuscript, Open-NeRF exhibits the capability to generate accurate 3D decomposition results for given open-vocabulary queries. This is visually demonstrated in Figure 2, which presents the decomposition results for the queries 'Lego excavator' within the *Kitchen* scene, and 'rabbit toy' within the *Toy-4* scene.

Further to this, Figure 5 in the main paper displays the outcomes of texture modifications; specifically, transitioning the textures of 'table' and 'table and vase' to a 'frozen' style[1], based on the results obtained from Open-NeRF. This illustrates the utility of our approach in modifying real-world textures within the 3D decomposition context.

In terms of NeRF decomposition, we employ a threshold strategy for relevancy scores, maintaining only the points that score higher than the predetermined threshold.

[1]For the implementation we followed CLIP-NeRF [7].

2.3. Qualitative Results of LSeg-based Methods

In the main manuscript, we present a qualitative and quantitative comparison between our proposed methodology, Open-NeRF, and LERF [3]. The comparative analysis excludes results from methodologies like FFD [4] and N3F [6]. This omission stems from the fact that N3F is predicated on DINO [2], a technique that lacks the capability to align image embeddings with text embeddings originating from open-vocabulary queries. Additionally, the FFD methodology, based on LSeg [5], is heavily dependent on the performance of 2D CLIP-LSeg. Regrettably, CLIP-LSeg struggles to yield satisfactory outcomes for scenes that contain novel objects or require the processing of open-vocabulary queries.

As Figure 4 illustrates, CLIP-LSeg can provide acceptable results for common objects within scenes, such as 'ta-

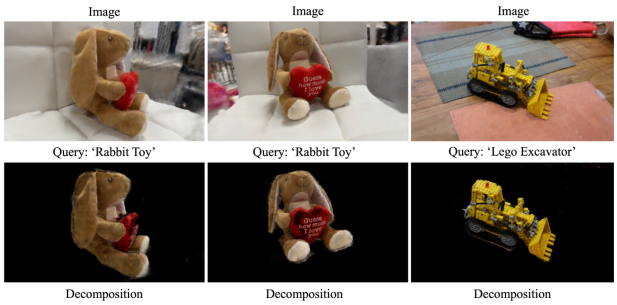

Figure 2. NeRF decomposition results from Open-NeRF with queries: 'rabbit toy' and 'Lego excavator'.

ble', 'vase', and 'grass' in the *'Garden'* scene sourced from Mip-nerf 360 [1]. However, it fails to adequately segment the 'football' within the scene. Moreover, in our collected scenes that comprise a multitude of novel objects such as 'Lava Monster', 'Harry Potter', and 'Lego', CLIP-LSeg consistently fails to yield usable results. It also struggles to handle open-vocabulary queries, providing no reasonable outcomes for queries such as 'Gucci', or 'Yellow Boxes'. Owing to its inability to yield satisfactory results in 2D, the 3D decomposition methodologies premised on LSeg likewise fail to deliver appropriate outcomes for NeRF decomposition.

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 1, 3

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021. 2

[3] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. *arXiv preprint arXiv:2303.09553*, 2023. 1, 2

[4] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022. 2

[5] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 1, 2

[6] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *2022 International Conference on 3D Vision (3DV)*, pages 443–453. IEEE, 2022. 2

[7] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields, 2022. 2

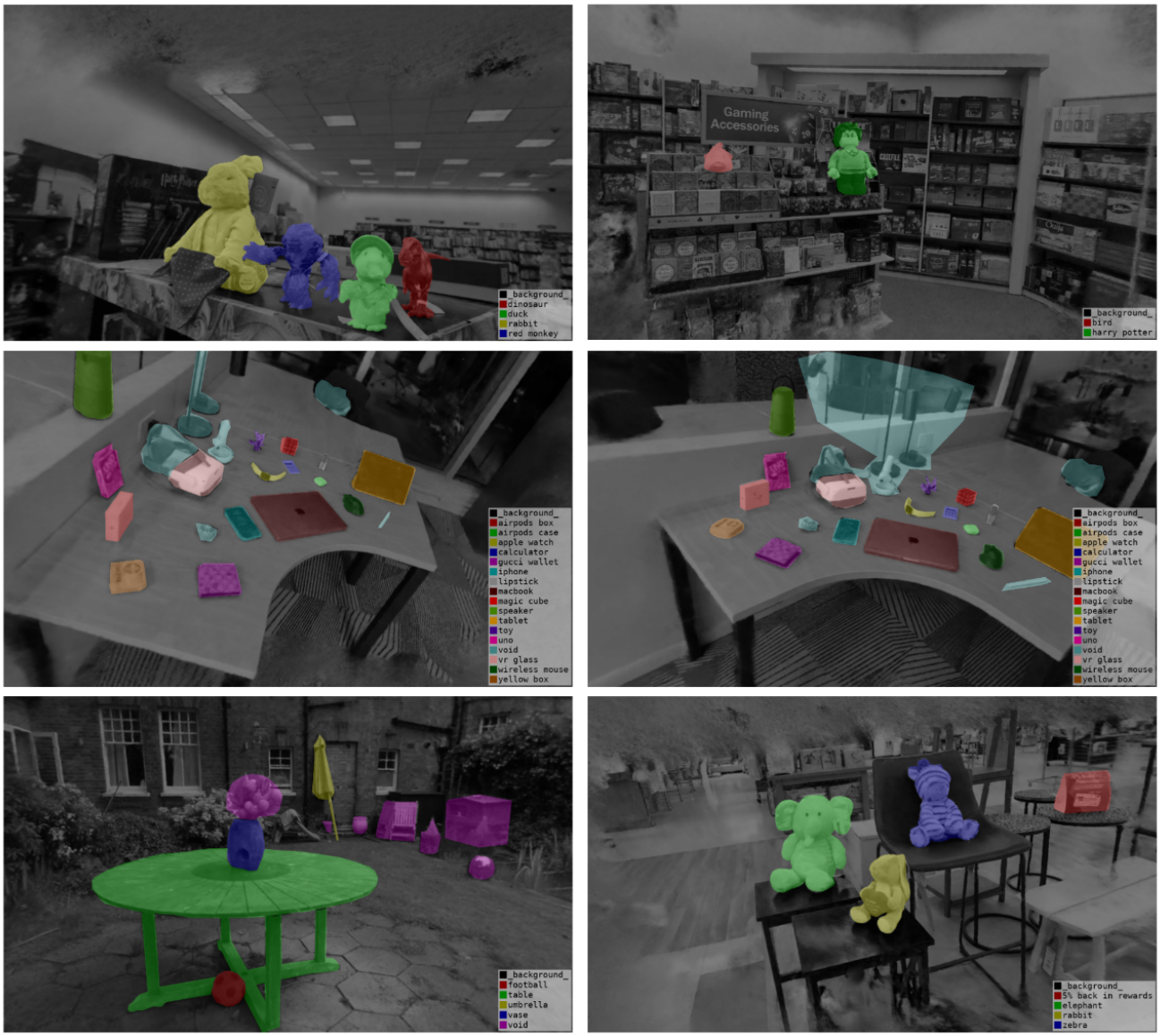

Figure 3. Examples of semantic annotations for the scenes.

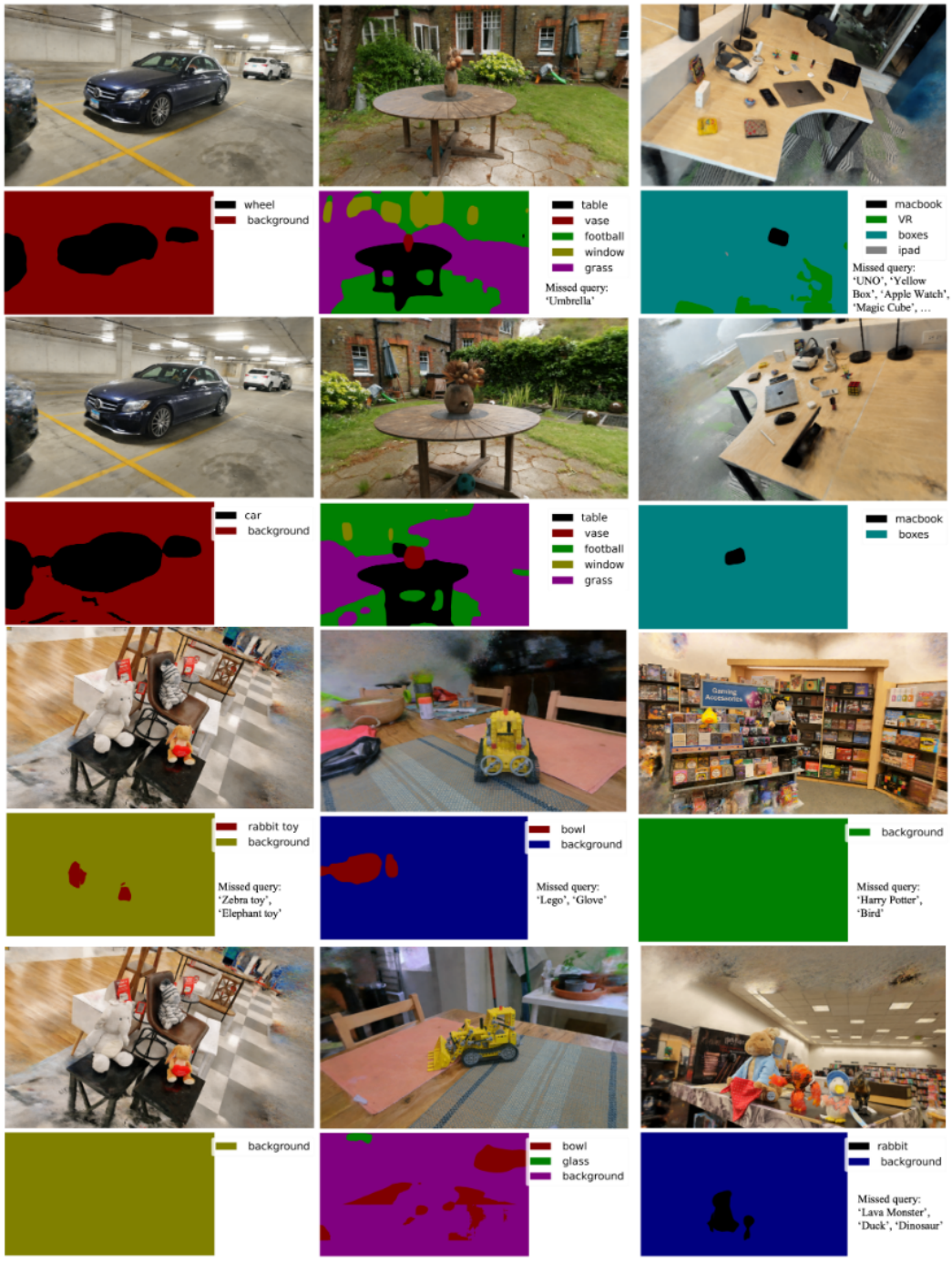

Figure 4. Results of LSeg in 2D open-vocabulary 2D segmentation on multiple scenes.