ReasonGrounder: LVLM-Guided Hierarchical Feature Splatting for Open-Vocabulary 3D Visual Grounding and Reasoning

Supplementary Material

1. DATASETS

This paper selects the LERF dataset, 3D-OVS dataset, and our proposed ReasoningGD dataset for both training and evaluation. The LERF and 3D-OVS datasets are widely used to evaluate the open-vocabulary 3D visual grounding performance of various methods. However, these datasets only provide explicit queries, lacking implicit, descriptive ones. For instance, they utilize a explicit query like "apple" rather than a implicit query such as "red nutrient-rich sweet fruit," which does not explicitly name the object. To enhance the utility of these two widely used datasets in evaluating our ReasonGrounder's ability to handle implicit queries, we have added additional annotations for implicit queries. Open-vocabulary 3D visual grounding and reasoning methods must also achieve amodal perception, meaning they should fully identify an object's structure and shape even when parts of it are occluded from a given angle of view. To address this, we introduce the ReasoningGD dataset, which includes over 10,000 scenes and more than 2 million modal and amodal annotations. These annotations cover both the visible mask area and the occluded portions of objects, enabling more comprehensive evaluation of amodal perception capabilities.

1.1. Widely used datasets

LERF dataset. As shown in Figure 7, LERF dataset comprises various scenes, encompassing both in-the-wild scenarios and posed long-tail scenes. These scenes are captured using the Polycam iPhone app and the featured images have a resolution of 994×738. Thanks to LangSplat's annotations of the four scenarios in this dataset, we can comprehensively evaluate the open-vocabulary 3D visual grounding capabilities.

- **(1)** ***Figurines***: The scene includes various toys and small objects arranged on a round table, such as a rubber duck, blue elephant figurine, Rubik's Cube, and other colorful items. A container with snack packets is also present.
- **(2)** ***Teatime***: The scene features two large plush toys, a bear and a sheep, seated at a wooden table as if hosting a tea party. The table is set with items like tea, an apple, cookies, and a mug.
- **(3)** ***Ramen***: The scene shows a bowl of ramen placed on a wooden table. The vibrant yellow bowl contains slices of pork, half-boiled eggs, narutomaki (fish cake), seaweed, and noodles. Chopsticks and a small cup, likely for soy sauce or sake, are set beside it.
- **(4)** ***Waldo kitchen***:The scene depicts a domestic kitchen with a vintage aesthetic, featuring white cabinets with metallic handles and a bright yellow countertop.

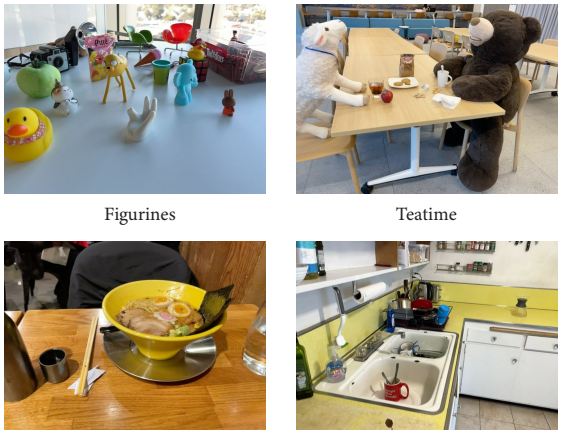

Figure 7. **Visualization of LERF dataset.** The LERF dataset comprises various scenes, encompassing both in-the-wild scenarios and posed long-tail scenes.

To further assess the ability to locate objects based on implicit queries, we have added additional implicit query annotations for these four scenarios. Each scene contains ten implicit queries for five objects, totaling 200 implicit queries across all four scenes. These annotations do not explicitly mention the name of the object to be queried but instead provide descriptive cues related to its characteristics. Below are examples of these scene annotations:

3D-OVS dataset. Each scene in the 3D-OVS dataset is accompanied by textual descriptions of objects, which guide the segmentation process. These descriptions are crucial for helping the model identify and segment objects in a scene based on their open-vocabulary labels. As a result, this widely used dataset is well-suited for evaluating open-vocabulary 3D visual grounding methods. As shown in Figure 8, this paper primarily conducts experiments based on five scenarios.

- **(1)** ***Bed***: The scene shows a selection of personal items arranged on a checkered fabric surface. Visible items include a red quilted handbag, a black loafer, a banana, and a digital camera held by a hand. The setup suggests a casual, stylish collection of fashion and everyday essentials.
- **(2)** ***Bench***: The scene displays a small collection of objects on a wooden surface, such as a blonde doll, a toy cat figurine, a miniature toy car, a bunch of green grapes, and an egg tart. The setting is outdoors, with a textured wall in the background.

Scene	Implicit Query
Old camera in **Figurines**	*It is a vintage device used for capturing photographs on film.*
Coffee mug in **Teatime**	*It is a sturdy vessel designed for holding hot beverages.*
Chopsticks in **Ramen**	*They are a pair of slender tools used for picking up food.*
Toaster in **Waldo kitchen**	*This is an appliance designed to brown slices of bread using heat.*

Table 8. **Examples of annotated implicit queries in LERF dataset.** These implicit annotations do not directly mention the name of the object to be queried; instead, they offer descriptive hints about the object's characteristics.

Figure 8. **Visualization of 3D-OVS dataset.** Each scene in the 3D-OVS dataset is accompanied by textual descriptions of objects, which guide the segmentation process.

- **(3) *Room***: The scene presents a playful arrangement of toys on a wooden table, including a rubber chicken in a wicker basket, a small rabbit figurine, a dinosaur toy, and a baseball.
- **(4) *Sofa***: The scene shows a variety of entertainment-related items on a gray surface. These include a plush toy in a festive costume, a stack of UNO cards, a pink gaming controller, a white Xbox controller, and a robot or mecha model.
- **(5) *Lawn***: The scene depicts a collection of items on a grassy surface, including a bottle of hand sanitizer, a red apple, a white baseball cap, a pair of black headphones, and a stapler.

To further assess the ability to localize objects based on implicit queries, we have added additional implicit query annotations for these five scenarios. Each scene contains ten implicit queries for four objects, totaling 200 implicit queries across all four scenes:

Scene	Implicit Query
Red bag in **bed**	*Search for a item typically filled with money and gifts.*
Orange cat in **Bench**	*It has a vibrant coat that ranges from light to dark orange.*
Base ball in **Room**	*It is a round object used in a popular bat-and-ball sport.*
UNO cards in **Sofa**	*It is typically played by two or more players.*
Stapler in **Lawn**	*It is a device used to fasten sheets of paper together.*

Table 9. **Examples of annotated implicit queries in 3D-OVS dataset.** These implicit annotations do not directly mention the name of the object to be queried; instead, they offer descriptive hints about the object's characteristics.

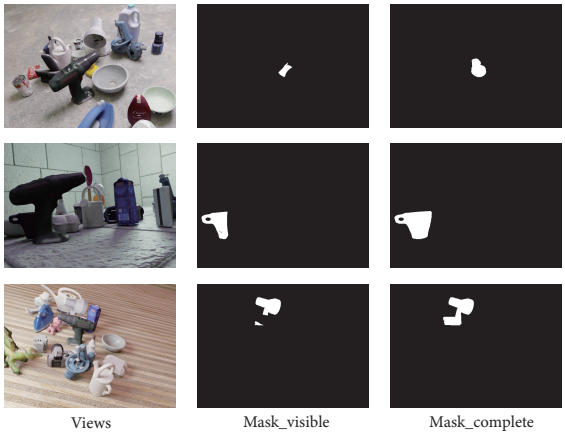

Figure 9. **Visualization of proposed ReasoningGD dataset.** ReasoningGD dataset encompasses a diverse range of occlusion scenarios and offers comprehensive, accurate annotations.

1.2. Proposed ReasoningGD dataset

Open-vocabulary 3D visual grounding and reasoning aim to localize objects in a scene based on implicit language descriptions, even when parts of the objects are occluded. However, the existing LERF and 3D-OVS datasets, which are primarily used to evaluate open-vocabulary 3D visual grounding, face limitations in assessing 3D reasoning capabilities. These include challenges in localization based on implicit instructions and identifying the full structural shape of occluded objects. These shortcomings stem from the lack of annotations for implicit instructions and the obscured portions of objects. As a result, these datasets lack the ground truth necessary for effectively evaluating open-

vocabulary 3D visual grounding and reasoning tasks.

This paper introduces the ReasoningGD dataset, which encompasses a diverse range of occlusion scenarios and offers comprehensive, accurate annotations. As shown in Figure 9, these annotations include both the visible and occluded parts of target objects from various perspectives, enabling more robust evaluation of reasoning capabilities in 3D visual grounding. The dataset comprises over 10K scenes, each featuring 10 to 15 objects. Each scene includes 100 viewing angles, with annotations provided for both the visible mask of each object at each angle and the full mask, which includes occluded parts. In total, the dataset contains over 2 million detailed annotations. The composition of each scene is primarily detailed in the Table 10.

Scene Composition	Description
Images	Stores the original training views
Points3d.ply	Includes 3D point cloud data of the scene
Mask_visible	The mask annotation of the visible parts of each object from different views
Mask_complete	The mask annotation of the full shape and structure of each object from different views
Transforms_test.json	Describes transformations for test data
Transforms_train.json	Describes transformations for training data

Table 10. **Composition and description of each scene in the ReasoningGD dataset.** The ReasoningGD dataset consists of diverse scenes designed to evaluate 3D visual grounding and reasoning capabilities under various conditions. Each scene is carefully structured with the key components.

2. IMPLEMENTATION DETAILS

Given training views with corresponding camera poses, we first construct a standard 3D Gaussian Splatting field. The training parameters used in this process align with those specified in the original paper. Across various scenes, ReasonGrounder employs consistent hyperparameters for uniformity. Each Gaussian in the field is assigned a 32-dimensional latent feature. This latent feature is subsequently mapped to hierarchical language features and hierarchical instance features. For object segmentation, we utilize the SAM ViT-H model to process the training views, generating object masks for each view. To supervise the hierarchical language features, we introduce the OpenCLIP ViT-B/16 model, extracting CLIP features of all object masks. The CLIP feature space has a dimensionality of 512. However, due to the computational burden posed by mapping the latent features of all Gaussians to 512 dimensions, we employ Principal Component Analysis (PCA) to reduce the dimensionality of the CLIP features to 64. These compressed 64-dimensional features are then used as supervision signals for the hierarchical language features of each Gaussian. The compression matrix obtained through PCA is retained, enabling decompression to restore the 64-dimensional hierarchical language features back to their original 512-dimensional space. This ensures efficient computation during training while maintaining fidelity for open-vocabulary 3D visual grounding during rendering.

For implicit queries, we introduce the Large Vision-Language Model (LVLM) to comprehend and reason about the target object. Specifically, we adopt the LLaVA-v1.5-7B model for this purpose. Using the inferred target object, the hierarchical language feature and instance feature are retrieved. To localize the target object, the hierarchical feature space is clustered using HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise). This approach determines the corresponding Gaussian group, effectively identifying the target object, even when occlusion is present in the current view. HDBSCAN is initialized with fixed parameters to ensure robust clustering:

- **min_cluster_size=10**: Clusters must contain at least 10 points to be valid.
- **cluster_selection_epsilon=0.01**: Ensures strict cluster boundaries for precision.
- **allow_single_cluster=False**: Prevents all points from being grouped into a single large cluster, treating poorly clustered points as noise instead.

This configuration enhances the clustering process, enabling accurate localization of objects while maintaining sensitivity to occlusion.

3. HIERARCHICAL FEATURE

ReasonGrounder is capable of performing fine-grained Gaussian grouping across multiple scales in diverse scene types, showcasing its versatility in handling complex environments. The hierarchical instance features learned in the Gaussian space are not only compact but also highly effective for decomposing scenes into meaningful components. Our method employs a hierarchical organization of the entire Gaussian field, structured according to the scale of the queried target object. This hierarchical structuring facilitates the adaptive grouping of Gaussians, ensuring scalability and precision. By isolating the appropriate Gaussian group corresponding to the target, ReasonGrounder enables accurate 3D localization and amodal perception, even for novel viewpoints that were not observed during training. To provide insights into the learned Gaussian features and their hierarchical structure, visualizations using Principal Com-

Method	*bed*	*bench*	*room*	*sofa*	*lawn*	**overall**
LSeg [22]	87.6	42.7	46.1	16.5	77.5	54.1
ODISE [35]	86.5	39.0	59.7	35.4	82.5	60.6
OV-Seg [25]	40.4	89.2	49.1	69.6	92.1	68.1
FFD [19]	86.9	42.8	51.4	9.5	82.6	54.6
LERF [16]	86.9	79.7	79.8	43.8	93.5	76.7
3D-OVS [27]	96.7	96.3	98.9	91.6	97.3	96.2
LangSplat [31]	99.2	98.6	99.3	97.9	99.4	98.9
ReasonGrounder	**99.2**	**99.1**	**99.4**	**98.2**	**99.4**	**99.1**

Table 11. **Localization Accuracy (%) on 3D-OVS dataset for open-vocabulary 3D visual grounding.** The first three methods target 2D visual grounding, whereas the remaining methods, including our ReasonGrounder, focus on 3D visual grounding.

ponent Analysis (PCA) are presented in Figure 10, 11 and 12. These visualizations illustrate the effectiveness of our method in representing and organizing Gaussian groups at varying scales.

4. MORE QUANTITATIVE RESULTS

To complement the mIoU score metric, we also evaluate the 3D-OVS dataset using the Localization Accuracy metric. Accordingly, we compare the performance of our method against other state-of-the-art approaches on this dataset, with the results presented in Table 11. It is evident from the table that our ReasonGrounder consistently achieves superior performance, further demonstrating the effectiveness and advantages of our approach.

5. MORE QUALITATIVE RESULTS

Qualitative results from scenes not featured in the main text or accompanying videos are presented in Figures 13, 14, 15, and 16. These additional examples showcase the versatility and robustness of the proposed ReasonGrounder framework across diverse and unseen environments. The experimental results further validate that ReasonGrounder effectively enables open-vocabulary 3D visual grounding and reasoning. Specifically, it demonstrates the capability to accurately localize objects and interpret relationships within complex 3D scenes, even when operating with an open-ended vocabulary. This highlights its potential for addressing a wide range of real-world tasks that require advanced scene understanding and reasoning.

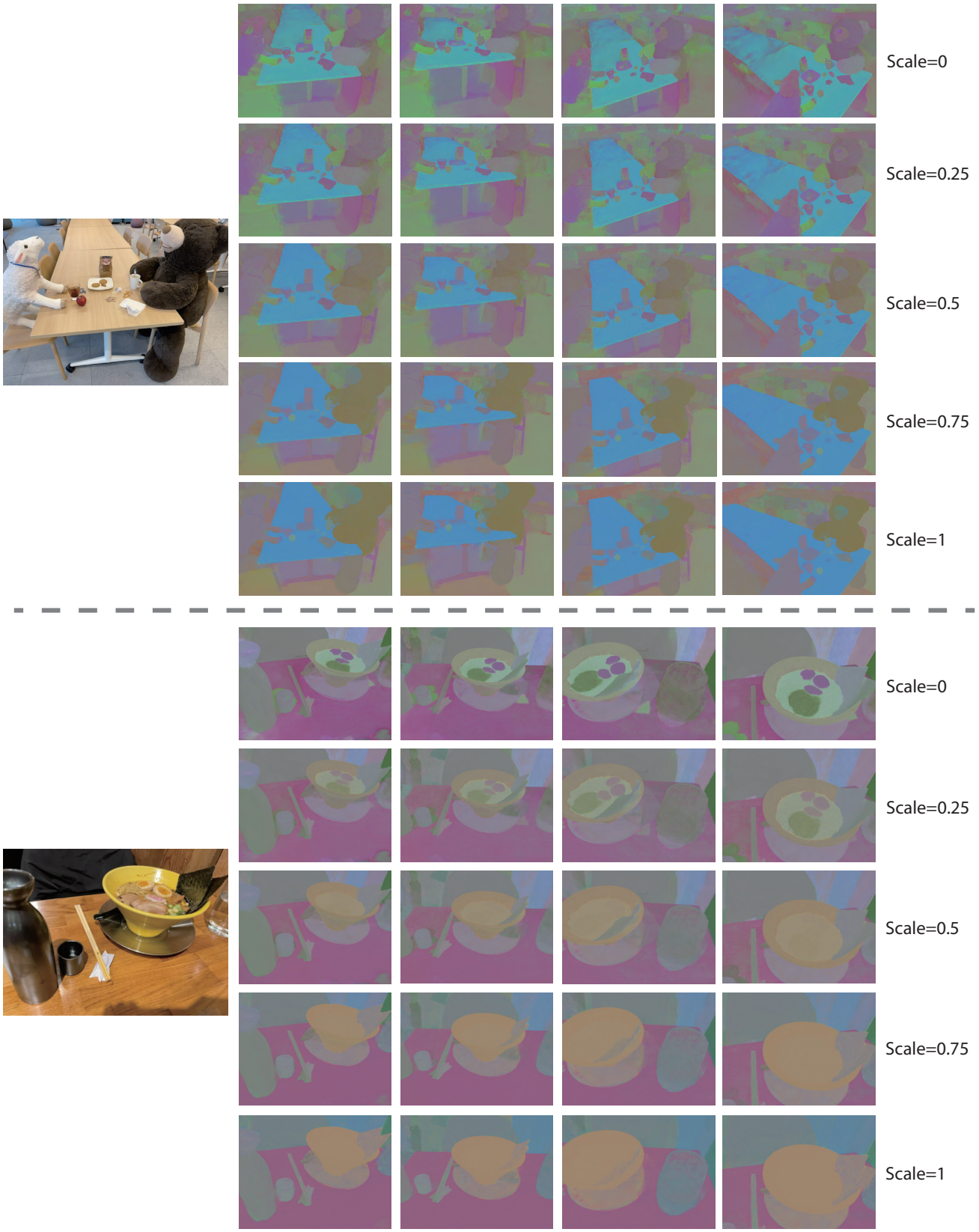

Figure 10. **Visualization of hierarchical feature.** The feature space dynamically adjusts its granularity in response to physical scale: finer granularity is employed at smaller scales to capture intricate details, while coarser granularity is utilized at larger scales to emphasize overarching structures. This adaptive mechanism ensures robust and effective representation across a wide range of scales.

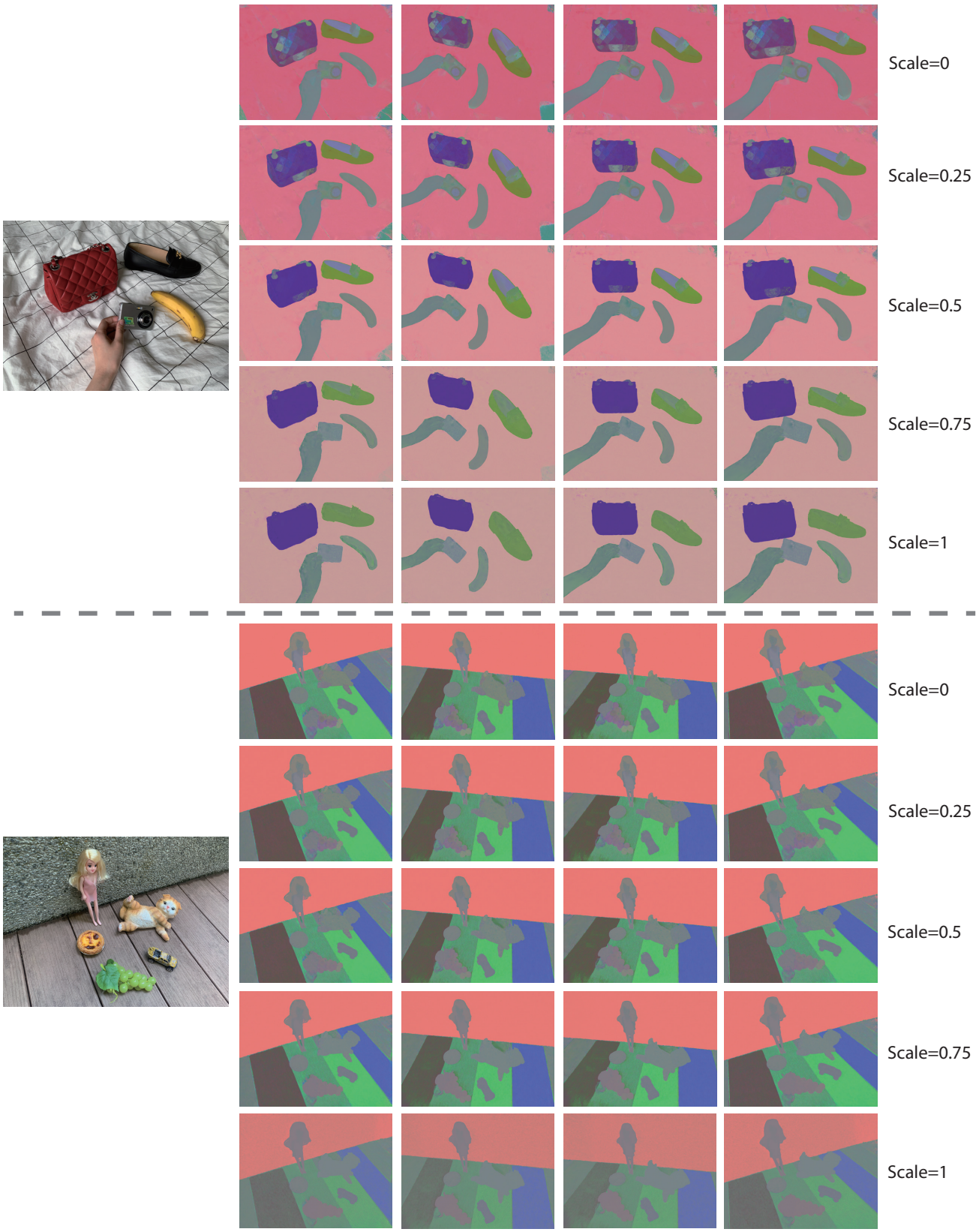

Figure 11. **Visualization of hierarchical feature.** The feature space dynamically adjusts its granularity in response to physical scale: finer granularity is employed at smaller scales to capture intricate details, while coarser granularity is utilized at larger scales to emphasize overarching structures. This adaptive mechanism ensures robust and effective representation across a wide range of scales.

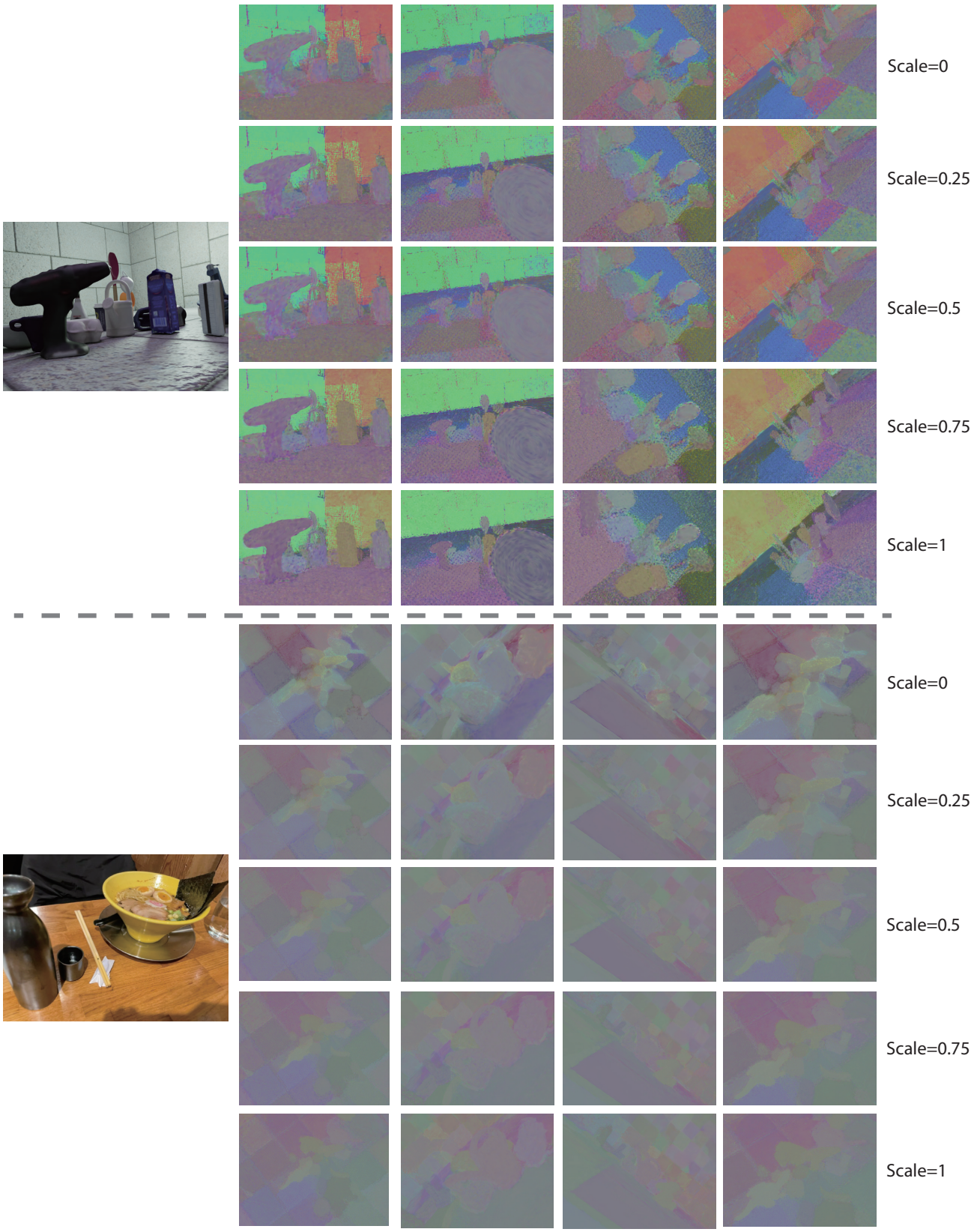

Figure 12. **Visualization of hierarchical feature.** The feature space dynamically adjusts its granularity in response to physical scale: finer granularity is employed at smaller scales to capture intricate details, while coarser granularity is utilized at larger scales to emphasize overarching structures. This adaptive mechanism ensures robust and effective representation across a wide range of scales.

Which item on the table is blue and resembles an animal with big ears and a long nose?

It is blue elephant. The blue elephant toy is a standing cartoon figure in light blue. It has a rounded body, large and cute ears, and a short, upturned trunk. The design is simple and carries an adorable cartoon style.



Which item on the table is yellow, he will often appear in the bath water, squeezing it will make a squawk?

The yellow duck toy is a bright yellow rubber duck with a classic design. It has a cheerful expression, with a prominent orange beak and round body. Around its neck is a pink beaded necklace, adding a playful touch.

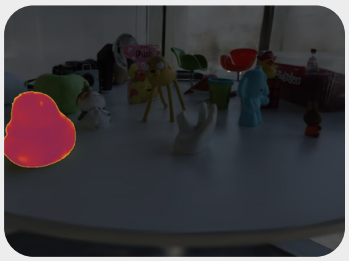

Figure 13. Qualitative comparisons on the ***Figurines*** scene of the LERF dataset.

If I run out of cookies on my desk, where do I get more cookies?

The cookie bag is a brown paper package with a simple design. It stands upright, and the top appears to be folded over. The front of the bag features a label with an illustration or logo, suggesting it contains cookies.



If I refill a cup of coffee, where should I put it when I'm done drinking it?

This is a rectangular, light wood-colored table with a smooth surface and a simple design. There are several wooden chairs around it, giving a modern and functional style, suitable for meetings or casual dining.

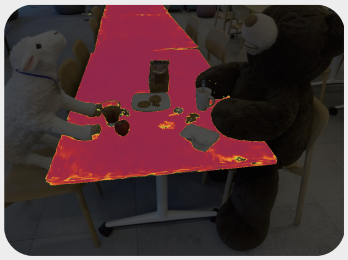
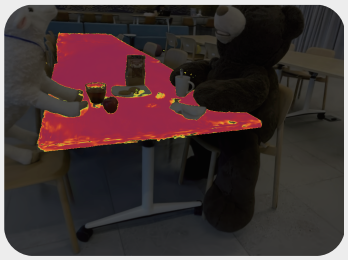
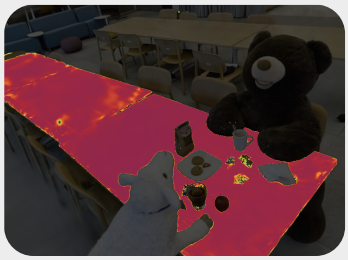

Figure 14. Qualitative comparisons on the ***Teatime*** scene of the LERF dataset.

Now I need to shoot some scenery, which object in the scene can be used for shooting records?

It is a camera, which is a common item used for taking photos and recording videos. In this image, the camera is placed next to a banana, a handbag, and a pair of shoes.

What's touching this camera right now?

The hand is holding a compact camera, positioned over a checkered fabric background. The fingers are gently wrapped around the camera, with a relaxed grip, and the overall appearance suggests a casual pose.

Figure 15. Qualitative comparisons on the ***Bed*** scene of the 3D-OVS dataset.

What has a vibrant coat that ranges from light to dark orange?

The cat has a vibrant coat that ranges from light to dark orange, making it stand out in the scene with the doll and the grapes.

Which is the sweet, round, egg-based dessert in the scene?

The egg tart is a small pastry with a golden-brown, flaky crust. Its filling is a creamy, custard-like mixture with a slightly caramelized top, giving it a rich and inviting appearance.

Figure 16. Qualitative comparisons on the ***Bench*** scene of the 3D-OVS dataset.