GARField: Group Anything with Radiance Fields

Supplementary Material

A. Additional Results

We show additional figures and videos using GARField for 1) hierarchical decomposition, 2) global clustering, and 3) interactive selection. All video visualizations use Gaussian Splatting [12], as described below.

A.1. Gaussian Splat Visualizations

We use Gaussian Splatting [12] to emphasize the 3D nature of GARField and its applications for 3D group extraction. Here, for simplicity, we do not optimize GARField directly with gaussians. Instead, we train a NeRF-based GARField and a Gaussian Splatting model separately. Then, we assign an affinity feature to every gaussian by querying the feature field at the gaussian's center point. We use these features to manipulate the 3D scene, *e.g.* clustering, selection, and filtering. All implementation described here will be made public. To visualize clusters in 3D, we override each gaussian's color parameters to the RGB color of the colormap.

We found gaussian centroids align well enough with the underlying feature field that querying only their centerpoints was sufficient. For larger gaussians, this approximation becomes less accurate, however we found this to not be an issue since large gaussians are explicitly culled early in training, and after training tend to reside primarily in the background of the scene, where the more important problem is geometry under-reconstruction.

A.2. 3D Hierarchical Decomposition

In the main text, we visualized hand-picked nodes from the resulting hierarchy in Main Paper Fig. 6. Here, we exhaustively visualize entire subtrees of selected scenes by selecting the primary region of interest (i.e. desk, dozer, bouquet).

A.2.1 Full Tree Visualizations

In Fig. 12 and in provided videos we visualize each layer of the resulting tree organized by node depth in different rows. Each node is shown colorized by the number of internal clusters, with the remainder of the tree drawn with low opacity to give context. Note that nodes at the same level do not necessarily correspond to the same scale because intermediate nodes are pruned.

One can see how each part is recursively broken into subparts in lower layers of the tree, for example the statue gets broken into the base and rest of the statue, followed by shield, torso, hair, and etc. Videos of trees showcase the view-consistency of 3D scene decomposition, with whole objects being clustered together like the bear or dozer, which can then be broken into coherent subparts. The lowest levels of the tree contain very fine details such as petals of flowers, or hooves of the sheep.

Additional Limitations: One consequence of scale-conditioning is that object parts of different sizes branch off the tree separately rather than all at once: multiple objects on the same table may appear at different levels of the tree. The tree generation in this work is a naive greedy algorithm, which can result in spurious small groups at deeper levels, as seen in the trees in the Supplement. Future work may explore more sophisticated ways of hierarchical clustering.

Note how some nodes can contain noise or partial clusters, for example the third row, last node of Fig. 12, where the red cluster is a spurious cluster which more suitably belongs to the base of the statue in the prior tree level. We believe artifacts like this happen for two main reasons: 1) affinity feature vectors within an object vary smoothly with scale, and sampling a scale within this changing region (i.e. between two modes of a segmentation of an object) can produce more ambiguous affinities between the two modes. Since the tree construction algorithm samples scales at a set interval, these scales could land on such a boundary region, producing spurious clusters. This effect can be especially severe at the boundaries of objects, where affinities between scales must drastically change. One potential approach to remedy this could be adapting the scale within a small window at each tree node to minimize cluster variance, which would bias towards tightly coupled clusters. 2) The tree construction algorithm is greedy, meaning any node split is final. Allowing the algorithm to search and backtrack while optimizing for metrics like minimizing cluster variance or tree size could prove beneficial.

Sometimes spurious background points are grouped together with the object of interest, a behavior which could be remedied by more strongly taking geometric proximity into account when constructing the tree. We hypothesize this may occur when distant points do not share enough common viewpoints, meaning there is no contrastive loss pushing their embeddings apart. This might be addressed by applying a small regularizing contrastive loss to all points in 3D whether observed together or not. Another failure mode is that when view coverage is insufficient, different sides of the same object can be grouped separately. For example, in rows 3 and 4 of Fig. 12 the two sides of the statue's face are grouped differently.

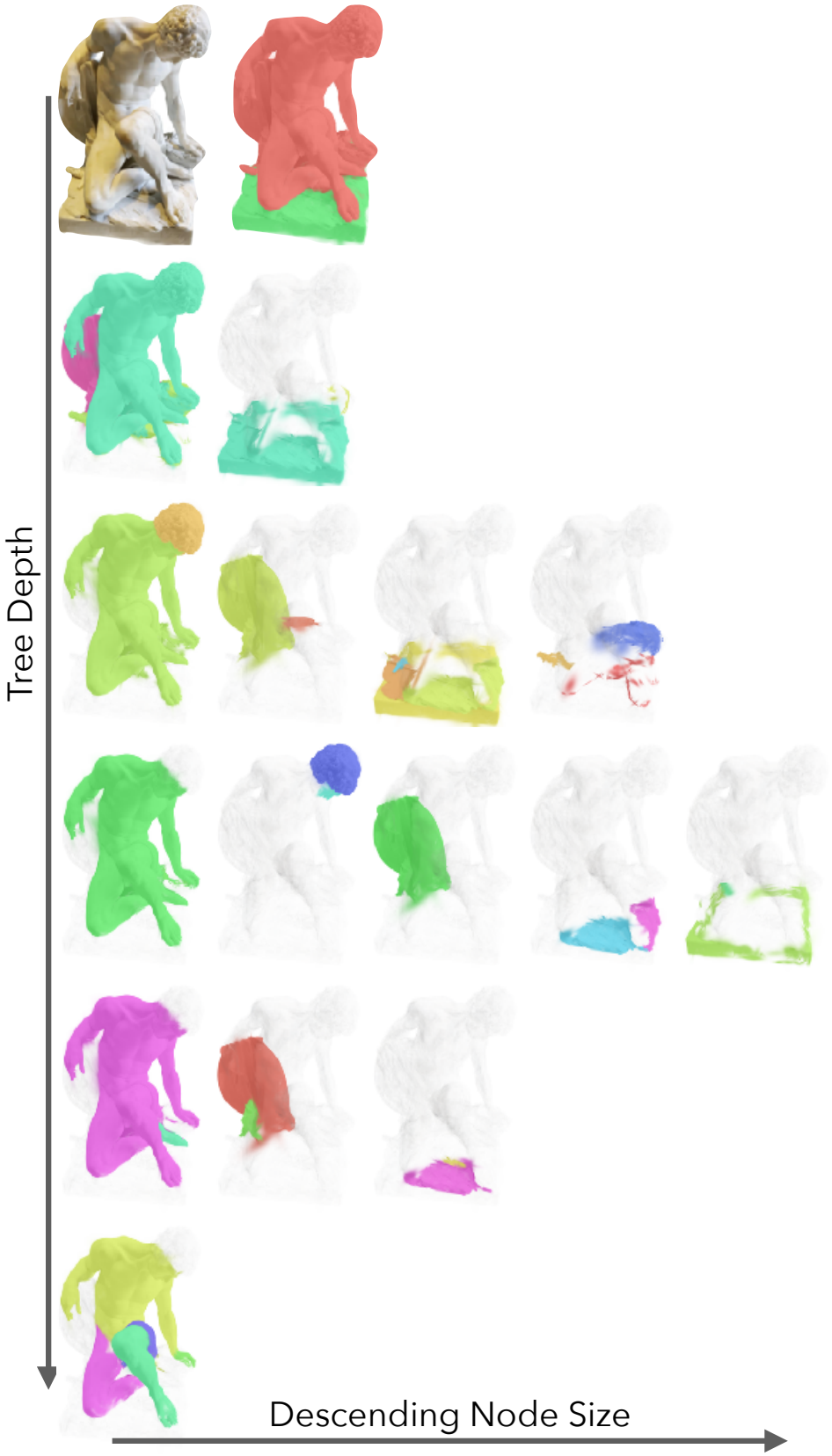

Figure 12. **Complete Tree**: A complete visualization of all layers and all nodes in the tree from Fig. 6. Colors illustrate different clusters within each node, and each row visualizes all the nodes at a given depth in the tree, sorted by size.

A.2.2 Compressed Tree Visualizations

We additionally provide videos of compressed trees, where each layer of the tree is merged into one visual by distinctly coloring all clusters. Leaf nodes at one layer are further propagated to deeper layers of the tree to visualize all clusters at the lowest level, corresponding to the most granular decomposition. Though these visualizations do not show hierarchy because they merge all nodes, they illustrate how lower layers of the scene decomposition correspond to semantically meaningful high granularity and higher levels correspond to coarser granularities.

A.3. Multi-Scale Clustering

We provide video versions of Main Paper Fig. 7 to showcase the view-consistency of the results shown in the images. These videos first show the global clustering of the scene, followed by video renderings of sub-object clusters.

A.4. Global Clustering

To emphasize that GARField can model scene-level groupings, we cluster GARField features globally *i.e.* all gaussians in a scene. Figures 14 through 20 show all scenes in Fig. 7 globally clustered at scales 0 to 1, at increments of 0.05.

We also include a video where the excavator scene in Main Paper Fig. 1 is globally clustered at three distinct scales. We find that GARField successfully groups together large group in the backgrounds, like the road or bushes on the sidewalk.

A.5. Interactive Selection

People can use clicks to interact with GARField and extract groups of different sizes, as shown in Fig. 5 of the main paper. User clicks are transformed into 3D points using projective geometry (visualized with a red sphere in the video). At a given scale, we select a set of 3D gaussians based on their affinity with the selected point. To retrieve multiple groups, we query GARField across a range of scales and merge groups with large overlap. In the video, a user can extract the excavator, crane, and scooper from Fig. 1 with a single click.

B. Experiment details

B.1. Hierarchical Decomposition

Once we select a cluster of interest, we construct a tree by recursively clustering with HDBSCAN. For this process we use an HDBSCAN cluster epsilon of 0.1 and a minimum cluster size of 40, fixed for all experiments. The tree is constructed greedily in a depth-first search, by recursing *only* on non-noise clusters (see Sec.). Note that because we add noise clusters back to the tree after constructing it, this can result in small disappearing regions, like in the lower levels of the succulent scene. These artifacts would better be addressed with a non-greedy tree construction, which we hope to address in future work.

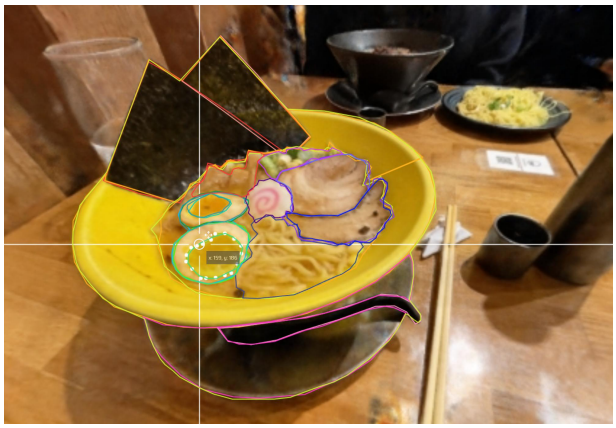

Figure 13. **Masks for 3D Completeness Experiments**: Overlapping masks (*egg*, *noodles*, *nori* masks inside *ramen* mask) model the desired hierarchical groupings. We labeled these polygonal masks using 'Make Sense' [30], an online tool for mask annotation.

To speed tree contruction, we first sub-sample the input gaussian splat with Open3D's voxel-downsampling to reduce the resolution of points to $0.01\times$ the scale being queried, for example an affinity of 0.1 scale downsamples to .001 voxel resolution. After tree construction, the resulting tree is pruned to remove chains of nodes with one child and one parent.

B.2. Treatment of Clustering Noise

One challenge to overcome is the fact that HDBSCAN can output 'noise' clusters, which do not get any cluster labels. These can arise because of gaussians which do not align well with NeRF geometry, features which are noisy because they lie on the boundary of two groups, or noise in the trained affinity field. To handle these noise clusters, we assign labels to gaussians considered noise with the label of the nearest *physical* clusters computed in the Euclidean space, as opposed to the feature space. We find this produces more cohesive results than soft clustering within the feature space itself. During global clustering (Figs. 11, 7) these noise clusters are assigned to clusters across the entire scene, and during tree decomposition (Fig. 6) these noise clusters are locally assigned from the clusters available at each node only.

B.3. 3D Completeness Experiment

B.3.1 Ground Truth Annotation

We annotate ground truth segmentation masks on a randomly selected novel view using the online tool 'Make Sense' [30], employing a polygon shape for the annotation. In Fig. 13, we present the visualization on our state during the data annotation process.

The annotation process begins with the assignment of a specific label point to each target object within a given view. Note that the selection of the view is randomized, involving zooming in, zooming out, or changing the angle to enhance the evaluation of view consistency effectively. These label points serve as the basis for the subsequent mask annotation, which are made at a varying level of granularity. As a case in Fig. 22, in the bouquet scene, considering the click points from different angles, we annotate the masks at different hierarchical levels: the petal of the flower (fine level), the individual flower (medium level) and the whole bouquet (coarse level). For ground truth masks in other scenes, we follow similar rules, building a mask hierarchy based on the semantic meaning, ranging from fine part of the object to coarse whole object. However, note that the number of mask levels may vary depending on the complexity and the nature semantics in the scene. For example, the bear's arm in the teatime scene, Fig. 21, is only annotated with two levels of hierarchy: the left hand and the whole bear.

B.3.2 Complete Visualizations

A comprehensive presentation of the evaluation results regarding to the view consistency of GARField is shown in Figs. 21, 22, 23, 24, 25. This includes all the scenes not shown in the main text. For each scene, we show the clicked label points for the annotated randomly selected views, ground truth masks at different hierarchical levels and the comparison of the closest masks obtained by SAM and GARField. We also provide the zoomed-in images of the results for better visualization.

B.4. Hierarchical Grouping Recall Experiment

B.4.1 Ground Truth Annotation

In this experiment, we annotate one novel view for each of the five scenes. For each novel view, we mark one or several objects which has a rich hierarchy. The ground truth masks are any parts, subparts, or the entire object of the scene that can be considered as groups by a human. Taking the ramen scene (Figs. 13, 26) as an example, the parts or subparts of the objects labeled include nori, egg, egg yolk, noodles, and so on. Additionally, the complete soup and the entire ramen bowel is also annotated as a group. Unlike the experiments on 3D completeness, this experiment aims to test whether the model can extract all the reasonable masks of the objects which contain rich hierarchy. Therefore, we did not stratify the level of the annotated masks.

B.4.2 Complete Visualization

In Fig. 26, We show the ground truth masks as well as all the methods masks at the finest masks. Note that all the ground truth masks are arranged in descending order of size. In our experiment, we systematically recover all the masks that corresponds to the annotated ground truth through different method. For each distinct method employed, which are SAM, GARField without scale condition, GARField without dense supervision, we sequentially showcase the masks that get the highest IOU score of the correspondence to the ground truth masks. We will release all the ground truth annotations for all experiments.

Figure 14. **Global Clustering Results ("Bouquet")**: Global clusters at smaller scales ($s = 0$) distinguish between different sections of the bouquet, as well as the two halves of the table. At a larger scale, the bouquet and table are considered whole.

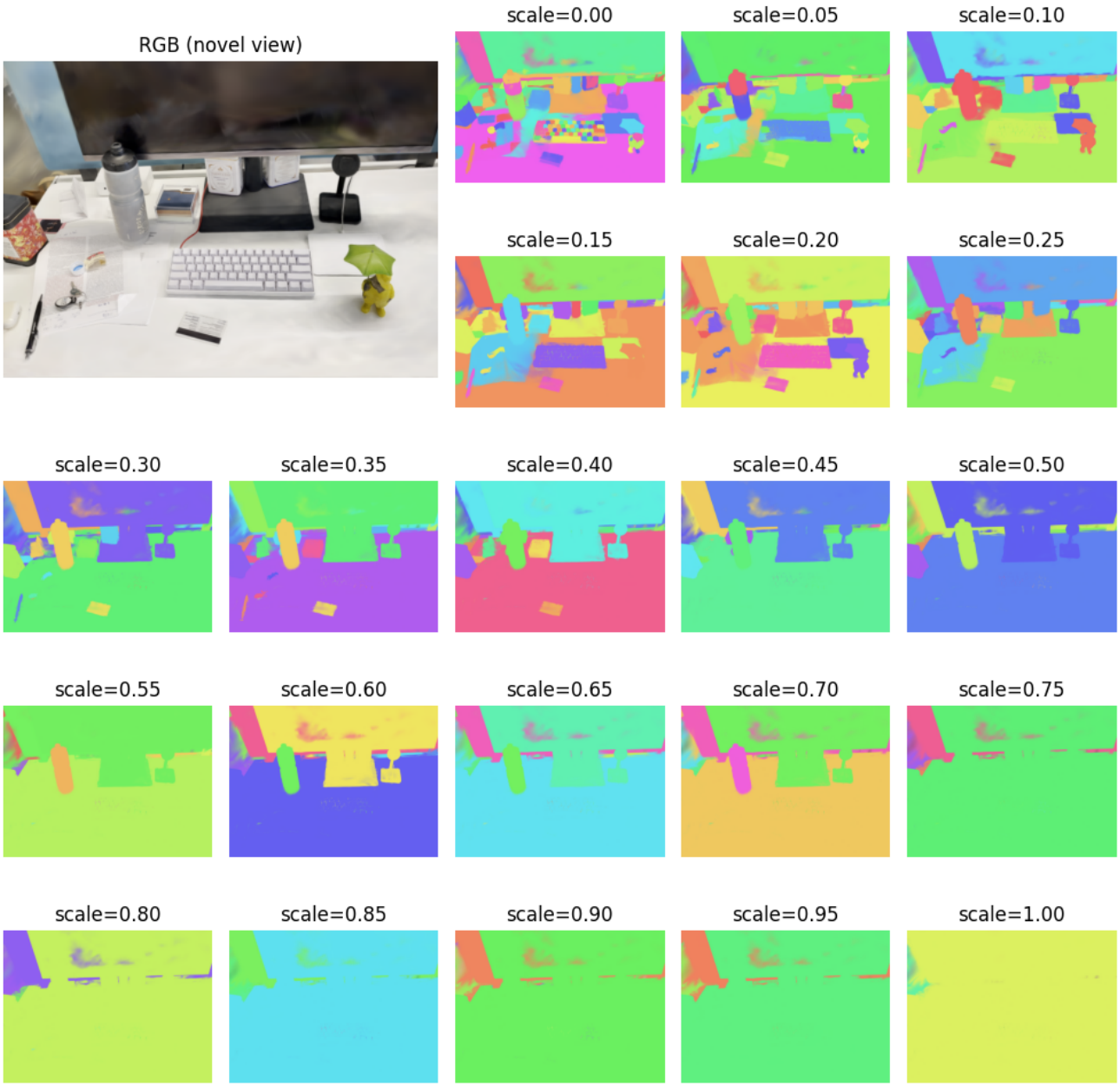

Figure 15. **Global Clustering Results ("Desk")**: At larger scales ($s = 0.5$), the desk is grouped together with the clutter on it *e.g.* keyboard, card, bird figurine).

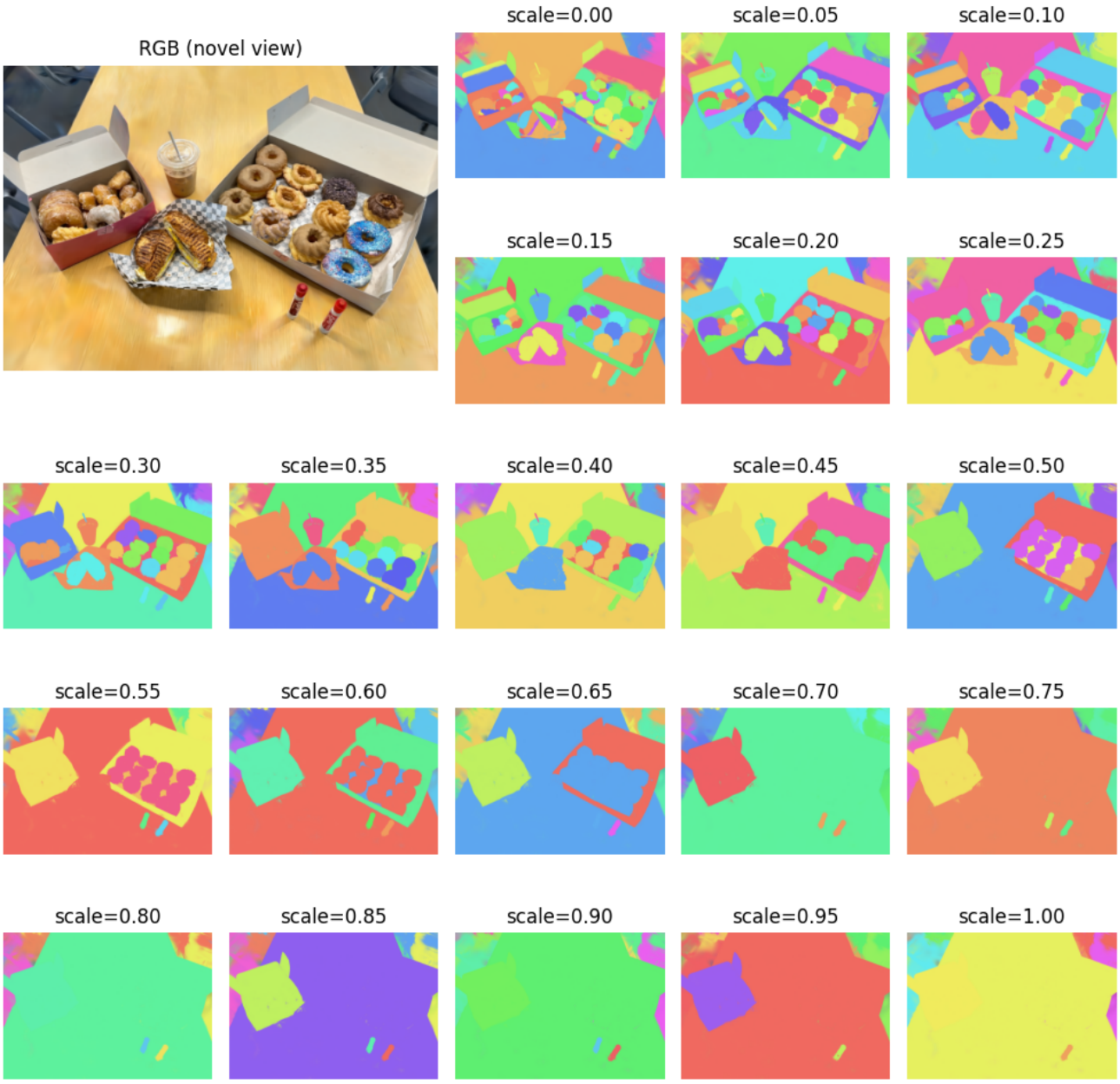

Figure 16. **Global Clustering Results ("Donuts")**: At a very small scale ($s = 0.0$), GARField can distinguish between different pieces of the breakfast sandwich in the middle of the scene. As scale increases, its grouping shifts quite noticably — into its two halves, or the full sandwich with the checkerboard packaging.

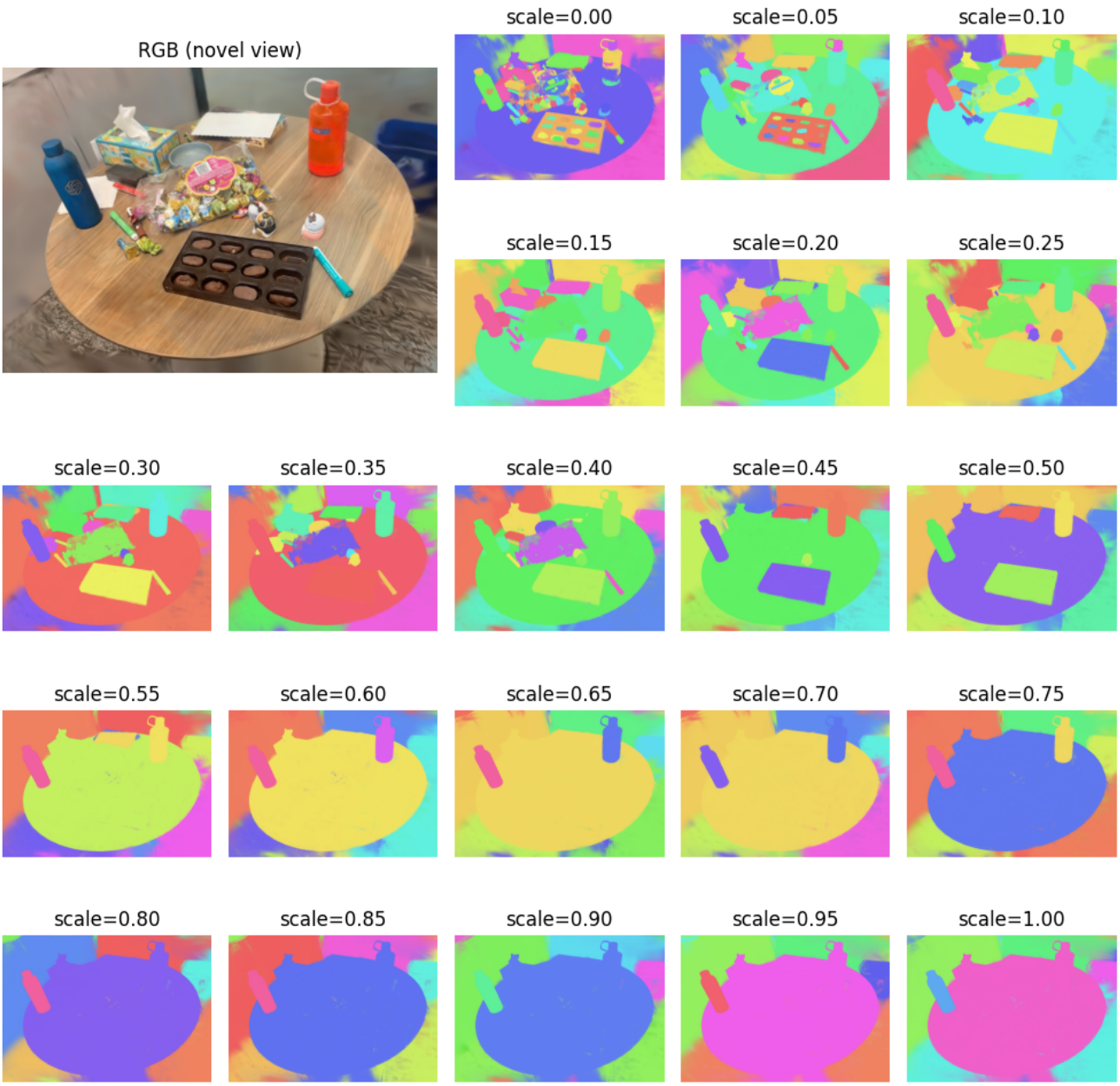

Figure 17. **Global Clustering Results ("Table")**: At the smallest scale $(s = 0.0)$, the global clusters highlight parts of objects *e.g.* labels on water bottles, pieces of chocolate.

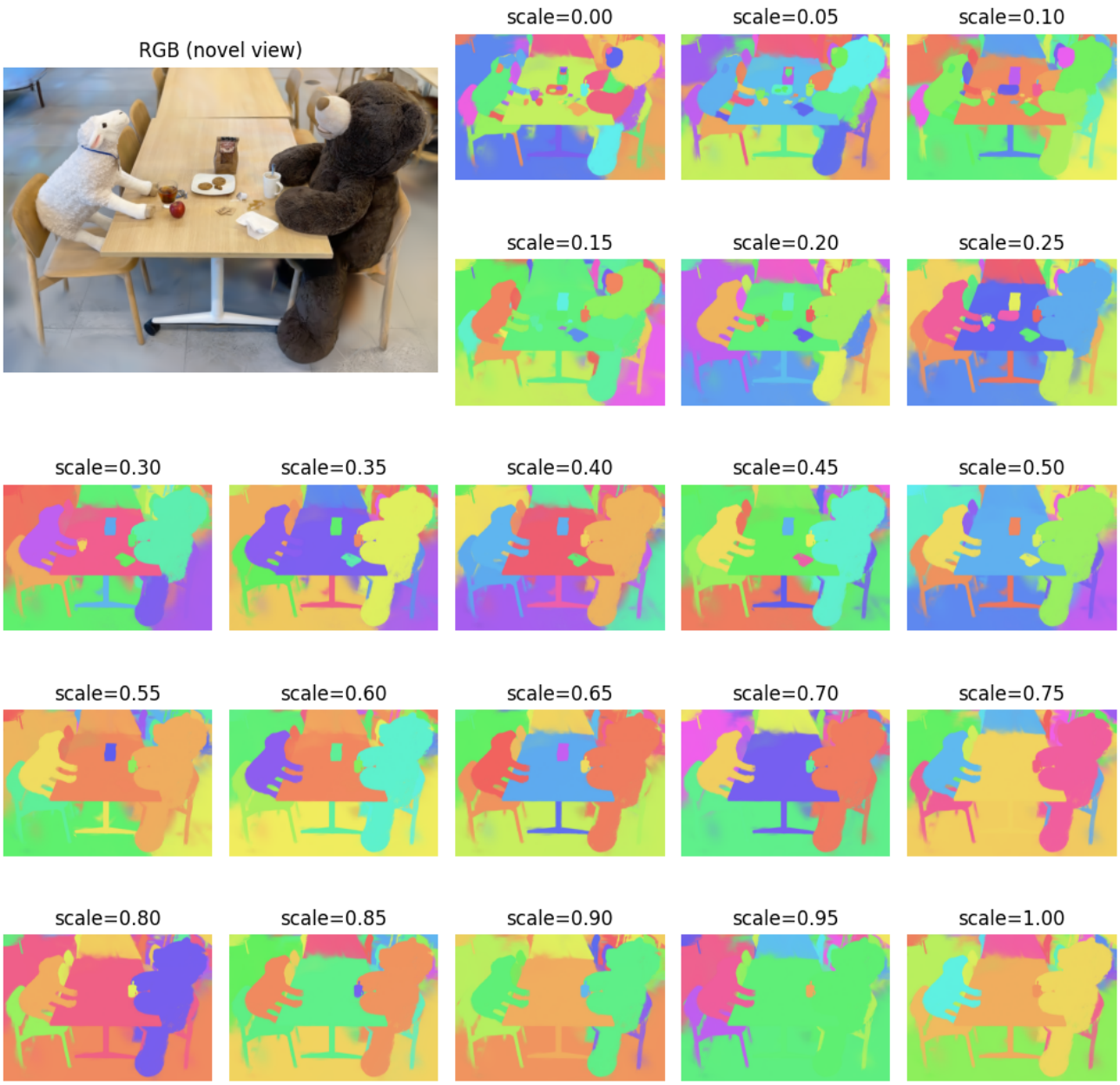

Figure 18. **Global Clustering Results ("Teatime")**: The food, utensils, and the table are included in different clusters at small scales, and the same cluster at larger scales. Parts of the stuffed animals (*e.g.* sheep hooves, bear nose) can also be seen at $s = 0.0$.

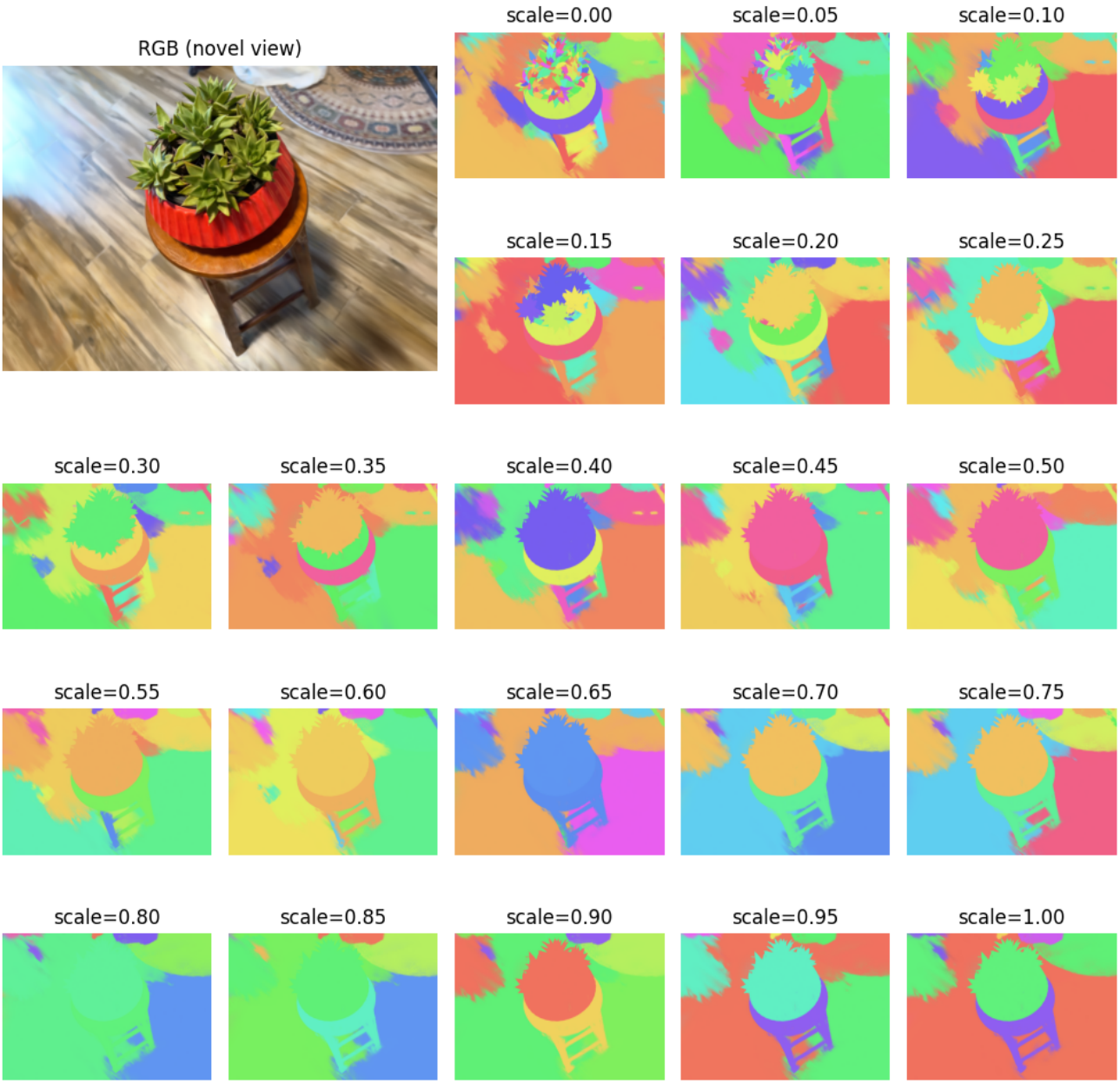

Figure 19. **Global Clustering Results ("Succulent")**: Global clusters at smaller scales ($s = 0.0$) distinguish between fine features like succulent leaves, while they are considered a single group at larger scales ($s = 1.0$).

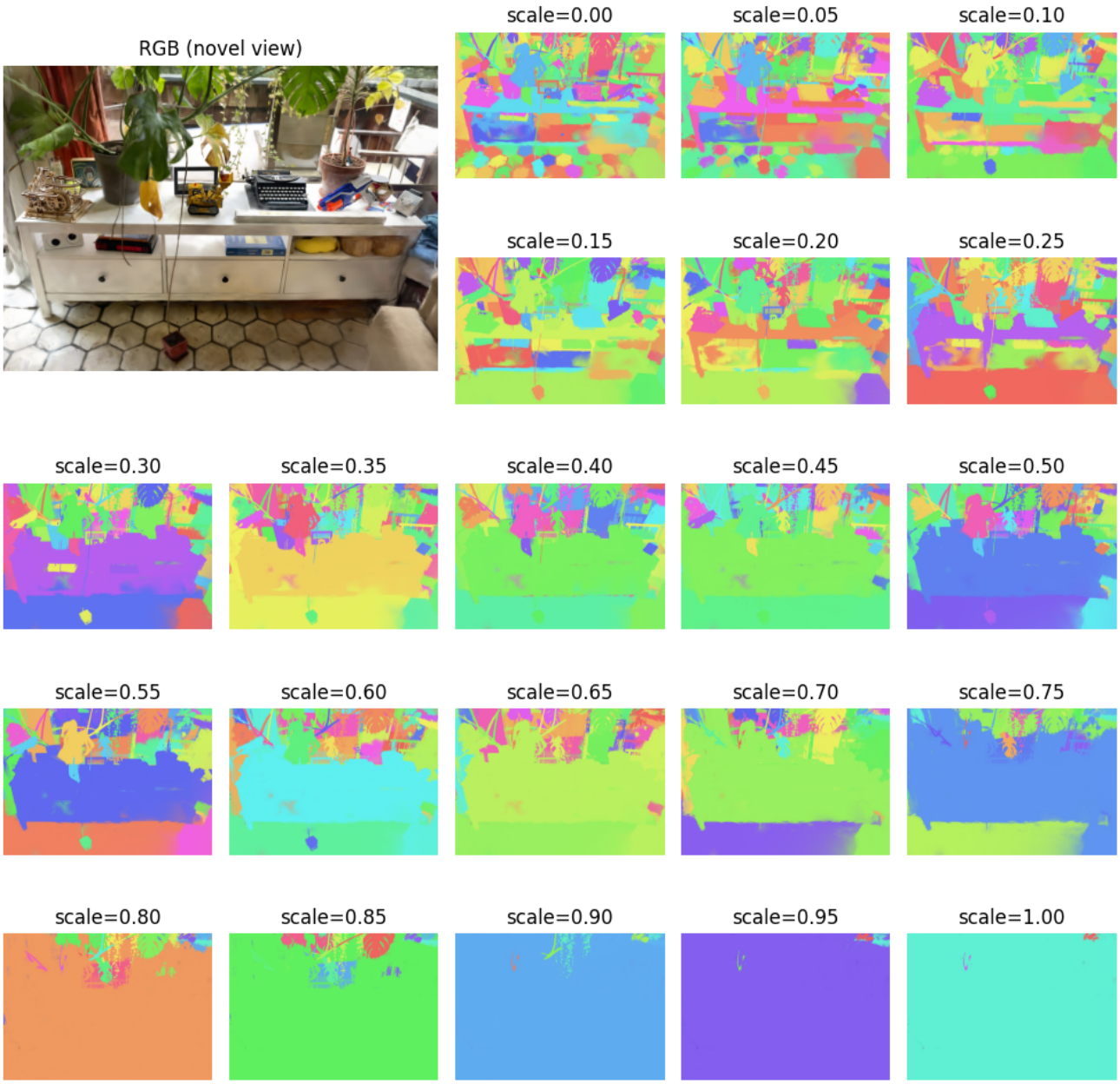

Figure 20. **Global Clustering Results ("Living Room"**: The individual hexagonal tiles on the floor may be grouped separately ($s = 0.0$) or together ($s = 0.5$).

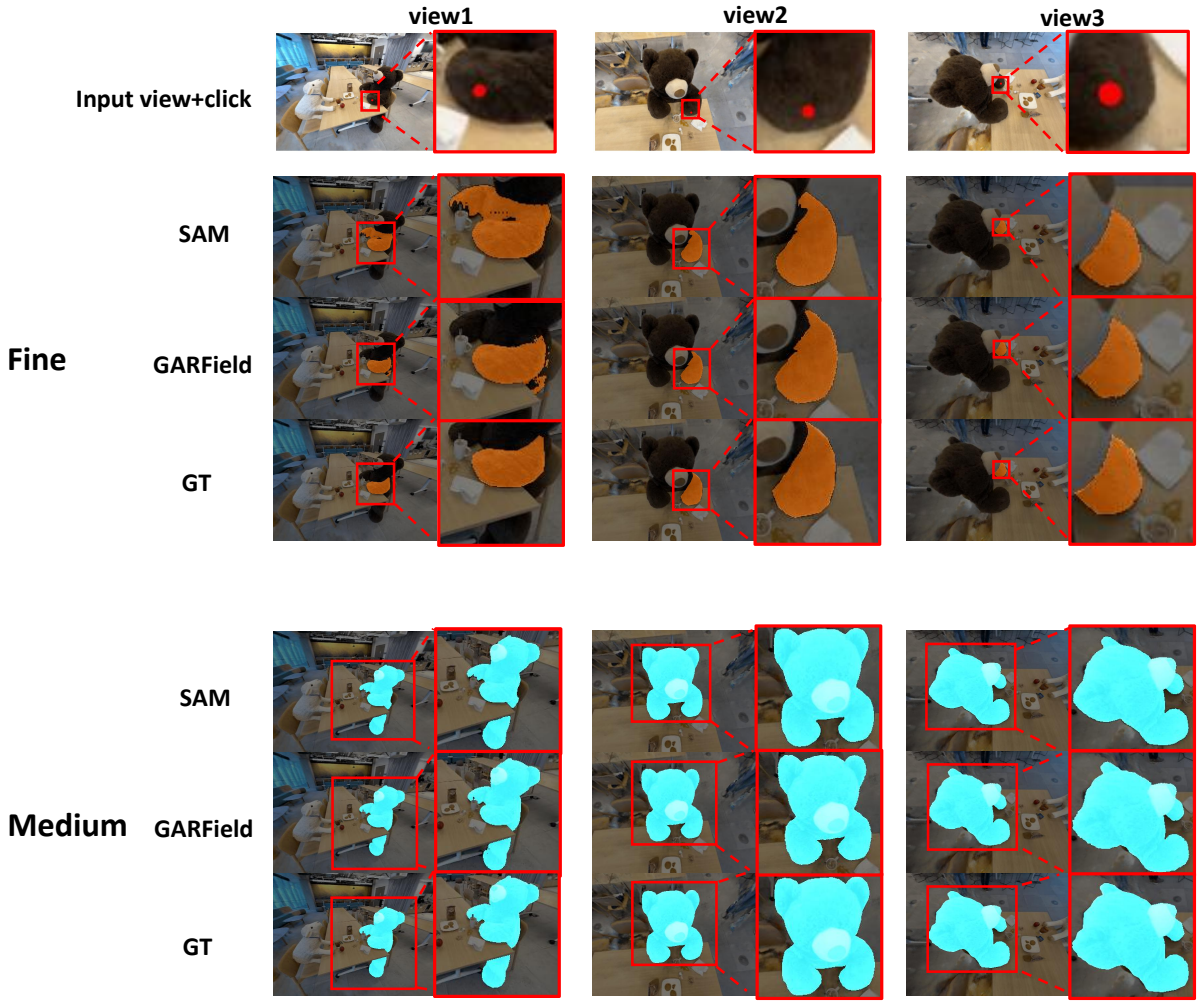

Figure 21. **View Consistency Experiment-Teatime**: We constructed two hierarchies, which are fine and medium. These correspond to the semantic meanings of the bear's left hand and the whole bear, respectively.

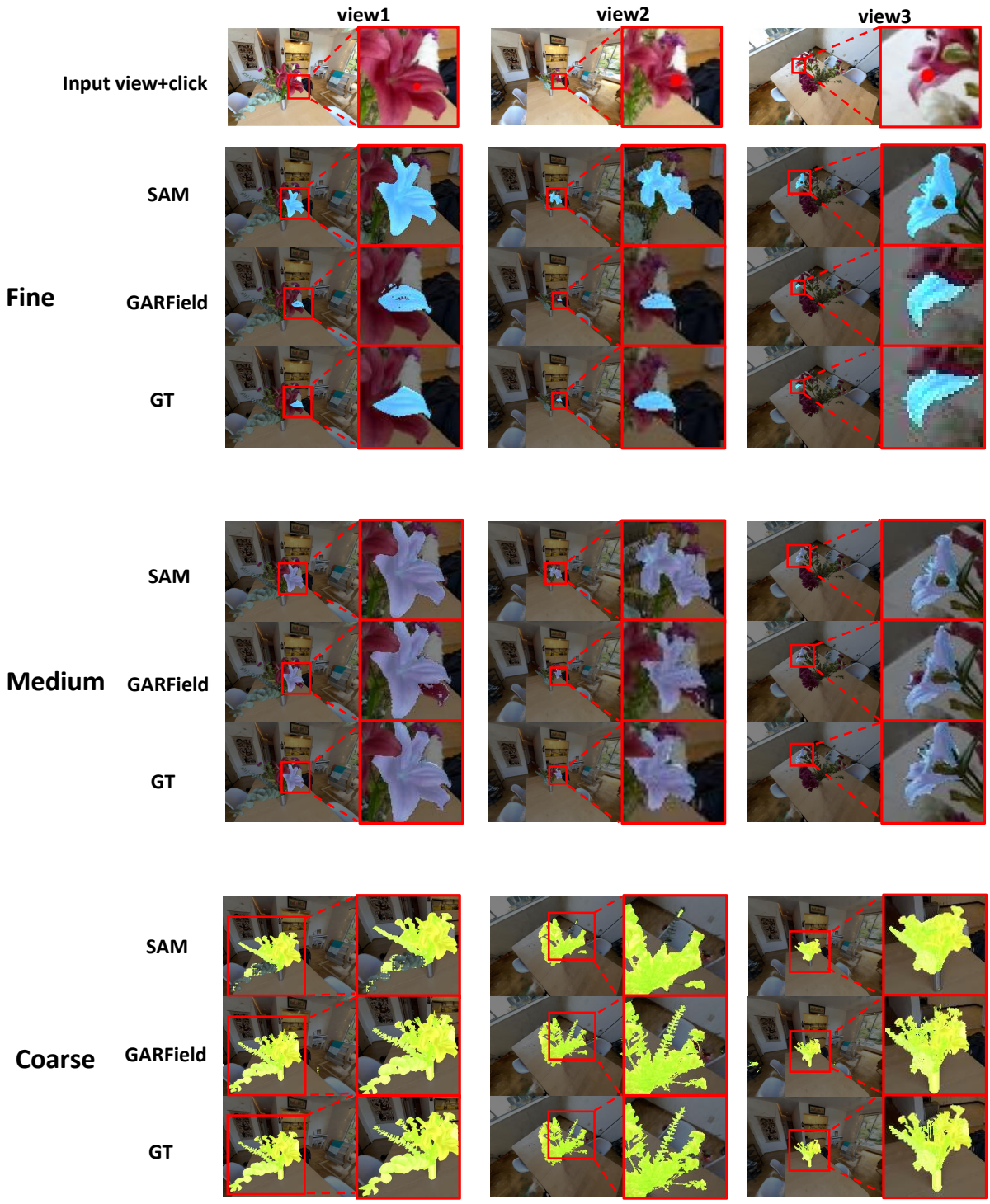

Figure 22. **View Consistency Experiment-Bouquet**: We constructed three hierarchies, which are fine medium and coarse. These correspond to the semantic meanings of the petal of the flower, the individual flower and the whole bouquet, respectively.

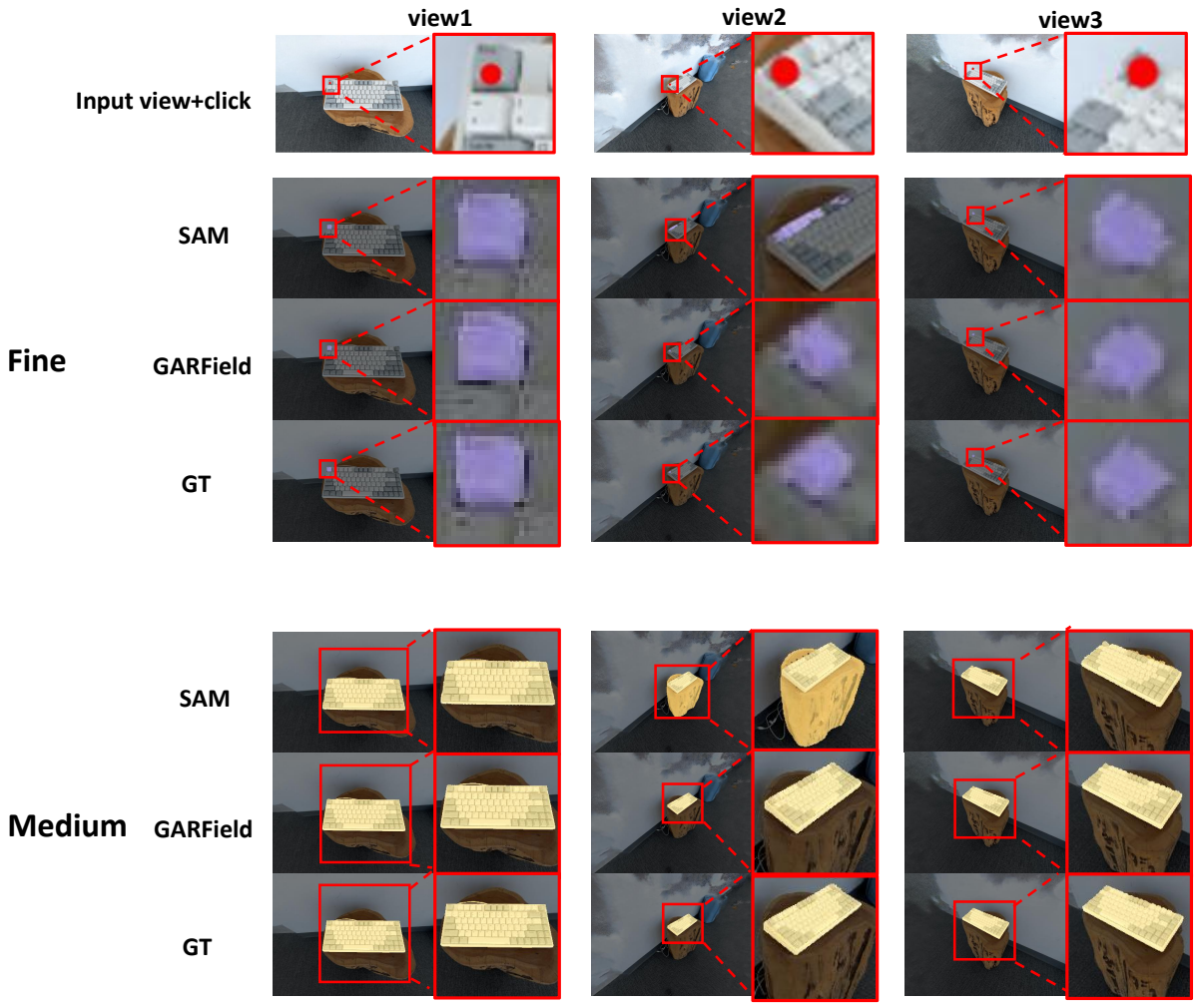

Figure 23. **View Consistency Experiment-Keyboard**: We constructed two hierarchies, which are fine and medium. These correspond to the semantic meanings of single key and the whole keyboard, respectively.

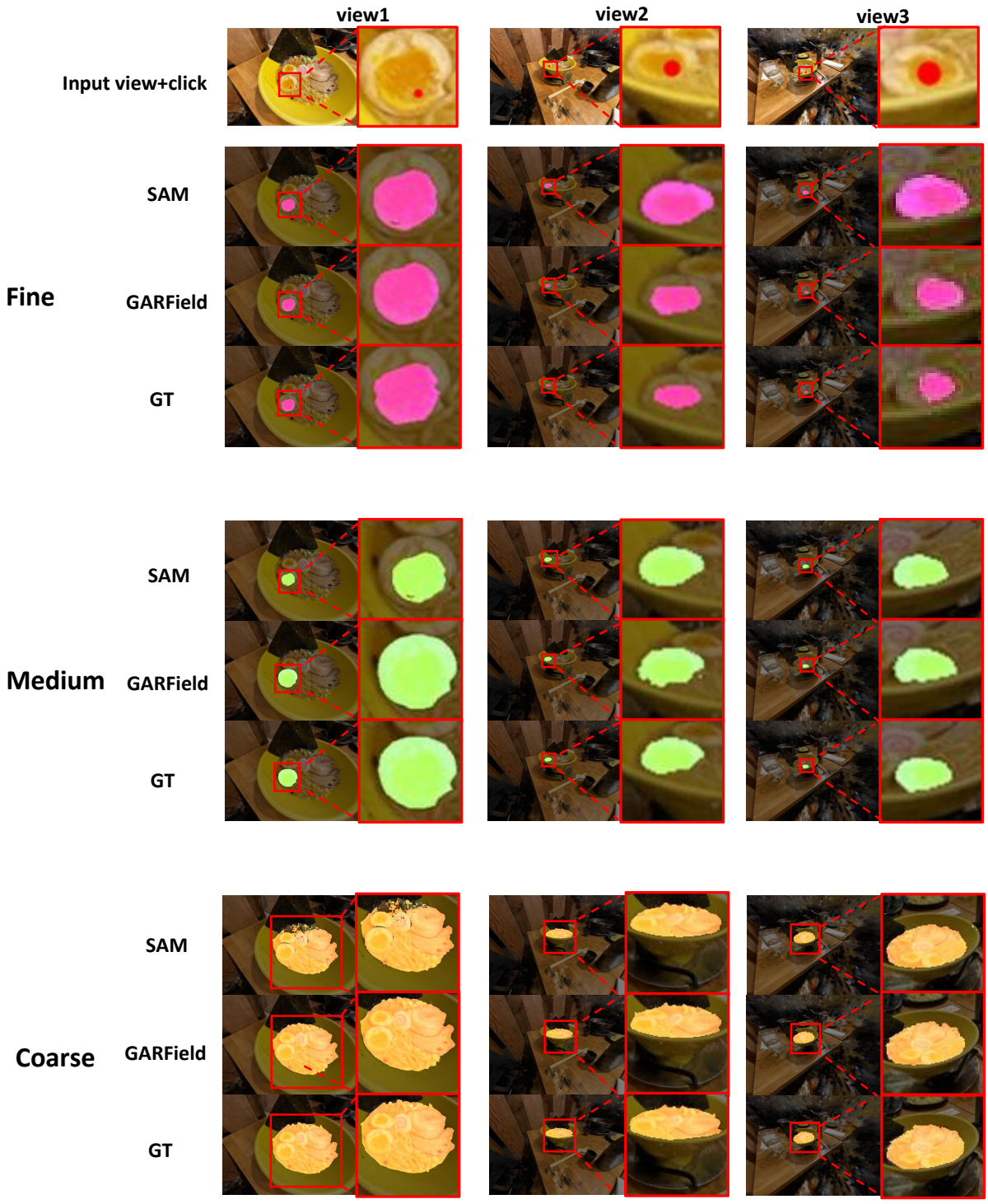

Figure 24. **View Consistency Experiment-Ramen**: We constructed three hierarchies, which are fine, medium and coarse. These correspond to the semantic meanings of egg yolk , one single egg and the whole soup area, respectively.

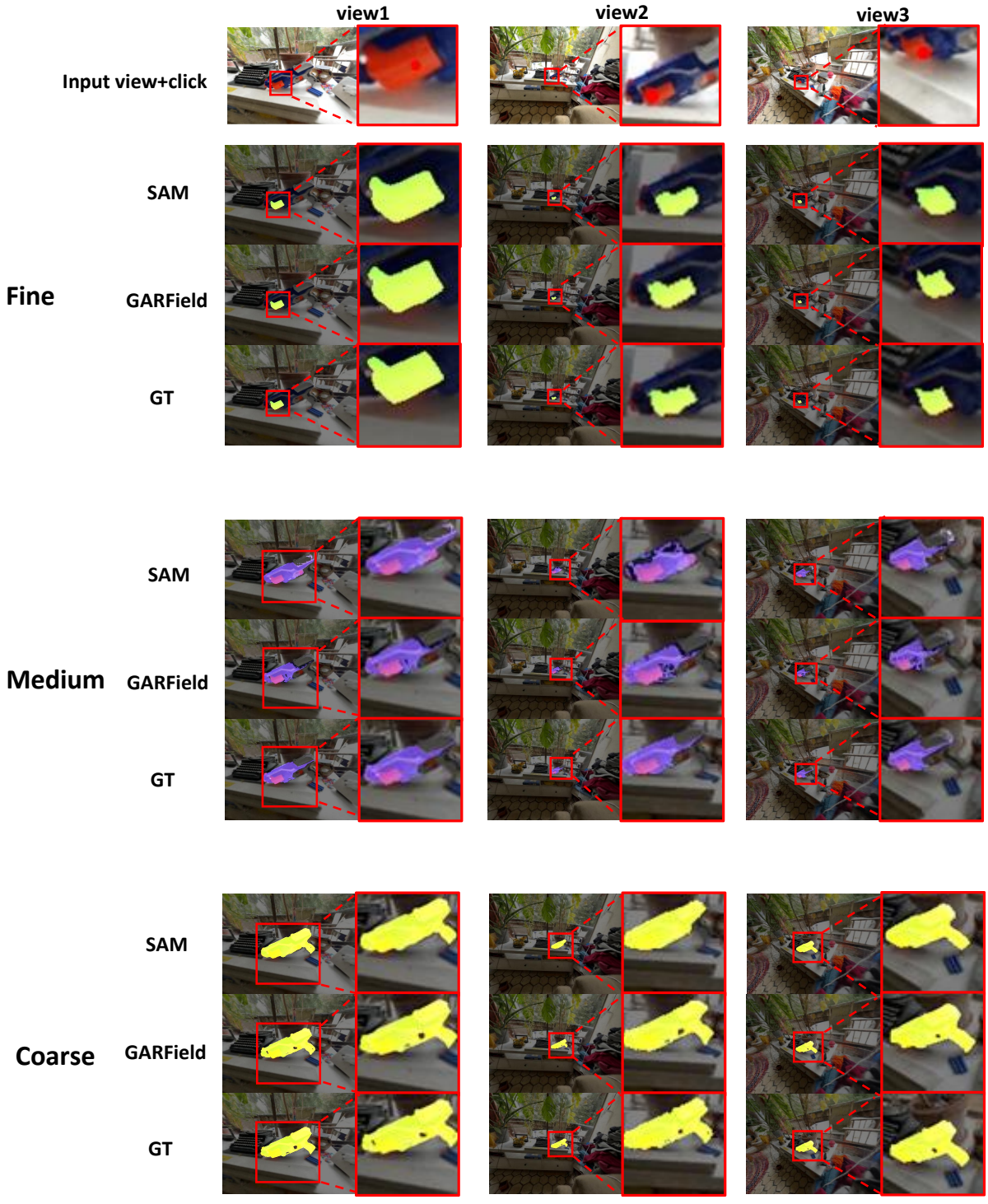

Figure 25. **View Consistency Experiment-Living room**: We constructed two hierarchies, which are fine medium and coarse. These correspond to the semantic meanings of the small orange part of the nerf gun, medium blue part of the nerf gun and the whole nerf gun, respectively.

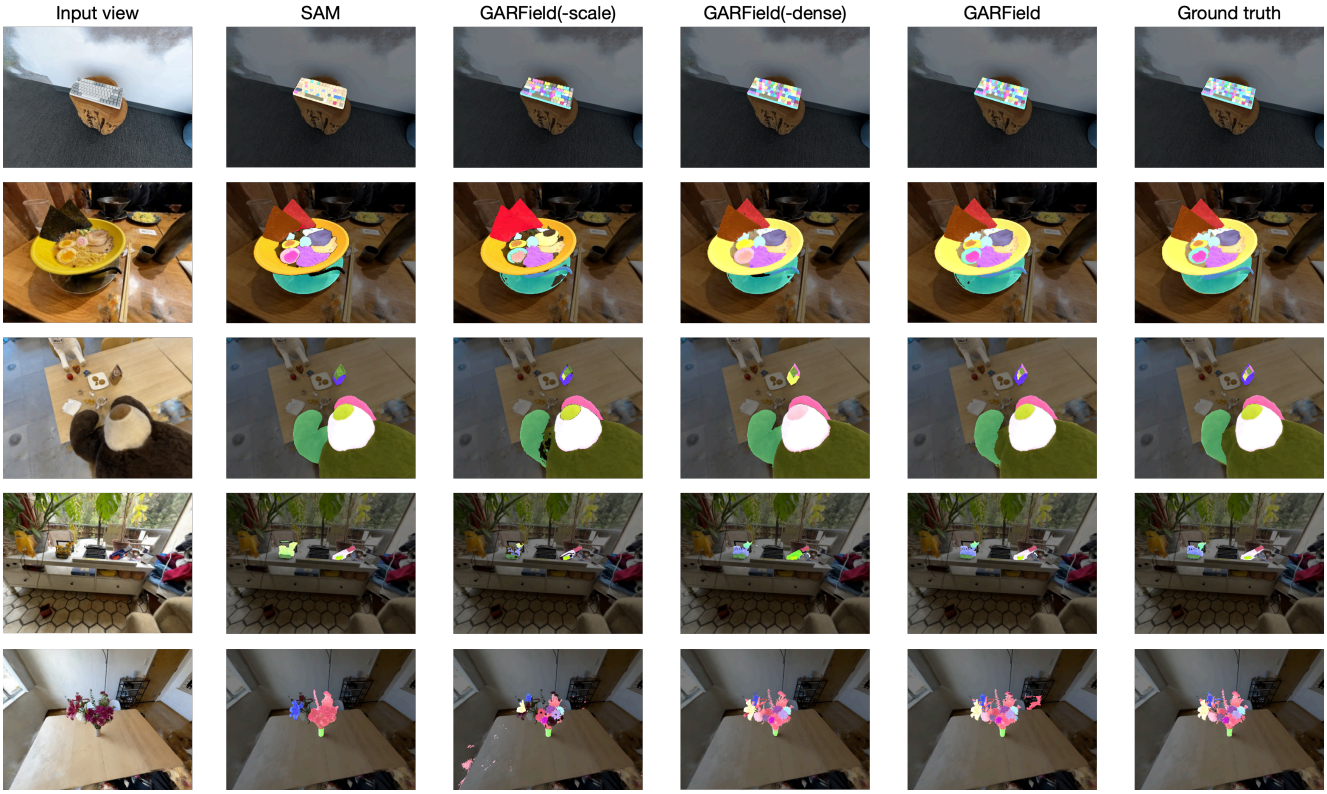

Figure 26. **Hierarchical Grouping Recall Experiments**: We concentrate on methods such as SAM and the ablation study of GARField. GARField outperforms SAM in obtaining finer, smaller masks (e.g. capturing all the tiny keys in a keyboard scene). Unlike GARField without hierarchy grouping, GARField achieves more layered grouping results (e.g. in the ramen scene, it successfully identifies the entire ramen mask through hierarchical clustering). Furthermore, compared to GARField without dense supervision, GARField provides more stable and thorough grouping outcomes (e.g. in the teatime scene, GARField more comprehensively identifies the small labels on the cookie bag).