DIV-FF: Dynamic Image-Video Feature Fields For Environment Understanding in Egocentric Videos Supplementary Material

Lorenzo Mur-Labadia
lmur@unizar.es

Josechu Guerrero
jguerrer@unizar.es

Ruben Martinez-Cantin
rmcantin@unizar.es

1. Annotation procedure for the Affordance Segmentation task

For the evaluation of the Dynamic Object Segmentation task, we reuse the object retrieval task labels from N3F [3]. In contrast, we manually annotate the affordance masks in the EPIC-Diff sequences to quantitatively evaluate the Affordance Segmentation task. For each scene, we identified five interactions occurring within the video. Following the procedure proposed by Nagarajan et. al [1] for interaction hotspots evaluation, we mark 6 points on each image where the interaction could feasibly occur. As Figure 1 shows, these annotations are localized on both the relevant parts of the object for the affordance and grounded in real human demonstration. We then transformed these annotations into a heatmap by centering a Gaussian distribution at each marked point, as shows Figure 1. The ground-truth affordance mask corresponds to the pixels where the heatmap is above 0.5. We show two examples of the binary masks in Figure 2. Overall, we collected 872 binary masks between all the sequences. The list of the evaluated affordances is shown in Table 1. We will release the affordance annotations and the code.

Figure 1. **Key-points marked for the affordance masks.**

2. Computational details

Training. For geometry reconstruction, we downscale the images from the EPIC Kitchens videos to a resolution of 128 × 228. SAM masks are extracted from higher-resolution images (228 × 456) to better capture small objects and fine details. Both EgoVideo patches and CLIP tiles are resized to 224 × 224 prior to feature extraction. We also train per-scene an auto-encoder to reduce the dimensionality of CLIP embeddings from 512 to 128 following [2].

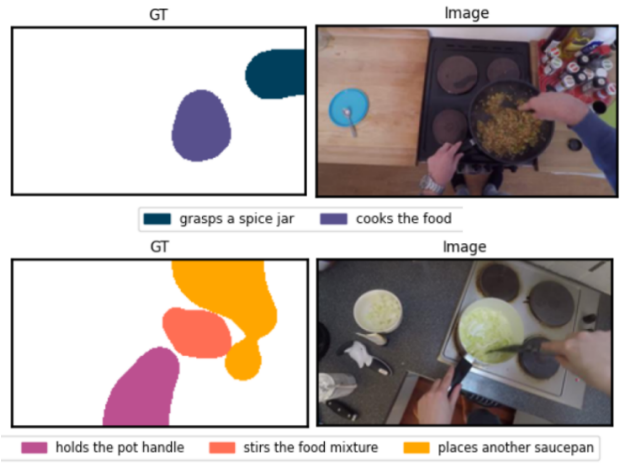

Figure 2. **Affordance Segmentation ground-truth masks.** From the marked points, we derived a Gaussian heatmap. The evaluation affordance masks corresponds where the heatmap is above 0.5.

Testing. Each scene requires approximately 25MB of disk space, compared to the 17MB needed to store only the geometry reconstruction. Since rendering the semantic features is very costly in memory terms, we infer sequentially the persistent, dynamic and actor feature fields during rendering to keep the memory constrains. Our experiments were conducted on an NVIDIA GeForce RTX 4090, where the rendering time for the LERF baseline is 1.48 seconds. During inference, LERF compares relevancy maps across 30 different scales to select the optimal one. In contrast, DIV-FF leverages SAM-generated masks to extract CLIP object-aligned features, eliminating the need for multi-scale inference. This reduces DIV-FF's inference time to 0.82 seconds. It is worth noting that DIV-FF renders both the

Scene	Dynamic Objects text queries	Affordance Segmentation text queries
P01_01	green cutting board, blue lid, cheese grater, saucepan, pot	grasp a spice jar, cook the food, wash a plate, cut vegetables, take a knife with right hand
P03_04	bowl, blue cutting board, knife, pot	cut the onion, wash a kitchen utensil, hold the pot handle, stir the food mixture, place another saucepan
P04_01	white spoon, transparent bottle, white bottle, pan	turns on the faucet, wash a kitchen utensil, cook the food, open a cabinet
P05_01	cup, electric kettle, banana, milk bottle	heat water with the kettle, toast the bread, pour water in the mug, take the bottle milk, heat food in the microwave
P06_03	pot, flour package, jug, orange bag	add ingredients to the mixture, wash kitchen utensils, take the water pitcher, open the bag, knead the ingredients
P08_01	frying pan, coffee cup, cutting board, plate	cook the ingredients, prepare the coffee, cut the ingredients, drink the coffee, spread the mixture on the toast
P09_02	spagueti package, white cutting board, saucepan, plate	hold the pan, heat water, control the stove, prepare the omelette, wash kitchen utensils
P13_03	plate, colander, pasta, scissors	drain the pasta, open a drawer, add sauce to the pasta, open the cheese bag, prepare the food
P16_01	pot, package, knife protector, cutting board	cut the ingredients, cook the ingredients, wash kitchen utensils, open the package, place a frying pan in an available hob
P21_01	white plate, plastic package, blue plate, paper bag	open the fridge, soak the tomatoes, manipulate the ingredients, wash kitchen utensils, take a knife with the right hand

Table 1. **Text queries used during the Dynamic Objects and Affordance Segmentation evaluation experiments.**

image and video feature fields, whereas LERF only extracts image-language features.

3. Extra qualitative results

We report extra qualitative results for the image language feature map of DIV-FF in Figure 3. We first show the PCA feature map of the rendered scene, followed by relevancy maps both for dynamic (*'green cutting board ´, 'spoon ´, 'food ´*...) and static (*'sink ´, 'gas cooktop ´, 'drainer ´*) objects. We also attach videos depicting the full egocentric video sequence, showcasing views from both the actor's and a static perspective (which is a novel view along all the different time-steps). Similarly, the video language of DIV-FF is illustrated in Figure 4, which highlights multiple affordance actions from the same novel viewpoint.

4. Limitations

The image-language field of DIV-FF inherits several limitations from SAM, notably in the excessive segmentation of objects that omits some of its parts. This is evident in cases such as the *'cup ´* in P04-01, *'plate ´* in P13-03 and *'sink ´* in P21-01 examples of Figure 3. The segmentation

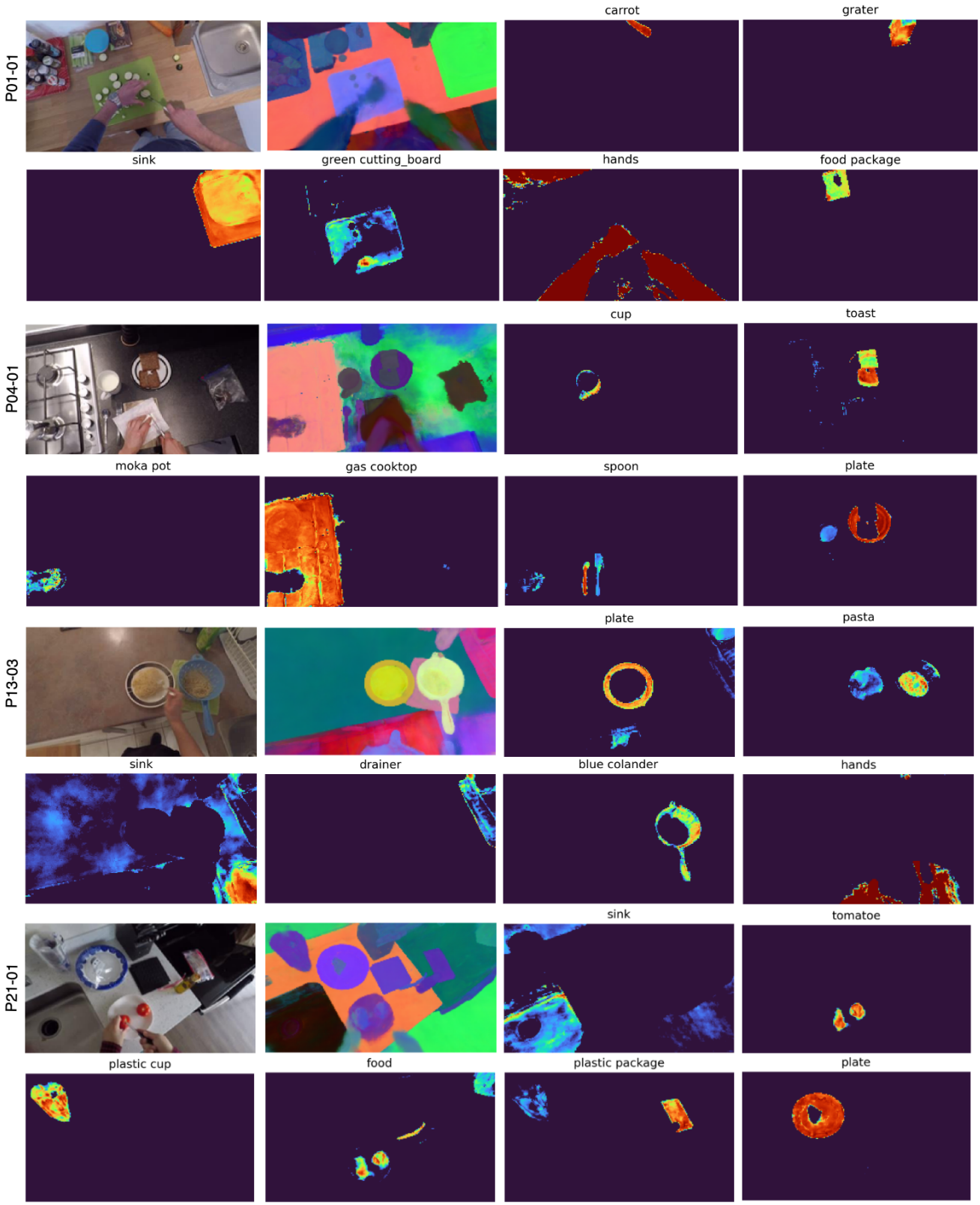

Figure 3. **Additional results of the DIV-FF Image Language relevancy map in novel views.** We visualize the ground-truth image, the PCA of the image-language features and different relevancy maps for different text queries.

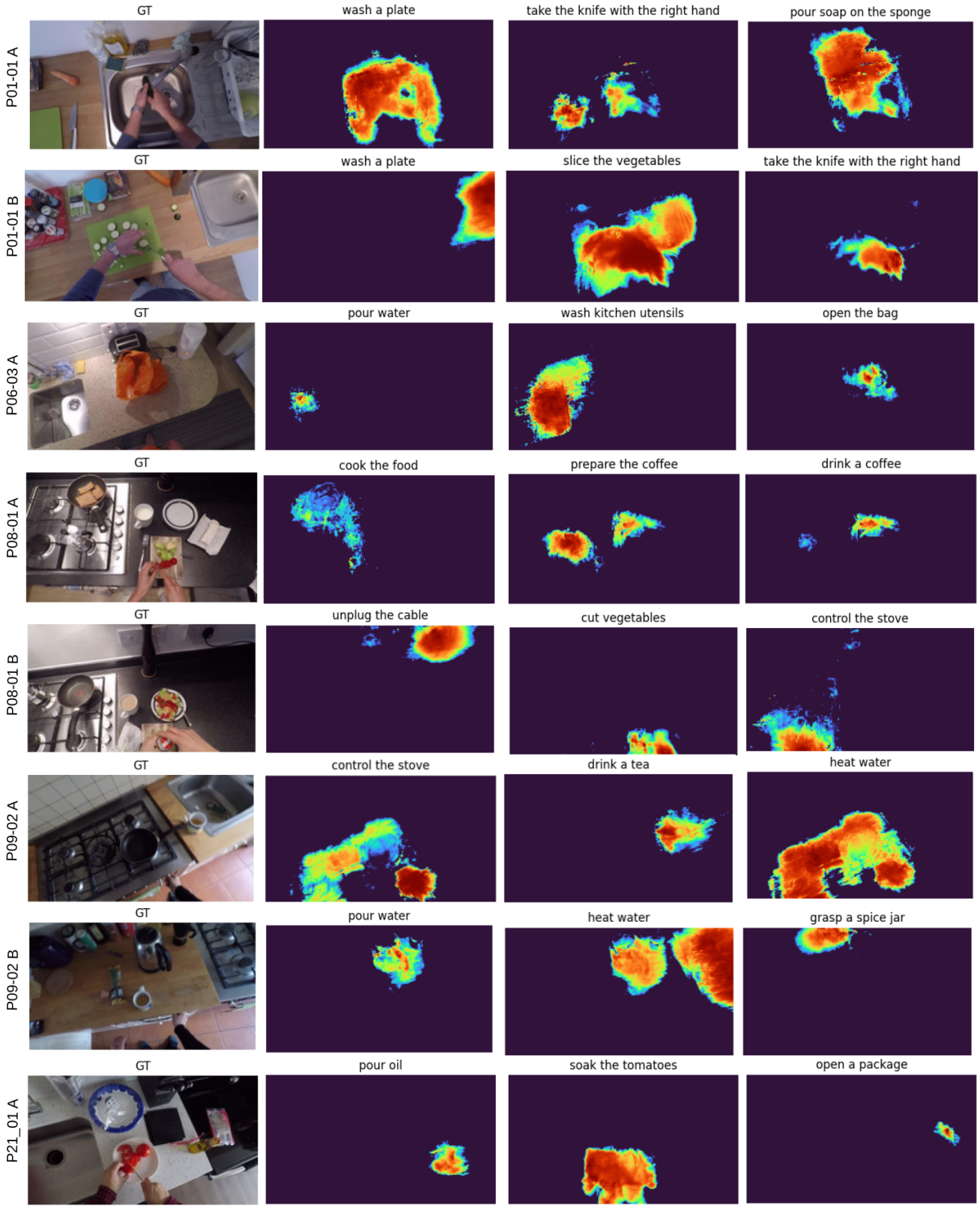

Figure 4. **Additional results of the DIV-FF Image Language relevancy map in novel views.** We visualize the ground-truth image and three different relevancy maps of the video-language feature field corresponding of affordable interactions.

produced by SAM either omits some parts of the objects or introduces artifacts such as holes.

Another limitation of DIV-FF is the degradation associated to the geometry, specially when rendering the actor's hands (scenes P01-01 and P13-03 in Figure 3). The actor's hands continuous movement and the biased top-view (egocentric) perspective of all the images pose a significant challenge in accurately rendering the hands in novel views, which is later reflected in the relevancy maps for the *'hands ´* text query.

Despite the inclusion of persistent, dynamic and actor streams in DIV-FF to enhance the capture of egocentric video, the rendering quality of objects in contact with the actor, such as the *'green cutting board ´* in P01-01 or *'pasta ´* in P13-03 in Figure 3, is compromised. This degradation is primarily due to frequent occlusions by the actor's hands, disrupting the continuity of views

The main limitation in the video-language feature field of DIV-FF is the rendering of diffuse relevancy maps, which introduce excessive noise (*'pour soap on the sponge ´* in P01-01 A or *'control the stove ´* in P09-02 A, both in Figure 4) in some cases.

References

[1] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8688–8697, 2019. 1

[2] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 1

[3] 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. 1