ConceptPose: Training-Free Zero-Shot Object Pose Estimation using Concept Vectors

Supplementary Material

6. Additional ablations

6.1. Backbone Ablation

To demonstrate ConceptPose's adaptability across different vision-language architectures, we evaluate five backbones on REAL275 (Table 1): three SigLIP2 [5] variants (giant-384, large-384, base-384), CLIP ViT-L/14@336px [2], and DINOv3-L/16 [4] with the dinotxt text grounding head [1]. Our baseline method is SigLIP2-giant-384. All experiments maintain identical pipeline configurations without voxelization to isolate the impact of backbone architecture.

SigLIP2 and CLIP (GradCAM-based). For SigLIP2-giant-384 (our baseline), SigLIP2-large-384, and SigLIP2-base-384, we use Hugging Face implementations with GradCAM [3] applied to the `post_layernorm` layer. All SigLIP2 variants use 384×384 input resolution. For CLIP ViT-L/14@336px, we target `visual.transformer.resblocks[-1].ln_1` (layer norm of the final transformer block) at its native 336×336 resolution. Both methods compute image-text similarity through normalized dot products, with GradCAM extracting gradient-weighted spatial activations from intermediate transformer features.

DINOv3 + dinotxt (direct patch similarity). DINOv3, as a vision foundation model (VFM) without native language alignment, requires the dinotxt text grounding head from the official repository's "Pretrained heads - Zero-shot tasks with dino.txt" configuration. The dinotxt head produces 2048-dimensional text embeddings partitioned into two complementary 1024-dim subspaces: the first half aligns with the class token (global semantics), while the second half is explicitly trained to align with patch tokens (spatial semantics). Critically, GradCAM cannot be effectively applied to DINOv3 + dinotxt because the text-patch alignment is learned during dinotxt head training and encoded directly in the feature space—there is no gradient path from text queries to spatial activations at inference time, as text features are pre-encoded and frozen. Instead, we follow the official approach of computing direct cosine similarity between image patch features and the patch-aligned text embeddings (second 1024-dim), which leverages the architecture's inherent spatial grounding without requiring backpropagation. DINOv3 uses 224×224 input resolution following its standard preprocessing.

Results. We evaluate the performance of ConceptPose with different backbones and report the results in Table 1. Our baseline SigLIP2-giant-384 achieves the best performance (72.0% ADD(-S), 60.9% BOP AR), followed by SigLIP2-large-384 (67.0%, 56.8%) and SigLIP2-base-384 (63.8%, 54.0%). DINOv3-L/16 + dinotxt reaches competitive performance (62.3%, 52.6%) despite using lower input resolution (224 vs. 384) and a fundamentally different saliency extraction method (direct patch similarity vs. GradCAM). CLIP ViT-L/14@336px shows the weakest results (54.1%, 46.3%), likely because its contrastive pre-training emphasizes global image-text matching rather than fine-grained spatial correspondence. The 17.9 percentage point gap between SigLIP2-giant-384 and CLIP ViT-L/14@336px (both using GradCAM with comparable resolutions) highlights the importance of pre-training objectives that encourage spatial grounding. These results reveal a clear positive correlation between VLM capacity and ConceptPose performance, demonstrating that our method benefits directly from stronger foundation models while remaining architecture-agnostic.

6.2. Concept Number Ablation

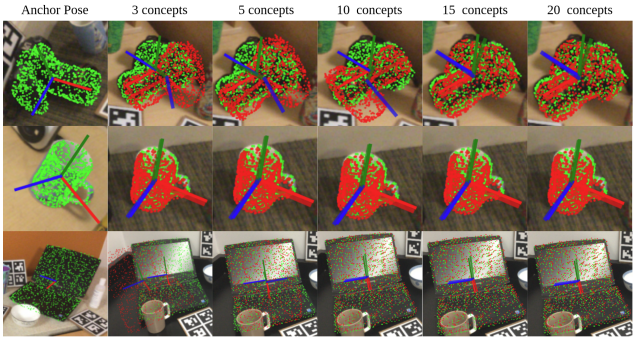

Figure 1. Qualitative visualization of performance changes across different numbers of concepts (L) on REAL275 dataset. The first column is the anchor pose, column 2-6 are estimated query poses with different numbers of concepts (L) and ground truth query pose.

To expand on our analysis of the impact of concept quantity in section 4.5, We present an addidtional evaluation on the number of concepts (L). As explain in section 4.1, we generate L=20 concepts per object type by default, however, in this experiment, we truncate the concept list to 3, 5, 10, 15, and 20 concepts per object to study the impact of using less concepts independent of the quality of the concepts. Figure 1 shows a qualitative visualization of the performance changes on different object types. All experiments use the same pipeline configuration without voxeliza-

Table 1. Ablation study of Vision-Language Model (VLM) backbones on REAL275 dataset. We compare SigLIP2 variants (giant-384, large-384, base-384) against CLIP ViT-L/14@336px and DINOv3-L/16 with dinotxt. † indicates our baseline method used throughout the main paper.

Backbone	ADD(-S)	BOP AR	ADD	ADD-S	MSSD AR	MSPD AR	VSD AR	10°/5cm	5°/2cm	3DmIoU	3DIoU50	3DIoU75
SigLIP2-giant-384†	**72.0**	**60.9**	**60.8**	**91.2**	**66.0**	**70.1**	**46.6**	**47.4**	**26.1**	**76.8**	**90.0**	**69.8**
SigLIP2-large-384	67.0	56.8	56.2	89.6	61.8	65.5	43.2	43.9	22.4	74.0	86.1	64.8
SigLIP2-base-384	63.8	54.0	53.2	88.6	59.1	62.2	40.6	40.9	24.5	72.8	84.8	61.3
CLIP ViT-L/14@336px	54.1	46.3	43.8	84.4	50.3	53.3	35.3	32.5	13.8	67.4	76.5	52.4
DINOv3-L/16 + dinotxt	62.3	52.6	52.0	87.8	57.6	60.8	39.3	40.1	18.9	71.5	82.8	59.7

Table 2. Ablation study on the number of concepts (L) on REAL275 dataset. We evaluate performance using 3, 5, 10, 15, and 20 concepts per object. All experiments use the same pipeline configuration without voxelization. † indicates our default baseline configuration.

#L	ADD(-S)	BOP AR	ADD	ADD-S	MSSD AR	MSPD AR	VSD AR	10°/5cm	5°/2cm	3DmIoU	3DIoU50	3DIoU75
20	**74.0**	**62.4**	**62.4**	**92.5**	**67.5**	**71.8**	**47.9**	**49.2**	**27.6**	**77.9**	**92.2**	**71.9**
15†	72.0	60.9	60.8	91.2	66.0	70.1	46.6	47.4	26.1	76.8	90.0	69.8
10	70.3	59.0	58.7	89.6	63.8	67.8	45.4	45.5	24.6	75.7	88.6	68.3
5	68.3	57.6	57.3	88.5	62.3	66.3	44.2	44.9	24.4	75.3	87.6	67.8
3	64.3	54.9	52.6	88.4	59.2	63.0	42.6	42.2	23.3	74.5	87.2	64.5

tion. In Table 2, the results show consistent performance improvements with more concepts, with $L = 20$ achieving the best overall performance. However, as our baseline configuration we used $L = 15$ in order to balance between performance and computational efficiency.

6.3. Correspondence Method Ablation

Table 3. Ablation study of correspondence methods for matching concept distributions between query and anchor point clouds on REAL275 dataset. We compare five different divergence and similarity measures. All methods use the same pipeline configuration.

Method	**ADD(-S)**	**BOP AR**
Bidirectional KL	72.0	**0.6107**
KL Divergence	72.0	0.6085
Reverse KL	72.0	0.6072
Asymmetric	72.1	0.6053
Cosine	72.1	0.6042

To evaluate the impact of different correspondence methods for matching concept vectors between query and anchor, we compare five divergence and similarity measures on REAL275 (Table 3). All methods achieve similar performance, with ADD(-S) 10cm ranging from 72.0% to 72.1% and BOP scores from 0.6042 to 0.6107. Bidirectional KL divergence achieves the best BOP score (0.6107), while Asymmetric and Cosine tie for the highest ADD(-S) at 72.1%. The minimal performance variation suggests that ConceptPose's effectiveness is robust to the choice of correspondence method, as all tested measures effectively capture the semantic similarity encoded in concept vectors.

7. Performance Analysis by Object

Table 4. Concepts used for REAL275 evaluation.

Cat.	#L	Concepts
bottle	20	cap, lid, spout, nozzle, pump, trigger, body, base, neck, shoulder, label, handle, grip, threads, tamper_evident_ring, overcap, sleeve, collar, punt, carry_loop
bowl	20	body, rim, bottom, foot, handle, lid, spout, ear, pedestal, inner_surface, outer_surface, rib, embossment, medallion, divider, perforation, knob, groove, notch, marking
camera	20	lens, camera_body, viewfinder, display_screen, grip, shutter_button, mode_dial, control_dial, flash, hot_shoe, lens_cap, lens_hood, eyecup, focus_ring, zoom_ring, battery_door, memory_card_door, strap_lug, lens_mount, directional_pad
can	20	body, lid, base, rim, pull_tab, rivet, score_line, shoulder, cap, nozzle, logo, barcode, nutrition_facts, ingredients_list, ridges, seam, net_content_statement, warning_label, recycling_symbol, product_image
laptop	20	screen, keyboard, touchpad, webcam, hinge, lid, bottom_case, port, vent, speaker_grille, logo, bezel, rubber_foot, power_button, keyboard_deck, indicator_light, fingerprint_sensor, screw, touchpad_button, security_slot
mug	20	body, handle, rim, base, interior, lid, sleeve, logo, decorative_element, handle_aperture, foot_ring, slider, sipper_hole, gasket, thumb_rest, spout, infuser, carrying_loop, vent_hole, spout_cover

We use the following default prompt to generate concepts. For practical efficiency, we generate one shared concept vocabulary per object type ($L = 20$ concepts), you can find the generated concepts in Table 4.

Table 5. Performance breakdown by object type on REAL275 dataset. We report results across all six object types (bottle, bowl, camera, can, laptop, mug) with the number of anchor-query pairs evaluated for each.

Obj. Type	#Pairs	ADD(-S)	BOP AR	VSD AR	MSSD AR	MSPD AR	ADD-S	ADD	10°/5cm	5°/2cm	3DmIoU	IoU@50	IoU@75
bottle	13	100.0	68.7	68.5	90.0	47.7	100.0	0.0	0.0	0.0	84.9	100.0	100.0
bowl	150	82.7	45.6	31.8	85.2	19.9	82.7	5.3	2.0	0.0	72.9	93.3	54.0
camera	482	68.9	72.5	45.0	96.6	76.0	95.9	68.9	32.2	1.7	76.6	93.8	67.4
can	144	85.4	72.4	48.5	97.2	71.5	85.4	23.6	6.2	0.0	74.8	85.4	66.7
laptop	651	76.8	65.0	45.7	81.9	67.5	83.3	76.8	76.3	60.7	75.4	78.8	73.7
mug	560	60.4	75.6	49.4	97.4	80.0	98.4	60.4	49.8	20.7	78.6	97.7	70.0

"Give me {num_labels} labels that describe different concepts for a {object_label}. I will be using these labels for localization, please make sure they are generalizable to different instances within the same category {object_label}, semantically orthogonal to each other, and must be visible from at least one external viewpoint. Also, please use the most common names and do not use positional descriptions. Please just give me all the labels as a python list, no additional explanations please."

Table 5 shows detailed performance metrics for each object type on REAL275 using our baseline configuration (SigLIP2-giant-384 with $L = 15$ concepts). Performance varies significantly across object types, revealing several patterns. Laptop achieves the strongest strict pose accuracy (5°/2cm: 60.7%), likely due to its distinctive geometric features and rich visual texture (keyboard, ports, logos) that provide reliable concept localization. Mug shows the highest BOP AR (75.6%) despite moderate ADD(-S) (60.4%), suggesting the method produces visually plausible poses even when metric accuracy is lower. Bowl demonstrates a large gap between ADD-S (82.7%) and ADD (5.3%), reflecting the expected behavior for highly symmetric objects where multiple orientations are geometrically valid. Camera and can both achieve strong BOP performance (72.5%, 72.4%), indicating robust pose estimation on objects with moderate geometric complexity. Bottle's small sample size (13 pairs) limits statistical significance despite perfect ADD(-S) recall.

8. Continuous Concept Vectors vs. Binary Masks

Our GradCAM-based saliency provides continuous activations, as opposed to binary masks produced by segmentation-based approaches. To validate the advantage of soft activations, we create binary concept vector masks by thresholding at 0.5 and evaluate on REAL275. As shown in Table 6, both ADD(-S) (71.5% → 30.5%) and BOP AR (60.4% → 28.2%) decrease strongly, indicating that continuous concept vectors are important for establishing accurate correspondences.

Table 6. Continuous concept vectors vs. binary masks (thresholded at 0.5) on REAL275.

	ADD(-S)	BOP AR	ADD	ADD-S	10°/5cm	5°/2cm
Concept vectors	**71.5**	**60.4**	**60.6**	**90.8**	**47.2**	**26.0**
Binary masks (thr.=0.5)	30.5	28.2	25.2	62.7	17.7	6.1

9. Effect of Concept Quantity on Correspondence Quality

Spatially extended and repeated parts are a strength of our method: a single spatially localized concept (e.g., car door) can anchor the object, but cannot resolve its scale or orientation. In contrast, a repeated concept (e.g., wheel) helps resolve the object's location, scale, and one additional DoF of orientation. This is analogous to a 2D line segment resolving 4 DoF while a point resolves only 2 DoF. Figure 2 shows this behavior with increasing number of concepts (3, 10, 15) on a car object.

10. Occlusion Robustness Analysis

We present a correlation analysis between the amount of occlusion and pose error (ADD-S) on all objects in LINEMOD-Occlusion (LM-O), sampling 20 anchor-query pairs randomly for each object. Figure 3 shows the results. Regression analysis shows $R^2 < 0.3$, indicating no linear correlation between occlusion and error. ConceptPose handles occlusion through several design choices: we crop the input image to the object's bounding box before VLM inference to alleviate potential contamination due to the patch nature of ViTs, backproject only pixels within the object mask to 3D to reduce background noise, and utilize RANSAC's robustness against partial observations by finding the largest geometrically consistent subset among visible points.

11. Cross-Instance Concept Generalization

In our task, concepts are inherently shared across instances within a category—not per instance. Figure 4 visualizes the same concepts (lens, grip, controls, hot_shoe) activated across different camera instances, demonstrating

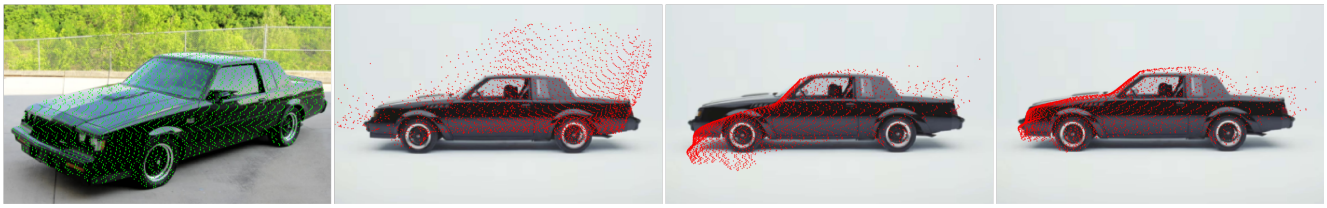

Figure 2. Pose estimation with increasing number of concepts (3, 10, 15) on a car object.

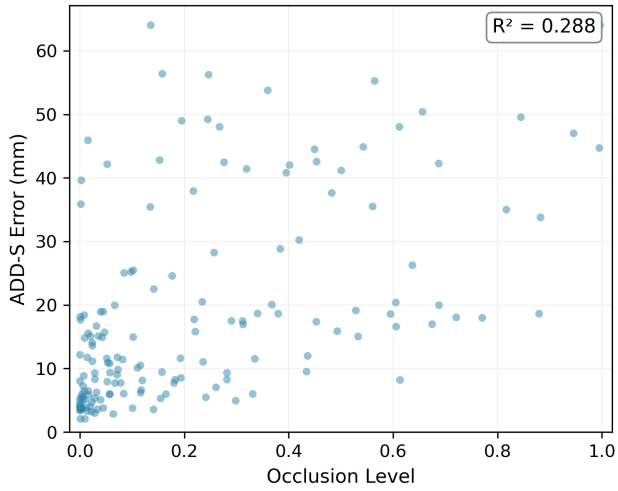

Figure 3. Correlation between occlusion ratio and ADD-S error on LINEMOD-Occlusion. Each subplot shows one object with 20 randomly sampled anchor-query pairs. Regression analysis shows $R^2 < 0.3$, indicating no linear correlation between occlusion and error.

our training-free approach's ability to generalize across instances, unlike learned geometric descriptors.

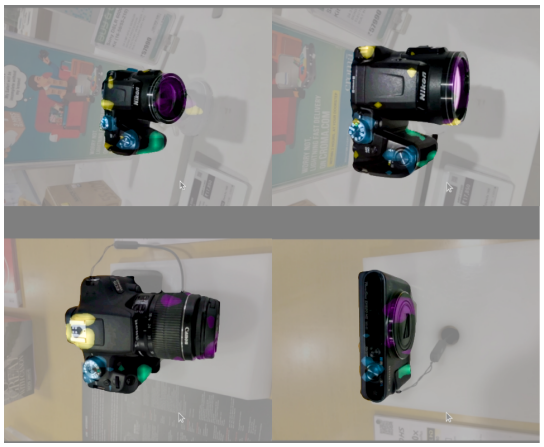

Figure 4. Cross-instance concept activation on different camera instances. The same category-level concepts [lens, grip, controls, hot_shoe] are consistently localized across diverse instances.

12. Extended Qualitative Results

Figures 5 and 6 present 10 extended qualitative visualizations of ConceptPose across all four evaluation datasets (REAL275, Toyota-Light, YCB-Video, LINEMOD), with saliency maps and estimated poses. It demonstrates ConceptPose's ability to generate semantically meaningful concept activations and accurate pose estimates across diverse object categories and imaging conditions.

References

[1] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. Dinov2 meets text: A unified framework for image- and pixel-level vision-language alignment, 2024. 1

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, 2021. 1

[3] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020. 1

[4] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. 1

[5] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, 2025. 1

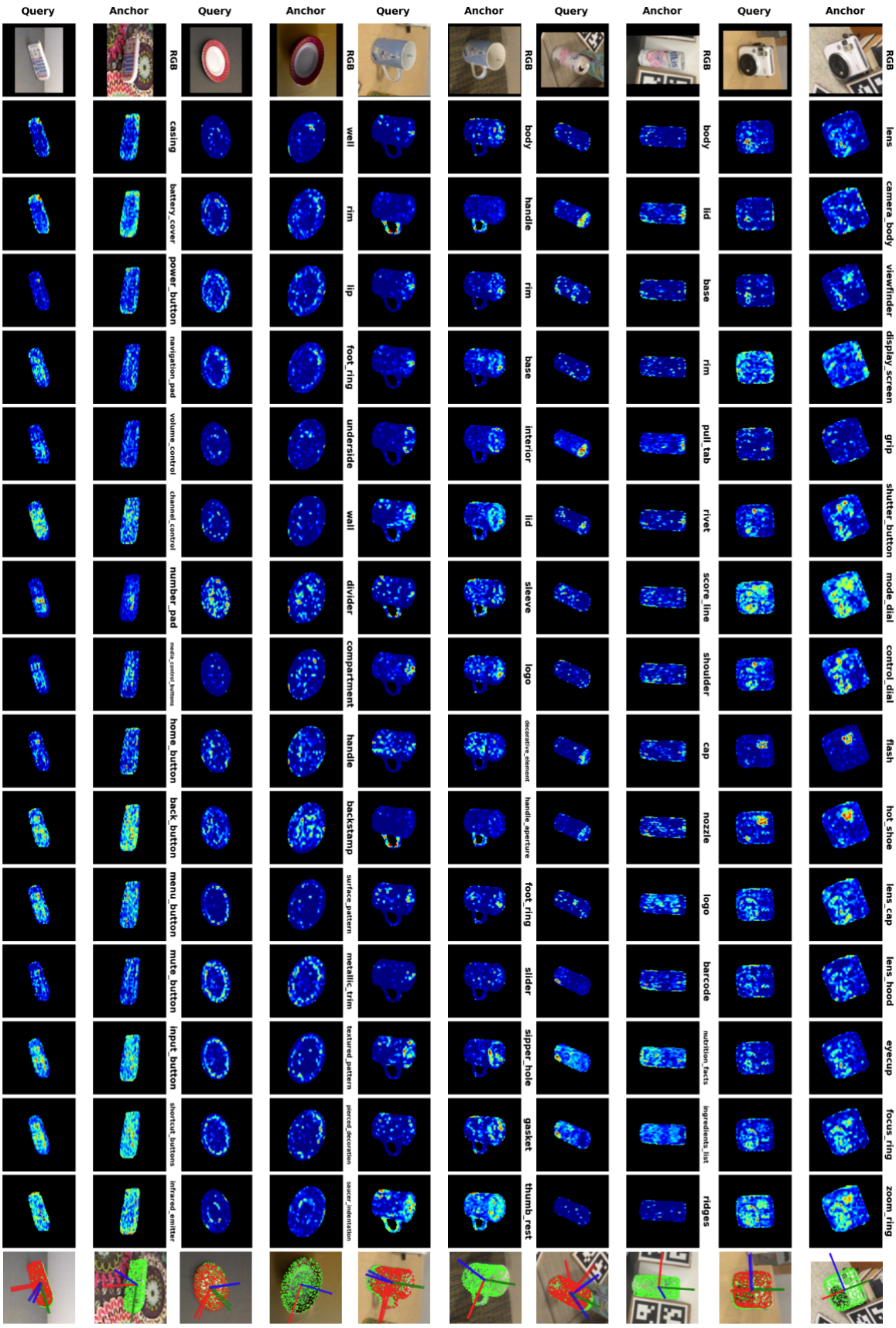

Figure 5. Qualitative results (1/2), estimated query pose, ground truth anchor/query pose.

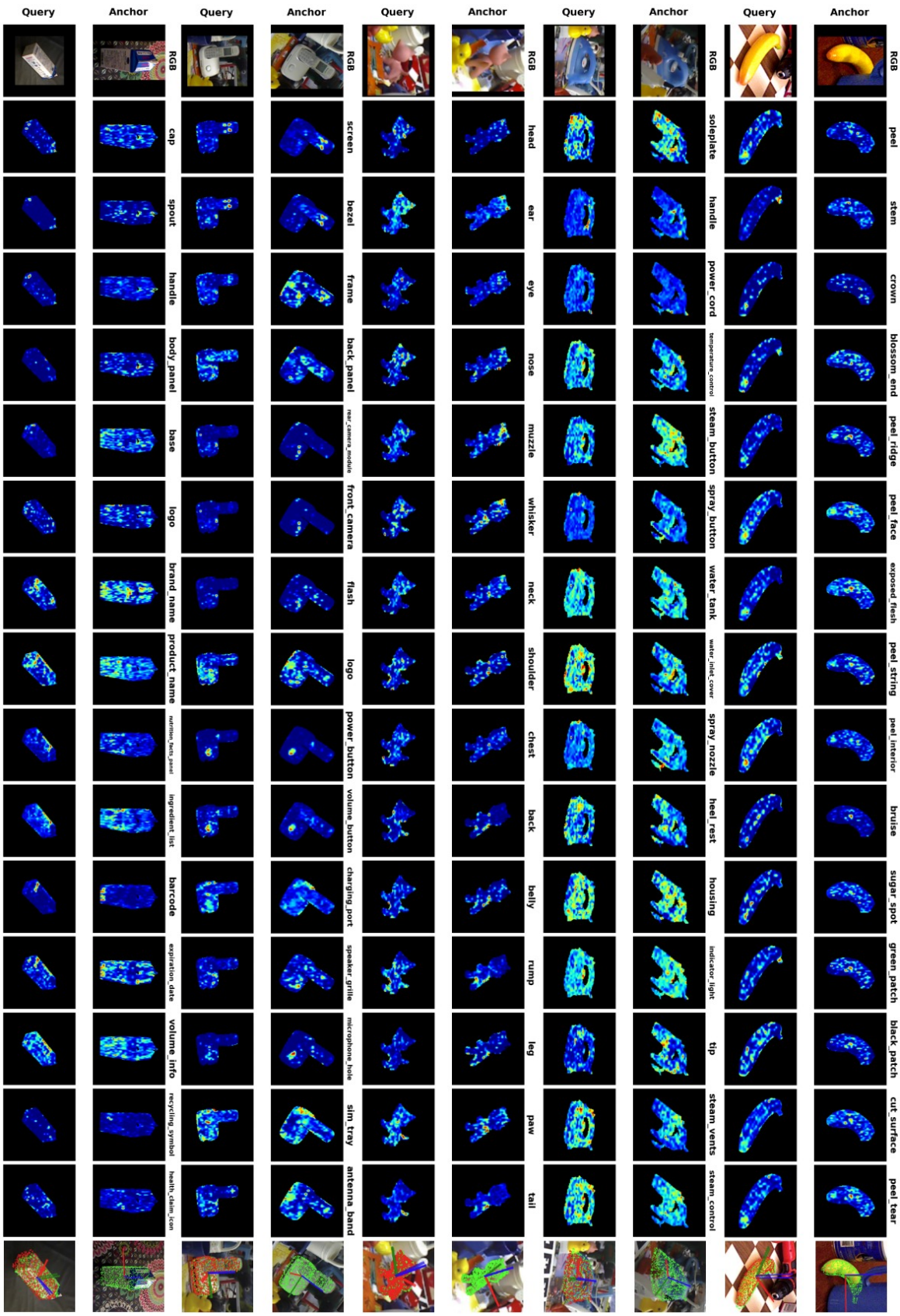

Figure 6. Qualitative results (2/2), estimated query pose, ground truth anchor/query pose.