A1. Additional Studies

Necessity of Open-world Object Detector. Challenging scenes, such as cluttered or complex backgrounds, occlusions, or variations in lighting conditions, can pose significant challenges for traditional two-view object detection and pose estimation (see the main draft). Whereas our proposed method utilize a open-world object detector, not limited to a specific group of classes, improves the zero-shot generalization by the retrieval-and-matching strategy. When retrieving using the global feature representation, which may mistakenly have large activations with non-related objects (Figure A1), results in the inaccurate 6DoF estimation in the later stage. The proposed hierarchical representation for object retrieval across viewpoints (Table A1), both improves the segmentation and retrieval accuracy, as well as benefits the subsequent pose estimation.

Quantitative Results on Each Instance We provide a comprehensive analysis of the average median error and pose accuracy across various thresholds. Specifically, we present per-instance metrics for two-view 6DoF object pose estimation, focusing on datasets with clustered backgrounds, namely LINEMOD [3] and YCB-Video [1]. The results are summarized in Table A2 and Table A3, revealing a significant improvement in both per-instance accuracy and overall accuracy. This observation highlights the effectiveness of our promptable approach in mitigating the negative impact of background clutter and substantially enhancing the estimation accuracy.

Furthermore, we present per-instance metrics for two-view 6DoF object pose estimation on datasets containing single object with rich textures [5] and poor textures [2] of each scebe. As depicted in Table A4 and Table A5, our method outperforms other two-view-based methods in terms of pose accuracy, with assistance of foreground object segmentation and retrieval.

Table A1. **Ablation Studies.** We conducted an analysis of the segmentation, retrieval, and relative pose estimation tasks to validate the model design. The correlation-based detector in Gen6D [4] often performs poorly in the clustered LINEMOD dataset when using only a single reference image (top row). The proposed framework, utilizing an *Open-world Detector* that relies on global representation (second row), shows slightly lower performance compared to our full model, which incorporates hierarchical representation (last row). The results are averaged over a subset comprising 1/10 of the LINEMOD dataset.

Method	Segmentation Acc.		Retrieval Acc.	Pose Acc.		
	mIoU (↑)	Accuracy(↑)	mAP↑	Med. Err↓	Acc30↑	Acc15↑
Gen6D [4]	0.087	0.102	0.067	44.644	0.369	0.106
Ours(Global,Top-1)	0.605	0.815	0.817	14.912	0.787	0.493
Ours(Hierarchical,Top-3)	**0.621**	**0.842**	**0.844**	**12.639**	**0.810**	**0.529**

Table A2. We conduct experiments on zero-shot two-view object pose estimation on LINEMOD dataset, and report Median Error and Accuracy at 30º, 15º averaged across all 13 scenes.

Metrics	Method	Per Instance													
		ape	benchvise	camera	can	cat	driller	duck	eggbox	glue	holepuncher	iron	lamp	phone	Avg
Acc15 (↑)	Gen6D	0.016	0.125	0.112	0.120	0.028	0.157	0.029	0.114	0.023	0.0670	0.108	0.241	0.105	0.096
	LoFTR	0.091	0.423	0.338	0.429	0.172	0.445	0.190	0.433	0.119	0.253	0.411	0.582	0.322	0.324
	Ours	0.439	0.450	0.493	0.531	0.444	0.47916	0.456	0.607	0.380	0.502	0.585	0.467	0.445	0.483
Acc30 (↑)	Gen6D	0.133	0.445	0.400	0.485	0.232	0.482	0.203	0.437	0.147	0.279	0.496	0.609	0.380	0.364
	LoFTR	0.291	0.663	0.608	0.7125	0.388	0.687	0.370	0.722	0.248	0.480	0.738	0.855	0.542	0.562
	Ours	0.789	0.710	0.764	0.826	0.732	0.765	0.758	0.840	0.686	0.809	0.857	0.733	0.743	0.770
Med. Err (↓)	Gen6D	79.705	32.504	35.970	30.407	54.468	30.665	57.292	31.781	88.044	45.288	30.094	25.551	39.392	44.855
	LoFTR	70.094	19.227	22.550	17.585	43.069	18.356	44.083	16.887	90.000	31.782	17.904	11.871	26.063	33.036
	Ours	16.716	17.762	15.102	12.699	17.921	15.926	17.641	10.530	19.144	14.779	13.157	16.203	16.929	15.731

Table A3. We conduct experiments on zero-shot two-view object pose estimation on YCB-Video dataset, and report Median Error and Accuracy at 30º, 15º averaged across all 10 scenes.

Metrics	Method	Per Instance										
		001	002	003	004	005	006	007	008	009	010	Avg
Acc15 (↑)	Gen6D	0.046	0.063	0.028	0.017	0.084	0.027	0.250	0.102	0.073	0.085	0.077
	LoFTR	0.483	0.539	0.297	0.245	0.457	0.298	1.000	0.4953	0.508	0.457	0.478
	Ours	0.441	0.547	0.401	0.457	0.521	0.381	0.937	0.738	0.524	0.492	0.544
Acc30 (↑)	Gen6D	0.204	0.190	0.140	0.108	0.253	0.138	0.562	0.308	0.221	0.192	0.232
	LoFTR	0.637	0.817	0.481	0.485	0.739	0.506	1.000	0.785	0.6885	0.721	0.686
	Ours	0.655	0.857	0.755	0.748	0.816	0.680	1.000	0.953	0.778	0.771	0.801
Med. Err (↓)	Gen6D	53.87	49.995	80.992	64.819	50.587	66.999	27.633	45.461	53.817	50.597	54.477
	LoFTR	17.198	13.484	36.942	31.474	17.832	28.999	2.038	15.359	14.613	17.475	19.541
	Ours	18.582	12.133	18.385	17.257	14.171	20.100	1.408	7.7875	14.156	15.428	13.941

Table A4. We conduct experiments on zero-shot two-view object pose estimation on OnePose dataset, and report Median Error and Accuracy at 30º, 15º averaged across all 10 objects.

Metrics	Method	Per Instance										
		aptamil	jzhg	minipuff	hlyormosiapie	brownhouse	oreo	mfmilkcake	diycookies	taipingcookies	tee	Avg
Acc15 (↑)	Gen6D	0.350	0.445	0.387	0.397	0.424	0.421	0.417	0.357	0.394	0.299	0.389
	LoFTR	0.872	0.931	0.964	0.897	0.984	0.957	0.947	0.822	0.975	0.834	0.918
	Ours	0.871	0.959	0.925	0.886	0.968	0.975	0.920	0.8	0.963	0.849	0.911
Acc30 (↑)	Gen6D	0.845	0.914	0.925	0.901	0.944	0.914	0.923	0.796	0.938	0.831	0.893
	LoFTR	0.945	0.982	0.982	0.978	0.992	0.992	0.969	0.878	0.993	0.918	0.963
	Ours	0.949	0.979	0.973	0.974	0.976	0.985	0.967	0.895	0.993	0.930	0.962
Med. Err (↓)	Gen6D	19.542	16.356	17.348	17.500	16.747	16.612	16.963	19.132	17.787	19.867	17.785
	LoFTR	5.407	4.182	3.978	3.869	3.555	3.938	4.077	5.041	4.147	5.312	4.351
	Ours	2.997	1.460	1.786	2.155	1.470	1.2033	1.765	2.769	2.147	3.799	2.155

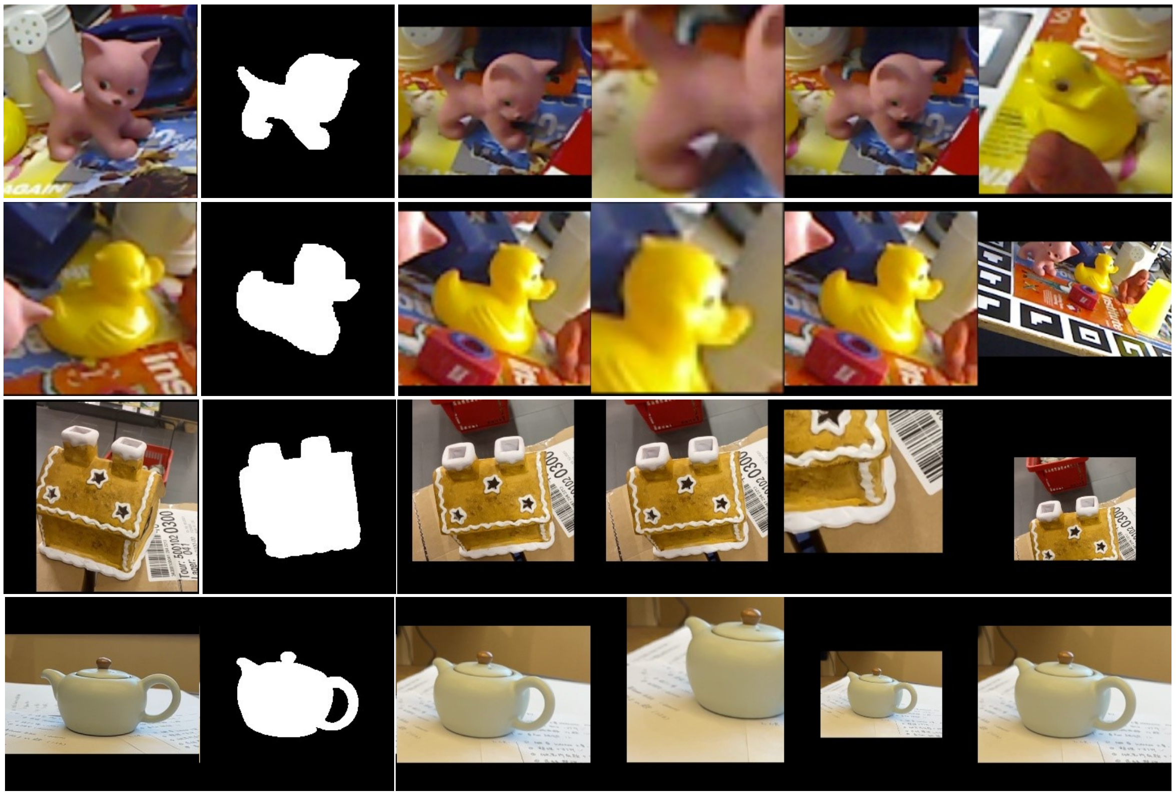

Figure A1. **Ablation study**. Visualizations of retrieved object masks and proposals, selected from the Top-3 proposals using the global [CLS] token similarity.

Table A5. We conduct experiments on zero-shot two-view object pose estimation on OnePose++ dataset, and report Median Error and Accuracy at 30º, 15º averaged across all 9 objects.

Dataset	Method	Per Instance									
		toyrobot	yellowduck	sheep	fakebanana	teabox	orange	greenteapot	lecreusetcup	insta	Avg
Acc15 (↑)	Gen6D	0.171	0.123	0.197	0.156	0.204	0.135	0.185	0.185	0.067	0.158
	LoFTR	0.794	0.676	0.772	0.68	0.782	0.685	0.783	0.708	0.443	0.703
	Ours	0.753	0.768	0.781	0.683	0.844	0.7	0.860	0.708	0.460	0.728
Acc30 (↑)	Gen6D	0.451	0.361	0.472	0.423	0.478	0.388	0.479	0.413	0.232	0.411
	LoFTR	0.912	0.901	0.922	0.893	0.903	0.855	0.969	0.928	0.738	0.891
	Ours	0.882	0.936	0.901	0.88	0.919	0.905	0.953	0.907	0.781	0.896
Med. Err (↓)	Gen6D	32.998	35.811	31.366	36.202	31.536	36.829	30.609	35.185	48.317	35.428
	LoFTR	6.368	9.773	7.336	8.751	6.488	9.439	7.348	8.472	17.136	9.012
	Ours	3.792	5.470	4.435	4.990	3.194	8.044	3.967	6.072	16.492	6.273

References

[1] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015. A1

[2] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *arXiv preprint arXiv:2301.07673*, 2023. A1

[3] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*, pages 548–562. Springer, 2013. A1

[4] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 298–315. Springer, 2022. A2

[5] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. A1