D^{3}Fields: Dynamic 3D Descriptor Fields for Zero-Shot Generalizable Robotic Manipulation

Yixuan Wang[1*], Mingtong Zhang[1*], Zhuoran Li[2,3*], Katherine Driggs-Campbell[1], Jiajun Wu[2], Li Fei-Fei[2], Yunzhu Li[1,2]

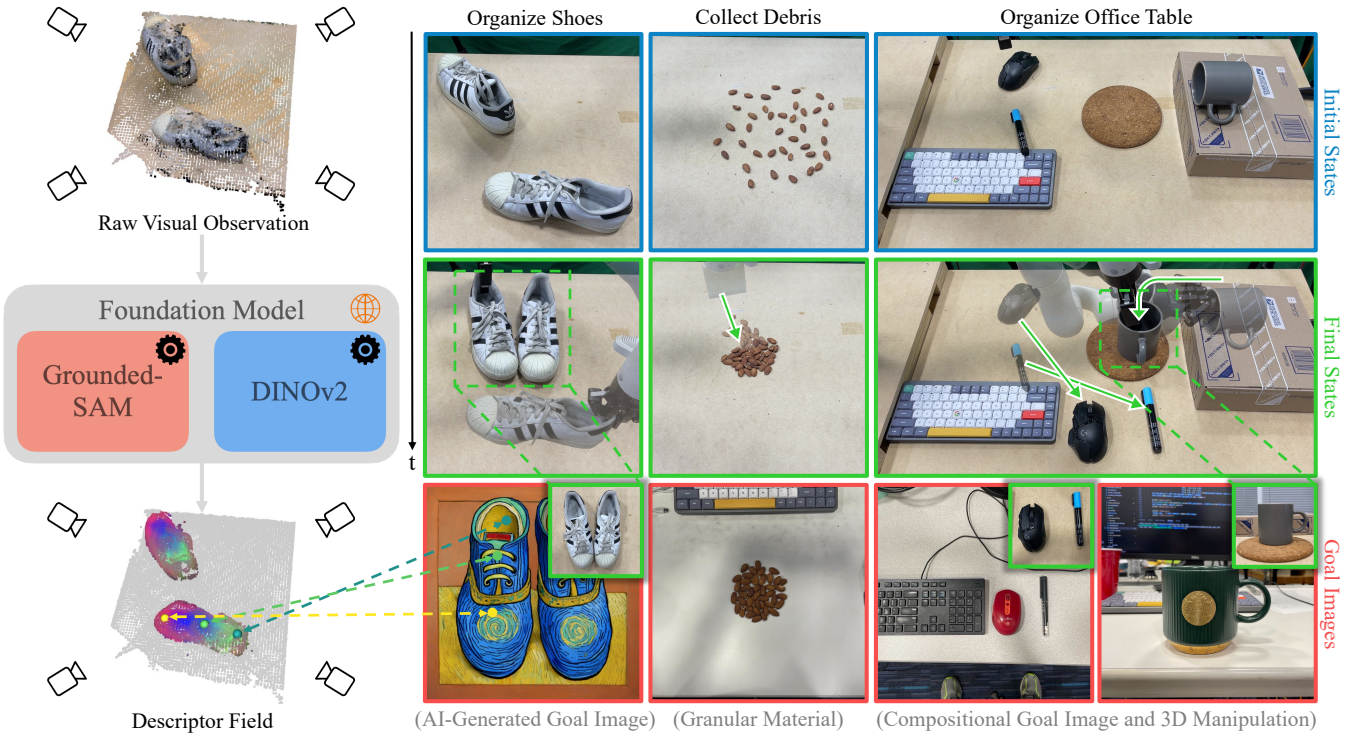


Fig. 1: **D^{3}Fields Representation and Application to Various Manipulation Tasks.** D^{3}Fields take in multi-view RGBD images and encode semantic features and instance masks using foundational models. The gray and colored points in the bottom left visualize background and semantic features mapped to RGB space using Principal Component Analysis (PCA), demonstrating consistency across instances. We use our representation for diverse tasks in a zero-shot manner. These tasks are defined by 2D goal images with diverse instances and styles. We address pick-and-place tasks such as shoe organization and tasks requiring dynamic modeling like collecting debris. We also demonstrate in the office table organization that our framework can accomplish 3D manipulation and compositional task specification.

***Abstract*— Scene representation has been a crucial design choice in robotic manipulation systems. An ideal representation should be 3D, dynamic, and semantic to meet the demands of diverse manipulation tasks. However, previous works often lack all three properties simultaneously. In this work, we introduce D^{3}Fields — dynamic 3D descriptor fields. These fields capture the dynamics of the underlying 3D environment and encode both semantic features and instance masks. Specifically, we project arbitrary 3D points in the workspace onto multi-view 2D visual observations and interpolate features derived from foundational models. The resulting fused descriptor fields allow for flexible goal specifications using 2D images with varied contexts, styles, and instances. To evaluate the effectiveness of these descriptor fields, we apply our representation to a wide range of robotic manipulation tasks in a zero-shot manner. Through extensive evaluation in both real-world scenarios and simulations, we demonstrate that D^{3}Fields are both generalizable and effective for zero-shot robotic manipulation tasks. In quantitative comparisons with state-of-the-art dense descriptors, such as Dense Object Nets and DINO, D^{3}Fields exhibit significantly better generalization abilities and manipulation accuracy. Project Page: https://robopil.github.io/d3fields/**

*Denotes equal contribution. https://robopil.github.io/d3fields/
[1]University of Illinois Urbana-Champaign [2]Stanford Univeristy [3]National University of Singapore

I. Introduction

The choice of scene representation is essential in robotic systems. An ideal representation for robotic manipulation tasks in everyday settings is expected to encompass three key attributes: 3D space structure, dynamic adaptability, and semantic richness. This inclusive approach ensures that robots can efficiently and accurately perform a wide range of manipulation tasks, adapting to the complexities and variability of real-world environments. However, previous

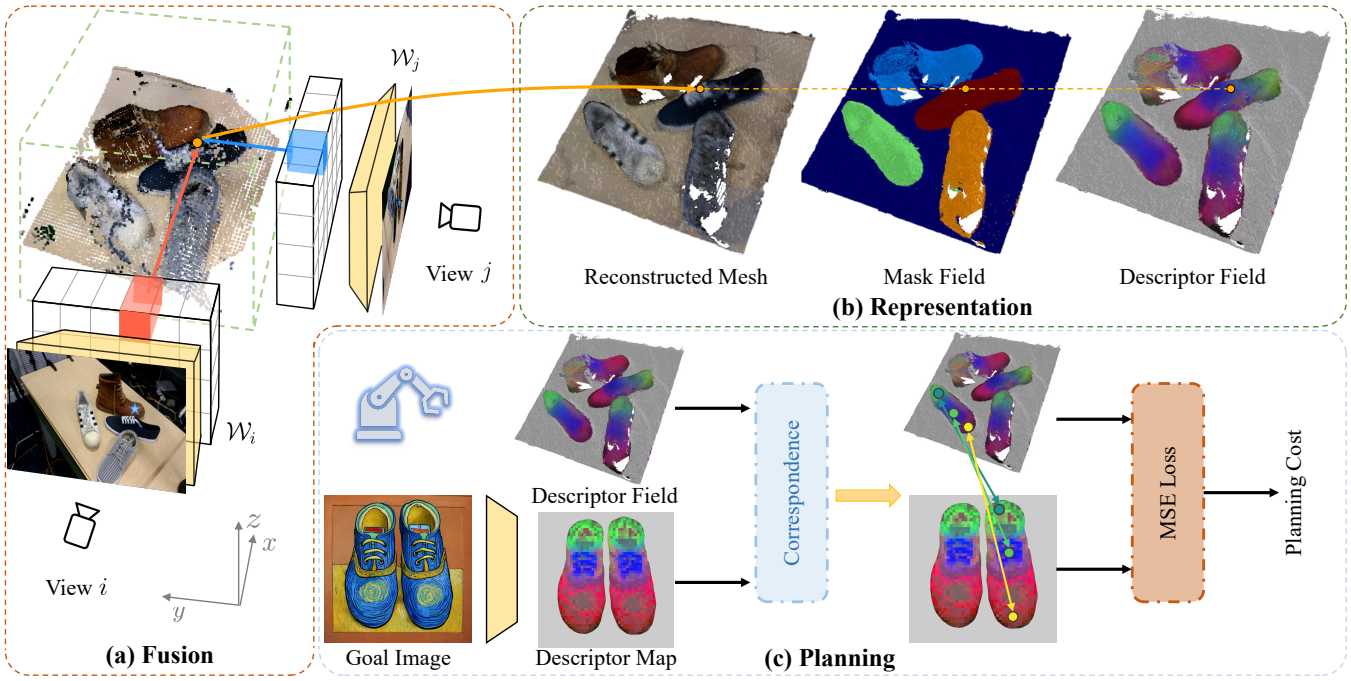


Fig. 2: **Overview of the Proposed Framework.** (a) The fusion process fuses RGBD observations from multiple views. Each view is processed by foundation models to obtain the feature volume $\mathcal{W}$. Arbitrary 3D points are processed through projection and interpolation. (b) After fusing information from multiple views, we obtain an implicit distance function to reconstruct the mesh form. We also have instance masks and semantic features for evaluated 3D points, as shown by the mask field and descriptor field in the top right subfigure. (c) Given a 2D goal image, we use foundation models to extract the descriptor map. Then we correspond 3D features to 2D features and define the planning cost based on the correspondence.

research on scene representations in robotics often does not encompass all three properties. Some representations exist in 3D space [1–4], yet they overlook semantic information. Others focus on dynamic modeling [5–8], but only consider 2D data, neglecting the role of 3D space. Some other works are limited by only considering semantic information such as object instance and category [9–13].

In this work, we aim to satisfy all three criteria and address these limitations by introducing D^{3}Fields, unified descriptor fields that are 3D, dynamic, and semantic. D^{3}Fields represent an approach that offers a comprehensive solution that encapsulates the spatial information, the temporal evolution of dynamic systems, and the understanding of semantic context.

D^{3}Fields processes arbitrary points within the 3D world coordinate frame, yielding both rich geometric and semantic information pertinent to these points. This includes the precise instance mask, dense semantic features, and the signed distance to the object surface. Notably, deriving these descriptor fields requires no training and is conducted in a zero-shot manner, utilizing large visual foundation models and vision-language models (VLMs). In our approach, we employ a set of advanced models. We first use Grounding-DINO [14], Segment Anything (SAM) [15], XMem [16], and DINOv2 [17] to extract information from multi-view 2D RGB images. Subsequently, we project the 3D points back to each camera's perspective, interpolating to compute representations from each viewpoint, and fuse the data to derive the descriptors for the associated 3D points, as shown in Fig. 1 (left). Leveraging the dense semantic features and instance mask of our representation, we achieve robust tracking 3D points specific to target object instances and then train the dynamics models. These learned dynamics models can be incorporated into a Model-Predictive Control (MPC) framework. This integration is pivotal for planning and executing complex manipulation tasks, demonstrating the practical applications and effectiveness of our approach in real-world scenarios.

Notably, The representations of D^{3}Fields are uniquely capable of handling goal specifications derived from 2D images, whether sourced from the internet, smartphones, or even generated by AI models. This capability addresses a significant challenge encountered in previous methods: the difficulty in managing goal images due to their varied styles, contexts, and object instances, which often differ markedly from the robot's workspace environment. Our proposed D^{3}Fields adeptly establishes dense correspondences between the robot workspace and the target configurations. These correspondences give us the task objective, enabling us to precisely plan the robot's actions with the learned dynamics model within the MPC framework. Remarkably, this task execution process does not require any further training, offering a highly flexible and convenient interface for humans to instruct robots.

We evaluate our method across a wide range of household robotic manipulation tasks in a zero-shot manner. These tasks include organizing shoes, collecting debris, and organizing office desks, as shown in Fig. 1 (right). Furthermore, we offer detailed quantitative comparisons between our method

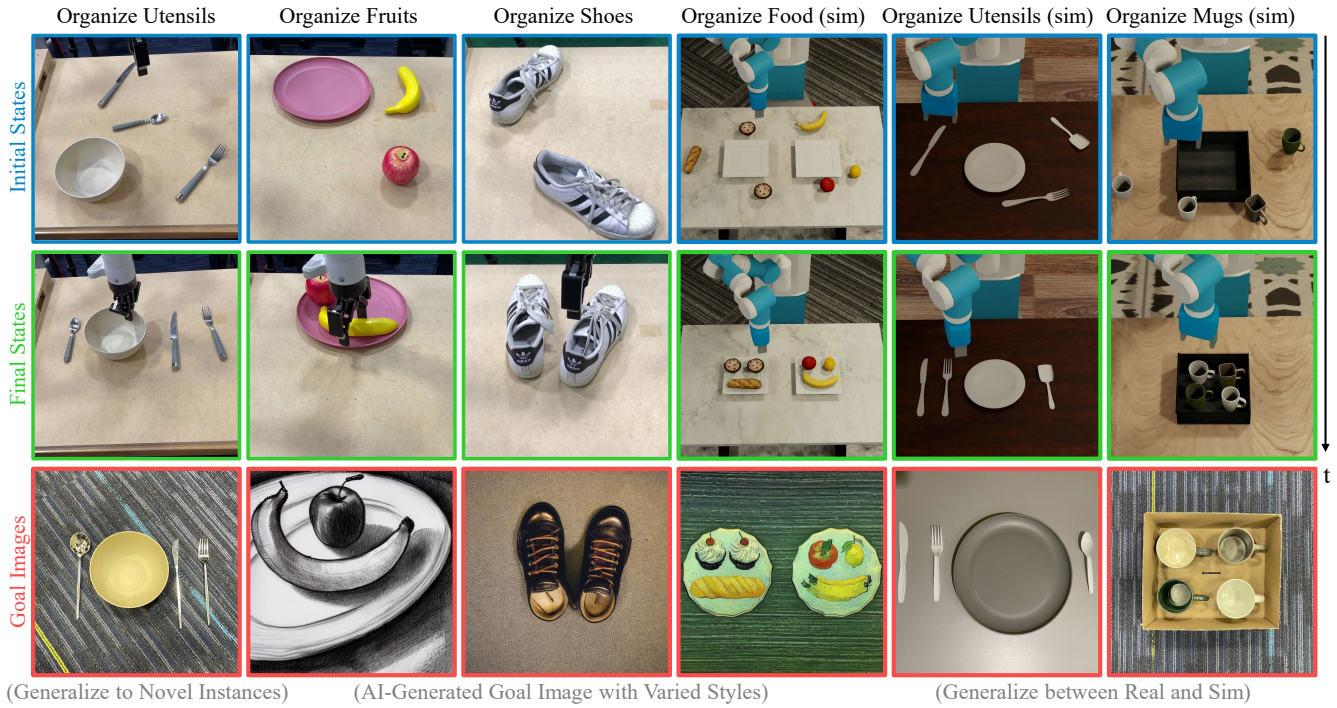


Fig. 3: **Qualitative Results.** We qualitatively evaluate our proposed framework on household manipulation tasks, both in the real world and in simulation, encompassing tasks such as organizing utensils, fruits, shoes, food, and mugs. The figure highlights that our representation can generalize across varied instances, styles, and contexts. For instance, in the organizing fruits example, the goal image, unlike the workspace, is styled as a sketch drawing. Because our representation can map bananas with varied styles and appearances to similar features, the banana in the workspace can correspond to the banana in the sketch. This allows the task to be successfully completed. This wide range of tasks showcases the generalization capabilities and manipulation precision of our framework.

and other state-of-the-art dense descriptor techniques. Our results indicate that our approach significantly outperforms in terms of generalizability and manipulation accuracy.

To summarize our contributions: (1) We introduce a superior representation, D^{3}Fields, that is **3D**, **dynamic**, and **semantic**. (2) We present a novel and flexible goal specification method using 2D images that incorporate a range of styles, contexts, and instances. (3) Our proposed robotic manipulation framework supports zero-shot generalizable manipulation applicable to a broad spectrum of household tasks, demonstrating its practical utility and adaptability.

II. Method

In this section, we introduce the problem formulation in Section V-B.1 and define camera transformation and projection notations in Section V-B.2. The construction of D^{3}Fields is detailed in Section V-B.3. Section V-B.4 discusses tracking keypoints and learning of dynamics associated with our representation, while Section V-C.3 showcases how our representation enables zero-shot generalizable manipulation skills, thus facilitating practical applications in the real world.

III. Experiments

In this section, we evaluate our representation across various manipulation tasks. across a diverse range of manipulation tasks. These tasks vary in terms of goal image styles, instances, and contexts, demonstrating the versatility of our approach. We visualize D^{3}Fields and present tracking results in Section V-C.2, offering insights into the effectiveness of our representation. Then, we highlight our framework's zero-shot generalizability in both real-world scenarios and simulated environments in Section V-C.3. Lastly, we provide a quantitative comparison with existing baselines in Section V-C.5. This comparison emphasizes our framework's superior capabilities in generalization and manipulation precision, marking a significant advancement in the field.

IV. Conclusion

In this work, we introduce D^{3}Fields, a novel approach that adeptly encodes 3D semantic features, 3D instance masks, and models the complex dynamics underlying various scenarios. A key focus of our work is on enabling zero-shot generalizable robotic manipulation tasks. These tasks are uniquely specified using 2D goal images that encompass a wide range of styles, contexts, and instances. Our framework demonstrates exceptional proficiency in performing an array of household manipulation tasks, effectively adapting to both simulated environments and real-world settings. Notably, it outperforms established baseline methods like Dense Object Nets and DINO, showcasing superior generalization capabilities and enhanced manipulation accuracy. This achievement marks a significant advancement in the field of robotic manipulation, highlighting the potential of D^{3}Fields in diverse and dynamic robotic manipulation tasks.

References

[1] L. Manuelli, Y. Li, P. Florence, and R. Tedrake, "Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning," in *Conference on Robot Learning (CoRL)*, 2020.

[2] Y. Li, S. Li, V. Sitzmann, P. Agrawal, and A. Torralba, "3d neural scene representations for visuomotor control," *arXiv preprint arXiv:2107.04004*, 2021.

[3] H. Shi, H. Xu, Z. Huang, Y. Li, and J. Wu, "Robocraft: Learning to see, simulate, and shape elasto-plastic objects with graph networks," *arXiv preprint arXiv:2205.02909*, 2022.

[4] Y. Ze, N. Hansen, Y. Chen, M. Jain, and X. Wang, "Visual reinforcement learning with self-supervised 3d representations," *IEEE Robotics and Automation Letters (RA-L)*, 2023.

[5] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International conference on machine learning*. PMLR, 2019, pp. 2555–2565.

[6] W. Yan, A. Vangipuram, P. Abbeel, and L. Pinto, "Learning predictive representations for deformable objects using contrastive estimation," in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 564–574.

[7] Y. Wang, Y. Li, K. Driggs-Campbell, L. Fei-Fei, and J. Wu, "Dynamic-Resolution Model Learning for Object Pile Manipulation," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[8] M. Minderer, C. Sun, R. Villegas, F. Cole, K. P. Murphy, and H. Lee, "Unsupervised learning of object structure and dynamics from videos," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[9] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Conference on Robot Learning*. PMLR, 2018, pp. 306–316.

[10] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield, "6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 13 081–13 088.

[11] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, G. Iyer, S. Saryazdi, T. Chen, A. Maalouf, S. Li, N. V. Keetha, A. Tewari, J. Tenenbaum, C. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba, "ConceptFusion: Open-set multimodal 3D mapping," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[12] K. Mazur, E. Sucar, and A. J. Davison, "Feature-realistic neural fusion for real-time, open set scene understanding," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8201–8207.

[13] W. Liu, Y. Du, T. Hermans, S. Chernova, and C. Paxton, "Structdiffusion: Language-guided creation of physically-valid structures using unseen objects," in *RSS 2023*, 2023.

[14] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu *et al.*, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[16] H. K. Cheng and A. G. Schwing, "XMem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *ECCV*, 2022.

[17] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.

[18] A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning*. PMLR, 2023, pp. 287–318.

[19] W. Huang, F. Xia, T. Xiao, H. Chan, J. Liang, P. Florence, A. Zeng, J. Tompson, I. Mordatch, Y. Chebotar *et al.*, "Inner monologue: Embodied reasoning through planning with language models," in *Conference on Robot Learning*. PMLR, 2023, pp. 1769–1782.

[20] J. Liang, W. Huang, F. Xia, P. Xu, K. Hausman, B. Ichter, P. Florence, and A. Zeng, "Code as policies: Language model programs for embodied control," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9493–9500.

[21] W. Huang, C. Wang, R. Zhang, Y. Li, J. Wu, and L. Fei-Fei, "Voxposer: Composable 3d value maps for robotic manipulation with language models," in *7th Annual Conference on Robot Learning*, 2023.

[22] Y. Zhu, Z. Jiang, P. Stone, and Y. Zhu, "Learning generalizable manipulation policies with object-centric 3d representations," in *7th Annual Conference on Robot Learning*, 2023.

[23] K. Mülling, J. Kober, O. Kroemer, and J. Peters, "Learning to select and generalize striking movements in robot table tennis," *The International Journal of Robotics Research*, vol. 32, no. 3, pp. 263–279, 2013.

[24] Y. Duan, M. Andrychowicz, B. Stadie, O. Jonathan Ho, J. Schneider, I. Sutskever, P. Abbeel, and W. Zaremba, "One-shot imitation learning," *Advances in neural in-*

formation processing systems, vol. 30, 2017.

[25] Y. Li, J. Wu, R. Tedrake, J. B. Tenenbaum, and A. Torralba, "Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids," in *ICLR*, 2019.

[26] W. Wang, A. S. Morgan, A. M. Dollar, and G. D. Hager, "Dynamical scene representation and control with keypoint-conditioned neural radiance field," in *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2022, pp. 1138–1143.

[27] W. Gao and R. Tedrake, "kpam 2.0: Feedback control for category-level robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 2962–2969, 2021.

[28] L. Manuelli, W. Gao, P. Florence, and R. Tedrake, "kpam: Keypoint affordances for category-level robotic manipulation," in *The International Symposium of Robotics Research*. Springer, 2019, pp. 132–157.

[29] W. Gao and R. Tedrake, "kpam-sc: Generalizable manipulation planning using keypoint affordance and shape completion," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6527–6533.

[30] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.

[31] X. Lin, C. Qi, Y. Zhang, Y. Li, Z. Huang, K. Fragkiadaki, C. Gan, and D. Held, "Planning with spatial-temporal abstraction from point clouds for deformable object manipulation," in *6th Annual Conference on Robot Learning*, 2022.

[32] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 892–909.

[33] T. Xiao, I. Radosavovic, T. Darrell, and J. Malik, "Masked visual pre-training for motor control," *arXiv:2203.06173*, 2022.

[34] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," *CoRL*, 2022.

[35] Z. Mandi, H. Bharadhwaj, V. Moens, S. Song, A. Rajeswaran, and V. Kumar, "Cacti: A framework for scalable multi-task multi-scene visual imitation learning," *arXiv preprint arXiv:2212.05711*, 2022.

[36] A. Stone, T. Xiao, Y. Lu, K. Gopalakrishnan, K.-H. Lee, Q. Vuong, P. Wohlhart, B. Zitkovich, F. Xia, C. Finn, and K. Hausman, "Open-world object manipulation using pre-trained vision-language models," 2023.

[37] Y. Yoon, G. N. DeSouza, and A. C. Kak, "Real-time tracking and pose estimation for industrial objects using geometric features," in *2003 IEEE International conference on robotics and automation (cat. no. 03CH37422)*, vol. 3. IEEE, 2003, pp. 3473–3478.

[38] M. Zhu, K. G. Derpanis, Y. Yang, S. Brahmbhatt, M. Zhang, C. Phillips, M. Lecce, and K. Daniilidis, "Single image 3d object detection and pose estimation for grasping," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 3936–3943.

[39] L. Zhu, A. Mousavian, Y. Xiang, H. Mazhar, J. van Eenbergen, S. Debnath, and D. Fox, "Rgb-d local implicit function for depth completion of transparent objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 4649–4658.

[40] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dexnerf: Using a neural radiance field to grasp transparent objects," in *5th Annual Conference on Robot Learning*, 2021.

[41] Y. Wi, P. Florence, A. Zeng, and N. Fazeli, "Virdo: Visio-tactile implicit representations of deformable objects," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 3583–3590.

[42] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, "Neural descriptor fields: Se (3)-equivariant object representations for manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6394–6400.

[43] D. Driess, J.-S. Ha, M. Toussaint, and R. Tedrake, "Learning models as functionals of signed-distance fields for manipulation planning," in *Conference on Robot Learning*. PMLR, 2022, pp. 245–255.

[44] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations," in *Proceedings of Robotics: Science and Systems*, Virtual, July 2021.

[45] T. Weng, D. Held, F. Meier, and M. Mukadam, "Neural grasp distance fields for robot manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[46] D. Driess, I. Schubert, P. Florence, Y. Li, and M. Toussaint, "Reinforcement learning with neural radiance fields," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[47] D. Shim, S. Lee, and H. J. Kim, "SNeRL: Semantic-aware neural radiance fields for reinforcement learning," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 31 489–31 503.

[48] L. Yen-Chen, P. Florence, J. T. Barron, T.-Y. Lin, A. Rodriguez, and P. Isola, "NeRF-Supervision: Learning dense object descriptors from neural radiance fields," in *IEEE Conference on Robotics and Automation (ICRA)*, 2022.

[49] Z. Tang, B. Sundaralingam, J. Tremblay, B. Wen, Y. Yuan, S. Tyree, C. Loop, A. Schwing, and S. Birchfield, "RGB-only reconstruction of tabletop scenes for collision-free manipulator control," in *ICRA*, 2023.

[50] A. Zhou, M. J. Kim, L. Wang, P. Florence, and C. Finn, "Nerf in the palm of your hand: Corrective augmentation for robotics via novel-view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 907–17 917.

[51] N. M. (Mahi)Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam, "CLIP-Fields: Weakly Supervised Semantic Fields for Robotic Memory," in *Proceedings of Robotics: Science and Systems*, Daegu, Republic of Korea, July 2023.

[52] Y. Wi, A. Zeng, P. Florence, and N. Fazeli, "Virdo++: Real-world, visuo-tactile dynamics and perception of deformable objects," *arXiv preprint arXiv:2210.03701*, 2022.

[53] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 8748–8763.

[54] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.

[55] J. Kerr, C. M. Kim, K. Goldberg, A. Kanazawa, and M. Tancik, "Lerf: Language embedded radiance fields," in *International Conference on Computer Vision (ICCV)*, 2023.

[56] S. Sharma, A. Rashid, C. M. Kim, J. Kerr, L. Y. Chen, A. Kanazawa, and K. Goldberg, "Language embedded radiance fields for zero-shot task-oriented grasping," in *7th Annual Conference on Robot Learning*, 2023.

[57] W. Shen, G. Yang, A. Yu, J. Wong, L. P. Kaelbling, and P. Isola, "Distilled feature fields enable few-shot manipulation," in *7th Annual Conference on Robot Learning*, 2023.

[58] Y. Ze, G. Yan, Y.-H. Wu, A. Macaluso, Y. Ge, J. Ye, N. Hansen, L. E. Li, and X. Wang, "Multi-task real robot learning with generalizable neural feature fields," in *7th Annual Conference on Robot Learning*, 2023.

[59] Y. Li, T. Lin, K. Yi, D. Bear, D. L. Yamins, J. Wu, J. B. Tenenbaum, and A. Torralba, "Visual grounding of learned physical models," in *International Conference on Machine Learning*, 2020.

[60] C. Li, R. Zhang, J. Wong, C. Gokmen, S. Srivastava, R. Martín-Martín, C. Wang, G. Levine, M. Lingelbach, J. Sun, M. Anvari, M. Hwang, M. Sharma, A. Aydin, D. Bansal, S. Hunter, K.-Y. Kim, A. Lou, C. R. Matthews, I. Villa-Renteria, J. H. Tang, C. Tang, F. Xia, S. Savarese, H. Gweon, K. Liu, J. Wu, and L. Fei-Fei, "BEHAVIOR-1k: A benchmark for embodied AI with 1,000 everyday activities and realistic simulation," in *6th Annual Conference on Robot Learning*, 2022.

[61] P. R. Florence, L. Manuelli, and R. Tedrake, "Dense object nets: Learning dense visual object descriptors by and for robotic manipulation," in *Proceedings of The 2nd Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., vol. 87. PMLR, 29–31 Oct 2018, pp. 373–385.

V. Supplementary

A. Related Works

1) Foundation Models for Robotics: Foundation models generally refer to those trained on extensive, diverse datasets, often employing self-supervision at scale, which can then be adapted (e.g., fine-tuned) to a wide range of specific downstream tasks. Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities for language. Robotics researchers have recently released a series of works that leverage LLMs, including SayCan [18] and Inner Monologue [19], to directly generate robot plans. Some later works have explored the use of LLMs as code generators for robotic actions: Code as Policies [20] uses 2D object detectors for perception, whereas VoxPoser [21] generates a 3D value map. Yet, their perception modules lack accurate modeling of the precise geometry and dynamics of objects. Our D^3Fields aim to address this by capturing detailed 3D geometry and dynamics, thus enhancing the perception capabilities in robotic manipulation.

Meanwhile, visual foundation models, such as SAM [15] and DINOv2 [17], have achieved remarkable success in zero-shot generalization capabilities across various vision tasks. However, their focus is primarily on 2D vision tasks. Adapting these models for dynamic 3D environments poses a significant challenge. The recent advancement in GROOT [22] showcases how to construct 3D object-centric representations using foundation models and exhibits notable few-shot generalization capabilities. Despite these advancements, GROOT does not fully address the complexities of learning about object dynamics, nor does it focus on attaining zero-shot generalization in robotic manipulation. The limitation highlights the urgent for further innovation in integrating visual foundation models with dynamic 3D object handling.

2) Representation for Visual Robotic Manipulation: Scene representation has been a pivotal component in robotic manipulation systems. Some early work relies on 2D representations, such as bounding boxes [23, 24]. Many recent methods have seen a shift towards constructing particle representations of the environment, utilizing learned dynamics to capture the system's underlying structure [3, 7, 8, 25–29]. They demonstrate impressive results in unstructured environments and with non-rigid objects. However, they are not semantic, which hinders their ability to generalize to new tasks and scenarios. Some research opts for a fixed-dimension latent vector derived from high-dimensional sensory data as the representation [2, 5, 6, 30–36], but such a representation struggles to scale effectively for complex manipulation tasks that require high precision and explicit scene structures. Other approaches adopt 6 DoF object poses

as their representation [9, 10, 37, 38], though focusing primarily on grasping tasks instead of more dynamic ones. In this work, we aim to overcome these limitations by introducing D^{3}Fields, a representation that models dynamic 3D environments across varying semantic levels.

3) Neural Fields for Robotic Manipulation: Researchers have explored a variety of approaches leveraging neural fields as a representation for robotic manipulation [39–41, 41–52]. Among them, Neural Descriptor Fields [42] are the most relevant to ours. These fields build neural feature fields that demonstrate generalizability across different instances with several demonstrations; but they focus on learning geometric, not semantic features, which restricts their ability to generalize across categories.

Recently, a series of works distilled neural feature fields using foundation models such as CLIP and DINO for supervision [53, 54]. LeRF distills neural feature fields to handle open-vocabulary 3D queries and develops task-oriented grasping based on it [55, 56]. Shen et al. [57] adopted a similar distilled feature field for the grasping task. Both methods require dense camera views to train the neural field. GNFactor attempts to address this by introducing a voxel encoder [58]. However, distilling foundation models to create neural feature fields faces significant drawbacks: (1) They often require dense camera views for a quality field, which is expensive and impractical for real-world scenarios. (2) Distilled neural fields necessitate retraining for new scenes, limiting their generalization and making them ineffective for dynamic scenes. In contrast, our D^{3}Fields requires no training for new scenes and is capable of working with sparse views and dynamic settings. The distinct capability of D^{3}Fields marks a significant advancement in the field of robotic manipulation.

B. Method

In this section, we introduce the problem formulation in Section V-B.1 and define camera transformation and projection notations in Section V-B.2. The construction of D^{3}Fields is detailed in Section V-B.3. Section V-B.4 discusses tracking keypoints and learning of dynamics associated with our representation, while Section V-C.3 showcases how our representation enables zero-shot generalizable manipulation skills, thus facilitating practical applications in the real world.

1) Problem Formulation: We formulate our problem as zero-shot robotic manipulation problem given a 2D goal image $\mathcal{I}$. We denote the workspace scene representation as $\boldsymbol{s}_{\text{goal}}$. The primary objective in this context is to determine an optimal sequence of actions $\{a^t\}$ to effectively minimize the task objective:

$$\begin{aligned} \min_{\{a_t\}} \quad & c(\boldsymbol{s}^T, \boldsymbol{s}_{\text{goal}}), \\ \text{s.t.} \quad & \boldsymbol{s}^t = g(\boldsymbol{o}^t), \quad \boldsymbol{s}^{t+1} = f(\boldsymbol{s}^t, a^t), \end{aligned} \tag{1}$$

where $c(\cdot,\cdot)$ is the cost function measuring the distance between the terminal representation $\boldsymbol{s}^T$ and the goal representation $\boldsymbol{s}_{\text{goal}}$. Representation extraction function $g(\cdot)$ takes in the current multi-view RGBD observations $\boldsymbol{o}^t$ and outputs the current representation $\boldsymbol{s}^t$. $f(\cdot,\cdot)$ is the dynamics function that predicts the future representation $\boldsymbol{s}^{t+1}$, conditioned on the current representation $\boldsymbol{s}^t$ and action a^t. The optimization aims to find the action sequence $\{a_t\}$ that minimizes the cost function $c(\boldsymbol{s}^T, \boldsymbol{s}_{\text{goal}})$.

Algorithm 1 Fusion Process

procedure FUSION($\mathbf{x}$) ▷ Input 3D point
  $\mathbf{u}_i, \mathbf{r}_i \leftarrow \text{Project}(\mathbf{x}, i)$ ▷ 3D Projection
  $\mathbf{r}'_i \leftarrow \mathcal{R}_i[\mathbf{u}_i]$
  $\mathbf{d}_i \leftarrow \mathbf{r}'_i - \mathbf{r}_i$ ▷ Depth difference
  $\mathbf{d}'_i \leftarrow \text{Truncate}(\mathbf{d}_i, \mu)$ ▷ Apply truncation
  $\mathbf{f}_i, \mathbf{p}_i \leftarrow \text{Interpolate}(\mathcal{W}_i^{\mathbf{f}}, \mathcal{W}_i^{\mathbf{p}}, \mathbf{u}_i)$
  $v_i, w_i \leftarrow \text{Weights}(\mathbf{x}, i)$ ▷ Weights for fusion
  $\mathbf{f}, \mathbf{p} \leftarrow \text{Fuse}(\mathbf{f}_i, \mathbf{p}_i, v_i, w_i)$ ▷ Fuse features

2) Notation: Camera Transformation and Projection: We assume all cameras' intrinsic parameters $\mathbf{K}$ and extrinsic parameters $\mathbf{T}$ are known. The camera i extrinsic parameters are defined as follows:

$$\mathbf{T}_i = \begin{bmatrix} \mathbf{R}_i & \mathbf{t}_i \\ 0^T & 1 \end{bmatrix} \in \mathbb{SE}(3), \tag{2}$$

where Euclidean group $\mathbb{SE}(3) := \{\mathbf{R}, \mathbf{t} \mid \mathbf{R} \in \mathbb{SO}^3, \mathbf{t} \in \mathbb{R}^3\}$. For a 3D point $\mathbf{x}$ in the world frame, we could obtain projected pixel $\mathbf{u}_i$ and distance to camera $\mathbf{r}_i$ as follows:

$$\mathbf{u}_i = \pi\left(\mathbf{K}_i\left(\mathbf{R}_i\mathbf{x} + \mathbf{t}_i\right)\right), \quad \mathbf{r}_i = [0,0,1]^T\left(\mathbf{R}_i\mathbf{x} + \mathbf{t}_i\right), \tag{3}$$

where π performs perspective projection, mapping a 3D vector $p = [x, y, z]^T$ to a 2D vector $q = [x/z, y/z]^T$.

3) D^{3}Fields Representation: To build the implicit 3D descriptor fields $\mathcal{F}^t(\cdot)$, We design the algorithm 1 to fuse observations $\mathbf{o}^t$ from multiple viewpoints. For simplicity, we represent $\mathbf{o}^t$ as $\mathbf{o}$, and $\mathcal{F}^t(\cdot)$ as $\mathcal{F}(\cdot)$ respectively in this subsection. The implicit 3D descriptor field $\mathcal{F}(\cdot)$ is defined as

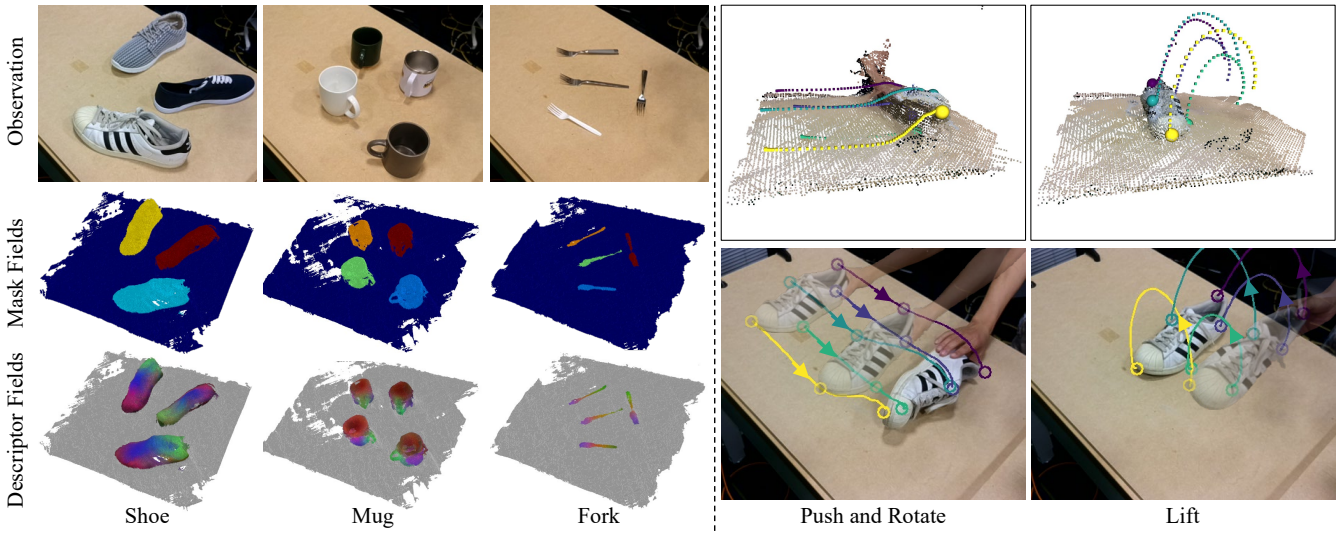
$$(\mathbf{d}, \mathbf{f}, \mathbf{p}) = \mathcal{F}(\mathbf{x}), \tag{4}$$

where $\mathbf{x}$ is an arbitrary 3D point in the world frame, and $(\mathbf{d}, \mathbf{f}, \mathbf{p})$ is the corresponding geometric and semantic descriptor. $\mathbf{d} \in \mathbb{R}$ is the signed distance from $\mathbf{x}$ to the surface. $\mathbf{f} \in \mathbb{R}^N$ represents the semantic information of N dimension. $\mathbf{p} \in \mathbb{R}^M$ denotes the instance probability distribution of M instances. M could be different across scenarios.

More precisely, we define a single-view RGBD observation from camera i as $\boldsymbol{o}_i = (\mathcal{I}_i, \mathcal{R}_i)$, where the RGB image $\mathcal{I}_i \in \mathbb{R}^{H\times W\times 3}$, and depth image $\mathcal{R}_i \in \mathbb{R}^{H\times W}$. To map an arbitrary 3D point $\mathbf{x}$ the to image space, we employ the projection Equation 3. Through bilinear interpolation, we then ascertain the corresponding depth value $\mathbf{r}'_i = \mathcal{R}_i[\mathbf{u}_i]$. Then the descriptors from camera i are obtained by

$$\begin{aligned} &\mathbf{d}_i = \mathbf{r}'_i - \mathbf{r}_i, \quad \mathbf{d}'_i = \max(\min(\mathbf{d}_i, \mu), -\mu), \\ &\mathbf{f}_i = \mathcal{W}_i^{\mathbf{f}}[\mathbf{u}_i], \quad \mathbf{p}_i = \mathcal{W}_i^{\mathbf{p}}[\mathbf{u}_i], \end{aligned} \tag{5}$$

where DINOv2 [17] extracts the semantic feature volume $\mathcal{W}_i^{\mathbf{f}} \in \mathbb{R}^{H\times W\times N}$ from RGB observation $\mathcal{I}_i$. $\mathcal{W}_i^{\mathbf{p}} \in$



(a) Representation Visualization **(b) Tracking Visualization**

Fig. 4: **Representation and Tracking Visualizations.** (a) To demonstrate that the representation is both 3D and semantic, we visualize the representation across different object categories. Mask fields distinctly color 3D points based on their instance masks, effectively distinguishing among different instances. Descriptor fields assign colors to 3D points by translating features into the RGB spectrum through PCA. This results in consistent color patterns within categories, exemplified by mug handles consistently appearing green across various mug instances. (b) To demonstrate that our representation is dynamic, we apply it to tracking tasks and showcase two tracking examples, both of which involve 3D motions and partial observations from single viewpoints. The robust 3D tracking results serve as evidence that our representation is 3D, dynamic, and semantic.

$\mathbb{R}^{H\times W\times M}$ is the instance mask volume using Grounded-SAM [14, 15]. The parameter μ specifies the truncation threshold for the Truncated Signed Distance Function (TSDF).

We fuse descriptors from all K views as follows:

$$v_i = H(\mathbf{d}_i + \mu), \quad w_i = \exp\left(\frac{\min\left(\mu - |\mathbf{d}_i|, 0\right)}{\mu}\right), \tag{6}$$

and then

$$\mathbf{d} = \frac{\sum_{i=1}^{K} v_i \mathbf{d}'_i}{\delta + \sum_{i=1}^{K} v_i}, \mathbf{f} = \frac{\sum_{i=1}^{K} v_i w_i \mathbf{f}_i}{\delta + \sum_{i=1}^{K} v_i}, \mathbf{m} = \frac{\sum_{i=1}^{K} v_i w_i \mathbf{m}_i}{\delta + \sum_{i=1}^{K} v_i}, \tag{7}$$

where H is the unit step function and δ is a small value to avoid numeric errors. $v_i = 0$ when $\mathbf{x}$ is not observable in camera i, because if $\mathbf{x}$ is occluded in camera i, it should not contribute to the descriptor of $\mathbf{x}$. In addition, we could only have a confident estimation when $\mathbf{x}$ is close to the surface. Therefore, w_i will decay as $|\mathbf{d}_i|$ increases. For $\mathbf{x}$ that is far away, $\mathbf{f}$ and $\mathbf{m}$ will degrade to 0^T.

4) Tracking Scene Changes and Dynamic Model Learning: In order to track the changes and model the dynamics of the descriptor fields, we convert the implicit field function $\mathcal{F}(\cdot)$ (an Eulerian representation) to a set of keypoints $\boldsymbol{s}$ (a Lagrangian representation). To do this, we create voxels $\mathbf{x} \in \mathbb{R}^{W\times L\times H\times 3}$ in the workspace and evaluate $(\mathbf{d}, \mathbf{f}, \mathbf{p}) = \mathcal{F}(\mathbf{x})$. We filter out $\mathbf{x}_i \in \mathbf{x}$ where $\mathbf{d}_i$ is large or $\mathbf{p}_i$ has a low probability to avoid empty space and the background. After obtaining filtered points $\mathbf{x}'$, we use farthest point sampling to find surface points $\boldsymbol{s} \in \mathbb{R}^{3\times n_s}$ of an instance.

We will then use the dynamic implicit 3D descriptor field $\mathcal{F}(\cdot)$ to track these keypoints and train the corresponding dynamics models. Without losing generalization, consider the tracking of a single instance $\boldsymbol{s}^t \in \mathbb{R}^{3\times n_s}$. For clarity, we denote $\mathbf{f}$ and $\mathbf{d}$ from $\mathcal{F}(\cdot)$ as $\mathcal{F}_{\mathbf{f}}(\cdot)$ and $\mathcal{F}_{\mathbf{d}}(\cdot)$. We formulate the tracking problem as an optimization problem:

$$\min_{\boldsymbol{s}^{t+1}} \quad ||\mathcal{F}_{\mathbf{f}}(\boldsymbol{s}^{t+1}) - \mathcal{F}_{\mathbf{f}}(\boldsymbol{s}^0)||_2. \tag{8}$$

As $\mathcal{F}(\cdot)$ is differentiable, we adopt a gradient-based optimizer. This method could be naturally extended to multiple-instance scenarios. We found that relying exclusively on features for tracking led to instability. To address this, we incorporate rigid constraints and distance regularization, which significantly enhance the stability and reliability of the tracking process.

We instantiate to represent the dynamics model $f(\cdot,\cdot)$ as graph neural networks (GNNs). We follow [59] to predict object dynamics. Please refer to [25, 59] for more details on how to train the GNN-based dynamics model. The optimized dynamics model plays a pivotal role in trajectory optimization, a process we discuss thoroughly in Section V-B.5.

5) Zero-Shot Generalizable Robotic Manipulation: As described in Section V-B.3, we denote initial tracked points and features as $\boldsymbol{s}^0$ and $\mathbf{f}^0$. We estimate $\boldsymbol{s}_{\text{goal}} \in \mathbb{R}^{2\times n_s}$ of goal image $\mathcal{I}_{\text{goal}}$ as follows:

$$\begin{aligned} \alpha_{ij} &= \exp\left(||\mathcal{W}^{\mathbf{f}}_{\text{goal}}[\mathbf{u}_i] - \mathbf{f}^0_j||_2\right), \\ w_{ij} &= \frac{\exp\left(s\alpha_{ij}\right)}{\sum_{i=1}^{H\times W} \exp\left(s\alpha_{ij}\right)}, \end{aligned} \tag{9}$$

then we have $\boldsymbol{s}_{\text{goal},j} = \sum_{i=1}^{H\times W} w_{ij}\mathbf{u}_i$, where $\mathcal{W}^{\mathbf{f}}_{\text{goal}}$ is the feature volume extracted from $\mathcal{I}_{\text{goal}}$ using DINOv2. s is the hyperparameter to determine whether the heatmap w_{ij} is more smooth or concentrating. Although Equation. 9 only

shows a single instance case, it could be naturally extended to multiple instances by using instance mask information.

Note that there is a fundamental difference in the spaces of the goal scene representation, $\boldsymbol{s}_{\text{goal}}$ which is in the image space, and the current state representation, $\boldsymbol{s}^t$ which exists in 3D space. To reconcile this discrepancy, we introduce a reference camera setup equipped with estimated intrinsic and extrinsic parameters, denoted as $\mathbf{K}'$ and $\mathbf{T}'$ respectively. Rather than generating images from the reference viewpoint, our approach emphasizes the projection of 3D keypoints onto 2D images. Consequently, we formulate the task cost function directly within the image space as follows, leveraging these projections to accurately define and evaluate the task objectives.

$$c(\boldsymbol{s}^t, \boldsymbol{s}_{\text{goal}}) = ||\pi\left(\mathbf{K}'\left(\mathbf{R}'\boldsymbol{s}^t + \mathbf{t}'\right)\right) - \boldsymbol{s}_{\text{goal}}||_2^2. \tag{10}$$

C. Experiments

In this section, we evaluate our representation across various manipulation tasks. across a diverse range of manipulation tasks. These tasks vary in terms of goal image styles, instances, and contexts, demonstrating the versatility of our approach. We visualize D^{3}Fields and present tracking results in Section V-C.2, offering insights into the effectiveness of our representation. Then, we highlight our framework's zero-shot generalizability in both real-world scenarios and simulated environments in Section V-C.3. Lastly, we provide a quantitative comparison with existing baselines in Section V-C.5. This comparison emphasizes our framework's superior capabilities in generalization and manipulation precision, marking a significant advancement in the field.

1) Experiment Setup: In our real-world experiments, we utilize four OAK-D Pro cameras positioned at the corners of the workspace, to capture RGBD observations and employ the Kinova® Gen3 robotic arm for action execution. We set these 4 cameras at the corners of the workspace, to fully capture the workspace.

In simulation, we leverage OmniGibson and deploy the Fetch robot for mobile manipulation tasks [60]. Our evaluations span a variety of tasks, including organizing shoes, collecting debris, tidying the office table, arranging utensils, and a wide array of other tasks.

In our baseline comparisons, we implement methods utilizing Dense Object Nets (DON) and DINO for feature extraction [54, 61]. We conduct a quantitative evaluation of these methods across five distinct object categories, focusing specifically on single-instance manipulation tasks in real-world settings. This approach allows us to thoroughly assess the performance and effectiveness of these methods in practical, tangible scenarios. The results and analysis are presented in Section V-C.5.

2) Descriptor Fields Visualization and Keypoints Tracking: D^{3}Fields demonstrates a robust capability in providing 3D semantic representations, as shown in Fig. 4(a). We first visualize the mask fields by we first color-code 3D points based on their most likely instance, as determined by the mask fields, and our visualization shows a distinct 3D instance segmentation. Additionally, we map the semantic features to RGB space using PCA, as with DINOv2 [17]. Visualization of the descriptor fields reveals that D^{3}Fields retain a dense semantic understanding of objects. In the provided shoe example, despite the diverse appearances and poses of the shoes, there is a consistent color pattern across the different instances: shoe heels are predominantly represented in green, while shoe toes appear in red. This pattern of semantic feature representation is consistent and observable in other objects as well, such as mugs and forks, underlining the robustness of our model in capturing and differentiating semantics.

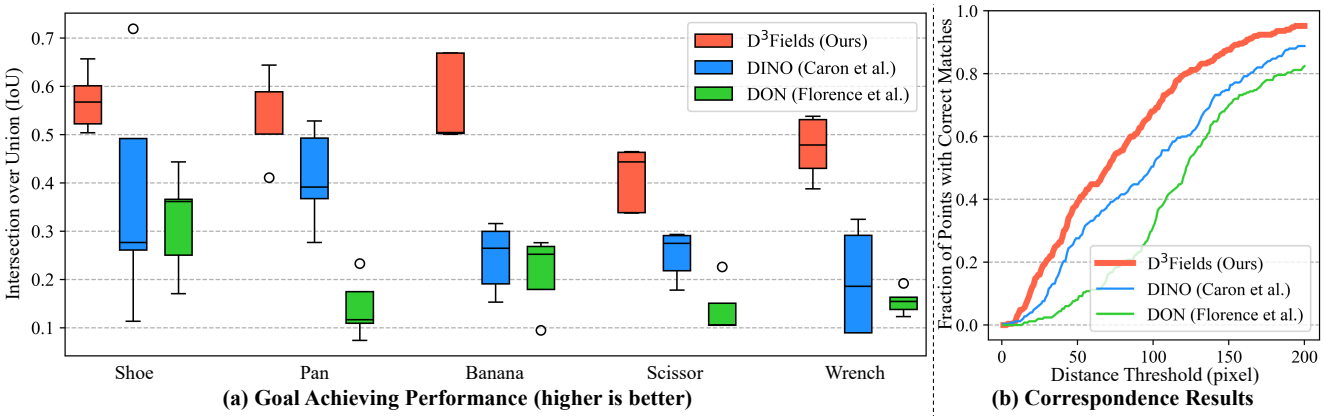
D^{3}Fields is not only adept at semantic representation but also excels in capturing complex dynamics. This capability is highlighted through our evaluation of the system's ability to track object keypoints. We show two examples of 3D keypoint tracking in Fig. 4(b). In the first scenario, we track a shoe as it is pushed and subsequently flipped. Despite only a part of the shoe being visible from the camera's perspective, our framework tracks its movement reliably. The second example demonstrates tracking a shoe that is lifted and then placed down. Our system robustly tracks the shoe in the 3D space, even when parts of it move out of the camera's view. These examples underscore the effectiveness of D^{3}Fields in maintaining accurate tracking in dynamic scenarios, a pivotal aspect for real-world applications.

3) Zero-Shot Generalizable Manipulation: We conduct a qualitative evaluation of D^{3}Fields in common household robotic manipulation tasks in a zero-shot manner, with partial results displayed in Fig. 1 and Fig. 3. We observed several key capabilities of our framework, which are as follows:

Generalization to AI-Generated Goal Images. In Fig. 1, we present an intriguing scenario where the goal image, artistically rendered in a Van Gogh style, features shoes that are distinctly different from those in the actual workspace. This example highlights a significant strength of D^{3}Fields: the ability to encode semantic information. Despite the variations in appearance, D^{3}Fields is able to categorize different shoes under similar descriptors. This capability allows our framework to effectively manipulate shoes in the workspace based on AI-generated goal images, which may vary in style and appearance yet share underlying semantic attributes.

Compositional Goal Images and 3D Manipulation. In the office desk organization task in Fig. 1, we showcase the practical application of our framework in a real-world task. Initially, the robot focuses on arranging items like the mouse and pen to match their positions in the goal image. Following this, the robot demonstrates its adaptability by repositioning the mug. It moves the mug from the top of a box to its designated spot on the mug pad, guided by a separate goal image depicting the mug in an upright position. This example illustrates the precision and versatility of our system in interpreting and executing tasks based on varying goal images, effectively demonstrating its capability to handle complex manipulation tasks.

Generalization across Instances and Materials. Our framework demonstrates a remarkable ability to generalize



(a) Goal Achieving Performance (higher is better) **(b) Correspondence Results**

Fig. 5: **Quantitative Evaluation.** We perform real-world quantitative evaluations by measuring final goal-achieving performance and keypoints correspondence accuracy. (a) We use IoU to measure goal-achieving performance. Results indicate that our method aligns with the goal configurations much better than DON and DINO across various object categories and scenarios. (b) We measure the keypoints correspondence accuracy according to the fraction of points with accurate matches, with correct matches determined by a distance threshold. Our method is consistently better at aligning with the goal image, regardless of the chosen threshold.

Fig. 6: **Object Set in Our Manipulation Tasks.** A diverse collection of objects utilized in work, showcasing an array of shoes, fruits, utensils, tools, and more, encompassing over **10** distinct categories in our tasks. This highlights the extensive range of items our framework is designed to handle, demonstrating its versatility and generalization capabilities across different categories and instances.

across various instances and materials, as shown in Fig. 6. It encompasses a wide range of items. Notably, granular objects, with their inherently more complex dynamics compared to rigid ones, present a significant challenge in manipulation tasks. Our framework adeptly addresses this complexity, as evidenced in the debris collection task in Fig. 1. Fig. 3 showcases the proficiency of our framework in instance-level generalization. In this scenario, the goal image features object instances that are distinct from those present in the workspace, yet our framework successfully adapts to these variations. This highlights its robust capability to interpret and respond to diverse instances, maintaining accuracy and effectiveness across varying scenarios.

Generalization across Simulation and Real World. We evaluated our framework on household tasks in the simulator, as shown in the utensil organization and mug organization examples in Fig. 3. In these scenarios, the system was provided with goal images sourced from real-world settings. Remarkably, our framework successfully manipulated the simulated objects to match these real-world goal configurations. This not only demonstrates its adaptability in handling different objects but also underscores its exceptional generalization capabilities, efficiently bridging the gap between simulated environments and real-world scenarios. Such cross-domain proficiency is crucial for practical robotic applications and highlights the robustness of our approach.

4) Cross-Domain Semantic Correspondence: To illustrate the semantic understanding and cross-instance correspondences achieved by our proposed D^{3}Fields, we have developed an interactive visualization of feature-level correspondences, as depicted in Fig. 7. Specifically, we extract the visual feature from the query point derived using DINOv2 and calculate its distance to the constructed D^{3}Fields. Highlighted in the image are regions close to the query point when measured in feature distance. As the visualization shows, the rim and internal region of the plate map to the corresponding plate regions in different scenarios with multiple semantically similar objects. Similarly, the tip and bar of the 2D drill image map to the analogous parts on different drills located in our real robot workspace. These results demonstrate that D^{3}Fields enable meaningful correspondences across instances and contexts, effectively handling different types of object symmetries and generalizing across simulation and real world.

5) Quantitative Comparisons with Baselines: In Fig. 5(a), we measure performance using the IoU between the mask of the goal image and the mask of the final state post-manipulation. Higher IoU values indicate a greater degree of alignment between the intended and achieved configurations. Our method demonstrates superior performance across five

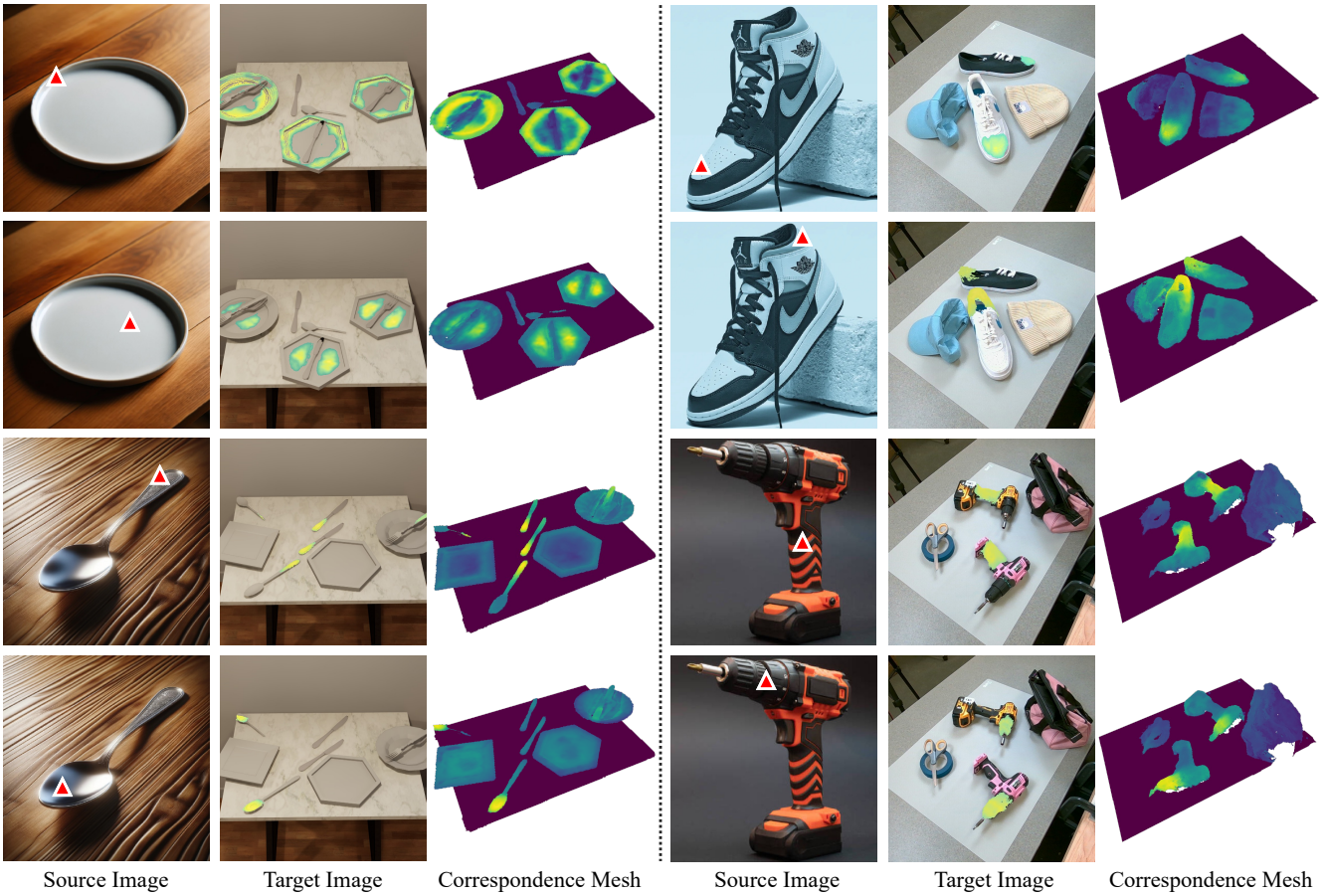


Fig. 7: **Cross-Domain Correspondence.** To effectively demonstrate the capability of our framework in establishing feature correspondence, We show the visualization in both simulation and real-world environments. We identify key points in the source images, marking them with red triangles. These corresponding features are then highlighted in both the 2D target image and the 3D mesh. By employing a diverse collection of source images showcasing everyday objects, such as plates, shoes and drills, we demonstrate the framework's versatility and precision in establishing feature correspondence across different settings. Moreover, our framework demonstrates exceptional granularity by achieving part-level precision within different targets of the same source image: When a specific part such as the spoon's tip in the source image is selected, the corresponding tip part is precisely highlighted. Similarly, targeting the handle of the spoon results in an exact highlight of the corresponding handle part.

distinct object categories, consistently outshining the baseline methods. For each category, we performed 5 experiments for the evaluation results. This not only highlights its exceptional manipulation accuracy but also its robust generalization capabilities. While the DINO model exhibits some struggles, particularly in distinguishing specific object components and consequently yielding less precise results, it still performs better than DON. Although DON shows commendable results with familiar objects and configurations, its performance dips in novel scenarios, revealing a lack of generalization. These results collectively emphasize the significant advantages of our method in diverse and accurate object manipulation.

In Fig. 5(b), we present the correspondence results. We label 10 corresponding keypoint pairs on both the goal image and the final manipulation result to sufficiently evaluate the correspondence accuracy. The accuracy of correspondence was determined by calculating the proportion of keypoints that were accurately matched, using a predefined distance threshold as the criterion. If the distance between corresponding keypoints exceeds this threshold, they are determined as unmatched. Our method shows superior performance across various thresholds, consistently outperforming the baseline models. DINO emerges as the second-best in terms of performance, exhibiting broad applicability but with a lower precision compared to our method. Meanwhile, DON lags in performance, primarily due to its struggles with generalization in novel scenarios. These results, in conjunction with those from Fig. 5(a), reiterate our method's outstanding capabilities in both generalization and accuracy. While DINO provides reasonable applicability, it lacks the precision of our approach, and the performance of DON is hindered by its limited adaptability.

D. Supplementary Results

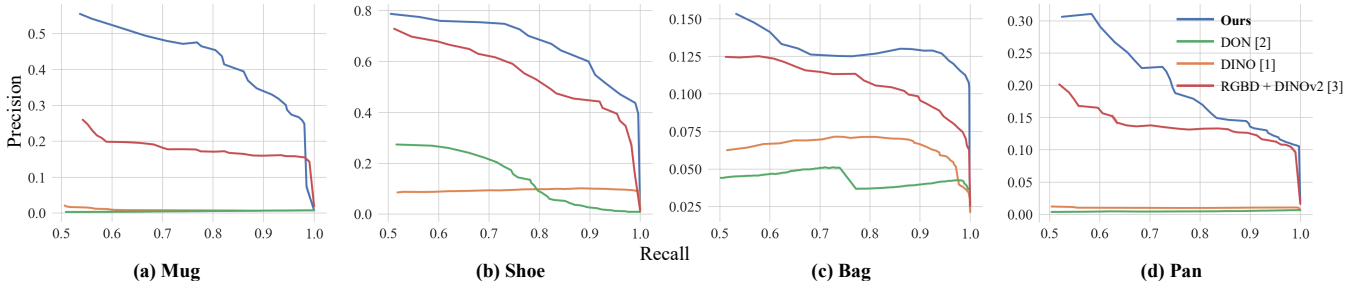


Fig. 8: **Precision-Recall of Various Thresholds for Different Instances.** The curves show how D^{3}Fields compares with 3 baseline methods in terms of matching quality, tested on 4 different instances: mug, bag, pan, and shoe. We use the precision-recall curve to measure the correspondence quality. Our method shows to consistently exceeds the performance of the baseline approaches, which demonstrates our method's capability to encode semantic information accurately and establish precise correspondences using the semantic information.

In our quantitative evaluation of correspondence quality, we compare our approach with 3 baseline methods: Dense Object Nets [61] (DON), DINO [54], and DINOv2 [17] incorporating RGBD aggregation. In our main paper, the quantitative comparison demonstrates that while DINO could encode semantic features for the whole object, it faces challenges in accurately distinguishing specific object components. This limitation results in less precise correspondence, though it still outperforms DON. Despite DON's can encode semantic features in seen environments and instances, its efficacy decreases in encounters with novel environments, highlighting its limited generalization capabilities. In this paper, we introduce an additional baseline method that directly maps DINOv2 features to 3D space using RGBD images. Compared with DINOv2 with RGBD, we offer a more thorough comparative analysis and emphasize the necessity of the fusion process in D^{3}Fields.

For each 2D source image, we manually label the points $\mathbf{x}_{2D}$, which are then paired with a set of corresponding 3D points $\mathbf{x}_{3D}$ on the point cloud. We are able to evaluate the correspondence accuracy utilizing the labeled data.

For each target vertex within the mesh, we project it onto the camera's image plane to derive its pixel coordinates, denoted as $\mathbf{u}_i$, along with its depth values, represented by $\mathbf{f}_i$. During the fusion process, we meticulously identify the most closely matching feature $\mathbf{f}_i$ for every 3D point, selecting based on the shortest distance within the feature space. This method ensures precise correspondence between 3D points and their corresponding features by leveraging spatial and depth information.

We compute the cosine similarity between features $\mathbf{f}src$ from the source image and features $\mathbf{f}tgt$ extracted from the target image. This similarity measure is pivotal for pinpointing matching points across disparate images. By establishing a predefined similarity threshold τ, ensure only those pairs with a similarity score higher than τ are preserved. This process effectively filters out less relevant matches, focusing on the most accurate correspondences based on feature similarity.

Our evaluation framework employs precision and recall, calculated based on the Intersection over Union (IoU) between matches identified by our system and the ground truth. Precision quantifies the accuracy of the system, represented as the proportion of correctly identified matches out of all matches flagged by the system. This metric focuses on the system's ability to avoid false positives. On the other hand, recall measures the system's capacity to capture all relevant matches, in line with the ground truth, highlighting its effectiveness in comprehensive match detection. This dual approach ensures a balanced evaluation of the system's performance, emphasizing both accuracy and completeness in identifying correspondences.

The precision and recall, essential for evaluating our system, are calculated based on the Intersection over Union (IoU) between the set of filtered correspondences post-thresholding and the manually annotated ground truth. Precision P and recall R are computed as follows:

$$P = \frac{|\mathbf{F}_\tau \cap \mathbf{G}|}{|\mathbf{F}_\tau|}, \tag{11}$$

$$R = \frac{|\mathbf{F}_\tau \cap \mathbf{G}|}{|\mathbf{G}|}, \tag{12}$$

where $\mathbf{F}_\tau$ denotes the set of correspondences filtered by the similarity threshold τ, and $\mathbf{G}$ represents the set of ground truth correspondences. The intersection between $\mathbf{F}_\tau$ and $\mathbf{G}$ identifies the matches that are correctly detected, while the cardinality refers to the size of each set. This method quantifies the accuracy of our system (precision) in identifying true positive matches, and its completeness (recall) in capturing all relevant correspondences defined by the ground truth.

As shown in Fig. 8, we generate a precision-recall curve by adjusting the cosine similarity thresholds. Our D^{3}Fields surpasses all three baselines across various instances. Compared with RGBD + DINOv2, our method shows to have more accurate correspondence results, highlighting the necessity of the effectiveness of D^{3}Fields. Both DINO and DON baselines exhibit low performance in all cases, with DINO showing promise but struggling to accurately differentiate specific object parts, leading to less precise results. Nonetheless, it surpasses DON in some cases, which, despite its commendable performance in familiar scenarios, suffers from reduced effectiveness in novel situations. This highlights its

constrained generalization capacity. Our quantitative analysis further corroborates these observations, affirming the superior performance and innovation of D^3Fields.