

Equivariant In-Context Learning for Grasp Adaptation

Rosa Wolf, Roman Freiberg, Loris Schneider, Rania Rayyes*, Gerhard Neumann*

Abstract—We combine in-context learning with equivariant grasp synthesis to enable zero-shot, geometrically aware adaptation from a few grasp examples. Using an existing equivariant gripper-scene encoder, both current scenes and context scenes are encoded. A context encoder then modulates the features of the current scene, steering predictions toward grasp modalities implied by the context. We show that this in-context feature modulation can guide grasps toward desired regions on the target object, supporting learning from prior rollouts and affordance-based grasping, without requiring retraining of the policy.

I. INTRODUCTION

The distribution of feasible grasps is inherently multimodal, as multiple distinct grasp configurations may exist for a single object [18], [3]. Modeling this distribution enables the generation of diverse and robust grasp candidates. Generative approaches, such as diffusion models and, more recently, flow matching, have demonstrated strong performance in learning such complex distributions, including for SE(3) grasp synthesis [21], [1], [20], [13], [23], [24]. However, most methods employ point cloud encoders, such as PointNet++ [1], [3], [24], Occupancy Networks [20], [19], or others [21], [13], [23]. While these capture geometric structure, they do not enforce consistency under object or scene rotations, which can only be learned implicitly from data, and data augmentation [15]. In contrast, equivariant neural networks enforce rotational and translational consistency by design, thus improving generalization and grasp precision [15], [9], [12], [8], [7], [24], [25]. However, in open environments, encountering objects or scenarios where even these methods fail is inevitable. For example, in a circular factory [14], where, in contrast to linear production, used and potentially deformed objects return to be remanufactured, resulting in a high variety of unfamiliar objects requiring manipulation. Retraining the policy on an updated dataset would be of limited value for objects that may only have been seen once; moreover, equivariant networks are typically memory-intensive and therefore expensive to train. To enable zero-shot adaptation abilities, in-context learning [16] has been recently adapted for robotics manipulation. In in-context learning, the prediction is steered by supplying the policy with a few examples. While some robotics methods [26], [5], [4] directly use the in-context learning abilities

Institute for Material Handling and Logistics (IFL), Karlsruhe Institute of Technology, 76131, Germany. Email: rosa.wolf@kit.edu. *equal supervision

This work is supported by the German Federal Ministry of Research, Technology, and Space (BMFTR) under the Robotics Institute Germany (RIG), the DFG SFB-1574-471687386 project, and the Ministry of Science, Research and Arts of the Federal State of Baden-Württemberg within the InnovationCampus Future Mobility.

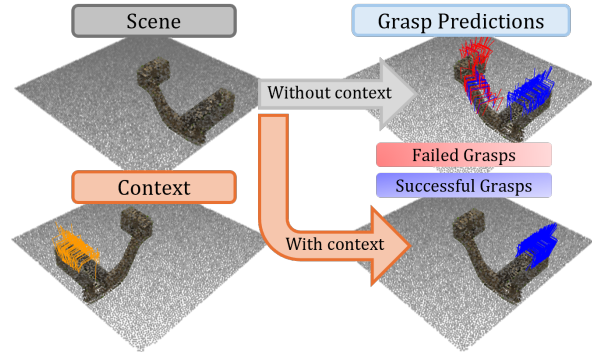


Fig. 1. Context grasps are provided to steer the policies predictions towards a specific grasp modality. Without context the policy predicts grasps covering the complete object. With a context the policy predicts grasps in the region indicated by the context. This behaviour can be used to increase the success rate when grasping an object.

of LLMs, for higher-level task planning, some [10], [11], [22] adapt the idea to the domain of trajectory-planning, by providing demonstration trajectories, to infer the task from, or to enable adaptation to OOD scenarios. In grasp-synthesis [10] and [11] use a context-aware feature modulation to capture relative relationships of cluttered objects in the same scene for collision-aware target-oriented grasp generation.

We develop an equivariant context-encoder that is integrated into an existing equivariant flow-based grasp-synthesis pipeline [8]. Context and current scene are encoded with the same SE(3)-equivariant backbone [8], and a context-attention module modulates current grasp-scene features at every flow step. Because conditioning is performed on relative grasp-scene geometry rather than absolute poses, the method can transfer contextual grasp preferences across different scenes and objects. This allows both demonstration-driven affordance steering and self-improvement from prior successful grasps without retraining.

II. PROBLEM FORMULATION

We consider the task of conditional grasp synthesis. Given a scene point cloud $s \in R^{N \times 3}$ and the gripper at pose $T = (R, p) \in SE(3)$, a learned encoding ϕ , relates the scene to the gripper at its current pose. Specifically, given a learned equivariant encoding of the scene and of the gripper, ϕ outputs a set of equivariant features, per joint query $\{q_i\}_{i=1}^{D_g+2}$ associated with the gripper with D_g degrees of freedom. To ensure a consistent reference frame for predicting rotational and translational velocities under arbitrary gripper kinematics, two additional query points are anchored at the origin of the gripper frame. Each query point $q_i = (p(q_i), \phi(q_i))$ is represented by its position $p(q_i)$ and feature vector $\phi(q_i) =$

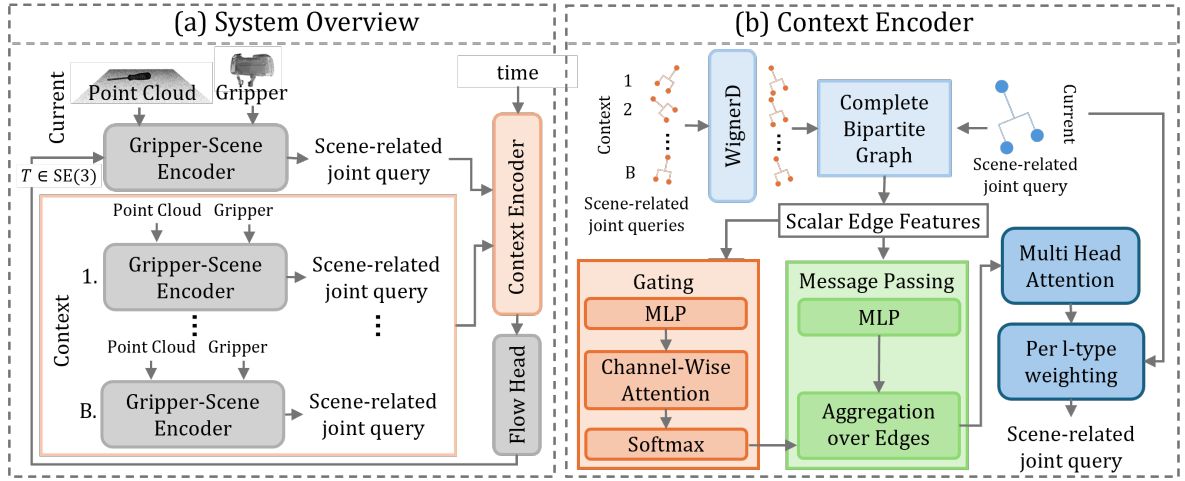


Fig. 2. We add a context encoder to the equivariant flow matching for grasp synthesis described in [8], to enable in-context learning. (a) Scenes from the context are encoded using the same equivariant gripper-scene encoder also used for encoding the current scene. The predicted current scene-related joint queries are modulated using the scene-related joint queries from all context scenes. The modulated features are given to a flow head for denoising the grasp pose. (b) The context encoder uses a graph-based attention mechanism with scalar features.

$\bigoplus_{\ell=0}^{\ell_{\max}} \phi^{\ell}(q_i)$, which is a direct sum of irreducible representations (irreps) in $SO(3)$ up to ℓ_{\max} . Given the learned scene-related joint queries, flow matching is applied to predict velocities for iteratively moving an initial randomly sampled pose T_0 to the target pose T_1 . [8]

To enable zero-shot adaptation, in-context learning is applied to modulate features based on experiences stored in the context, thus steering the flow towards the grasp modalities implied by the context. For this a context $\mathcal{D}^{\text{con}} = \{(s^{\text{con}_b}, T^{\text{con}_b} = (R^{\text{con}_b}, p^{\text{con}_b}), \{q_i^{\text{con}_b}\}_{i=1}^{D_g+2})\}_{b=1}^B$, containing scene point clouds s^{con_b} , grasp poses T^{con_b} , and scene-related joint queries $q_i^{\text{con}_b}$, for B context scenes is provided. By encoding the context scene-gripper relation with the same encoder ϕ as the current scene, modulation of scene-related joint queries can be executed entirely in the equivariant feature space. In the next section, the in-context feature modulation will be described. An overview of the system is given in Fig. 2. As gripper-scene encoder ϕ we use the equivariant encoder from [8].

III. METHOD

A. Feature modulation via context

Encoding the current scene and context scenes, results in current scene-related joint queries $\{q_i\}_{i=1}^{D_g+2}$, and scene related joint queries for the context $\{\{q_j^{\text{con}_b}\}_{j=1}^{D_g+2}\}_{b=1}^B$. The current joint queries are modulated using a graph-based attention mechanism over the context joint queries, which we call the context encoder. Afterwards, the modulated features are then given to the flow head for denoising.

In the context encoder, all context features are compared to current features. To this end, a complete bipartite graph is constructed, connecting all current query points with all context query points. To capture feature similarity, channel-wise scalar edge features are computed:

$$v_{i,b,j} = \left[\sum_{\ell=0}^{\ell_{\max}} W_c^{\ell} \sum_{m=-\ell}^{\ell} \phi_{c,m}^{\ell}(q_i) \phi_{c,m}^{\ell}(\hat{q}_j^{\text{con}_b}) \right]_{c=1}^C, \quad (1)$$

where W are learned weights and $c \in \{1, \dots, C\}$ are the channels. To ensure consistency across gripper frames, we move to a common local gripper frame for information exchange. As later velocities are predicted on the current gripper frame, context features are first mapped to that frame using Wigner D-matrices:

$$\phi(\hat{q}_j^{\text{con}_b}) = \bigoplus_{\ell=0}^{\ell_{\max}} \langle D^{\ell}(R) D^{\ell} \rangle (R^{\text{con}_b})^{-1}, \phi_{q_i^{\text{con}_b}}^{\ell}. \quad (2)$$

The positions of context query points are also mapped to the current gripper frame. Based on these aligned positions, additional scalar features are computed to capture relative distances and directions between query points. These features encode local correspondences; when modulating the features, it should be taken into account that query points in same positions and arrangement should have more similar features than those farther apart. Specifically we define distance $d_{i,b,j}$ and directional features $s_{i,b,j}$:

$$d_{i,b,j} = \text{enc}(\|p_i - p_j^{\text{con}_b}\|) \quad (3)$$

$$s_{i,b,j} = \left[\sum_{\ell=0}^{\ell_{\max}} M_c^{\ell} \sum_{m=-\ell}^{\ell} \phi_{c,m}^{\ell}(q_i) Y_m^{\ell}(q_i - q_j^{\text{con}_b}) \right]_{c=1}^C, \quad (4)$$

where Y^{ℓ} are the real spherical harmonics, M are learned weights, and enc is a sinusoidal position embedding. We provide a time encoding as an additional feature, as a lower feature similarity during earlier stages of the denoising process is to be expected.

The scalar edge features are passed to a gating mechanism, which identifies the most relevant context query points for each current query point q_i , both in terms of feature similarity and spatial alignment. Additionally, the scalar features are processed by a message-passing MLP to enrich the feature representation. The enriched features are then aggregated to each current query point q_i , using the attention weights computed by the gating mechanism. After an additional multi-head attention refinement, the scene-related joint queries

Dataset	Context Size					
	0	2	4	10	20	40
ID	0.66 (0.40)	0.83 (0.31)	0.84 (0.29)	0.86 (0.27)	0.88 (0.25)	0.90 (0.22)
OOD	0.63 (0.40)	0.90 (0.24)	0.89 (0.24)	0.90 (0.22)	0.90 (0.22)	0.90 (0.22)
Real World [17]	0.69 (0.35)	0.75 (0.34)	0.76 (0.33)	- (-)	- (-)	- (-)

TABLE I

MEAN PERCENTAGE (STANDARD DEVIATION) OF GRASP PREDICTIONS THAT FALL WITHIN OBJECT TARGET REGIONS ACROSS CONTEXT SIZES.

are modulated with the resulting attention-weights w_i , by modulating each irreducible representation ℓ uniformly over its $2\ell + 1$ components:

$$\hat{\phi}_{q_i} = \bigoplus_{\ell} \left(w_i^{\ell} \otimes I_{2\ell+1} \right) \phi_{q_i}^{\ell}. \quad (5)$$

B. Training data generation

Similar to the concept followed by [22], training data is generated to enable the policy to learn using arbitrary context without the intention of capturing real-world plausibility. For this we randomly sample a target and a non-target region on the object point cloud and allocate pre-generated grasps to both regions. From the target region, we sample one target grasp and a random number of other grasps for the context, for training. The context encoder is trained end-to-end with the complete network, using the flow-matching loss.

IV. EXPERIMENTS

We evaluate the policy’s ability to use the provided context by measuring the percentage of predicted grasps that fall within a specified target region, given a context with grasps from that region. We show that:

- A) Using the context grasps from the target region, grasp predictions are steered towards that region,
- B) Using the context encoder does not impact the overall grasp success rate.

A. Target Region Hit-Rate

a) Evaluation setup with Simulation Data: We generate 395 scenes using 78 objects from the YCB [2] and Google Scanned Objects [6] datasets. Each scene contains an object with a pose and target region not seen during training. Of these objects, 70 were used during training, and 8 are novel objects. Target regions are sampled randomly (see sec. III-B). The context is sampled from the same object, with identical target and non-target regions, but under a different object pose, to demonstrate robustness towards object rotation and translation.

b) Evaluation setup with Real-world Data: Additionally, we evaluate the policy, trained entirely on simulated data, on the TaskGrasp [17] dataset. The dataset contains grasps defined on real-world recorded multi-view point clouds of household objects. The grasps are assigned affordance labels for a total of 56 tasks. We identify affordance-based target regions by collecting the object points in the

proximity of the contact points of grasps suitable for a given task. After filtering and discarding target regions that span over 90% of the object point cloud and object-task combinations that yield fewer than four suitable grasps, we evaluate the policy on 151 objects and 756 object-task combinations from the TaskGrasp [17] dataset. Due to the limited number of grasps in the dataset, we only evaluate up to a context size of four.

c) Results: Table I shows that already with a small context size, the target region hit rate is improved, compared to the baseline, where the policy is not provided with a context. Larger context sizes further increase the hit rate, as they are more likely to cover the full target region compared to small context sizes, where many grasps may lie at the border from target and non-target region.

B. Overall Grasp Success Rate

Datasets	wo	0	2	4	10	20	40
id	0.98	0.97	0.96	0.96	0.96	0.96	0.96
ood	0.93	0.97	0.85	0.86	0.85	0.85	0.85

TABLE II

MEAN SUCCESS RATES OF ALL GRASP PREDICTIONS, FROM OUR POLICY USING CONTEXT SIZES 0 TO 40, WITH THE BASELINE (WO) [8], WHICH DOES NOT CONTAIN A CONTEXT ENCODER.

Grasp success rate is evaluated by simulating all predicted grasps in a MuJoCo environment. A grasp is considered successful if the object can be lifted to a predefined height without failure. As shown in Table II, the policy incorporating a context encoder achieves a success rate comparable to the baseline [8], which does not contain a context encoder. Slightly lower success rates for some context sizes could stem from the random sampling of target regions, which does not take into account that grasps on some parts of an object might be more prone to failure.

V. OUTLOOK: IMPROVEMENT OF GRASP SUCCESS RATE

Steering grasps towards a specific modality, implied by the context, has the potential to improve grasp success rates. As visualized in fig 1, successful grasps from the policy execution can be fed back as context to steer grasp prediction towards stable grasp regions. A current limitation is posed by grasp modalities equally containing successful and unsuccessful grasps, requiring a more informed selection of the context.

VI. CONCLUSIONS

We introduce in-context equivariant grasp learning to steer policy predictions at each step of the flow-matching forward process toward grasp regions implied by the context. This attention-based modulation operates solely on relative grasp–scene structure, without access to absolute grasp poses in the context, which enables generalization across scenes captured in different configurations. We showed that the context can be used to guide the policy towards specific grasp modalities. In subsequent studies, we will further investigate the potential of this method for grasp refinement.

REFERENCES

- [1] K. R. Barad, A. Orsula, A. Richard, J. Dentler, M. A. Olivares-Mendez, and C. Martinez, "Graspldm: Generative 6-dof grasp synthesis using latent diffusion models," *IEEE Access*, vol. 12, pp. 164 621–164 633, 2024.
- [2] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set," *IEEE Robotics Automation Magazine*, vol. 22, no. 3, pp. 36–52, 2015.
- [3] J. Carvalho, A. T. Le, P. Jahr, Q. Sun, J. Urain, D. Koert, and J. Peters, "Grasp diffusion network: Learning grasp generators from partial point clouds with diffusion models in so(3)xr3," 2024. [Online]. Available: <https://arxiv.org/abs/2412.08398>
- [4] A. L. Chandra, I. Nematollahi, C. Huang, T. Welschehold, W. Burgard, and A. Valada, "Diwa: Diffusion policy adaptation with world models," in *Proceedings of The 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Lim, S. Song, and H.-W. Park, Eds., vol. 305. PMLR, 27–30 Sep 2025, pp. 3378–3400. [Online]. Available: <https://proceedings.mlr.press/v305/chandra25a.html>
- [5] K. Chen, Z. Shen, Y. Zhang, L. Chen, F. Wu, Z. Bing, S. Haddadin, and A. Knoll, "Lemmo-plan: Llm-enhanced learning from multimodal demonstration for planning sequential contact-rich manipulation tasks," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 11 972–11 978.
- [6] L. Downs, A. Francis, N. Koenig, B. Kinman, R. Hickman, K. Reymann, T. B. McHugh, and V. Vanhoucke, "Google scanned objects: A high-quality dataset of 3d scanned household items," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 2553–2560.
- [7] R. Freiberg, A. Qualmann, N. A. Vien, and G. Neumann, "Diffusion for multi-embodiment grasping," *IEEE Robotics and Automation Letters*, vol. 10, no. 3, pp. 2694–2701, 2025.
- [8] —, "Towards a multi-embodied grasping agent," 2025. [Online]. Available: <https://arxiv.org/abs/2510.27420>
- [9] N. Funk, J. Urain, J. Carvalho, V. Prasad, G. Chalvatzaki, and J. Peters, "Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching," *arXiv preprint arXiv:2409.04576*, 2024.
- [10] D.-C. Hoang, A.-N. Nguyen, C.-M. Nguyen, A.-B. Phi, Q.-T. Duong, K.-D. Tran, V.-A. Trinh, V.-D. Tran, H.-N. Pham, P.-Q. Ngo, D.-Q. Vu, T.-U. Nguyen, V.-D. Vu, D.-T. Tran, and V.-T. Nguyen, "Collision-free grasp detection from color and depth images," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 11, pp. 5689–5698, 2024.
- [11] D.-C. Hoang, J. A. Stork, and T. Stoyanov, "Context-aware grasp generation in cluttered scenes," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 1492–1498.
- [12] B. Hu, X. Zhu, D. Wang, Z. Dong, H. Huang, C. Wang, R. Walters, and R. Platt, "Orbitgrasp: Se (3)-equivariant grasp learning," in *8th Annual Conference on Robot Learning*, 2024. [Online]. Available: <https://openreview.net/forum?id=clqzoCruLY>
- [13] D. Huang, W. Dong, C. Tang, and H. Zhang, "Hgdifuser: Efficient task-oriented grasp generation via human-guided grasp diffusion models," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 19 538–19 545.
- [14] G. Lanza, B. Deml, S. Matthiesen, M. Martin, O. Brützel, and R. Hörsting, "The vision of the circular factory for the perpetual innovative product," *at-Automatisierungstechnik*, vol. 72, no. 9, pp. 774–788, 2024.
- [15] B. Lim, J. Kim, J. Kim, Y. Lee, and F. C. Park, "Equigraspflow: Se(3)-equivariant 6-dof grasp pose generative flows," in *8th Annual Conference on Robot Learning*, 2024.
- [16] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 11 048–11 064. [Online]. Available: <https://aclanthology.org/2022.emnlp-main.759/>
- [17] A. Murali, W. Liu, K. Marino, S. Chernova, and A. Gupta, "Same object, different grasps: Data and semantic knowledge for task-oriented grasping," in *Conference on Robot Learning*, 2020.
- [18] R. Newbury, M. Gu, L. Chumbley, A. Mousavian, C. Eppner, J. Leitner, J. Bohg, A. Morales, T. Asfour, D. Kragic, D. Fox, and A. Cosgun, "Deep learning approaches to grasp synthesis: A review," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3994–4015, 2023.
- [19] G. Singh, S. Kalwar, M. F. Karim, B. Sen, N. Govindan, S. Sridhar, and K. M. Krishna, "Constrained 6-dof grasp generation on complex shapes for improved dual-arm manipulation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, pp. 7344–7350.
- [20] P. Song, P. Li, and R. Detry, "Implicit grasp diffusion: Bridging the gap between dense prediction and sampling-based grasping," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, P. Agrawal, O. Kroemer, and W. Burgard, Eds., vol. 270. PMLR, 06–09 Nov 2025, pp. 2948–2964. [Online]. Available: <https://proceedings.mlr.press/v270/song25b.html>
- [21] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki, "Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 5923–5930.
- [22] V. Vosylius and E. Johns, "Instant policy: In-context imitation learning via graph diffusion," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.
- [23] Z. Weng, H. Lu, D. Kragic, and J. Lundell, "Dexdiffuser: Generating dexterous grasps with diffusion models," *IEEE Robotics and Automation Letters*, vol. 9, no. 12, pp. 11 834–11 840, 2024.
- [24] Z. Zhang, L. Zhou, C. Liu, Z. Yuan, S. Guo, R. Zhao, M. H. Ang Jr, and F. E. Tay, "Dexgrasp-diffusion: Diffusion-based unified functional grasp synthesis method for multi-dexterous robotic hands," *arXiv preprint arXiv:2407.09899*, 2024.
- [25] T. Zhong and C. Allen-Blanchette, "Gagrasp: Geometric algebra diffusion for dexterous grasping," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*, 2025, pp. 6771–6778.
- [26] H. Zhou, Y. Lin, L. Yan, and H. Min, "Human-in-the-loop learning for adaptive robot manipulation using large language models and behavior trees," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2025, pp. 19 039–19 046.