Compositional Neural Field Movement Primitives

Ahmet Tekden[1] Yasemin Bekiroglu[1,2]

***Abstract*— Compositionality enables scalable and data-efficient robot learning by representing complex behavior as combinations of simpler elements. Building on this principle, we propose a learning-from-demonstration framework that connects perception and motion through shared object-level representations. Scenes are modeled using object-centric neural representations with latent-conditioned deformations, while motion is generated through a temporal mixture-of-experts (MoE) framework that combines object-conditioned movement primitives over time. This spatiotemporal compositionality preserves the efficiency of movement primitives while grounding motion in visual scene structure. Experiments in simulation and on real robots demonstrate strong generalization, data efficiency, robustness to scene variations, and compatibility with both language-based segmentation and 3D scene representations.**

I. Introduction

Learning from Demonstration (LfD) [1], [2] is a widely used approach for teaching robots task-specific skills from human demonstrations. A central challenge is learning diverse, long-horizon behaviors from limited data. Compositionality offers a natural way to address this challenge by decomposing tasks into simpler, reusable components. Such formulations improve data efficiency and generalization compared to monolithic approaches [3]. Prior work has explored compositional motion representations, for example, through funnel-based approaches to guide sequential behaviors [4]. Motivated by these observations, representing both scene and motion in a compositional manner provides a principled framework for learning structured robot behaviors.

Movement primitives (MPs) [5], [6], [7], [8], [9] are a standard representation for encoding and generalizing demonstrated skills. They provide smooth, compact trajectory models that can be learned from few demonstrations, but typically rely on hand-crafted low-dimensional task parameters and manually specified compositional structure. Image-conditioned alternatives [7] reduce feature engineering but often require substantially more data.

Neural field movement primitives (NFMP) [8] extend MPs by jointly modeling scenes and motions, but operate at the scene level and become increasingly difficult to scale as the number of objects or task parameters grows. To address these limitations, we extend NFMP by incorporating compositional structure into the joint modeling of scenes and motions. Specifically, we introduce a framework that connects perception and motion through shared object-level latent representations. Scenes are modeled using a spatial Mixture-of-Experts (MoE) with object-centric representations, while motion is generated through a temporal MoE that combines object-conditioned movement primitives over time. Object-centric structure is encouraged through soft mask-based importance sampling, allowing each expert to specialize without requiring precise object segmentation. Temporal gating enables multiple experts to contribute within the same motion segment, capturing the relational nature of manipulation. This is illustrated in Fig. 1, which visualizes object interactions and corresponding scene reconstructions.

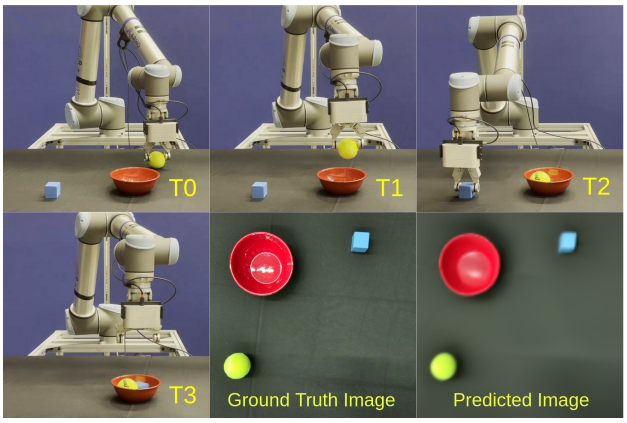

Fig. 1. Sequence of object interactions (T0–T3) and corresponding ground-truth and reconstructed images. The compositional formulation captures object-level structure and enables consistent motion generation across task stages.

II. Method

We consider learning robot motion from demonstration in settings where task variation is induced by changes in object configurations. Each demonstration consists of a scene observation, coarse object masks used only during training to roughly localize objects, and a corresponding time-aligned trajectory. Our goal is to learn object-centric latent scene representations that capture these variations and use them to generate the appropriate motion for a new scene.

A. Preliminaries: NFMP

Neural Field Movement Primitives (NFMP) [8] jointly model scenes and robot motions through a shared latent representation. A neural field encodes the scene as a continuous function conditioned on this latent variable, while a motion

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation, the Chalmers AI Research Center (CHAIR), and the Chalmers Gender Initiative for Excellence (Genie). [1]Department of Electrical Engineering, Chalmers University of Technology, SE-412 96 Gothenburg, Sweden. [2]Department of Computer Science, University College London, WC1E 6BT London, U.K. Email: tekden@chalmers.se

[1]https://youtu.be/_rpfihCIMfg

model generates trajectories from the same embedding. This is formally expressed as

$$s(x) = F(x \mid z), q(t) = G(t \mid z) \tag{1}$$

where z denotes the shared latent variable, F and G are the scene and motion functions, $s(x)$ represents the scene signal at spatial coordinate x (e.g. RGB or occupancy), and $q(t)$ denotes the motion signal, such as joint angles or end-effector position. This formulation enables generalization across scene variations but operates at the scene level and does not explicitly capture object-level or compositional structure.

B. Compositional Scene and Motion Generation

To introduce compositional structure, we extend NFMP by incorporating object-centric scene representations and temporally varying motion conditioning. Rather than modeling the scene as a single latent entity, we decompose it into object-level components and associate each with a latent representation that can be selectively used during motion generation. This enables structured generalization across variations in object configurations and task compositions.

We model each scene using an object-centric spatial Mixture-of-Experts (MoE):

$$\hat{I}(x|\{z_i\}_{i=1}^N) = \sum_{i=1}^{N} M_i(x_i)\, F_i(x_i) + M_{\text{bg}}\, F_{\text{bg}}(x), \tag{2}$$

with deformed coordinates $x_i = x - \Delta_i(x \mid z_i)$, where F_i is the appearance field of object i, Δ_i is a latent-conditioned deformation field, and F_{bg} models the background. The masks $M_i(x)$ are learned through a soft competition mechanism over object and background components, encouraging each expert to specialize without requiring explicit segmentation. This yields a smooth object-centric representation in which changes in object position or geometry correspond to structured variations in latent space.

Given object-level latents $z_i{}_{i=1}^N$, we model compositional motion by forming a time-varying latent representation:

$$z(t) = \sum_{i=1}^{N} w_i(t)\, \hat{z}_i, \qquad w_i(t) = \frac{\exp(s_i(t))}{\sum_{j=1}^{N} \exp(s_j(t))}, \tag{3}$$

where $w_i(t)$ are time-dependent weights obtained from learnable relevance scores. This temporal gating allows different objects to influence different phases of the trajectory, while also enabling multiple objects to contribute within the same segment. The resulting latent is used to condition a trajectory generator that produces robot motion over time. This allows the model to adapt motion generation to different scene configurations by selectively attending to relevant objects over time.

Object-Centric Training: During training, we use coarse object masks to guide the specialization of spatial experts without requiring precise segmentation. Sampling is biased toward regions associated with each object, encouraging each component to focus on a single object while maintaining flexibility in ambiguous regions. This weak supervision promotes stable object-centric decomposition and improves robustness to variations in object appearance and position.

Latent Relabeling: To obtain a compact and structured latent space, we represent each object using a small set of endpoint latent vectors that span the observed variations. Demonstration-specific latents are expressed as convex combinations of these endpoints, enabling smooth interpolation across configurations. This parameterization reduces the effective degrees of freedom while preserving flexibility.

Implementation Details: All fields are implemented as coordinate-based MLPs, with latent conditioning realized through FiLM layers [10]. We apply Lipschitz regularization [11] to encourage smooth latent-to-deformation mappings. For orientation trajectories, we follow [12] and model rotations in tangent-space axis-angle form. The same formulation can be extended from 2D appearance fields to 3D occupancy fields [13] for 3D scene modeling.

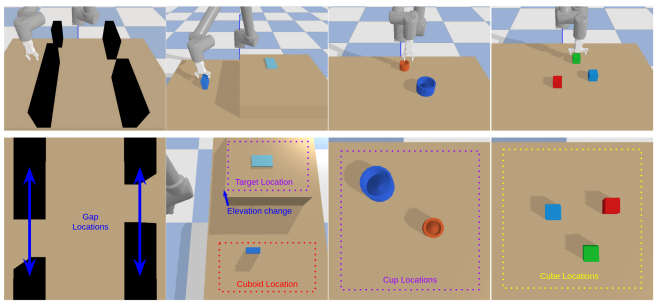

Fig. 2. Simulation tasks with corresponding scene variations. The bottom row shows the camera images used by the methods. In these images, arrows highlight geometric changes and rectangles mark possible object locations.

TABLE I

SIMULATION RESULTS (MED/ACCURACY).

Method (data regime)	Wall Avoidance	Incline Pick-and-Place	Cup Stacking	Cube Stacking
CNN-CNMP (Low)	2.66/0.63	3.98/0.11	9.03/0.03	13.44/0.00
CNN-CNMP (High)	1.33/0.97	2.12/0.68	2.81/0.23	5.87/0.00
NFMP (Grid)	0.65/0.99	NA/NA	NA/NA	NA/NA
Ours w/o gating (Low)	0.66/**1.00**	0.85/**1.00**	1.06/0.81	1.39/0.79
Ours (Low)	**0.48/1.00**	**0.73/1.00**	**0.60/1.00**	**0.66/0.97**

* Low/High training sizes per task: Wall Avoidance (10/30), Incline Pick-and-Place (20/50), Cup Stacking (20/50), and Cube Stacking (30/100).
* Bold numbers indicate the best performance in each task under the low- and all-data regimes, respectively.
* Lower MED and higher accuracy indicate better performance.

III. EXPERIMENTS

We evaluate the proposed framework in simulation and on real-robot manipulation tasks. The simulation studies assess generalization and data efficiency, while the real-world experiments demonstrate robustness to perception noise, category-level generalization, and compatibility with both 2D and 3D scene representations.

A. Simulation Experiments

We first evaluate the method on four simulated manipulation tasks: Wall Avoidance, Incline Pick-and-Place, Cup Stacking, and Cube Stacking (Fig. 2). These tasks vary in

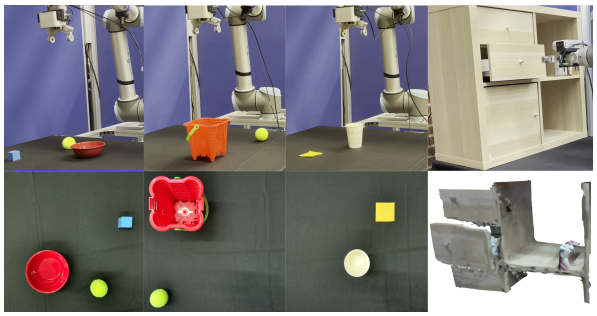

Fig. 3. Real-world tasks and model inputs. For the 3D task, representative mesh is visualized.

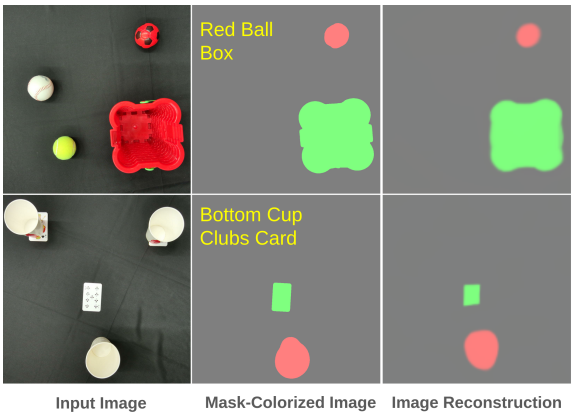

Fig. 4. Language-conditioned manipulation examples. Prompts specify the target object, enabling task execution in cluttered multi-object scenes.

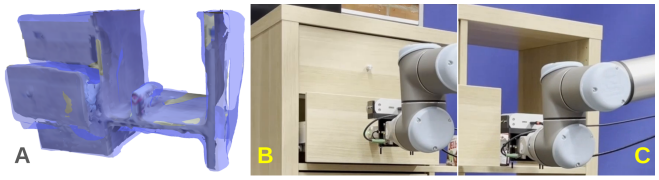

Fig. 5. Input mesh and reconstructed surface (A), and corresponding robot executions at key grasp poses (B, C). The reconstructed surface (shown in blue) captures drawer opening and box position correctly.

object geometry, placement, and scene structure. We compare against CNN-CNMP [7], a modified Neural Field Movement Primitive baseline (NFMP) [8] (adapted to use FiLM conditioning instead of hypernetworks), and an ablation without temporal gating. NFMP is only evaluated on Wall Avoidance, as its grid-based training scales poorly with the number of task parameters.

Performance is measured using the mean Euclidean distance (MED) between predicted and demonstrated trajectories and the task success rate. Results are reported under low- and high-data regimes for baselines, while the proposed method is evaluated only in the low-data regime. As shown in Table I, the proposed method consistently outperforms all baselines in the low-data setting and remains competitive with, or better than, baselines trained with substantially more demonstrations. The temporal gating mechanism further improves performance, particularly in more complex tasks.

B. Real-Robot Experiments

We further evaluate the framework on four real-world manipulation tasks (Fig. 3), including tabletop pick-and-place, category-level generalization, and language-conditioned multi-object scenarios. Demonstrations are collected via kinesthetic teaching, except for the final task, where motion planning is used, and trajectories are temporally aligned using gripper-state transitions.

In the first task, the robot performs pick-and-place under varying object locations. Using 30 demonstrations, the system successfully completes all 25 test trials and remains robust to clutter, succeeding in 5 additional trials with distractor objects.

The second and third tasks evaluate category-level generalization using masks generated by a vision-language segmentation model [14]. In the second task, the robot picks up a ball and drops it into a box; in the third task, it picks up a cup and places it on a sheet of paper. Trained on only 20–30 demonstrations, the system achieves 14/15 and 15/15 successes, respectively, including evaluations on unseen object instances within the same semantic category. We further evaluate a language-conditioned setting where a user prompt specifies the target object in a cluttered scene. Across multiple trials and prompt variations, the system successfully completes all executions (Fig. 4).

We evaluate 3D scene generalization on a drawer manipulation task, where a carton box must be placed inside a drawer under varying openings and object positions. The task requires coordinated reasoning over articulated geometry and object pose. Scene representations are constructed from occupancy fields derived from multi-view scans [15]. The model is trained on four demonstrations and evaluated on 10 novel configurations, achieving successful execution in all cases. Figure 5 shows a representative reconstruction and the corresponding key task poses for the given configuration.

IV. Conclusion

This paper presents a compositional framework for motion generation from visual demonstrations that couples object-centric scene representations with temporally structured motion modeling. The approach learns smooth, object-level latent representations from a small number of demonstrations and uses them to generate trajectories through time-varying expert aggregation. Experiments in both simulation and real-world settings demonstrate that object-centric compositionality yields interpretable representations, strong data efficiency, and systematic generalization across scene variations. Despite these advantages, the current formulation relies on coarse object masks during training to encourage object-centric specialization, and performance can degrade in cluttered scenes with visually similar distractor objects. Future work will investigate reducing these supervision requirements and incorporating richer semantic representations to improve robustness and broader, scene-level generalization.

REFERENCES

[1] S. Schaal, "Learning from demonstration," *Adv. Neural Inf. Process. Syst.*, vol. 9, 1996.

[2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robot. Auton. Syst.*, vol. 57, no. 5, pp. 469–483, 2009.

[3] Y. Du and L. P. Kaelbling, "Position: Compositional generative modeling: A single model is not all you need," in *Int. Conf. Mach. Learn. (ICML)*, 2024.

[4] R. R. Burridge, A. A. Rizzi, and D. E. Koditschek, "Sequential composition of dynamically dexterous robot behaviors," *Int. J. Robot. Res.*, vol. 18, no. 6, pp. 534–555, 1999.

[5] A. J. Ijspeert, J. Nakanishi, H. Hoffmann, P. Pastor, and S. Schaal, "Dynamical movement primitives: learning attractor models for motor behaviors," *Neural computation*, vol. 25, no. 2, pp. 328–373, 2013.

[6] A. Paraschos, C. Daniel, J. R. Peters, and G. Neumann, "Probabilistic movement primitives," *Adv. Neural Inf. Process. Syst.*, vol. 26, 2013.

[7] M. Y. Seker, M. Imre, J. H. Piater, and E. Ugur, "Conditional neural movement primitives." in *Robot.: Sci. Syst. (RSS)*, vol. 10, 2019.

[8] A. Tekden, M. P. Deisenroth, and Y. Bekiroglu, "Neural field movement primitives for joint modelling of scenes and motions," in *IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS)*, 2023, pp. 3648–3655.

[9] Y. Yildirim and E. Ugur, "Conditional neural expert processes for learning movement primitives from demonstration," *IEEE Robot. Autom. Lett.*, 2024.

[10] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *AAAI Conf. Artif. Intell. (AAAI)*, vol. 32, no. 1, 2018.

[11] H.-T. D. Liu, F. Williams, A. Jacobson, S. Fidler, and O. Litany, "Learning smooth neural functions via lipschitz regularization," in *ACM SIGGRAPH Conf.*, 2022, pp. 1–13.

[12] A. Ude, B. Nemec, T. Petrić, and J. Morimoto, "Orientation in cartesian space dynamic movement primitives," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2014, pp. 2997–3004.

[13] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4460–4470.

[14] L. Medeiros, "Lang segment anything," GitHub repository, 2023, https://github.com/luca-medeiros/lang-segment-anything.

[15] A. Millane *et al.*, "nvblox: Gpu-accelerated incremental signed distance field mapping," in *IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024, pp. 2698–2705.