# Semantic–Geometric Task Representations for Bimanual Manipulation from Human Demonstrations to Robot Action Planning

Franziska Herbert[1,2], Vignesh Prasad[1,2,3], Han Liu[1,2], Dorothea Koert[1] and Georgia Chalvatzaki[1,2,3]

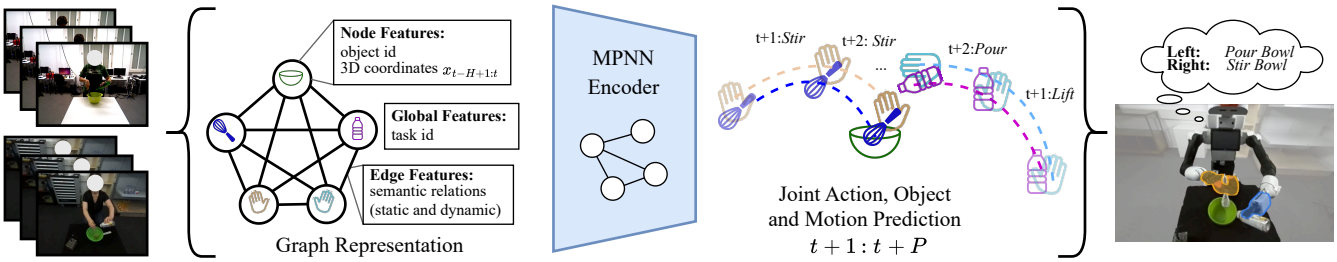


**Fig. 1:** Overview of our approach. We learn graph representations from bimanual human demonstrations of manipulation tasks. A Message Passing Neural Network is used to learn the underlying concepts of the task, which can subsequently be used to replicate the task on other agents or environments.

***Abstract*—Learning structured task representations from human demonstrations is essential for bimanual manipulation, where action ordering, object involvement, and interaction geometry vary significantly across executions. A key challenge lies in jointly capturing the discrete semantic task structure and the temporal evolution of object-centric geometric relations in a form that supports reasoning over task progression. We introduce a semantic–geometric graph-based task representation that jointly encodes object identities, inter-object semantic relations, and per-object motion histories, via a Message Passing Neural Network (MPNN) encoder and a Transformer-based decoder. The encoder operates solely on the temporal scene graph, producing structured representations decoupled from action labels. The decoder then conditions on action-context to forecast future actions, associated objects, and object motions. This decoupling learns task-agnostic representations, enabling encoder reuse across embodiments through decoder-only finetuning on a small robot dataset. Across eleven bimanual tasks from two human datasets, we find that the benefit of structured semantic–geometric representations over simpler sequence-based models grows with task variability in action ordering and object involvement. At deployment, a planner couples the action and motion predictions of the learned human task representations with learned Probabilistic Movement Primitives, achieving full task success on two real-robot bimanual tasks and outperforming graph ablations, Transformer, decoder-only, and finetuned vision-language model baselines.**

## I. Introduction

Bimanual manipulation tasks exhibit substantial variability: for example when clearing a cluttered table into a box, the order of object selection, hand assignment, and interaction geometry all vary across demonstrations. Learning generalizable task representations from human demonstrations requires capturing both *what* is happening (the semantic structure of actions, objects, and their interactions) and *how* it happens (the geometric evolution of the scene). Without both, a representation either memorizes sequences without generalizing, or tracks motion without understanding task structure.

Scene graphs [1], [2] capture semantic and geometric structure naturally, and GNNs [3], [4] aggregate this information via message passing. However, existing scene-graph approaches to task understanding [5]–[7] typically emphasize either semantic structure or geometric evolution, rarely integrating both. As a result, they remain at the level of frame-wise action recognition [7], motion forecasting [6], or single-step unimanual predictions [5], without demonstrating robotic decision-making.

We introduce a semantic-geometric task graph representation and GNN encoder that jointly encodes object identities, inter-object semantic relations, per-object motion histories, and global task context into coupled node, edge, and global embeddings. Our decoder jointly forecasts future action sequences, associated objects, and continuous object motions from a shared graph embedding. At deployment, a planner scores candidate robot skills by combining predicted action likelihoods with motion-consistency under learned ProMPs [8], exploiting both prediction heads at test time.

Our key contributions are: **(1)** a semantic-geometric task graph representation jointly encoding object identities, inter-object relations, and per-object motion histories for bimanual manipulation from human demonstrations; **(2)** a decoupled graph-encoder/transformer-decoder architecture separating scene representation from action-conditioned forecasting, enabling encoder reuse across embodiments via decoder-only finetuning; **(3)** an empirical characterization of when structured relational inductive biases are beneficial — across eleven bimanual tasks, the advantage over sequence-based baselines grows with task variability in action ordering and object involvement; **(4)** a real-robot action planner coupling semantic and geometric predictions at deployment, achieving full task success on two bimanual tasks and outperforming graph ablations, Transformer, decoder-only, and finetuned VLM baselines under similar data budgets.

[1] Computer Science Dept., TU Darmstadt, Germany. [2] Hessian.AI. [3] Robotics Institute Germany (RIG).
This research is funded by EU Horizon Europe Projects MANiBOT (101120823), ARISE (101135959), the DFG Emmy Noether Programme (CH 2676/1-1), the ERC project "SIREN" (101163933) and the BMFTR Projects "RIG" (16ME1001) and "IKIDA" (01IS20045).
The authors gratefully acknowledge the computing time provided on the high-performance computer Lichtenberg II at TU Darmstadt, funded by the BMFTR and the State of Hesse.
Contact: franziska.herbert@tu-darmstadt.de

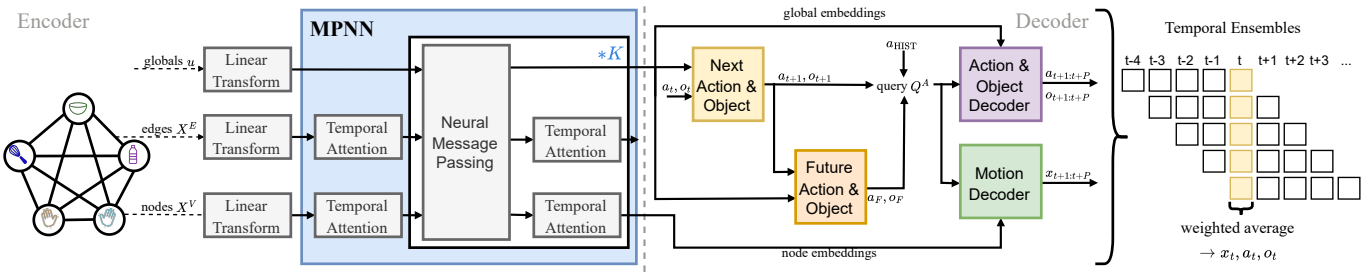


**Fig. 2:** Model architecture: the graph encoder transforms features into embeddings via the MPNN, and the decoders forecast actions, objects, and motions.

## II. Related Work

Graph-based action and motion prediction has focused largely on skeleton-based representations [9]–[12], which are less suited to tabletop manipulation where structure lies in object-level relations. Scene graphs [1], [2], [13], [14] offer a natural relational framework and have been applied to Human-Object Interaction [15], [16], typically for frame-level recognition without modeling task progression. Dreher et al. [7] use semantic relation graphs for frame-wise action classification in bimanual demonstrations but do not forecast future actions or motions. Razali et al. [6] learn motion prediction from geometric scene graphs in a zero-shot, action-conditioned manner without online adaptation. Lagamtzis et al. [5] combine action recognition, future action prediction, and motion forecasting, but neglect semantic edge features and global context, and are limited to single-step unimanual predictions. Beyond graph approaches, task representations range from symbolic formalisms [17] to LLMs/VLMs [18], [19] and end-to-end policies [20], [21], which require large pretraining corpora and do not naturally extend to continuous multi-object motion forecasting. In contrast, we jointly forecast actions, objects, and motions from a unified semantic-geometric graph, retaining data efficiency while transferring to a physical bimanual robot for online action planning.

## III. Learning Semantic-Geometric Task Graph-Representations

We propose an approach for learning semantic-geometric graph-based task representations from bimanual human demonstrations. Each task is modeled as a sequence of actions spanning multiple time steps, in order to predict future task progression by modeling past object movements and semantic relationships. We formalize each demonstration as a spatio-temporal graph $\mathcal{G}^{(t)} = (\mathcal{V}, \mathcal{E}, u)$ whose node, edge, and global features evolve over $H$ past frames. Using an extended message-passing formulation [22], our encoder jointly learns node embeddings $h_v$, edge embeddings $f_{vw}$, and global embeddings $g_u$ through iterative information exchange; the decoder then forecasts future actions, objects, and motions over a prediction horizon $P$ (Figure 1).

For each frame $t$, we construct a fully-connected bidirectional spatial-temporal graph encoding the scene over $H$ historical frames, annotated with bimanual actions $a_t = (a_t^R, a_t^L)$ and objects $o_t = (o_t^R, o_t^L)$. **Nodes** represent objects and hands, using one node per object with temporal features to reduce complexity. The node feature matrix concatenates one-hot object ID $c_n$ with 3D coordinates over $H$ past frames at stride $S$, giving $X^V \in \mathbb{R}^{N \times H \times d_V}$. **Edges** encode semantic spatial and dynamic relations [23] via multi-hot encodings over the same history, $X^E \in \mathbb{R}^{M \times H \times d_E}$. A **global** variable $u \in \mathbb{R}^{d_U}$ holds a one-hot task ID shared across all nodes and edges.

Node, edge, and global features are linearly projected into a shared hidden dimension $d_{\text{MP}}$ and encoded with RoPE [24] over their temporal dimension. An MPNN iteratively refines all embeddings over $K$ rounds, updating edge, node and global embeddings using per-iteration learnable weights $W_k^E$. Message passing alternates with temporal self-attention over the history dimension to capture cross-timestep dependencies. The resulting node embeddings $h_v^K$ and global embeddings $g_u^K$ (temporally averaged to $\bar{g}_u^K$) are passed to the decoder.

The decoder predicts future actions, objects, and object motions as object-level 3D trajectories over horizon $P$ using Transformer decoders attending to node embeddings. Since actions evolve at coarser time scales than motions, we adopt a multi-stage decoder. First, an MLP predicts the immediate next action-object pair $a_{t+1}, o_{t+1}$ from the current pair $a_t, o_t$ and temporally averaged global embeddings $\bar{g}u^K$, capturing short-term task progression. Second, another MLP predicts future semantic pairs $a_F, o_F$ — the next high-level action after the current one — using $a_{t+1}, o_{t+1}$ and $\bar{g}_u^K$. This decouples long-term task sequencing from action duration and separates perceptual encoding from action-conditioned reasoning.

To predict sequences over horizon $P$, we construct action-object queries that encode both temporal and semantic context. Task progression depends on motion history (in graph embeddings) and high-level action-object history. We track the $n_{\text{past}}$ most recent pairs per hand (left/right), denoting pairs as $a_i = (a_i, o_i)$. For each hand $i \in R, L$, we form a query $a_{i,Q} = [a_{i,\text{HIST}} | a_{i,t+1} | a_{i,F}]$, concatenating past, predicted next, and future semantic pairs. The queries from both hands are concatenated into a final query vector $Q^A$. Two Transformer decoders [25] operate on $Q^A$: one attends to global embeddings $g_u^K$ to predict action-object sequences, while the other attends to node embeddings $h_v^K$ to predict object motion.

The model is trained jointly on action and motion predictions via a combined loss of weighted cross-entropy (compensating for class imbalance) for action and object classifiers and MSE for motion predictions. At each step, action chunking and temporal ensembling with exponential decay [26] is performed

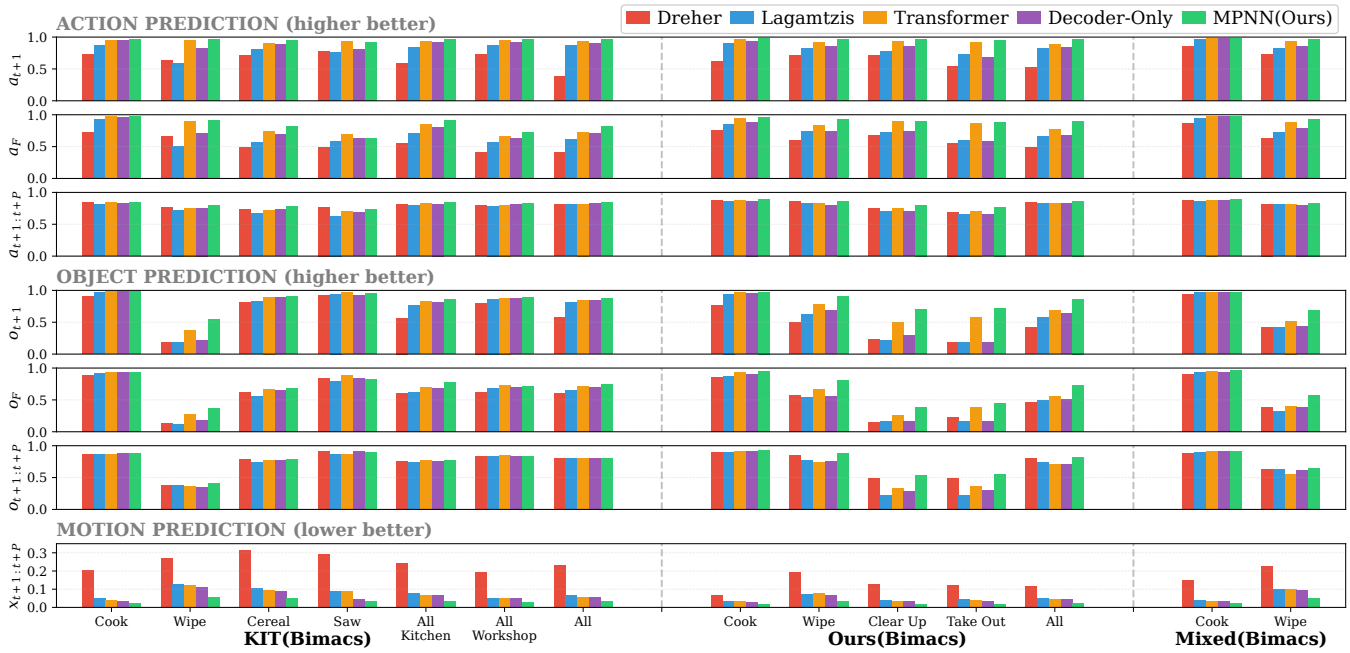


**Fig. 3:** Cross validation results on selected tasks from the KIT(Bimacs), from Ours(Bimacs) and a mixed dataset, including also multi-task models. Action prediction accuracies (next $a_{t+1}$, future $a_F$ and horizon $a_{t+1:t+P}$) are shown on top, object predictions in the middle (next $o_{t+1}$, future $o_F$ and horizon $o_{t+1:t+P}$) and motion prediction ($x_{t+1:t+P}$) RMSE on the bottom.

on overlapping predictions $a_{t+1:t+P}$, $o_{t+1:t+P}$ and $x_{t+1:t+P}$, assigning higher weights to older predictions and producing smooth final estimates as weighted averages.

### A. Coupled Online Action Planning

At deployment, the model operates within an online planning framework that couples semantic and geometric outputs into a single decision. High-level actions are executed via parameterizable ProMPs [8]—one per action type—learned over 6D task-space poses and gripper openings from teleoperated demonstrations. After each primitive, the planner selects the next action-object pair as follows. Decoder outputs yield categorical action and object probabilities, symbolic precondition mask removes infeasible pairs, and for the top-$k$ candidates a trajectory score is computed: model motion predictions are rolled out and compared to the corresponding ProMP via Mahalanobis distance, with an exponential penalty for large deviations from the current state. The final score is the product of action-object probability and trajectory score, coupling *what* to do with geometric consistency. Candidates with plausible symbolic but inconsistent motion predictions are down-weighted, and vice versa. A final workspace feasibility check is applied before execution. Left- and right-hand actions are scored independently to handle asynchronous bimanual boundaries.

## IV. Experiments and Results

We evaluate whether semantic-geometric graph representations learned from human demonstrations capture long-horizon task structure and generalize across tasks, subjects, and embodiments. Specifically, we assess (i) the benefit of structured relational inductive biases, and (ii) transfer to a physical bimanual system.

**Datasets.** We use KIT(Bimacs) [7]: RGB-D recordings of five cooking and four workshop tasks by six subjects (10 trials each), with hand actions, 3D object bounding boxes, and object relations. We annotate action-object labels, remove background detections, smooth trajectories, standardize demonstrations, and augment via mirroring and temporal resampling. We collect Ours(Bimacs): four tabletop tasks (*cooking*, *wiping*, *clear up*, *take out*), each performed 10 times by four subjects.

**Baselines.** We compare our GNN encoder to four baselines sharing the same decoder. Dreher [7] is an MPNN with semantic edges but no geometry. Lagamtzis [5] is an RGCN using object IDs and geometry but no edge or global features. A Transformer encoder [25] operates on sequential node features without graph structure and a Decoder-Only model omits the encoder. Dreher and Lagamtzis isolate the role of semantic and geometric features, while the Transformer and Decoder-Only baselines evaluate explicit graph structure impact.

### A. Prediction Results on Human Demonstrations

Figure 3 reports action accuracy and accuracy, and motion RMSE across models trained on relevant tasks from KIT(Bimacs), Ours(Bimacs), and a mixed setting, using cross-validation averaged over 4 seeds.

**KIT(Bimacs).** For semantic next and future action predictions, MPNN and Transformer achieve the highest accuracies, while Decoder-Only underperforms on complex tasks, revealing the limits of sequence memorization alone. Dreher [7] and Lagamtzis [5] perform poorly on semantic predictions, confirming the need to combine geometric and semantic information. For horizon action prediction, differences are smaller as models can achieve high accuracy by repeating the current action; MPNN nonetheless benefits most under high action and

**TABLE I:** Cross-Validation Results of training the MPNN (Ours) model on robot demonstrations, evaluating mixed models trained on human demonstrations on the robot dataset, and finetuning those models on robot data.

| | Method | Accuracy (↑) | | | | | | RMSE [m] (↓) |
|---|---|---|---|---|---|---|---|---|
| | | $a_{t+1}$ | $a_F$ | $a_{t+1:t+P}$ | $o_{t+1}$ | $o_F$ | $o_{t+1:t+P}$ | |
| Cooking | Robot | 0.7460±0.0501 | 0.6594±0.0362 | 0.6443±0.0453 | 0.8452±0.0317 | 0.8538±0.0087 | 0.7840±0.0148 | 0.0547±0.0013 |
| | Not Finetuned | 0.9279±0.0235 | 0.9253±0.0075 | 0.7661±0.0108 | 0.8044±0.0106 | 0.9185±0.0057 | 0.7950±0.0062 | 0.0304±0.0002 |
| | Finetuned | **0.9454±0.0249** | **0.9635±0.0165** | **0.8815±0.0166** | **0.8878±0.0107** | **0.9459±0.0146** | **0.8664±0.0059** | **0.0212±0.0009** |
| Clear Up | Robot | 0.9213±0.0554 | 0.8610±0.0889 | 0.6784±0.0797 | 0.5358±0.0792 | 0.2830±0.0596 | 0.2052±0.0612 | 0.0352±0.0088 |
| | Not Finetuned | 0.9515±0.0091 | 0.8873±0.0188 | 0.7434±0.0210 | 0.6671±0.0401 | **0.4508±0.0232** | 0.4870±0.0432 | 0.0352±0.0021 |
| | Finetuned | **0.9636±0.0105** | **0.9407±0.0213** | **0.8347±0.0231** | **0.6922±0.0367** | 0.4278±0.0282 | **0.5573±0.0197** | **0.0235±0.0010** |

**TABLE II:** Results of online action planning on the real-robot *cooking* and *clear up* tasks with finetuned models. The table shows the overall success rates, the planner infeasibility rates, and motion prediction RMSE.

| | Method | Success Rate (↑) | Infeasibility Rate (↓) | RMSE [m] (↓) |
|---|---|---|---|---|
| Cooking | Dreher [7] | **1.0000±0.0000** | 0.0429±0.1020 | 0.1569±0.0229 |
| | Lagamtzis [5] | 0.7250±0.1750 | 0.0890±0.1741 | 0.0854±0.0036 |
| | Transformer | **1.0000±0.0000** | **0.0000±0.0000** | 0.0509±0.0043 |
| | Decoder-Only | **1.0000±0.0000** | **0.0000±0.0000** | 0.0544±0.0022 |
| | VLM | 0.0000±0.0000 | 1.0000±0.0000 | 88.1708±8.4930 |
| | MPNN (Ours) | **1.0000±0.0000** | **0.0000±0.0000** | **0.0469±0.0022** |
| Clear Up | Dreher [7] | 0.7990±0.2481 | 0.4449±0.2206 | 0.1681±0.0236 |
| | Lagamtzis [5] | 0.3320±0.3430 | 0.6605±0.2604 | 0.1873±0.0169 |
| | Transformer | 0.7320±0.2265 | 0.3899±0.2673 | 0.0622±0.0058 |
| | Decoder-Only | 0.7390±0.2165 | 0.3298±0.3184 | 0.0540±0.0046 |
| | VLM | 0.0000±0.0000 | 1.0000±0.0000 | 18.3042±14.5292 |
| | MPNN (Ours) | **1.0000±0.0000** | **0.1079±0.1353** | **0.0440±0.0039** |

object variability tasks (e.g. *wiping*). MPNN achieves the best motion prediction across all tasks; Dreher [7] underperforms here due to lacking geometric features.

**Ours(Bimacs).** Results on *cooking* and *wiping* mirror KIT(Bimacs) findings. On the high object-variability tasks (*clear up*, *take out*), MPNN achieves the highest object prediction accuracy, with Decoder-Only clearly underperforming, confirming the benefit of reasoning over past object motions.

**Multi-task and Mixed(Bimacs).** Multi-task models match single-task observations, demonstrating that the MPNN encoder generalizes across tasks. Training on Mixed(Bimacs) — combining KIT(Bimacs) and Ours(Bimacs) demonstrations of *cooking* and *wiping* — does not decrease MPNN accuracy and improves most objectives, suggesting the representation is robust to environment variation.

### *B. Transferring Representations for Online Action Planning*

We evaluate transfer of learned task representations to a bimanual robot within an online action planning framework [8] (Section III-A). We consider *cooking* (KIT(Bimacs)) and the high object-variability *clear up* task (Ours(Bimacs)). Models pretrained on human demonstrations are evaluated before and after decoder-only finetuning on small robot datasets (10 demos for *cooking*, 15 for *clear up*), and compared to robot-only training (Table I). Pretraining on human data substantially improves accuracy and reduces RMSE over robot-only training, demonstrating effective transfer. Finetuning yields further gains, likely from embodiment and motion differences. The effect is strongest on *clear up*, where human pretraining significantly improves object prediction under high variability. All real-robot experiments therefore use finetuned models.

**Online Action Planning.** We integrate all models into the online planner (Section III-A, $k = 4$) and evaluate 10 trials per task. Metrics include sub-task completion, motion RMSE, and infeasibility rate (fraction of planner queries rejected by downstream feasibility checks, triggering requery). Trials fail after 10 consecutive rejections. We compare to a finetuned *Florence-2-Base* [27] VLM on the same robot dataset, which takes RGB input, action history, task description, and object positions. On *cooking*, all models except Lagamtzis [5] achieve full success, reflecting its low variability; our model attains the lowest RMSE. On the more challenging *clear up* task, only our model achieves full success, with the lowest infeasibility rate and RMSE. Transformer and Decoder-Only baselines produce more infeasible (geometrically unsafe) plans, while the VLM fails entirely, unable to adapt actions mid-execution and yielding high motion error.

## V. Discussion

Results on human demonstrations show that semantic-geometric graph representations improve long-horizon task modeling, particularly for motion forecasting and object-centric reasoning. The MPNN encoder benefits from jointly encoding semantic relations and geometric evolution, while weaker relational inductive biases struggle under high action-object variability, where sequence-based reasoning alone is insufficient. This highlights task variability as a key factor: simpler architectures suffice for low-variability settings, but explicit semantic-geometric structure becomes critical as interaction complexity increases. This transfers to robot experiments, where learned representations transfer (with finetuning) to a bimanual robot, with our MPNN the only model achieving full success on the high-variability *clear up* task.

## VI. Conclusion

We introduced a semantic-geometric task graph representation unifying object identities, inter-object relations, and motion histories, with a decoupled GNN encoder-decoder separating scene representation from action-conditioned forecasting. This design enables encoder reuse via lightweight decoder finetuning and supports transfer from human demonstrations to robot execution. A downstream planner combines semantic action-object predictions with geometric predictions scored against learned ProMPs for online action planning. Across eleven tasks, our approach outperforms GNN ablations [5], [7], Transformers, Decoder-Only models, and a finetuned VLM, with gains increasing under higher action-object variability.

Limitations and future work include extending the representation to full 6D object poses [28], integrating VLMs as perceptual front-ends, and coupling semantic-geometric predictions with vision-language-action policies to guide low-level generative control. Such integration could provide a structured, interpretable, and data-efficient substrate for complex bimanual manipulation.

## REFERENCES


[1] I. Armeni *et al.*, "3d scene graph: A structure for unified semantics, 3d space, and camera," in *IEEE/CVF international conference on computer vision*, 2019.

[2] P. Gay *et al.*, "Visual graphs from motion (vgfm): Scene understanding with object geometry reasoning," in *Asian Conference on Computer Vision*, 2018.

[3] Z. Wu *et al.*, "A comprehensive survey on graph neural networks," *IEEE transactions on neural networks and learning systems*, 2020.

[4] M. M. Bronstein *et al.*, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," *preprint arXiv:2104.13478*, 2021.

[5] D. Lagamtzis *et al.*, "Graph neural networks for joint action recognition, prediction and motion forecasting for industrial human-robot collaboration," in *56th International Symposium on Robotics*, 2023.

[6] H. Razali *et al.*, "Action-conditioned generation of bimanual object manipulation sequences," in *AAAI conference on artificial intelligence*, 2023.

[7] C. R. Dreher *et al.*, "Learning object-action relations from bimanual human demonstration using graph networks," *IEEE Robotics and Automation Letters*, 2019.

[8] A. Paraschos *et al.*, "Probabilistic movement primitives," *Advances in neural information processing systems*, 2013.

[9] B. Xiaohan Nie, C. Xiong, and S.-C. Zhu, "Joint action recognition and pose estimation from video," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] Q. Li *et al.*, "Directed acyclic graph neural network for human motion prediction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.

[11] M. Li *et al.*, "Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction." IEEE, 2021.

[12] J. Tao *et al.*, "Scene-perception graph convolutional networks for human action prediction," in *International Joint Conference on Neural Networks*, 2021.

[13] U.-H. Kim *et al.*, "3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents," *IEEE transactions on cybernetics*, 2019.

[14] J. Wald *et al.*, "Learning 3d semantic scene graphs from 3d indoor reconstructions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[15] R. Morais *et al.*, "Learning asynchronous and sparse human-object interaction in videos," in *IEEE/CVF conference on computer vision and pattern recognition*, 2021.

[16] J. Wu *et al.*, "Hiergat: hierarchical spatial-temporal network with graph and transformer for video hoi detection," *Multimedia Systems*, 2025.

[17] C. R. Garrett *et al.*, "Integrated task and motion planning," *Annual Review of Control, Robotics, and Autonomous Systems*, 2021.

[18] M. Ahn *et al.*, "Do as i can, not as i say: Grounding language in robotic affordances," in *Conference on Robot Learning (CoRL)*, 2022.

[19] J. Liang *et al.*, "Code as policies: Language model programs for embodied control," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 9493–9500.

[20] M. J. Kim *et al.*, "OpenVLA: An open-source vision-language-action model," in *Conference on Robot Learning (CoRL)*, 2024.

[21] K. Black *et al.*, "$\pi_0$: A vision-language-action flow model for general robot control," in *Robotics: Science and Systems (RSS)*, 2025.

[22] P. W. Battaglia *et al.*, "Relational inductive biases, deep learning, and graph networks," *preprint arXiv:1806.01261*, 2018.

[23] F. Ziaeetabar *et al.*, "Recognition and prediction of manipulation actions using enriched semantic event chains," *Robotics and Autonomous Systems*, 2018.

[24] J. Su *et al.*, "Roformer: Enhanced transformer with rotary position embedding," *Neurocomputing*, 2024.

[25] A. Vaswani *et al.*, "Attention is all you need," *Advances in neural information processing systems*, 2017.

[26] T. Z. Zhao *et al.*, "Learning fine-grained bimanual manipulation with low-cost hardware," *Robotics: Science and Systems*, 2023.

[27] B. Xiao *et al.*, "Florence-2: Advancing a unified representation for a variety of vision tasks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[28] J. Brandstetter *et al.*, "Geometric and physical quantities improve e (3) equivariant message passing," in *International Conference on Learning Representations*, 2022.