PatchAlign: Fine-Grained Semantic–Geometric Alignment for Robotic Manipulation

Yanglan Dong[1,2], Jiaming Guo[2*], Yunkai Gao[3], Siming Lan[3], Shaohui Peng[4], Zihao Zhang[3], Rui Zhang[2], Xing Hu[2]
[1]University of Science and Technology of China, Hefei, China
[2]State Key Lab of Processors, Institute of Computing Technology, CAS, Beijing, China
[3]Institute of AI for Industries, Nanjing, China
[4]Intelligent Software Research Center, Institute of Software, CAS, Beijing, China

***Abstract*—Robot manipulation policies that generate actions from visual observations and language instructions have achieved promising progress, yet robust scene understanding remains a major challenge, especially for precise manipulation tasks. Methods relying solely on 2D visual representations capture high-level semantics but lack spatial awareness, while existing 3D-enhanced approaches fail to preserve fine-grained geometric details and local semantic–geometric correspondences. This limitation suggests that effectively combining semantic information from pretrained visual encoders with detailed local 3D geometry is crucial for enabling precise and reliable manipulation. To address these limitations, we propose PatchAlign, a patch-level cross-modal fusion framework that explicitly aligns RGB image patches with local 3D geometric regions. Semantically rich image patches are extracted using a frozen pretrained visual encoder, while 3D geometry is captured via a lightweight clustering strategy. By establishing patch-level semantic–geometric correspondences, our method enables precise local reasoning without compromising the generalization of pretrained vision-language models. We evaluate our method on multiple manipulation benchmarks across diverse tasks. Experimental results demonstrate that the proposed framework consistently improves manipulation performance and robustness, particularly in scenarios requiring fine-grained spatial reasoning and precise object interactions. We believe this work presents an effective framework for fine-grained cross-modal fusion, offering new insights into leveraging 3D geometry in vision-language-action learning.**

***Index Terms*—Vision-Language-Action Models, Robotic Manipulation, 3D Scene Representation**

I. Introduction

Integrating visual observations with language instructions for action generation has emerged as a promising approach for building multi-task robotic manipulation policies. Recent methods [1]–[7] have achieved significant progress by leveraging large-scale pretrained visual encoders. However, while these 2D-centric representations excel at capturing high-level semantic information for instruction following, they inherently lack the precise depth and spatial awareness required for fine-grained object interaction. This deficiency in capturing 3D geometric nuances often leads to performance bottlenecks in complex manipulation tasks that demand sub-centimeter accuracy. Consequently, effectively bridging semantically rich 2D features with explicit 3D geometric structure is essential for enabling robust and precise robotic manipulation.

Building upon this motivation, recent research on manipulation policies has explored various strategies to integrate 3D geometric information. One line of work preserves pretrained visual encoders by projecting 3D information into two-dimensional representations and predicting actions via 2D heatmaps [8], but the indirect use of geometry limits their ability to fully exploit spatial structure. Other methods inject 3D positional encodings into vision-language backbones [9]–[11], enabling spatial awareness without major architectural changes, yet often operating at a coarse level and sacrificing fine-grained geometric detail. Another line of research treats 3D data as an additional modality [12], [13], typically integrating global point cloud features through feature fusion or direct injection into the action expert. While these approaches incorporate explicit 3D cues, they fail to capture fine-grained local geometry and precise semantic–geometric correspondences, motivating the question: how can we exploit fine-grained 3D structure while maintaining alignment with pretrained 2D semantic representations?

We observe that effective integration of RGB and 3D modalities hinges on the spatial granularity at which semantic representations are organized. Pretrained visual encoders naturally structure semantic information at the image-patch level, where each patch serves as a minimal and meaningful unit for semantic abstraction. This suggests that patch-level representations provide a natural interface for semantic–geometric fusion.

Guided by this insight, we introduce PatchAlign, a patch-level cross-modal fusion strategy that aligns RGB semantics and 3D geometry at a shared spatial granularity. PatchAlign leverages a frozen pretrained visual encoder to extract semantically rich image patch features, while a point cloud encoder captures detailed three-dimensional geometric representations through localized regions. By explicitly associating each image patch with its corresponding local 3D geometric features, the proposed approach enables precise semantic–geometric interaction at the patch level, without compromising the generalization capabilities of pretrained vision-language models. As a result, PatchAlign overcomes the limitations of global

* Corresponding author

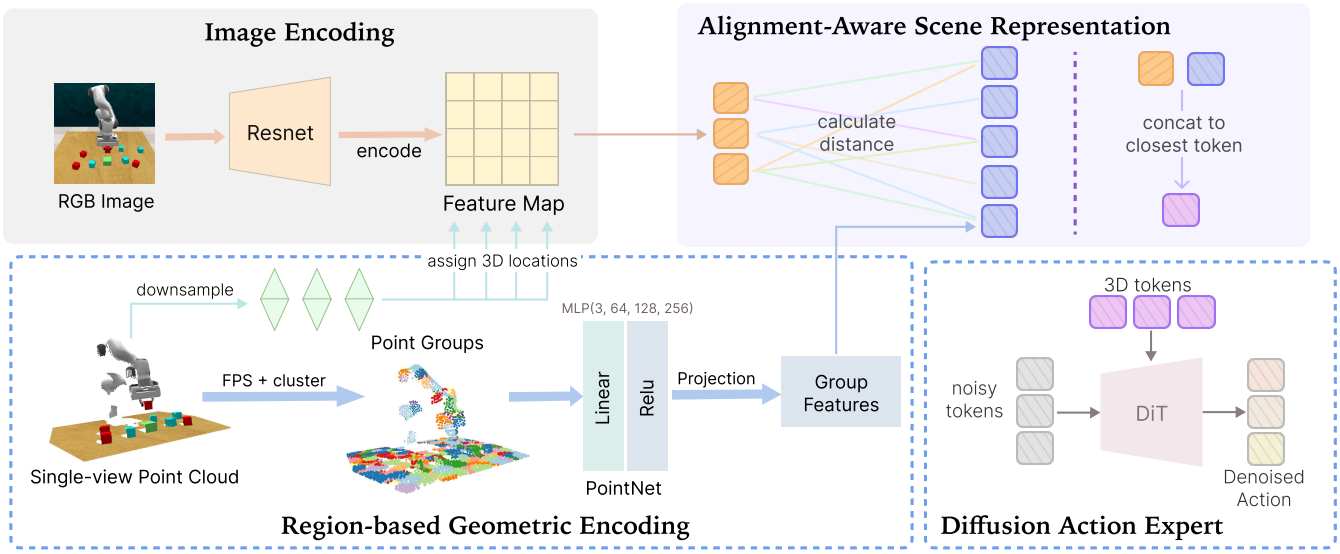

Fig. 1: Network Architecture. RGB images and 3D point clouds are encoded via CLIP and PointNet, with point clouds processed through KNN-based local grouping and spatially aligned with image features via nearest-neighbor fusion, producing multi-modal scene tokens that are fed into a diffusion transformer for action prediction.

or coarse-grained fusion and provides a principled mechanism for integrating 2D semantics with fine-grained 3D geometry in manipulation policies.

We evaluate the proposed approach across multiple simulated environments. On the multi-task RLBench [14] benchmark, our method achieves an average success rate of 88%, yielding a 10% absolute improvement over a strong 3D diffusion-based baseline [11], demonstrating its effectiveness in complex multi-task scenarios. On the CALVIN [15] benchmark, our method consistently outperforms the baseline and improves the average success rate by 11.3%, highlighting its robustness across diverse manipulation settings. Extensive ablation studies further confirm the effectiveness of our design, showing that both geometric clustering and alignment-aware fusion are essential to the final performance. These results indicate that fine-grained alignment leads to more reliable and spatially grounded action prediction.

II. Related Work

A. Robot Manipulation Policies

Early works like CLIPort [16] leveraged pretrained semantic priors for pick-and-place tasks. Subsequent studies, such as Octo [6], scaled performance by training diffusion policies on massive robot datasets. While powerful, these data-driven approaches incur high costs and rely on extensive robot trajectories. Recent methods incorporate large VLMs as perceptual backbones [2], [17]–[20] to enable open-vocabulary reasoning. However, they remain heavily dependent on large-scale pretraining and extensive fine-tuning. Overall, existing vision-language-conditioned policies typically rely on large-scale pretraining or massive robot datasets and often struggle in data-limited settings. Motivated by this limitation, we focus on improving learning efficiency by developing more effective visual–geometric fusion strategies.

B. 3D Scene Representations for Robotic Manipulation

Early robot manipulation policies relied on 2D images for action prediction, limiting their geometric reasoning ability. To incorporate explicit 3D structure, various approaches have been explored. For example, PerAct [21] discretizes space into 3D grids using voxel representations, while Act3D [22] leverages 3D query points for focus. RVT [23] instead adopts multi-view RGB-D reprojection to infer 3D poses. 3D Diffusion Policy [24] processes point clouds with a PointNet-based encoder and demonstrates strong multi-task performance, but lacks rich semantic priors from large-scale RGB pre-training. To address this limitation, recent methods integrate 3D geometry into pre-trained vision-language models. For instance, LLaVA3D [9] augments a VLM backbone with 3D positional encodings, which often require substantial retraining on 3D embodied data to realign visual-language representations.

C. Point Cloud Integration in Action Policy Learning

Instead of modifying the visual backbone, some works treat point clouds as a new modality and focus on fusion techniques. PointVLA [12] injects point cloud features into a frozen action expert. This preserves the expert's original performance but limits adaptation to 3D geometry. Other methods such as GeoVLA [13] encode RGB and point clouds separately and fuse global features. However, they often lack fine-grained spatial alignment, struggling with complex scenes precise manipulation. In summary, existing approaches lack a mechanism to integrate 3D geometry with 2D semantics at a sufficiently detailed level. In contrast, our model introduces a patch-level fusion method which enables a more precise and spatially-grounded reasoning for manipulation tasks.

III. METHODOLOGY

A. Policy Architecture

We consider a diffusion-based visuomotor policy architecture for action generation. As illustrated in Figure 1, the policy consists of two main components: a multi-modal tokenization stage that converts heterogeneous inputs into a unified token sequence, and an action expert that predicts robotic actions through an iterative denoising process. Specifically, the model takes two types of visual observations as input: RGB images and point clouds. These observations, along with gripper history, are tokenized to produce tokenized representations, while language instructions are encoded using a pretrained language encoder. The resulting tokens, together with a noisy trajectory, are fed into a diffusion transformer, which integrates all modalities through cross-attention layers to predict the robot's actions. The transformer predicts the noise component added to the trajectory during the forward diffusion process. The denoised tokens are aggregated via self-attention heads and mapped to an 8-dimensional action vector, consisting of a 7-DoF end-effector pose and a 1-dimensional gripper openness value.

B. Patch-Level Semantic-Geometric Alignment

Although modern visuomotor policies are trained end-to-end on robotic datasets, their semantic understanding largely stems from pretrained visual and language encoders. These encoders usually provide strong semantic priors but lack explicit 3D structural reasoning.

A common strategy is to encode the entire point cloud into a single global feature and inject it into the policy network [12], [13]. While widely adopted, this approach has inherent limitations. First, aggressive spatial pooling discards local geometric variations critical for object-level interaction. Second, global features lack explicit spatial correspondence with image representations, hindering fine-grained semantic–geometric alignment for precise manipulation.

Inspired by recent advances in 3D representation learning [25], which structure point clouds into localized geometric groups and encode them as region-level tokens, we argue that 3D information should be incorporated in a spatially localized and alignment-aware manner. Specifically, we represent point clouds as local geometric regions and align them with patch-level visual features, yielding a unified representation that preserves semantic priors from pretrained 2D models while enhancing fine-grained geometric awareness.

C. Region-Based Geometric Encoding

A key challenge in incorporating 3D point clouds lies in how to represent raw geometry in a form that is both informative for manipulation and compatible with 2D-centric semantic encoders. To address this issue, we construct a set of localized geometric regions that explicitly preserve fine-grained spatial structure while remaining lightweight and scalable.

Specifically, given an input point cloud $P \in \mathbb{R}^{N\times 3}$, we first apply downsample it to N' points, ensuring uniform spatial coverage of the scene. From the downsampled set, we use Farthest Point Sampling (FPS) to select K center points ${c_k}_{k=1}^{K}$, each serving as an anchor for a local geometric region. For each center c_k, we retrieve its M nearest neighbors using K-Nearest Neighbors (KNN), forming a localized point group

$$\mathcal{N}_k = \{p_1^{(k)}, p_2^{(k)}, \ldots, p_M^{(k)}\}.$$

Each local group is then encoded independently using a lightweight PointNet to produce a region-level geometric feature $f_k^{\text{pc}} \in \mathbb{R}^{d_g}$. Compared to global point cloud descriptors, this region-based representation retains local geometric details such as surface orientation, spatial extent, and relative layout, which are essential for precise object manipulation.

D. Alignment-Aware Scene Representation

While pretrained visual encoders provide strong semantic priors, their outputs are inherently spatial and patch-based. Therefore, effective multi-modal fusion requires establishing correspondence between local image features and localized 3D geometry.

We employ a CLIP [26] pre-trained ResNet-50 [27] as the visual feature extractor. Given an input image $I \in \mathbb{R}^{256\times 256\times 3}$, the encoder outputs a spatial feature map $F^{\text{img}} \in \mathbb{R}^{32\times 32\times d_v}$, where d_v is the feature dimension. During training, we freeze all parameters of the visual encoder to preserve its pretrained semantic priors and ensure stable cross-modal alignment.

We perform multi-modal fusion via nearest-neighbor feature concatenation, which has been shown to be more effective than attention-based alternatives in similar settings [28]. To establish spatial correspondence, the 3D point cloud is spatially aligned with the image feature map through resolution matching. Specifically, the point cloud is downsampled and associated with the spatial resolution of the visual feature map, such that each image patch feature at location (i, j) is assigned a corresponding 3D coordinate $p_{ij} \in \mathbb{R}^3$. For each p_{ij}, We compute the Euclidean distance to all point cloud centers c_k and concatenate the image feature with the nearest geometric feature:

$$t_{ij} = \left[F_{ij}^{\text{img}} \parallel f_{k^*}^{\text{pc}}\right], \quad k^* = \arg\min_k \|p_{ij} - c_k\|_2,$$

where $\parallel$ denotes concatenation.

This yields a set of spatially aligned scene tokens ${t_{ij}}_{i,j=1}^{32}$, which serve as structured inputs to the downstream diffusion-based action model.

E. Implementation Details

To effectively model multi-modal visuomotor policies, we adopt a diffusion-based framework that has shown strong performance in 3D manipulation tasks [11]. Building on this foundation, we extend the original architecture by replacing the RGB tokens with our proposed scene representation, where each token encodes patch-level semantic information and spatially aligned local 3D geometry. This modification enables the policy to incorporate fine-grained geometric cues directly into the decision-making process, enhancing its spatial reasoning and action precision.

TABLE I: Evaluation on RLBench. Success rates (%) for different manipulation tasks.

Method	Success rate (%)										Avg. (%)
	close jar	open drawer	sweep to dustpan	meat off grill	turn tap	slide block	put in drawer	drag stick	push buttons	stack blocks	
GNFactor	25.3	76.0	28.0	57.3	50.7	20.0	0.0	37.3	18.7	4.0	31.7
Act3D	52.0±5.7	84.0±8.6	80.0±9.8	66.7±1.9	64.0±5.7	**100.0±0.0**	54.7±3.8	86.7±1.9	64.0±1.9	0.0±0.0	65.3
3DDA	82.7±1.9	89.3±7.5	94.7±1.9	88.0±5.7	80.0±8.6	92.0±0.0	77.3±3.8	**98.7±1.9**	69.3±5.0	12.0±3.7	78.4
Ours	**93.3±4.6**	**98.7±2.3**	**98.7±2.3**	**96.0±4.0**	**84.0±8.0**	84.0±10.6	**97.3±4.6**	94.7±9.2	**94.7±2.3**	**46.0±4.0**	**88.7**

TABLE II: Evaluation on CALVIN. Success rates (%) for different manipulation tasks.

Method	Success rate (%)						Avg. (%)
	lift slider	open drawer	move slider right	stack block	rotate block	push block	
3DDA	**36**	56	84	4	12	**44**	39.3
Ours	32	**80**	**100**	**28**	**24**	40	**50.6**

TABLE III: Ablation Study. We ablate key design choices in both *point cloud representation* and *fusion method*.

Method	Point Cloud Representation		Fusion Method			Avg. (%)
	Clustering	Patching	Neighboring Fusion	Global Tokenization	Global Injection	
Ex0		✓	✓			85.3 (↓**3.4**)
Ex1	✓			✓		78.2 (↓**10.5**)
Ex2	✓				✓	80.6 (↓**8.1**)
Ours	✓		✓			88.7

For geometric region construction, we adopt a lightweight clustering strategy that balances representational fidelity and computational efficiency. Unlike object-level point clouds commonly used in 3D recognition, our input consists of surface observations captured from a single viewpoint, without access to interior geometry. Following prior point cloud processing methods [29], we first downsample the point cloud. We then select region centers with a stride of 4, and construct local neighborhoods by grouping the 32 nearest points for each center. This configuration provides sufficient local geometric context while keeping computation manageable, leading to stable training and strong performance across tasks, as validated in our ablation studies.

The model is trained on four NVIDIA A100 GPUs using the Adam optimizer with a learning rate of 1×10^{-4} and a batch size of 160. This configuration provides stable optimization and is used consistently across all experiments.

IV. Experiments

A. Evaluation on RLBench

We evaluate our method in RLBench [14], a widely used robotic benchmark built on the CoppeliaSim [30] simulator and manipulated with a Franka Panda robotic arm. In this environment, the policy predicts the next key end-effector pose. This predicted pose is then passed to the built-in BiRRT [31] motion planner for trajectory generation and execution.

Baseline. We adopt 3D Diffuser Actor as the primary baseline, which is a representative visuomotor policy that integrates 3D representations with diffusion models. We also report the success rates of GNFactor [32] and Act3D [22]. The reported baseline results are obtained from their respective original papers.

Settings. All methods, including the baseline, are trained in a multi-task learning setup. To specifically assess the effectiveness of point cloud utilization, we restrict the input to a single forward-facing RGB-D camera. We follow the single-view benchmark protocol introduced in 3D Diffuser Actor and adopt the same set of ten manipulation tasks for both training and evaluation. For each task, we use 100 expert demonstrations for training and report the average success rate on 25 evaluation trials. The overall performance is summarized by the mean success rate in all 10 tasks, enabling a comprehensive comparison across methods.

Results. As shown in Table I, our method achieves the highest average success rate of 88.7% on RLBench under the single-view RGB-D setting, surpassing the best baseline (3DDA) by 10.3% and demonstrating clear improvements over Act3D (65.3%) and GNFactor (31.7%). Performance gains are evident in tasks that require precise geometric reasoning and accurate spatial alignment, where leveraging detailed 3D structure is critical for successful execution. These results demonstrate the importance of explicitly encoded, spatially aligned 3D geometry in precise manipulation.

B. Evaluation on Calvin

We also evaluate our method on the CALVIN [15] benchmark, a simulated tabletop manipulation environment built upon the PyBullet [33] physics engine. In CALVIN, since no

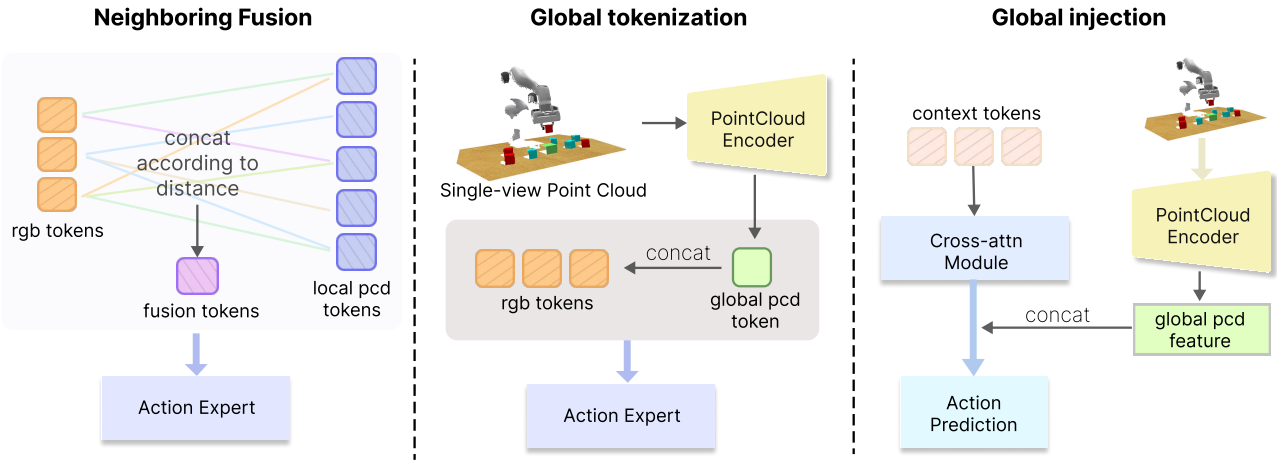

Fig. 2: Different point cloud fusion strategies. We compare two global feature integration methods against our fusion method.

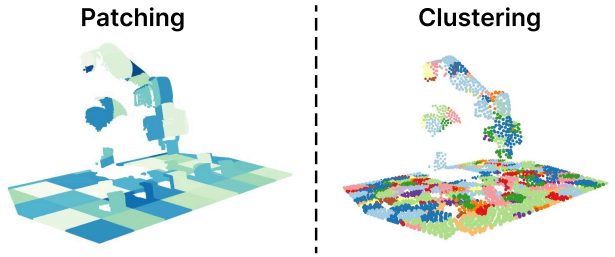

Fig. 3: Clustering and patch-based pointcloud representation.

motion planner is provided, we directly predict end-effector pose trajectories from inputs.

Settings. To reduce computational cost, we select 6 distinct tasks from CALVIN for multi-task training and evaluation. We use the labeled demonstration data provided in the CALVIN dataset, with 150 expert trajectories per task, and report the success rate for 25 evaluation trials.

Results. As shown in Table II, our method achieves an overall success rate of 50.6%, surpassing the 3DDA baseline by 11.3%. Our method consistently outperforms the 3DDA baseline on most of the evaluated tasks, and demonstrates substantial advancements in those tasks that require more precise geometric reasoning, such as *stack_block* and *rotate_block*. These results suggest that patch-level semantic-geometric alignment contributes to more robust and accurate manipulation.

C. Ablation Study

We conduct a comprehensive ablation study to dissect the contribution of our proposed point cloud representation and fusion strategy. Since our method is built upon 3D Diffuser Actor (3DDA), the variant without point cloud fusion is equivalent to the original 3DDA and is treated as the baseline.

Point Cloud Representation. We first ablate how point clouds are structured and encoded within the network. Beyond our cluster-based approach, we evaluate the following:

Ex0 Patch-based incorporation. As illustrated in Fig. 3, instead of geometry-aware clustering, we adopt fixed grid partitioning aligned with the image plane. The point cloud is segmented into patches that correspond directly to the 8×8 pixel regions of the RGB feature map, treating 3D geometry as a set of spatially fixed 2.5D windows.

Fusion Strategy. Given a structured representation, we next ablate how to integrate it into the policy. We use two different approaches, which is shown in Fig. 2.

Ex1 Global tokenization. We extract a global point cloud feature via max-pooling, assign it with the 3D position of the end-effector, and concatenate it with other input tokens as an additional 3D token.

Ex2 Global injection. The global point cloud feature is concatenated with the intermediate features before the self-attention layers within the action expert's diffusion transformer.

Results. Table III presents the ablation study. Our patch-based approach shows clear advantages over the baseline, highlighting the value of incorporating 3D geometry. However, fixed-grid partitioning can produce mixed patches spanning multiple objects, which is suboptimal for PointNet encoders relying on local structure.

For feature fusion, simply appending global 3D features (Ex1) offers limited improvement, while injecting features into the action expert (Ex2) improves geometric awareness but can hurt visual-semantic reasoning, as seen in tasks like *slide_blocks*. These results suggest that effective 3D integration requires not only adding geometry but also preserving alignment with existing visual features.

V. Conclusion

In this work, we present PatchAlign, a unified framework that integrates point cloud representations with RGB features for robust 3D perception and decision making. Unlike prior methods relying on global feature fusion or 2D projections, our approach achieves fine-grained semantic-geometric alignment

by establishing correspondences between image patch features and locally clustered point cloud features.

Extensive experiments on both RLBench and CALVIN benchmarks demonstrate that our method consistently outperforms strong 3D diffusion baselines, achieving significant improvements in tasks requiring precise spatial reasoning. Through additional ablation studies, we show that our method achieves better results than its ablated alternatives, validating the effectiveness of jointly optimizing representation and fusion strategies. This demonstrates that fine-grained 3D-2D alignment is a key enabler for robust and precise visuomotor control in language-conditioned tasks.

Looking ahead, PatchAlign opens avenues for multi-view 3D observations and real-world robotic deployment, bridging semantic understanding and geometric reasoning for robust manipulation in open-world environments.

ACKNOWLEDGMENT

This work is partially supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grants No.XDB0660300), NSF of China (Grants No.62302481, U22A2028), CAS Project for Young Scientists in Basic Research (YSBR-029) and Youth Innovation Promotion Association CAS.

REFERENCES

[1] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn *et al.*, "Rt-1: Robotics transformer for real-world control at scale," in *Proceedings of Robotics: Science and Systems*, 2023, pp. 1–11.

[2] B. Zitkovich, T. Yu, S. Xu, P. Xu, T. Xiao, F. Xia, J. Wu, P. Wohlhart, S. Welker, A. Wahid *et al.*, "Rt-2: Vision-language-action models transfer web knowledge to robotic control," in *Conference on Robot Learning*. PMLR, 2023, pp. 2165–2183.

[3] S. Reed, K. Żołna, E. Parisotto, S. G. Colmenarejo, A. S. Novikov, G. Barth-Maron *et al.*, "A generalist agent," *Transactions on Machine Learning Research*, pp. 28 091–28 114, 2022.

[4] E. Jang, A. Irpan, M. Khansari, D. Kappler, F. Ebert, C. Lynch, S. Levine, and C. Finn, "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*, 2022, pp. 991–1002.

[5] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, and A. Jain, "Open x-embodiment: Robotic learning datasets and rt-x models," in *Proceedings of the IEEE International Conference on Robotics and Automation*, 2024, pp. 6892–6903.

[6] Octo Model Team, D. Ghosh, H. Walke, K. Pertsch, K. Black, O. Mees *et al.*, "Octo: An open-source generalist robot policy," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024.

[7] H. Liu, L. Lee, K. Lee, and P. Abbeel, "Instruction-following agents with jointly pre-trained vision-language models," *arXiv:2210.13431*, 2022.

[8] P. Li, Y. Chen, H. Wu, X. Ma, X. Wu, Y. Huang *et al.*, "Bridgevla: Input-output alignment for efficient 3d manipulation learning with vision-language models," *arXiv:2506.07961*, 2025.

[9] C. Zhu, T. Wang, W. Zhang, J. Pang, and X. Liu, ""llava-3d: A simple yet effective pathway to empowering lmms with 3d capabilities"," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 4295–4305.

[10] D. Qu, H. Song, Q. Chen, Y. Yao, X. Ye, Y. Ding *et al.*, "Spatialvla: Exploring spatial representations for visual-language-action models," *arXiv:2501.15830*, 2025.

[11] T. W. Ke, N. Gkanatsios, and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," in *Conference on Robot Learning*, 2025, pp. 1949–1974.

[12] C. Li, J. Wen, Y. Peng, Y. Peng, F. Feng, and Y. Zhu, "Pointvla: Injecting the 3d world into vision-language-action models," *arXiv:2503.07511*, 2025.

[13] L. Sun, B. Xie, Y. Liu, H. Shi, T. Wang, and J. Cao, "Geovla: Empowering 3d representations in vision-language-action models," *arXiv:2508.09071*, 2025.

[14] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark and learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[15] O. Mees, L. Hermann, E. Rosete-Beas, and W. Burgard, "Calvin: A benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7327–7334, 2022.

[16] M. Shridhar, L. Manuelli, and D. Fox, "Cliport: What and where pathways for robotic manipulation," in *Conference on Robot Learning*, 2022, pp. 894–906.

[17] M. J. Kim, K. Pertsch, S. Karamcheti, T. Xiao, A. Balakrishna, S. Nair *et al.*, "Openvla: An open-source vision-language-action model," in *Conference on Robot Learning*, 2025, pp. 2679–2713.

[18] K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn *et al.*, "π_0: A vision-language-action flow model for general robot control," *arXiv:2410.24164*, 2024.

[19] Q. Li, Y. Liang, Z. Wang, L. Luo, X. Chen, M. Liao *et al.*, "Cogact: A foundational vision-language-action model for synergizing cognition and action in robotic manipulation," *arXiv:2411.19650*, 2024.

[20] X. Shi, B. Ichter, M. Equi, L. Ke, K. Pertsch, Q. Vuong *et al.*, "Hirobot: Open-ended instruction following with hierarchical vision-language-action models," *arXiv:2502.19417*, 2025.

[21] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Conference on Robot Learning*, 2023, pp. 785–799.

[22] T. Gervet, Z. Xian, N. Gkanatsios, and K. Fragkiadaki, "Act3d: 3d feature field transformers for multi-task robotic manipulation," in *Conference on Robot Learning*, 2023, pp. 3949–3965.

[23] A. Goyal, J. Xu, Y. Guo, V. Blukis, Y.-W. Chao, and D. Fox, "Rvt: Robotic view transformer for 3d object manipulation," in *Conference on Robot Learning*, 2023, pp. 694–710.

[24] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," in *Proceedings of Robotics: Science and Systems*, Delft, Netherlands, 2024, pp. 288–297.

[25] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 313–19 322.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, and G. Krueger, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[28] X. Jia, Q. Wang, A. Wang, H. A. Wang, B. Gyenes, E. Gospodinov *et al.*, "Pointmappolicy: Structured point cloud processing for multi-modal imitation learning," *arXiv:2510.20406*, 2025.

[29] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, 2017.

[30] E. Rohmer, S. P. Singh, and M. Freese, "V-rep: A versatile and scalable robot simulation framework," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 1321–1326.

[31] J. J. Kuffner and S. M. LaValle, "Rrt-connect: An efficient approach to single-query path planning," in *IEEE International Conference on Robotics and Automation*, 2000, pp. 995–1001.

[32] Y. Ze, Y. Ge, Y.-H. Wu, A. Macaluso, Y. Ge, J. Ye, N. Hansen, L. E. Li, and X. Wang, "Gnfactor: Multi-task real robot learning with generalizable neural feature fields," in *Conference on Robot Learning*. PMLR, 2023, pp. 284–301.

[33] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.

[34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 652–660.

[35] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *International Conference on Machine Learning*, 2010, pp. 807–814.