2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

VO-DP: Semantic-Geometric Adaptive Diffusion Policy for Vision-Only Robotic Manipulation

Zehao Ni[1,2,5,6*], Yonghao He[1*,†], Lingfeng Qian[1*], Jilei Mao[1], Fa Fu[1], Wei Sui[1], Hu Su[4], Junran Peng[3‡], Zhipeng Wang[2,5,6], Bin He[2,5,6‡]

[1] D-ROBOTICS
[2] National Key Laboratory of Autonomous Intelligent Unmanned Systems
[3] University of Science and Technology Beijing
[4] State Key Laboratory of Multimodal Artificial Intelligence System (MAIS) Institute of Automation of Chinese Academy of Sciences
[5] Frontiers Science Center for Intelligent Autonomous Systems
[6] Shanghai Institute of Intelligent Science and Technology, Tongji University
[*]Equal contribution [†]Project lead [‡]Corresponding author

***Abstract*— In the context of imitation learning, visuomotor-based diffusion policy learning is one of the main directions in robotic manipulation. Most of these approaches rely on point clouds as observation inputs and construct scene representations through point clouds feature learning, which enables them to achieve remarkable accuracy. However, the existing literature lacks an in-depth exploration of vision-only solutions that have significant potential. In this paper, we propose a Vision-Only and single-view Diffusion Policy learning method (VO-DP) that leverages pretrained visual foundation models to achieve effective fusion of semantic and geometric features. We utilize intermediate features from VGGT incorporating semantic features from DINOv2 and geometric features from Alternating Attention blocks. Features are fused via cross-attention and spatially compressed with a CNN to form the input to the policy head. Extensive experiments demonstrate that VO-DP not only outperforms the vision-only baseline DP significantly but also exhibits distinct performance trends against the point cloud-based method DP3: in simulation tasks, VO-DP achieves an average success rate of 64.6%—on par with DP3 64.0% and far higher than DP 34.8%, while in real-world tasks, it reaches 87.9%, outperforming both DP3 67.5% and DP 11.2% by a notable margin. Further robustness evaluations confirm that VO-DP remains highly stable under varying conditions including color, size, background, and lighting. Lastly, we open-source a robotic manipulation training library: it supports multi-machine/multi-GPU parallel training and mixed precision, is compatible with visuomotor policies (DP, DP3) and the RoboTwin simulator, and will be released upon publication.**

I. INTRODUCTION

Visuomotor policy learning has emerged as an important paradigm in robotic manipulation, leveraging visual observations to guide the generation of action sequences in an end-to-end manner. Current mainstream visuomotor methods can be broadly categorized into non-vision-only approaches and vision-only approaches. Vision-only approaches rely on RGB image inputs to achieve joint optimization of perception and action. Such methods depend on implicit 3D scene understanding and closely align with biological perception-action systems. In contrast, non-vision-only approaches rely on explicit 3D representations, such as point clouds (e.g., DP3 [1], 3D Diffuser Actor [2]) or RGB-D images (e.g., SEM [3], H^{3}DP [4]) as inputs to decouple the 3D modeling process. Benefiting from precise low-dimensional 3D representations, these methods have significantly improved the accuracy of robotic manipulation. However, non-vision-only methods heavily depend on high-cost hardware and exhibit limitations in complex scenes. First, acquiring RGB-D or point clouds requires expensive sensors such as depth cameras or LiDAR, and model performance is constrained by sensor accuracy. In comparison, RGB cameras provide significantly lower costs and higher practicality: their hardware costs are reducible by orders of magnitude, and further, system complexity arising

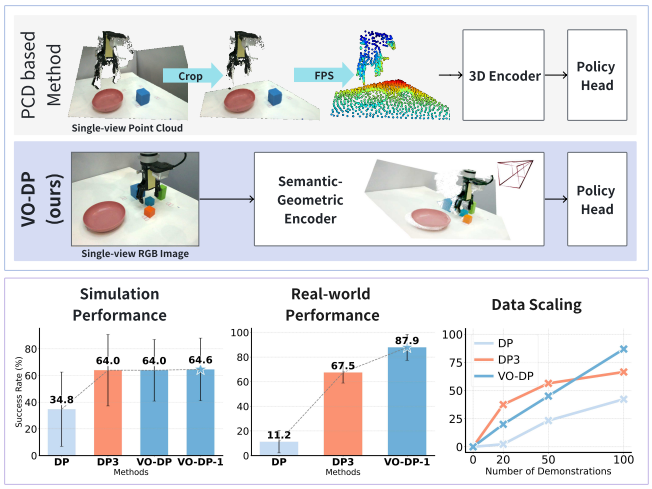

Fig. 1: VO-DP is a vision-only method for visuomotor robotic manipulation: it takes single-view RGB images as input, uses large vision models to extract semantic and geometric features from observations, and provides high-quality conditional inputs for the policy head. Experiments show it matches point cloud-based DP3's accuracy in simulation, outperforms it significantly in real-world tasks, and notably boosts vision-only method accuracy.

from multi-sensor calibration is avoided. Second, experimental results demonstrate that sparse 3D inputs are inadequate for semantic-intensive tasks and complex scenarios, which in turn leads to performance degradation.

We argue that the academic community lacks in-depth research and exploration under vision-only settings, particularly regarding how to learn effective representations for robotic manipulation when relying solely on RGB images. Currently, vision-only methods have not yet demonstrated performance superior to that of point cloud-based methods in robotic manipulation. This is largely attributed to underdeveloped representation learning modules in existing methods. To further unlock the potential of vision-only approaches, we propose VO-DP, a method that integrates and compresses both semantic and geometric features extracted from single-view image as input to a downstream policy head. Specifically, we leverage intermediate-layer features from the pretrained 3D reconstruction model VGGT [5], including semantics-aware features from DINOv2 [6] and geometry-aware features from the Alternating Attention network. We then design a cross-attention-based fusion module to adaptively inject semantics-aware features into geometry-aware features according to task-specific information preferences. Finally, we introduce a spatial feature compression module, which is based on CNN, to distill essential representations from the scene.

In summary, our contributions are four-fold:

- We demonstrate that vision-only visuomotor policies hold substantial performance potential in robotic manipulation, even achieving an accuracy level on par with point cloud-based methods.
- We propose VO-DP, a novel vision-only, single-view representation learning visuomotor method for robotic manipulation that adaptively fuses semantic and geometric information.
- We conduct a detailed evaluation and analysis of the VO-DP method on the RoboTwin 1.0 [7] simulation benchmark and real-world tasks, and it achieves state-of-the-art performance in both simulation and real-world experiments.
- We open-source a general training framework for robotic manipulation. Built on Accelerate[8], it supports multi-node multi-GPU training and multi-GPU evaluation with the RoboTwin simulator, offers mixed-precision training (bf16/fp16), and maintains compatibility with visuomotor methods such as DP and DP3.

II. RELATED WORK

A. Vision-Only Methods

Vision-only methods refer to those approaches that take RGB images as observation inputs. Some typical examples of such methods, including DP[9],ACT[10] and others [11], [9], [12], [10], [13], [14], have demonstrated exceptional performance in robotic manipulation tasks. However, it has been observed that while these methods deliver reasonably good performance in in-distribution scenarios, they show notable sensitivity to environmental variations during real-world deployment. Even minor changes in background, camera pose, or lighting conditions can trigger severe degradation in model performance[15]. Within image-based approaches, the impact of representation learning has received relatively little attention in existing studies. While DP[9], one of our baseline methods, does explore how different backbones influence success rates, its analysis remains confined to conventional image backbones such as ResNet[16] and ViT[17], with no extension to specialized or advanced representation learning architectures. We contend that, for robotic manipulation tasks, learning appropriate representations from RGB image-based inputs plays a pivotal role in enhancing the robustness and generalization capabilities of vision-only models.

Recent years have witnessed the rapid development of visual foundation models[18], [19], [20]. In particular, visual models with spatial perception capabilities, such as VGGT[5], can directly extract geometric information from RGB images, thereby providing rich feature options for vision-only methods. OV-DP leverages the semantic and geometric features provided by VGGT, fully unlocking the potential of vision-only methods without additional preprocessing that 3D-based methods frequently rely on.

B. Non-Vision-Only Methods

Non-vision-only methods refer to those that take 3D signals, such as depth information and point clouds, as observation inputs. PerAct[21] voxelizes point clouds data into tokens, which are then fused with text tokens in a transformer architecture. ACT3D[22] utilizes the CLIP[23] model to extract image features and further aggregates them with depth information to enhance model performance. 3D Diffuser Actor[2] converts 2D image features into 3D tokens using a depth map and employs a denoising network to generate trajectories. RVT[24] projects RGB-D data on three orthogonal planes to form virtual images, which are subsequently used for action prediction. RVT-2[25] optimizes the prediction head by leveraging heatmaps for trajectory generation, enabling more precise manipulation. DP3[1] takes single-camera point clouds as input and generates actions through a sequence of preprocessing steps—including point clouds filtering, clustering, feature extraction, and denoising. Notably, this point clouds preprocessing pipeline is not only complex but also relies on high-precision RGB-D cameras. To underscore the superiority of our proposed method, we thus select DP3 as one of our baseline methods.

Non-vision-only methods exhibit strong performance in both success rate and few-shot learning. However, they not only depend on intricate point clouds preprocessing pipelines but also often require accurate camera extrinsic calibration for reprojection, two factors that complicate real-world deployment and place high demands on sensor consistency and environmental stability. By contrast, our approach relies solely on image input, yet delivers accuracy on par with these non-vision-only methods.

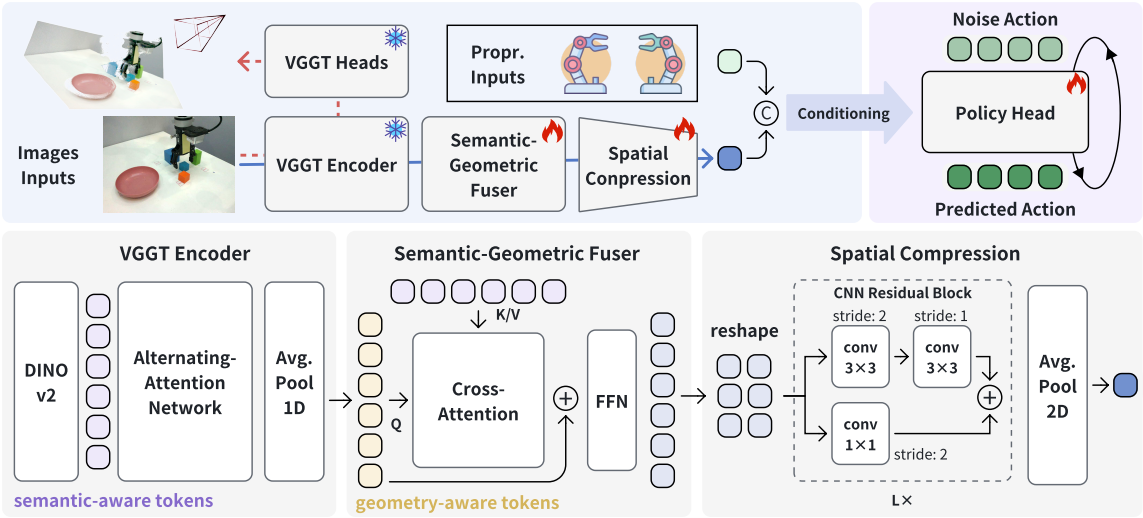

Fig. 2: Overall architecture of VO-DP. VO-DP has four core modules: 1) VGGT Encoder extracts semantic features from patchified images via DINOv2 and generates geometric features through its AA network; 2) Semantic-Geometric Fuser fuses per-frame geometric and semantic features using residual cross-attention and an FFN; 3) Spatial Compression module reshapes fused features, downsamples them with a lightweight ResNet, and concatenates the compressed spatial features with proprioceptive observations to form compact scenario representations; 4) Vision-Only Conditioned Action Generation module employs a DDPM-based policy head to generate actions using the scenario representations.

III. METHOD

We define the vision-only imitation learning task as follows: given a small set of expert demonstrations containing both video and robotic action trajectories, learn a visuomotor policy π that maps observations $O_t \in \mathcal{O}$ at time step t to actions $A_t \in \mathcal{A}$. The observation O_t, as shown in (1), consists of an RGB image history of size T and a sequence of T states of J joints, and the output $A_t \in \mathbb{R}^{N\times J}$ represents a predicted action trajectory of length N,

$$O_t = \{I_t \in \mathbb{R}^{T\times H\times W\times 3}, S_t \in \mathbb{R}^{T\times J}\}. \tag{1}$$

Notably, our method uses only single-view RGB images. The overall architecture is illustrated in 2, which consists of four sub-modules: a pretrained visual encoder III-A incorporating geometric priors for encoding observation images into semantic and geometric features; a semantic-geometric feature fusion module III-B for adaptive modality selection tailored to specific tasks; a scenario representation compressor III-C for distilling key information of scene; and a policy head III-D that predicts action chunks conditioned on the scenario features.

A. Geometry Prior-based Visual Encoder

Benefiting from its concise architecture and robust generalization capability, VGGT [5] is employed as the visual encoder in our method. Pretrained on a variety of 3D reconstruction tasks, VGGT can extract essential geometric features directly from one or a few input images and predict comprehensive 3D attributes of a scene — including camera parameters, point clouds, depth maps, and 3D point tracks. Specifically, VGGT is implemented as a large transformer [26], where each input image is first patchified into a set of tokens via DINOv2 [6]. The combined image tokens from all input frames are then processed through an Alternating-Attention (AA) network consisting of 24 AA blocks. Each AA block is designed with a frame-wise self-attention layer followed by a global self-attention layer. The features output by the AA network are further fed into prediction heads to estimate 3D attributes.

Therefore, we posit that the features processed by the AA network encapsulate rich 3D geometric information, which can enhance the precision of vision-only visuomotor policies. In our implementation, the image history I_t is patchified via DINOv2 into a set of $T \times P$ tokens, denoted as $\mathbf{h_t^{sem}} \in \mathbb{R}^{T\times P\times C}$, where C is the feature dimension. $\mathbf{h_t^{sem}}$ serves as semantic features of the observation. These tokens are then fed into the pretrained AA network, and the output from the 24-th AA block is adopted. Notably, within each AA block, VGGT concatenates features derived from both the frame-wise and global self-attention layers prior to feeding them into the prediction heads, thus enabling the integration of local and global information. Accordingly, we use these concatenated features as the geometric feature representation $\mathbf{h_t^{geo}} \in \mathbb{R}^{T\times P\times 2C}$.

B. Semantic-Geometric Feature Fuser

To effectively leverage both semantic features and geometric information, we fuse the per-frame features $\mathbf{g} = \mathbf{h_t^{geo}}[\mathbf{i}] \in \mathbb{R}^{P\times 2C}$ and $\mathbf{s} = \mathbf{h_t^{sem}}[\mathbf{i}] \in \mathbb{R}^{P\times C}$ using residual cross-attention, where i is the frame index, $\mathbf{g}$ serves as the query and $\mathbf{s}$ as the key-value pair, as follows:

$$\begin{aligned} \mathbf{h'} &= \mathrm{AvgPool}(\mathbf{g}), \\ \mathbf{h''} &= \mathbf{h'} + \mathrm{CrossAttn}(\mathbf{h'W_Q}, \mathbf{sW_K}, \mathbf{sW_V}), \end{aligned} \tag{2}$$

where $\mathrm{AvgPool}(\cdot)$ performs feature compression using 1D average pooling with a kernel size of 2 and stride of 2 along the feature dimension, $\mathbf{W_Q} \in \mathbb{R}^{C\times C}$, $\mathbf{W_K} \in \mathbb{R}^{C\times C}$, and $\mathbf{W_V} \in \mathbb{R}^{C\times C}$ are projection matrices. The features after cross-attention are further projected through a Feed-Forward Network (FFN) layer:

$$\mathbf{h_t^{sg}[i]} = \mathbf{h''} + \mathrm{FFN}(\mathbf{h''}), \tag{3}$$

where $\mathbf{h_t^{sg}[i]}$ is the fused features with semantic and geometric information.

C. Scenario Representation Compression

We then encode all observation tokens into compact scenario representations with a lightweight ResNet [16], as shown in Fig. 2. The layout of $\mathbf{h_t^{sg}}$ is first reshaped to $\mathbb{R}^{T\times C\times H_P\times W_P}$, where H_P, W_P denote the height and width of patch grid. We then apply three basic residual blocks, each with a kernel size of 3 and a stride of 2, to downsample the feature maps. An adaptive 2D average pooling layer compresses the remaining patches into a spatial feature $\mathbf{h_t^{sp}}$, which is projected into a low-dimensional space C' and concatenated with the proprioceptive observation $\mathbf{S_t}$, yielding the scene representation $\mathbf{h_t^{sc}} \in \mathbb{R}^{T\times(C'+J)}$,

$$\mathbf{h_t^{sc}} = [\mathrm{MLP}(\mathbf{h_t^{sp}}), \mathbf{S_t}]\,. \tag{4}$$

D. Vision-Only Conditioned Action Generation

For the policy head, we follow the original DP implementation [9], training it with a vision-only conditioned Denoising Diffusion Probabilistic Model (DDPM) [27] to approximates the conditional distribution $p(A_t|O_t)$ introduced in [28]. The denoising process follows:

$$A_t^{k-1} = \alpha(A_t^k - \gamma\varepsilon_\theta(h_t^{sc}, A_t^k, k) + \mathcal{N}(0, \sigma^2 I)) \tag{5}$$

where γ is the learning rate, α and σ are scalar coefficients predefined by a noise scheduler, and ε_θ is the noise prediction network with parameters θ which predicts the noise of the trajectory A_t^k conditioned on the scene feature h_t^{sc}. We train it with MSE loss as follows:

$$\mathcal{L}(\theta) = \mathrm{MSE}(\varepsilon^{\mathrm{k}}, \varepsilon_\theta(\mathrm{h_t^{sc}}, \mathrm{A_t^k}, \mathrm{k})). \tag{6}$$

IV. SIMULATION EXPERIMENTS

A. Experimental Setup

Benchmark. Simulation experiments are conducted using the RoboTwin [7] benchmark, which is built upon the SAPIEN [1] simulator and comprises 14 bimanual manipulation tasks, as shown in Fig. 3. RoboTwin serves as a challenging benchmark that requires manipulation policies to comprehend both semantic intent and geometric structure in visually complex environments. RoboTwin employs an open-source Cobot Magic platform[2], controlled via an action vector $J = 14$ to operate a pair of 6-DoF simulated robotic arms equipped with grippers. Visual input consists of RGB-D images and point clouds captured by an Intel RealSense D-435 camera at a resolution of 240×320.

[1]SAPIEN: https://sapien-sim.github.io/docs/

[2]Platform Introduction: https://global.agilex.ai/products/cobot-magic

Fig. 3: Simulation benchmark – 14 bimanual manipulation tasks. Left: Top-view RGB image of the task. Right: Reconstructed point clouds via VGGT.

Data Collection. Training data are collected from 100 valid scenes initialized randomly starting from seed 0 for each task. During testing, evaluation is performed over 100 valid scenes initialized from seed 10000 per task, with each scene repeated three times. The success rate is determined based on satisfying target pose constraints upon task completion and maintaining collision-free execution throughout the trajectory.

Training Details. The proposed method learns a mapping from an observation O_t to an action trajectory A_t of length $N = 8$. To maintain consistency with DP and DP3, the observation adopts a history length of $T = 3$ (denoted as VO-DP); additionally, the method is evaluated under an ablated setting with $T = 1$ (denoted as VO-DP-1). All models are trained for 300 epochs. During training, all samples are generated using the same random seed to ensure consistency. The models are trained on 8 NVIDIA A100 GPUs using the bfloat16 precision.

B. Simulation Performance

By comparing our method with DP (a traditional vision-only method) and DP3 (a native point cloud-based method), we validate the effectiveness of the pretrained fusion perception approach for vision-only robotic manipulation. Compared to DP, our method achieves a substantial performance improvement. Furthermore, compared to DP3, it achieves comparable manipulation accuracy at lower hardware costs, as shown in Table I.

Compared to DP, VO-DP and VO-DP-1 achieve dramatic performance improvements across all tasks, with particularly notable gains in several key scenarios. For instance, in the *Pick Apple Messy* task, their success rates rise substantially from the baseline of 31.0% to over 80.0%. In more complex tasks, such as *Block Hammer Beat* and *Blocks Stack (Easy)*, performance also improves sharply: starting from very low baselines of 0.7% and 3.7%, VO-DP and VO-DP-1 reach success rates of 85.0% and 69.3%, respectively.

These results strongly demonstrate that by integrating pretrained visual foundation models and a feature fusion-based perception strategy, VO-DP-1 significantly enhances the understanding of complex scenes and objects as well

TABLE I: RoboTwin Benchmark Results. DP: Diffusion policy[9], DP3: 3D diffusion policy[1], VO-DP: $T = 3$, VO-DP: $T = 1$

Method	Block Hammer Beat	Block Handover	Bottle Adjust	Container Place	Empty Cup Place
DP	0.7±0.9	77.7±4.5	39.3±0.5	14.0±6.9	69.3±2.5
DP3	79.3±1.2	**97.7±1.2**	**85.3±0.5**	**83.7±1.7**	**88.7±1.7**
VO-DP	**85.0±1.4**	89.7±0.5	63.3±1.2	43.0±3.7	82.0±2.2
VO-DP-1	78.7±5.2	**94.7±0.5**	69.3±2.5	31.3±2.6	77.3±1.7
Method	**Pick Apple Messy**	**Put Apple Cabinet**	**Dual Bottles Pick (Easy)**	**Dual Bottle Pick (Hard)**	**Diverse Bottles Pick**
DP	31.0±0.8	63.6±1.9	73.7±1.2	63.3±0.5	7.3±1.2
DP3	18.7±2.9	84.7±0.5	83.3±0.5	64.0±0.8	**60.7±0.5**
VO-DP	**80.0±0.8**	**94.3±2.3**	**88.3±0.9**	**67.3±3.3**	32.3±3.3
VO-DP-1	**81.7±0.9**	**98.0±0.8**	**86.3±0.5**	60.3±1.2	31.3±1.7
Method	**Shoe Place**	**Dual Shoes Place**	**Tool Adjust**	**Blocks Stack (Easy)**	**AVG. (↑)**
DP	19.3±1.2	4.7±0.5	20.0±2.9	3.7±1.2	34.8
DP3	**56.3±1.7**	13.7±1.7	**58.3±0.5**	22.0±2.2	**64.0**
VO-DP	43.0±0.8	**17.0±0.8**	**58.3±3.9**	**52.3±2.5**	**63.9**
VO-DP-1	52.0±0.8	**19.3±0.9**	55.3±2.6	**69.3±2.5**	**64.6**

as the precision of manipulation. This effectively overcomes the inherent limitations of traditional vision-only methods, specifically their inadequacies in perceptual accuracy and generalization.

Compared to DP3, which relies on raw 3D point cloud inputs, VO-DP only requires a lower-cost monocular camera, yet achieves comparable or even superior overall performance. In terms of average success rate (AVG), VO-DP (63.9%) achieves near-parity with DP3 (64.0%), while its single-frame variant—VO-DP-1 (64.6%), outperforms DP3. Importantly, VO-DP achieves top or joint-top performance across multiple key tasks, such as *Block Hammer Beat*, *Put Apple Cabinet*, and *Blocks Stack (Easy)*. This indicates that VO-DP effectively bridges the performance gap with 3D perception-based methods, relying solely on image data, by leveraging advanced visual models. In the *Pick Apple Messy* task, our method boosts the success rate by 63.0% relative to DP3, demonstrating the advantages of pretrained implicit spatial representations for perceiving complex scenes.

Although VO-DP slightly underperforms DP3, a raw point cloud-based baseline, in some structured tasks or those requiring precise geometric information (e.g., *Diverse Bottles Pick*), it still maintains high performance with only RGB image input, thereby significantly lowering hardware sensor requirements.

Comparison between 3 frame and 1 frame variants. VO-DP and VO-DP-1 exhibit comparable overall performance, with each demonstrating distinct strengths across different scenarios. Given that VO-DP-1 achieves slightly superior performance to VO-DP, VO-DP-1 is selected as the method for subsequent real-world experiments.

C. Ablation Study

We select five tasks to conduct ablation studies: *Pick Apple Messy* (PAM), *Block Hammer Beat* (BHB), *Dual Bottles Pick (Easy)* (DBPE), *Put Apple Cabinet* (PAC), and *Blocks Stack (Easy)* (BSE) from RoboTwin. From a semantic standpoint, these tasks cover a spectrum of robotic manipulation types, involving varied objects and diverse scenarios. From a geometric perspective, these tasks necessitate the robot to resolve spatial relationships—underscoring the core challenges in spatial perception and motion coordination.

TABLE II: Ablation study on different modality features.

Module	PAM	BHB	DBPE
w/o geo.	44.3±0.9	59.3±0.5	**95.3±0.9**
w/o sem.	38.7±1.7	60.7±4.9	81.3±0.5
VO-DP	**80.0±0.8**	**85.0±1.4**	**88.3±0.9**
Module	**PAC**	**BSE**	**AVG. (↑)**
w/o geo.	**98.0±0.8**	**58.7±0.9**	71.12
w/o sem.	93.7±2.0	45.3±2.5	63.9
VO-DP	**94.3±2.3**	**52.3±2.5**	**80.0**

Comparison of different modality features. VO-DP leverages both semantic and geometric features to enable spatial understanding. To evaluate the contribution of each feature modality, we perform an ablation study on the fusion module, retaining only semantics-aware features (denoted as "w/o geo.") or geometry-aware features (denoted as "w/o sem."), with results summarized in Table II. Results show that the full VO-DP model achieves the best overall performance, with an average success rate evidently higher than that of either ablated variant, confirming the effectiveness of our multimodal feature fusion design. Specifically, in tasks demanding strong semantic understanding (e.g., *Pick Apple Messy*, *Block Hammer Beat*), performance drops substantially when semantic features are removed (w/o sem.), underscoring the critical role of semantic priors in object recognition and task reasoning. For structurally complex tasks (e.g., *Dual Bottles Pick (Easy)*), the removal of geometric features (w/o geo.) leads to noticeable performance degradation, highlighting the importance of spatial structure for bimanual coordination. Notably, certain tasks (e.g., *Put Apple Cabinet*) achieve relatively high performance with only a single modality, indicating that perceptual demands vary across different tasks. Nevertheless, VO-DP consistently outperforms both ablated models in most scenarios, which demonstrates that our fusion mechanism robustly generalizes

to diverse manipulation requirements.

Comparison of Downsampling Strategies for Geometric Features. When integrating VGGT geometric features, we explore two downsampling strategies: average pooling and an MLP-based projection, to reduce the dimensionality of geometry-aware tokens to 1024 dimensions, with details summarized in Table III. We observe that increasing parameter counts via dimensional projection did not yield significant improvements in overall model performance. Consequently, we opt for the average pooling strategy for feature downsampling.

TABLE III: Ablation on different strategy for geometry token downsampling

Strategy	PAM	BHB	DBPE
mlp	**82.0±1.6**	66.3±1.7	88.7±1.2
pool	80.0±0.8	**85.0±1.4**	**88.3±0.9**
Strategy	**PAC**	**BSE**	**AVG. (↑)**
mlp	**99.3±0.9**	**62.3±3.3**	79.7
pool	94.3±2.3	52.3±2.5	**80.0**

Efficient scaling with demonstrations. As shown in Fig. 4, VO-DP exhibits high data efficiency and strong scaling capability. Compared with baselines (DP and DP3), it delivers more substantial performance gains as training demonstrations increase from 20 to 100, especially in high-complexity scenarios. For example, in *Pick Apple Messy*, its success rate jumps from 3.0% to 80.0%, outperforming DP and DP3 markedly. A similar trend appears in *Block Hammer Beat*: VO-DP's success rate rises from 4.7% to 85.0%, while DP3 only improves modestly and DP gains little from more data. These results confirm that the strong prior knowledge in its pretrained visual encoder enables VO-DP to learn and generalize more effectively with limited demonstration data.

V. REAL WORLD EXPERIMENTS

A. Experiment Setup

Real robot benchmark. We evaluate VO-DP on four real-world tasks and four robustness tests. As illustrated in Fig. 6, we use a Realman RM65-B robot equipped with an Inspire EG2-4C2 gripper, one RealSense L515 camera to capture real-world visual observations (containing both RGB images and point clouds) with robot states, and a controllable flashlight with adjustable color and frequency as an environmental disturbance source. All objects used in the experiments are shown in Fig. 6, which include multiple blocks and containers of varying shapes, sizes, and colors. We now briefly describe the four spatial tasks:

- **Pick&Place Small Cube (PPSC).** Grasp a 3 cm cube and place it at the center of the plate.
- **Pick&Place Big Cube (PPBC).** Grasp a 5 cm cube and place it at the center of the plate.
- **Cover Cuboid (CC).** Pick up a cup from the plate and move it to cover an upright 3cm×3cm×6cm cuboid.
- **Stack Cubes (SC).** Stack a blue 3 cm cube on top of an orange 3 cm cube.

All the tasks are visualized in Fig. 5.

Additionally, we design four robustness tests based on the *Pick&Place Small cube* task: **Size robustness.** Train: 3 cm cube; Test: cubes of 2.5 cm, 3 cm and 5 cm. **Appearance robustness.** Train: orange cubes; Test: cubes of all colors. **Illumination robustness.** Train: normal ambient lighting; Test: normal ambient lighting and strobe lighting. **Background robustness.** Train: a standard desktop surface; Test: a standard desktop surface and ones covered with colored papers.

Data Collection. We collect 200 demonstrations per task using the teleoperation device provided with the Realman robot. The operational area is uniformly partitioned into multiple grids. During data collection, the target object is sequentially placed at random positions within each grid. For instance, in the *Stack cubes* task, the orange cube is placed in distinct grids one after another, while the blue cube is positioned in all remaining feasible grids, thus ensuring coverage of all possible combinations. During testing, the same strategy is employed to ensure a uniform spatial distribution for evaluation.

Training Details. As in the simulation experiments, we select DP and DP3 as our baselines. Given the color sensitivity of the *Stack Cubes* task, we use the color-variant version of DP3 for comparison. Based on the simulation results, we choose the single-frame VO-DP-1 for real-world evaluation (unless otherwise specified below, VO-DP refers to VO-DP-1 by default). It is noteworthy that the point clouds processing in DP3 depends on a manually defined operational region—an approach that proves impractical for real-world robotic manipulation scenarios.

TABLE IV: Real-world Performance. In the four real-world tasks, it can be observed that VO-DP significantly outperforms the other two methods.

Method	PPSC	PPBC	CC	SC	AVG. (↑)
DP	23.3	16.7	3.3	1.7	11.2±9.1
DP3	73.3	68.3	75.0	53.3	67.5±8.5
VO-DP-1	**96.7**	**91.6**	**93.3**	**70.0**	**87.9±10.5**

B. Real-world Performance

Results for our real robot tasks are given in Table IV. Our experimental results demonstrate that VO-DP achieves strong performance and generalization capability in real-world physical environments. It attains an average success rate of 87.9% across all four tasks, significantly outperforming the point cloud-based method DP3 (67.5%) and the conventional vision-only approach DP (11.2%). Specifically, VO-DP achieves the highest performance in *Cover cuboid* and *Pick&Place cube*. This demonstrates that the visual encoder of VO-DP possesses the capability to extract robust, task-discriminative features—thereby facilitating accurate perception of object geometry, spatial relationships, and manipulation intent. The results confirm that VO-DP achieves effective transfer from simulation to complex real-world environments. Notably, it outperforms the method DP3 dependent on expensive depth sensors, while only utilizing a low-cost RGB camera. We attribute the degradation of

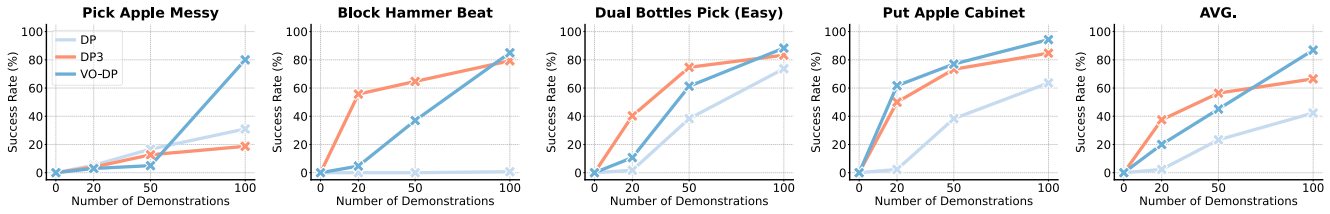

Fig. 4: Data efficiency and scaling capability comparison of VO-DP with baseline methods (DP and DP3) across four tasks. The table presents success rate changes as training demonstrations scale from 20 to 100, highlighting VO-DP's more substantial performance improvements.

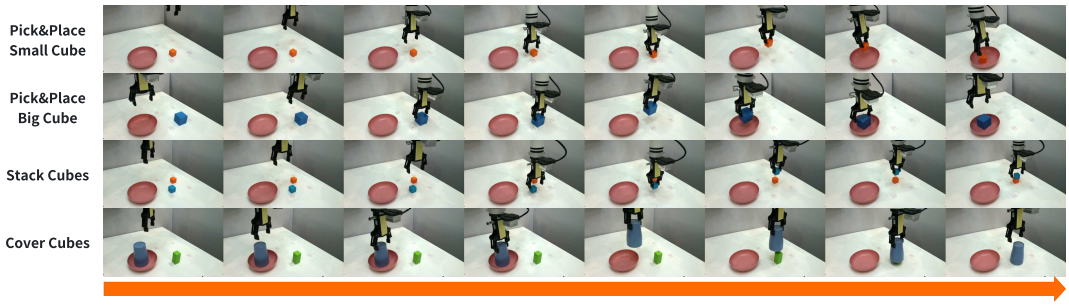

Fig. 5: Visualize of 4 real-world tasks: Pick&Place Small Cube (PPSC), Pick&Place Big Cube (PPBC), Cover Cuboid (CC), Stack Cubes (SC).

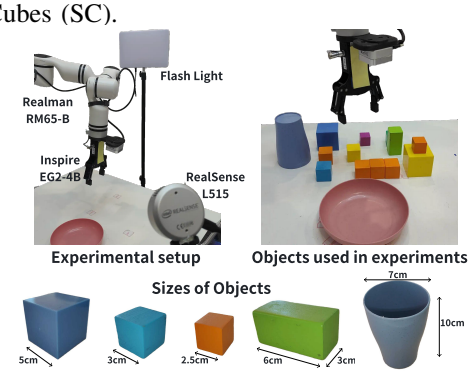

Fig. 6: Information on the equipment used in real-world experiments and the objects involved in tasks.

DP3's real-world performance, relative to its performance in simulations, to the lack of idealized sensing conditions. In real-world scenarios, depth sensors are inherently affected by noise, calibration inaccuracies, viewpoint dependence, and artifacts introduced during point clouds preprocessing. All these factors collectively lead to the performance gap of DP3 between simulation and real-world scenarios.

C. Robustness Test

Size Robustness. To evaluate the geometric robustness of VO-DP, we test the model trained on $3.0\,cm$ cubes using $2.5\,cm$, $3.0\,cm$ and $5.0\,cm$ cubes, as illustrated in Fig. 6. The evaluation involves randomized tests across 20 grids in the central operational area, with results summarized in Table V. Experimental results show that VO-DP generalizes geometrically to unseen object sizes, maintaining robust performance across both smaller and larger objects with an average success rate of 65.0%. This indicates that the visual encoder captures geometric and spatial representations that exhibit scale-invariant generalization.

TABLE V: Size Robustness. Train: 3.0 cm cubes; Test: 2.5 cm, 3.0 cm and 5 cm cubes.

3.0 cm	2.5 cm	5.0 cm	AVG.
85.0	60.0	50.0	**65.0±14.7**

Appearance Robustness. We evaluate the robustness of VO-DP to varying object appearances by testing the model trained on orange cubes using blue, green, and yellow cubes. The results are presented in Table VI. VO-DP achieves strong performance on the yellow cube, which is chromatically close to the training color demonstrating its capacity for color generalization. However, performance declines markedly on distantly colored objects such as blue and green, suggesting that semantic color understanding remains partially constrained by the training distribution.

TABLE VI: Appearance Robustness. Train: ■ cubes; Test: ■, ■, ■, ■ cubes.

■	■	■	■	AVG.
85.0	50.0	40.0	90.0	**66.3±21.6**

Illumination Robustness. We validate the robustness of VO-DP to varying lighting conditions by adjusting flashlight settings. In the *Light Switch* test, both intensity and color temperature are randomly configured for each evaluation position. During the *Blinking* test, a low-frequency blinking mode continuously alters the ambient illumination. Results are presented in Table VII. The results indicate that VO-DP maintains strong robustness under challenging illumination variations. Under extreme conditions such as stochastic light switching and low-frequency blinking, it achieves performance comparable to that under standard lighting, with an

average success rate of 83.3%. This confirms the ability of the method to extract illumination-invariant visual representations and its notable resilience to variations in color, brightness, and dynamic lighting interference.

TABLE VII: Illumination Robustness. Train: Normal; Test:Nomal, Light Switch, Blinking.

Normal	Light Switch	Blinking	AVG.
85.0	80.0	85.0	**83.3±2.4**

Background Robustness. Background generalization presents a significantly greater challenge for methods relying on RGB image inputs. To evaluate this, we cover the operational area with white, pink, and blue paper respectively during testing. The results are shown in Table VIII. The results indicate that VO-DP generalizes effectively across substantial variations in background appearance. The model achieves high manipulation success rates under diverse background colors demonstrating robust performance despite visual domain shifts.

TABLE VIII: Background Robustness. Train: desktop surface; Test: desktop surface, ■, ■ and ■ surface.

desktop surface	■	■	■	AVG.
85.0	90.0	80.0	95.0	**87.5±5.6**

VI. CONCLUSION

This paper addresses the underexplored potential of vision-only approaches in visuomotor diffusion policy learning for robotic manipulation, proposing a single-view, vision-only method (VO-DP) that bridges the performance gap between vision-only and point cloud-based baselines. VO-DP leverages pretrained visual foundation models to fuse semantic and geometric features effectively via targeted extraction, cross-attention fusion, and CNN compression, supporting downstream policy learning. Extensive experiments validate its efficacy: on the RoboTwin benchmark, it significantly outperforms vision-only baseline DP and matches point cloud-based DP3, while achieving the highest average success rate in real-world tasks with strong robustness across conditions. Overall, VO-DP demonstrates that vision-only approaches can achieve high accuracy and robustness in robotic manipulation without expensive depth sensors, highlighting their potential for cost-effective, scalable real-world deployment. Future work may extend VO-DP to multi-view settings or more complex dynamic manipulation tasks.

REFERENCES

[1] Y. Ze, G. Zhang and K. Zhang et al., "3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations," *arXiv preprint arXiv:2403.03954*, 2024.

[2] T.-W. Ke, N. Gkanatsios and K. Fragkiadaki, "3d diffuser actor: Policy diffusion with 3d scene representations," *arXiv preprint arXiv:2402.10885*, 2024.

[3] X. Lin, T. Lin and L. Huang et al., "Sem: Enhancing spatial understanding for robust robot manipulation," *arXiv preprint arXiv:2505.16196*, 2025.

[4] Y. Lu, Y. Tian and Z. Yuan et al., "H ^{3} dp: Triply-hierarchical diffusion policy for visuomotor learning," *arXiv preprint arXiv:2505.07819*, 2025.

[5] J. Wang, M. Chen and N. Karaev et al., "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5294–5306, 2025.

[6] M. Oquab, T. Darcet and T. Moutakanni et al., "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[7] Y. Mu, T. Chen and S. Peng et al., "Robotwin: Dual-arm robot benchmark with generative digital twins (early version)," in *European Conference on Computer Vision*, pp. 264–273, Springer, 2024.

[8] S. Gugger, L. Debut and T. Wolf et al., "Accelerate: Training and inference at scale made simple, efficient and adaptable," *Computer software]. Hugging Face. https://github. com/huggingface/accelerate*, 2022.

[9] C. Chi, Z. Xu and S. Feng et al., "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, p. 02783649241273668, 2023.

[10] T. Z. Zhao, V. Kumar and S. Levine et al., "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[11] A. Brohan, N. Brown and J. Carbajal et al., "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[12] E. Jang, A. Irpan and M. Khansari et al., "Bc-z: Zero-shot task generalization with robotic imitation learning," in *Conference on Robot Learning*, pp. 991–1002, PMLR, 2022.

[13] N. M. Shafiullah, Z. Cui and A. A. Altanzaya et al., "Behavior transformers: Cloning k modes with one stone," *Advances in neural information processing systems*, vol. 35, pp. 22955–22968, 2022.

[14] J. Mao, J. Guan and Y. Tang et al., "Omnid: Generalizable robot manipulation policy via image-based bev representation," *arXiv preprint arXiv:2508.11898*, 2025.

[15] A. Xie, L. Lee and T. Xiao et al., "Decomposing the generalization gap in imitation learning for visual robotic manipulation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3153–3160, IEEE, 2024.

[16] K. He, X. Zhang and S. Ren et al., "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[17] A. Dosovitskiy, L. Beyer and A. Kolesnikov et al., "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[18] Y. Hong, H. Zhen and P. Chen et al., "3d-llm: Injecting the 3d world into large language models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 20482–20494, 2023.

[19] R. Fu, J. Liu and X. Chen et al., "Scene-llm: Extending language model for 3d visual understanding and reasoning," *arXiv preprint arXiv:2403.11401*, 2024.

[20] C. Zhu, T. Wang and W. Zhang et al., "Llava-3d: A simple yet effective pathway to empowering lmms with 3d-awareness," *arXiv preprint arXiv:2409.18125*, 2024.

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

[22] T. Gervet, Z. Xian and N. Gkanatsios et al., "Act3d: 3d feature field transformers for multi-task robotic manipulation," *arXiv preprint arXiv:2306.17817*, 2023.

[23] A. Radford, J. W. Kim and C. Hallacy et al., "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[24] A. Goyal, J. Xu and Y. Guo et al., "Rvt: Robotic view transformer for 3d object manipulation," in *Conference on Robot Learning*, pp. 694–710, PMLR, 2023.

[25] A. Goyal, V. Blukis and J. Xu et al., "Rvt-2: Learning precise manipulation from few demonstrations," *arXiv preprint arXiv:2406.08545*, 2024.

[26] A. Vaswani, N. Shazeer and N. Parmar et al., "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[27] J. Ho, A. Jain and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[28] M. Janner, Y. Du and J. B. Tenenbaum et al., "Planning with diffusion for flexible behavior synthesis," *arXiv preprint arXiv:2205.09991*, 2022.