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

HeRO: Hierarchical 3D Semantic Representation for Pose-aware Object Manipulation

Chongyang Xu[1,3*], Shen Cheng[3†], Haipeng Li[2,3†], Haoqiang Fan[3], Ziliang Feng[1] and Shuaicheng Liu[2✉]

***Abstract*— Imitation learning for robotic manipulation has progressed from 2D image policies to 3D representations that explicitly encode geometry. Yet purely geometric policies often lack explicit *part-level* semantics, which are critical for pose-aware manipulation (e.g., distinguishing a shoe's "toe" from "heel"). In this paper, we present HeRO, a diffusion-based policy that couples geometry and semantics via hierarchical *semantic fields*. HeRO employs dense semantics lifting to fuse discriminative, geometry-sensitive features from DINOv2 with the smooth, globally coherent correspondences from Stable Diffusion, yielding dense features that are both fine-grained and spatially consistent. These features are processed and partitioned to construct a global field and a set of local fields. A hierarchical conditioning module conditions the generative denoiser on global and local fields using permutation-invariant network architecture, thereby avoiding order-sensitive bias and producing a coherent control policy for pose-aware manipulation. In various tests, HeRO establishes a new state-of-the-art, improving success on *Place Dual Shoes* by 12.3% and averaging 6.5% gains across six challenging pose-aware tasks. Code is available at https://github.com/Chongyang-99/HeRO.**

I. INTRODUCTION

Imitation learning for robotic manipulation [1], [2], [3], [4], [5] has advanced significantly, with policies evolving from 2D image-based methods [6], [7] to 3D representations [8], [9], [10], [11], [12], [13], [14]. While early image-based approaches achieve notable success, they often falter in tasks requiring precise spatial reasoning, as cameras inherently flatten the 3D world and lose crucial geometric information. To address this limitation, 3D imitation learning has emerged, leveraging representations like point clouds or voxels to explicitly model geometry. A pioneering example is the 3D Diffusion Policy (DP3) [8], which conditions its action generation on a compact 3D visual representation. By processing sparse point clouds, this approach can better capture spatial relationships for precise control, leading to improved performance.

However, despite their geometric strengths, these 3D methods often lack explicit semantic understanding. This limitation becomes a critical bottleneck in *pose-aware* manipulation scenarios, where success hinges on identifying and reasoning about specific object parts. For instance, a task like placing shoes, as illustrated in Fig. 1 (top), requires more than just successfully moving them to a location; it demands precise alignment based on functional parts like the *"toe"* and *"heel"*. Methods that rely purely on geometry cannot disambiguate between these semantically distinct parts, often leading to task failure.

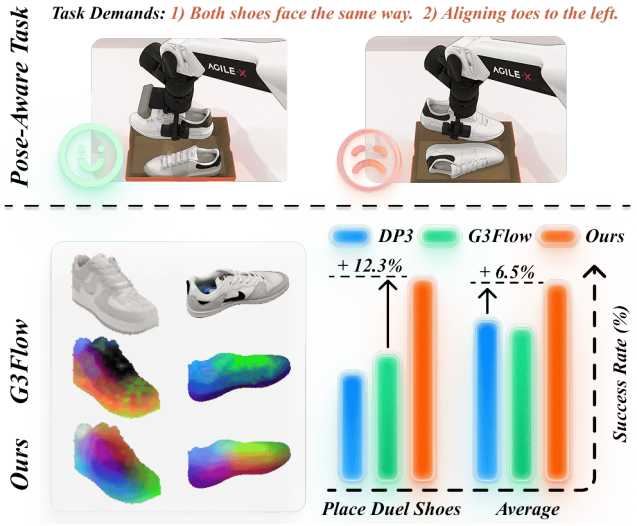

Fig. 1. **Pose-Aware Manipulation with Semantic Understanding. Top:** Many manipulation tasks are pose-aware (e.g., placing shoes with toes aligned left), demanding semantic part perception from policies. **Bottom Left:** Our dense semantic fields are smoother and more consistent than the baseline G3Flow [13]. **Bottom Right:** Our method achieves a 12.3% higher success rate on the dual shoe place task and a 6.5% higher average success rate of 6 challenging tasks.

To address it, recent work has focused on enriching 3D representations with semantic features [15], [11]. Pioneering approaches like G3Flow [13] construct a semantic field by leveraging powerful foundation models [16], marking a step forward in semantic-aware manipulation. However, despite its effectiveness, this method might yield a *holistic* semantic representation (Fig.2 top), causing distinct part-level semantics to become indistinguishable. For example, as visualized in Fig. 1 (bottom left), features for a shoe's "toe" and "heel" become similar. Consequently, for pose-aware manipulation tasks that depend on differentiating these parts, the policy struggles to achieve the required precision.

In this work, we present **HeRO** (**H**ierarchical **Se**mantic **R**epresentation for **O**bject manipulation) for part-level semantic perception of objects. Our central motivation is that *pose-aware manipulation requires dense, fine-grained representations with strong spatial semantic coherence*. Motivated by recent advances in dense correspondence from foundation models [17], [18], [19], we first propose Dense Semantic Lifting module to construct a semantic field that is *denser and more discriminative* than G3Flow (Fig. 1, bottom-left). Concretely, we fuse features from DINOv2 [16], which are

* This work was done during Chongyang Xu's internship at Dexmal.
† Corresponding Project Leaders.
[1] College of Computer Science, Sichuan University, [2] School of Information and Communication Engineering, University of Electronic Science and Technology of China, [3] Dexmal.
✉ Corresponding Author: liushuaicheng@uestc.edu.cn.

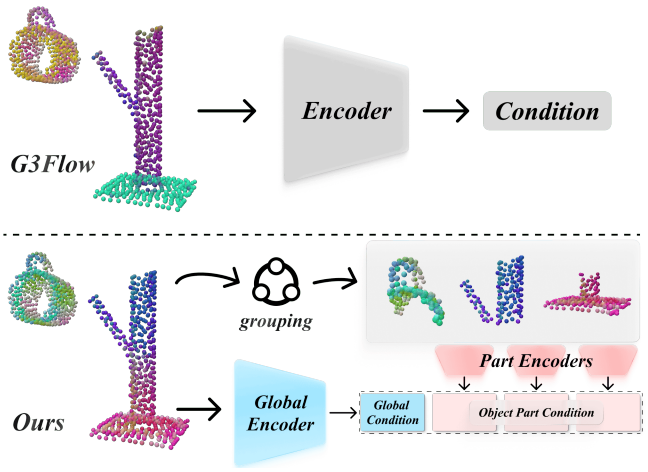

Fig. 2. **Comparison of Conditioning Mechanisms. Top:** The baseline employs a holistic conditioning approach, encoding the entire object point cloud into a single global vector, which lacks part-level details. **Bottom:** Our method uses a hierarchical approach. A global encoder captures overall context, while additional encoders extract complementary local features for fine-grained details. The resulting *Hierarchical Condition* provides both global and local information, enabling more precise manipulation.

discriminative and geometrically precise for sparse correspondences, with features from Stable Diffusion (SD) [20], which yield smooth, globally coherent correspondences. This complementary fusion preserves geometric accuracy while enforcing semantic consistency, enabling robust part-level correspondences for manipulation.

Given object point clouds and RGB-D observations, we apply **Dense Semantic Lifting** to extract a global semantic field $\mathcal{F}^G$ and a set of local semantic fields $\mathcal{F}^L = \{\mathcal{F}^{L,1}, \ldots, \mathcal{F}^{L,K}\}$. Specifically, we first extract complementary 2D features from DINOv2 and Stable Diffusion, fuse them through learnable weights, then lift these fused features to 3D by projecting each point onto the image plane and sampling the corresponding features. For $\mathcal{F}^G$, we create a temporally consistent global semantic field that combines geometric precision with semantic understanding through pose estimation between different timesteps. For $\mathcal{F}^L$, we partition $\mathcal{F}^G$ into K sub-parts using PCA-based grouping to obtain semantically coherent local features (Fig. 2, bottom).

The dense features $\mathcal{F}^G$ and $\mathcal{F}^L$ are then fed into a Hierarchical Conditioning Module (HCM), where they serve as conditions for the generative process. Unlike conventional conditioning that concatenates conditions with the denoiser features, thereby introducing an order-sensitive inductive bias, the set $\mathcal{F}^L$ is inherently unordered (e.g., $\mathcal{F}^{L,1}$ may correspond to either "toe" or "heel" across different shoes), which can confuse learning. We therefore adopt a permutation-invariant conditioning scheme [21], [22]: the HCM performs cross-attention between $\mathcal{F}^L$ and the denoising features without positional embeddings. As a result, the HCM effectively leverages the extracted hierarchical features, applies conditioning, and feeds them to the policy network for pose-aware object manipulation.

Through extensive experiments, HeRO sets a new state-of-the-art (SOTA) in pose-aware robotic manipulation. As shown in Fig. 1 (bottom right), it improves success rate by 12.3% over the prior best method, G3Flow [13], on the challenging *Place Dual Shoes* task, and achieves an average gain of 6.5% across six challenging pose-aware benchmark tasks. Overall, our main contributions are as follows:

- We present HeRO, a framework for part-level semantic perception that employs Dense Semantic Lifting to construct fine-grained 3D semantic fields by fusing complementary features from DINOv2 and Stable Diffusion, thereby preserving geometric precision while ensuring semantic coherence.
- We propose a Hierarchical Conditioning Module (HCM) for diffusion-based policies, which integrates global context and a set of permutation-invariant, part-aware features, overcoming the limitations of holistic global conditioning.
- We provide extensive validation in both simulation and the real world, demonstrating that our method establishes a new state-of-the-art on challenging pose-aware manipulation benchmarks.

II. RELATED WORK

A. Diffusion Models for Imitation Learning

Imitation learning [1], [2], [23], [24], [25] enables robots to acquire skills from expert demonstrations. Diffusion-based visuomotor policies [6], [7] generate actions from 2D observations, producing smooth and temporally consistent trajectories, while flow-matching methods [12], [26] further improve training stability. Extensions to second-order flows [27] incorporate acceleration and jerk for even smoother motion. Most existing work, however, focuses on simple end-effectors and low-DoF control, with only a few addressing dexterous manipulation [28], [29], [30], where stability and precision remain challenging.

The major limitation is their reliance on 2D features, which cannot fully capture spatial relationships and object-level semantics required for pose-aware manipulation. Recent studies show that integrating vision foundation models helps by lifting 2D semantic features into 3D representations, providing richer geometric context and semantic grounding that improve manipulation accuracy and robustness.

B. Vision Foundation Models

Vision foundation models have recently advanced dense correspondence learning along two complementary lines. Self-supervised Vision Transformers such as DINOv2 [16] provide discriminative features well suited for semantic matching, while generative diffusion models [31], [32], [18], [19] yield dense and spatially coherent correspondences. Prior work has shown these representations to be complementary [17], yet their synergy has not been fully explored in robotic manipulation. We leverage this complementarity by fusing DINO- and SD-derived features to construct 3D semantic fields that are both geometrically precise and semantically consistent, enabling robust part-level reasoning for pose-aware manipulation.

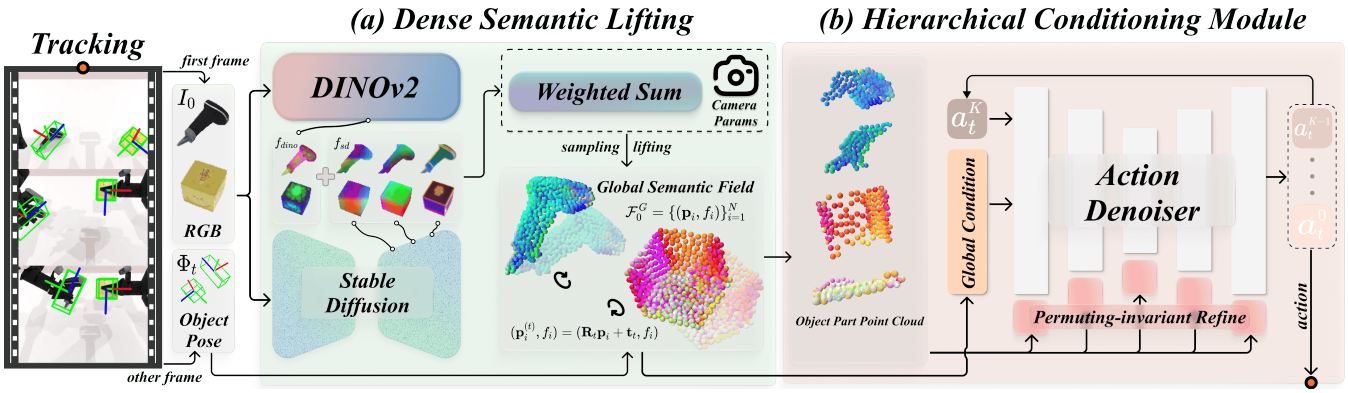

Fig. 3. **Method Overview.** Our framework generates precise, pose-aware actions in a two-stage process. **(a) Dense Semantic Lifting:** We track object 6D poses from sequential frames and lift fused 2D features from DINOv2 (semantic) and Stable Diffusion (geometric) into an object-centric *Dense Semantic Point Cloud*. **(b) Hierarchical Conditioning Module:** This point cloud is abstracted into a hierarchical *Object Part Point Cloud*, which conditions the diffusion policy via two pathways: as a *Global Condition* for the action denoiser and as fine-grained local guidance injected by a *Permutation-invariant Refine* module at each denoising step. This dual mechanism enables precise, pose-aware action generation.

C. Semantic-Aware 3D Perception

3D-based policies [8], [9], [10] improve spatial reasoning in robotic manipulation by lifting two-dimensional features from pre-trained vision foundation models into 3D space, providing both geometric structure and semantic context. Some methods [15], [11] create dense 3D descriptor fields from multi-view features, while others [9], [10] leverage lifted two-dimensional semantics to condition policies. Additional approaches [13] generate continuous semantic flows to handle dynamic interactions and occlusions.

Despite these advances, most approaches emphasize global representations and lack fine-grained, part-level precision required for pose-aware manipulation. Our method addresses this by fusing DINOv2 [16] and Stable Diffusion [20] features to construct dense global and local semantic fields. This fusion preserves geometric accuracy while enforcing semantic consistency, enabling precise part-level correspondences for manipulation.

III. METHOD

Our framework enhances diffusion-based robotic manipulation through three key components, as illustrated in Fig. 3.

First, we create rich semantic representations via **Dense Semantic Lifting** (§III-A), which combines geometric point clouds with semantic features from visual foundation models to form Global and Local Semantic Fields. Second, we design a **Hierarchical Conditioning Module** (§III-B) that provides the diffusion policy with both global scene context and fine-grained part-level information through permutation-invariant conditioning. Finally, we train a **Diffusion Policy** (§III-C) that generates precise manipulation actions conditioned on these hierarchical semantic representations.

We begin by reconstructing 3D object geometry from multi-view RGB-D observations, following G3Flow [13], yielding dense point clouds $\mathcal{P}$. We downsample each point cloud using Farthest Point Sampling to $N = 1024$ points: $\mathcal{P}_0 = \text{FPS}(\mathcal{P}, N)$.

A. Dense Semantic Lifting

While geometric point clouds provide spatial structure, effective manipulation requires understanding semantic properties like object parts, affordances, and material properties. Our Dense Semantic Lifting process enriches the geometric representation $\mathcal{P}_0$ with dense semantic features extracted from visual foundation models, creating a unified representation that combines both geometric precision and semantic understanding.

Feature Extraction: We leverage two complementary foundation models to capture different aspects of visual understanding for the first RGB frame I_0. DINOv2 [16] provides discriminative, fine-grained visual features $f_{\text{dino}} \in \mathbb{R}^{H\times W\times d_v}$. In parallel, Stable Diffusion [20] offers globally coherent semantic priors $f_{\text{sd}} \in \mathbb{R}^{H\times W\times d_s}$ by concatenating intermediate features from layers 2, 5, and 8, which encode rich semantic understanding developed through large-scale generative training. These complementary features capture both local visual details and global semantic context.

Feature Fusion: To unify these heterogeneous feature representations, we first apply PCA dimensionality reduction to project both feature maps to a common dimension d, yielding $f'_{\text{dino}} \in \mathbb{R}^{H\times W\times d}$ and $f'_{\text{sd}} \in \mathbb{R}^{H\times W\times d}$. The reduced features are then combined through a learnable weighted fusion:

$$f_{\text{fused}} = \alpha f'_{\text{dino}} + \beta f'_{\text{sd}}, \tag{1}$$

where α and β are learnable parameters that adaptively balance the contribution.

3D Lifting: We transfer the fused 2D features to the 3D domain by projecting each point $\mathbf{p}_i \in \mathcal{P}_0$ onto the image plane using camera intrinsics and sampling the corresponding fused feature f_i through bilinear interpolation:

$$\mathcal{F}_0^G = \{(\mathbf{p}_i^{(0)}, f_i)\}_{i=1}^N. \tag{2}$$

This process creates the initial Global Semantic Field that combines precise geometric structure with rich semantic information, providing a foundation for downstream manipulation reasoning.

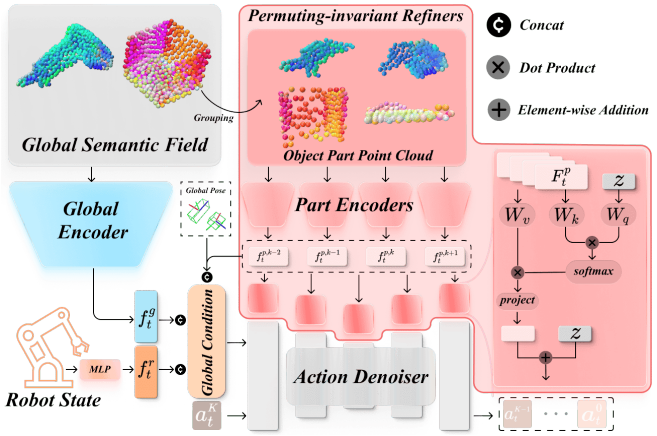

Fig. 4. **Hierarchical Conditioning Module Architecture.** Our model uses a dual-pathway design to guide the Action Denoiser. The **Global Path** processes the entire point cloud into a single global condition for high-level context. The **Local Path** partitions the point cloud into semantic parts, which are encoded by Permutation-invariant Refiners into a set of fine-grained embeddings. These local conditions are injected into the denoiser, enabling actions that are both globally consistent and locally precise.

Temporal Propagation: As objects move during manipulation, we maintain temporal consistency of the semantic field by tracking each object's 6D pose trajectory $\Phi_t = (\mathbf{R}_t, \mathbf{t}_t) \in SE(3)$, at each timestep t. The semantic field is updated by applying rigid-body transformations to point positions while preserving their associated semantic features, which represent intrinsic object properties independent of pose:

$$(\mathbf{p}_i^{(t)}, f_i) = (\mathbf{R}_t \mathbf{p}_i^{(t)} + \mathbf{t}_t, f_i), \quad i = 1, \ldots, N. \tag{3}$$

This transformation yields the updated Global Semantic Field at timestep t:

$$\mathcal{F}_t^G = \{(\mathbf{p}_i^{(t)}, f_i)\}_{i=1}^N. \tag{4}$$

Through this temporal propagation strategy, we construct the complete Global Semantic Field $\mathcal{F}_G = \{\mathcal{F}_t^G\}_{t=0}^T$ that maintains both geometric and semantic consistency across the entire manipulation sequence.

B. Hierarchical Conditioning Module

While $\mathcal{F}_G$ provides comprehensive scene understanding, effective manipulation requires fine-grained reasoning about individual object parts and their relationships. For instance, grasping different parts of a shoe (heel vs. toe) requires distinct manipulation strategies. We address this by conditioning the diffusion policy through dual pathways: global scene context and part-level information, as illustrated in Fig. 4.

Local Field Construction: At each timestep t, we decompose $\mathcal{F}_t^G$ into Local Semantic Fields that capture semantically meaningful object parts: $\mathcal{F}_t^L = \{\mathcal{F}_t^{L,k}\}_{k=1}^K$ where $K = 8$ represents the number of part-level clusters. We create an augmented representation for each point by concatenating spatial coordinates with semantic features: $\mathbf{x}_i = [\mathbf{p}_i^{(t)}; f_i] \in \mathbb{R}^{3+d}$ where $\mathbf{p}_i^{(t)} \in \mathbb{R}^3$ captures geometric structure and $f_i \in \mathbb{R}^d$ encodes semantic properties.

We apply PCA to identify the dominant structural variation across all points. Points are sorted along the first principal component, which typically aligns with the object's main elongation axis, and evenly partitioned into K clusters. This PCA-based approach naturally discovers part boundaries while ensuring each local field $\mathcal{F}_t^{L,k}$ represents a spatially and semantically coherent object region.

1) Global Conditioning: For effective diffusion-based action generation, the policy requires comprehensive contextual information integrating environmental understanding and robot state awareness. We construct this context through three complementary feature streams:

Scene Features: We encode the global semantic field $\mathcal{F}_t^G$ using a PointNet encoder to extract scene feature f_t^g. This captures overall geometric structure, semantic context, and spatial relationships between objects in the scene.

Robot Features: We encode the robot's joint states through a multi-layer perceptron (MLP) to produce robot feature f_t^r. This provides information about the manipulator's current configuration, pose constraints, and kinematic limitations that influence action feasibility.

Part Features: We individually encode each local field $\mathcal{F}_t^{L,k}$ using PointNet encoders and aggregate them using current object poses Φ_t to generate part feature f_t^p. This preserves fine-grained part-level distinctions while incorporating object-specific geometric context.

The three feature streams are concatenated to form a comprehensive global conditioning vector:

$$f_t^{\text{global}} = \text{Concat}\big(f_t^g, f_t^r, f_t^p\big). \tag{5}$$

This unified representation integrates scene-level semantics, robot kinematics, and part-level object understanding, providing contextual information for informed action generation.

2) Permutation-Invariant Part Conditioning: The local fields $\mathcal{F}_t^L$ represent unordered collections of object parts where the assignment of parts to indices varies unpredictably across different objects (e.g., $\mathcal{F}_t^{L,1}$ could be either heel or toe for different shoes). Conventional conditioning approaches that rely on concatenation or positional encodings would introduce order-sensitive biases. To address this challenge, we design a permutation-invariant conditioning pipeline that processes part features through attention mechanisms without positional embeddings, as shown in the Fig. 4 (right).

Part Feature Extraction: We encode each local field $\mathcal{F}_t^{L,k}$ using PointNet to extract part-specific features:

$$F_t^p = \{f_t^{p,k}\}_{k=1}^K.$$

Inter-Part Reasoning: We apply self-attention without positional embeddings to enable information exchange between different object parts while preserving permutation invariance across part orderings:

$$\hat{F}_t^p = \mathcal{A}_{\text{self}}(F_t^p). \tag{6}$$

The absence of positional embeddings ensures that the attention mechanism remains invariant to arbitrary permutations of the part set.

Cross-Attention Conditioning: The refined part features are injected into the diffusion U-Net through cross-attention

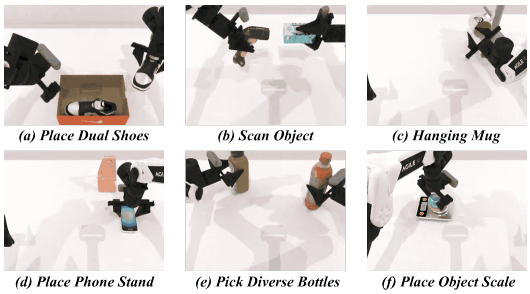

Fig. 5. **Simulated Manipulation Tasks.** We evaluate our method on six challenging tasks from the RoboTwin 2.0 benchmark. These tasks necessitate precise pose estimation and a nuanced understanding of object-part semantics to facilitate successful interaction.

layers, where U-Net features serve as queries and part features serve as keys and values:

$$z_{\text{next}} = z + \mathcal{A}_{\text{cross}}(z, \hat{F}_t^p). \tag{7}$$

This cross-attention mechanism allows the diffusion model to condition on detailed part-level information while maintaining permutation invariance.

This three-stage pipeline enables robust part-level reasoning without sensitivity to arbitrary part orderings, ensuring consistent performance across diverse object configurations.

C. Diffusion Policy Learning

We train a diffusion model to generate robot actions. Given expert actions A_i^*, the model learns to predict noise ϵ added at diffusion timestep t.

The network ϵ_θ uses all extracted features as conditions. The loss is:

$$\mathcal{L} = \mathbb{E}_{A_i^*, c_i, \epsilon, t}\left[\left\|\epsilon - \epsilon_\theta(a_t, t, c_i)\right\|^2\right], \tag{8}$$

where a_t is the noisy action. During inference, we iteratively denoise random Gaussian noise to generate actions conditioned on the extracted features.

IV. EXPERIMENTS

We conduct a series of experiments to validate our proposed method. Through quantitative benchmarks and qualitative analysis, we aim to answer the following questions:

(1) **Semantic Requirements:** How critical is semantic understanding for pose-aware manipulation tasks?

(2) **Semantic Representation Quality:** Why does our semantic representation outperform G3Flow?

(3) **Fine-grained Object Perception:** What is the specific contribution of fine-grained geometric perception to policy performance in pose-sensitive scenarios?

A. Experimental Setup

Tasks. We build upon the multi-object tasks(Place Dual Shoes, Pick Diverse Bottles) from G3Flow [13] and further select several challenging tasks from RoboTwin 2.0 [33] that require precise object pose alignment or involve complex object interactions, as shown in Fig. 5. We also selected three of these tasks for cross-object generalization experiments, real-world experiments validation, and ablation.

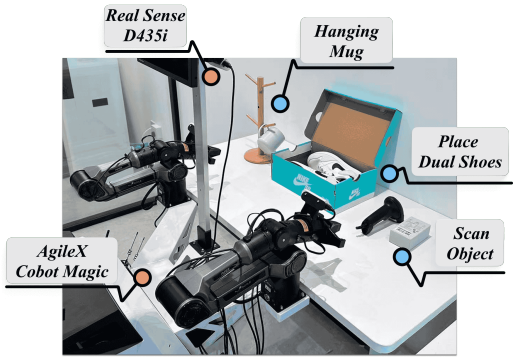

Fig. 6. **Real-World Experimental Setup.** Our experimental setup consists of an AgileX Cobot Magic dual-arm robot equipped with a RealSense D435i head-mounted camera.

Baselines. We compare our method with several state-of-the-art approaches to validate the effectiveness: **G3Flow** [13] is a foundation model-driven approach that constructs semantic flow by integrating 3D generative models, vision foundation models, and pose tracking, yielding dynamic object-centric representations. **Diffusion Policy (DP)** [6] formulates visuomotor policy learning as a conditional denoising diffusion process, predicting actions from image inputs for robotic manipulation. Building on this, **3D Diffusion Policy (DP3)** [8] introduces a lightweight MLP encoder for sparse point clouds to enable efficient 3D representation learning, with a variant (**DP3 w/ color**) that further incorporates RGB features projected onto the point clouds.

Training and Evaluation Details. For fair comparison, we reproduce G3Flow and retrain all baseline methods in the official RoboTwin 2.0 benchmark codebase. Following best practices from prior work rather than the default RoboTwin 2.0 leaderboard setting (50 demonstrations), we use 100 expert demonstrations for training in simulation enviorment.

All methods are trained with their recommended hyperparameters: 3000 epochs with batch size 256 for our method, DP3 and G3Flow, and 600 epochs with batch size 128 for DP. For standard benchmark tasks, we use all official assets for both training and evaluation. For cross-object generalization, we train on two-thirds of the provided assets and evaluate on the remaining one-third. Following G3Flow [13], we use GroundedSAM [34] for object segmentation and FoundationPose [35] for 6D object tracking. We extract and fuse features from DINOv2 [16] and finetuned Stable Diffusion [19] to construct the semantic point cloud. For real-world experiments, we use an AgileX Cobot Magic dual-arm embodiment equipped with RealSense D435i cameras as illustrated in Fig. 6. We collect 50 expert demonstrations for each task using teleoperation.

Evaluation Protocol. To ensure fair evaluation, we fix random seeds across training and testing. Each method is run with multiple seeds and evaluated on 100 test episodes per seed, reporting mean success rates and standard deviations. In simulation, object positions and poses are randomized following the benchmark settings. For real-world evaluation, we conduct 20 episodes with fixed object positions but randomized rotations. All Inference runs on a single 4090D.

TABLE I

COMPARATIVE EVALUATION ON THE 6 TASKS IN ROBOTWIN 2.0 BENCHMARK WITH STANDARD SETTING.

Method	*Place Dual Shoes*	*Scan Object*	*Hanging Mug*	*Place Phone Stand*	*Pick Diverse Bottles*	*Place Object Scale*	*Average*
DP [6]	7.0 ±3.6	10.3 ±2.1	15.3 ±2.6	37.7 ±6.2	18.3 ±3.1	2.7 ±1.2	15.2
DP3 [8]	17.7 ±3.4	23.3 ±1.2	23.3 ±6.3	52.0 ±2.9	27.7 ±7.0	10.7 ±3.9	25.8
DP3 w/ color [8]	16.3 ±2.9	14.6 ±1.2	21.7 ±5.6	41.7 ±6.1	17.0 ±5.1	4.7 ±1.2	19.3
G3Flow [13]	20.7 ±3.4	22.0 ±3.7	26.7 ±2.5	45.3 ±3.3	32.0 ±8.0	7.7 ±0.9	25.7
Ours	**33.0** ±5.7	**26.7** ±0.5	**31.0** ±1.4	**55.3** ±1.9	**34.7** ±3.8	**11.7** ±1.2	**32.3**

TABLE II

CROSS-OBJECT GENERALIZATION RESULTS (SUCCESS RATES IN %).

Method	*Place Dual Shoes*	*Scan Object*	*Hanging Mug*	*Average*
DP [6]	7.3 ±2.1	4.0 ±2.2	5.0 ±2.2	5.4
DP3 [8]	11.3 ±3.4	11.3 ±1.2	22.7 ±1.9	15.1
DP3 w/ color [8]	10.0 ±1.4	8.3 ±1.2	19.7 ±4.2	12.7
G3Flow [13]	16.3 ±2.9	11.7 ±2.6	25.0 ±3.7	17.7
Ours	**28.3** ±5.3	**14.3** ±0.5	**30.7** ±4.2	**24.4**

TABLE III

REAL-WORLD EXPERIMENT RESULTS (SUCCESS RATES IN %).

Method	*Place Dual Shoes*	*Scan Object*	*Hanging Mug*	*Average*
DP3 [8]	5	15	0	6.7
G3Flow [13]	10	30	10	16.7
Ours	**25**	**35**	**20**	**26.7**

B. Comparision of the State-of-the-Art

We conduct three distinct experiments against state-of-the-art baselines: a standard benchmark evaluation, a cross-object generalization test and real-world validation.

Standard Benchmark Performance. We first evaluation assesses performance under a standard benchmark protocol, wherein all methods are trained and tested on the same set of object assets. This "closed-set" configuration measures a policy's ability to master tasks within a known environment. As presented in Table I, our method establishes a new state-of-the-art, achieving an impressive average success rate of 32.3%. This result represents a significant 6.6% improvement over G3Flow, the strongest baseline. The superiority of our approach becomes particularly evident in tasks demanding precise semantic alignment. For instance, in *Place Dual Shoes* and *Hanging Mugs*, our method surpasses G3Flow by a substantial margin of 12.3% and 4.3%, respectively. This large performance gap highlights the efficacy of our hierarchical semantic conditioning, which excels at capturing and leveraging fine-grained, part-level object details. In contrast, while DP3-based methods show reasonable performance on geometrically simpler tasks like *Place Phone Stand*, their success rates decline sharply on more complex, pose-sensitive scenarios. This outcome confirms our hypothesis that relying solely on 3D geometric information is insufficient; explicit semantic reasoning is indispensable for achieving robust and precise manipulation.

Generalization to Unseen Objects. More rigorous evaluation specifically tests zero-shot generalization to novel objects. In this "open-set" protocol, policies are evaluated on object instances that were explicitly held out from the training set. This setup is a critical test of a model's ability to abstract and transfer knowledge beyond mere memorization. As shown in Table II, our method again demonstrates superior capabilities, consistently outperforming all baselines with an average success rate of 24.4%, which is a notable 6.7% lead over G3Flow. This success stems from our hierarchical semantic representation, which learns to abstract functional and geometric properties rather than overfitting to the specific visual characteristics of the training instances. For example, in the challenging *Place Dual Shoes* task, our method achieves a 12% higher success rate than G3Flow. This result underscores its ability to comprehend and align abstract semantic parts, such as the toe of one shoe to the heel of another, even on previously unseen models. Conversely, the performance of all baseline methods deteriorates significantly in this setting. Their inability to handle variations in object geometry and appearance highlights a critical reliance on memorizing the training data, rather than acquiring a truly generalizable manipulation skill.

Real-World Validation. We also conducted experiments entirely in the real-world setup where policies were trained using data collected directly from our robotic embodiment via teleoperation, and subsequently evaluated on the same hardware. As detailed in Table III, our method demonstrates strong and reliable performance, achieving the highest success rates across all evaluated tasks. This outcome is particularly significant as it validates that our hierarchical semantic representation is not only effective in theory but also robust enough to handle the nuances of a non-simulated environment. The superior performance compared to the baselines underscores our model's ability to learn meaningful and actionable policies from real-world data.

C. Analysis with visualization.

To address our second research question regarding the superior quality of our semantic representation, we conduct a two-part qualitative analysis. We first compare the execution of our policy against G3Flow on representative tasks, then delve into a detailed examination of the underlying semantic feature fields to diagnose the performance gap.

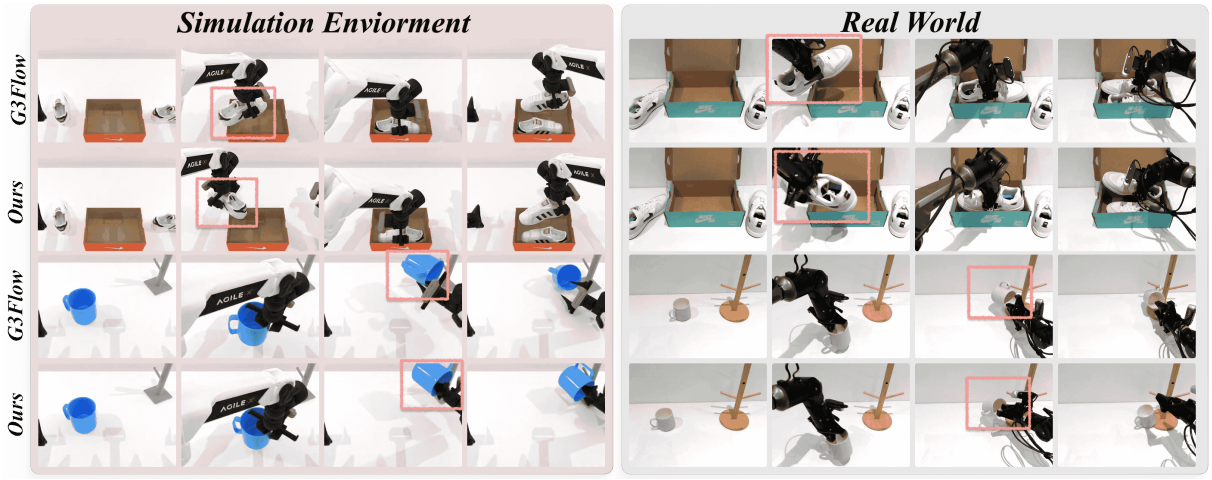

Fig. 7. **Qualitative comparison of policy execution.** For *Place Dual Shoes* (left), G3Flow executes an erroneous rotation, resulting in task failure, whereas our method achieves correct alignment. For *Hanging Mug* (right), G3Flow fails to align the handle with the rack. In contrast, our policy's fine-grained semantic representation enables precise handle orientation and successful task completion, both in simulated environments and on the physical robot.

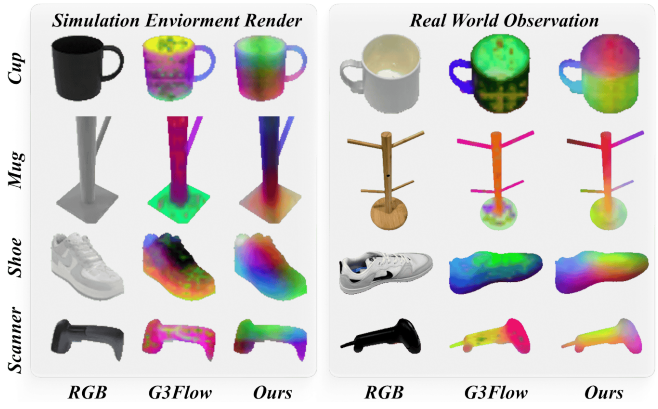

Fig. 8. **Visualization of object semantics field.** We visualize the semantics field of G3Flow and our method in simulation versus real-world data. G3Flow exhibit significant noise and geometric inconsistency. In contrast, our method generates smooth, coherent semantic fields that robustly delineate functional parts across both domains, providing a more stable representation for the policy.

Qualitative Policy Performance. We first visualize policy rollouts on challenging, pose-aware tasks in Fig. 7. The comparison reveals G3Flow's critical limitations, which stem from its inability to explicitly recognize part-level semantics. For instance, in the *Place Dual Shoes* task, G3Flow's holistic representation fails to distinguish between the shoe's toe and heel, leading to an erroneous rotation and an incorrect final placement that results in task failure. Similarly, in the *Hanging Mug* task, G3Flow is unable to isolate the handle as a key functional part, causing the policy to grasp the mug's body and fail to securely hook it onto the rack. In stark contrast, our policy consistently succeeds by leveraging a fine-grained understanding of object geometry and semantics. It correctly identifies the distinct parts of the shoe for precise alignment and isolates the mug handle for a successful grasp and hang, demonstrating a significantly enhanced capability for precise, real-world manipulation.

Analysis of Semantic Representation. To further diagnose the performance gap, we analyze the underlying semantic representations that drive policy decisions in Fig. 8. The quality of these learned dense object semantic field is crucial, as they form the basis for the policy's understanding of object structure and orientation. The visualizations reveal that G3Flow's semantic field are often noisy and inconsistent. The color gradients, representing semantic features, appear fragmented and do not align coherently with the object's underlying geometry, indicating a confused representation. This instability is exacerbated in real-world conditions, where lighting and texture variations lead to further degradation. Our method, however, produces significantly smoother and more geometrically consistent semantics. The color gradients transition logically across surfaces, clearly delineating functional parts like the cup handle or shoe toe with uniform and distinct feature representations, both in rendered simulation objects and real-world images. This stability and accuracy provide our policy with a reliable and unambiguous foundation for executing precise manipulation, directly explaining its superior performance.

TABLE IV

ABLATION STUDY OF MODEL COMPONENTS. WE EVALUATE THE CONTRIBUTION OF EACH KEY MODULE BY MEASURING THE AVERAGE SUCCESS RATE (%) ACROSS 3 BENCHMARK TASKS.

Dense Semantic	Global Pose-aware Condition	Part-aware Geometry Refine	Average (%)
			23.1
✓			23.7
	✓		24.3
		✓	27.6
✓	✓	✓	**30.2**

D. Ablation Study

To answer our third research question regarding the importance of fine-grained perception, we conducted an ablation study as shown in Table IV. The results clearly indicate that the *Part-aware Geometry Refinement* module is the most critical component. Adding it to the baseline model yields the largest performance gain, from 23.1% to 27.6%. This directly validates that a fine-grained understanding of object

geometry is essential for precise, pose-aware manipulation. While other components like semantic features and pose conditioning provide benefits, it is the combination of all modules that achieves the highest success rate of 30.2%, demonstrating a powerful synergy between them.

V. CONCLUSIONS

We introduced **HeRO**, a framework for part-level semantic perception in pose-aware object manipulation. By fusing discriminative DINOv2 features with globally coherent Stable Diffusion features, HeRO constructs dense 3D semantic fields that maintain both geometric precision and semantic consistency. The Hierarchical Conditioning Module effectively integrates global context with permutation-invariant, part-aware features, enabling diffusion-based policies to leverage structured hierarchical information for precise manipulation. Extensive experiments demonstrate HeRO outperforms prior methods, achieving state-of-the-art performance.

ACKNOWLEDGMENT

This work was supported in part by National Natural Science Foundation of China under grant No.62372091 and in part by Hainan Province Science and Technology Special-Fund under grant No. ZDYF2024(LALH)001.

REFERENCES

[1] A. Goyal, V. Blukis, J. Xu, Y. Guo, Y.-W. Chao, and D. Fox, "Rvt2: Learning precise manipulation from few demonstrations," *Robotics: Science and Systems*, 2024.

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

[3] Y. Wang and E. Johns, "One-shot dual-arm imitation learning," in *IEEE International Conference on Robotics and Automation*, 2025.

[4] K. Wu, N. Liu, Z. Zhao, D. Qiu, J. Li, Z. Che, Z. Xu, and J. Tang, "Learning from imperfect demonstrations with self-supervision for robotic manipulation," in *IEEE International Conference on Robotics and Automation*, 2025.

[5] Y. Dai, J. Lee, N. Fazeli, and J. Chai, "Racer: Rich language-guided failure recovery policies for imitation learning," in *IEEE International Conference on Robotics and Automation*, 2025.

[6] C. Chi, Z. Xu, S. Feng, E. Cousineau, Y. Du, B. Burchfiel, R. Tedrake, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *The International Journal of Robotics Research*, 2023.

[7] S. Liu, L. Wu, B. Li, H. Tan, H. Chen, Z. Wang, K. Xu, H. Su, and J. Zhu, "RDT-1B: a Diffusion Foundation Model for Bimanual Manipulation," in *International Conference on Representation Learning*, 2025.

[8] 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 *Robotics: Science and Systems*, 2024.

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

[10] A. Wilcox, M. Ghanem, M. Moghani, P. Barroso, B. Joffe, and A. Garg, "Adapt3R: Adaptive 3D scene representation for domain transfer in imitation learning," *arXiv preprint arXiv:2503.04877*, 2025.

[11] Y. Wang, G. Yin, B. Huang, T. Kelestemur, J. Wang, and Y. Li, "GenDP: 3D semantic fields for category-level generalizable diffusion policy," in *Conference on Robot Learning*, 2024.

[12] Q. Zhang, Z. Liu, H. Fan, G. Liu, B. Zeng, and S. Liu, "Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[13] T. Chen, Y. Mu, Z. Liang, Z. Chen, S. Peng, Q. Chen, M. Xu, R. Hu, H. Zhang, X. Li, *et al.*, "G3flow: Generative 3d semantic flow for pose-aware and generalizable object manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[14] Y. Jia, J. Liu, S. Chen, C. Gu, Z. Wang, L. Luo, X. Li, P. Wang, Z. Wang, R. Zhang, and S. Zhang, "Lift3d policy: Lifting 2d foundation models for robust 3d robotic manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[15] Y. Wang, M. Zhang, Z. Li, T. Kelestemur, K. Driggs-Campbell, J. Wu, L. Fei-Fei, and Y. Li, "D^3fields: Dynamic 3d descriptor fields for zero-shot generalizable rearrangement," in *Conference on Robot Learning*, 2024.

[16] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, *et al.*, "Dinov2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2023.

[17] J. Zhang, C. Herrmann, J. Hur, L. Polania Cabrera, V. Jampani, D. Sun, and M.-H. Yang, "A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence," *Advances in Neural Information Processing Systems*, 2023.

[18] L. Tang, M. Jia, Q. Wang, C. P. Phoo, and B. Hariharan, "Emergent correspondence from image diffusion," *Advances in Neural Information Processing Systems*, 2023.

[19] N. Stracke, S. A. Baumann, K. Bauer, F. Fundel, and B. Ommer, "CleanDiFT: Diffusion features without noise," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[20] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2022.

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

[22] Y. Wang, J. Zhou, H. Zhu, W. Chang, Y. Zhou, Z. Li, J. Chen, J. Pang, C. Shen, and T. He, "π^3: Scalable permutation-equivariant visual geometry learning," *arXiv preprint arXiv:2507.13347*, 2025.

[23] W.-D. Chang, F. Hogan, S. Fujimoto, D. Meger, and G. Dudek, "Generalizable imitation learning through pre-trained representations," in *IEEE International Conference on Robotics and Automation*, 2025.

[24] S. Xia, H. Fang, C. Lu, and H.-S. Fang, "Cage: Causal attention enables data-efficient generalizable robotic manipulation," in *IEEE International Conference on Robotics and Automation*, 2025.

[25] A. Xie, L. Lee, T. Xiao, and C. Finn, "Decomposing the generalization gap in imitation learning for visual robotic manipulation," in *IEEE International Conference on Robotics and Automation*, 2024.

[26] J. Sheng, Z. Wang, P. Li, and M. Liu, "MP1: MeanFlow tames policy learning in 1-step for robotic manipulation," *arXiv preprint arXiv:2507.10543*, 2025.

[27] T. Nguyen, Z. Wang, and E. Kyrkjebø, "Second-order flow matching for smooth trajectory generation," in *IEEE International Conference on Robotics and Automation*, 2025.

[28] Z. Weng, H. Lu, D. Kragic, and J. Lundell, "Dexdiffuser: Generating dexterous grasps with diffusion models," *IEEE Robotics and Automation Letters*, 2024.

[29] Z. Liang, Y. Mu, H. Ma, M. Tomizuka, M. Ding, and P. Luo, "Skilldiffuser: Interpretable hierarchical planning via skill abstractions in diffusion-based task execution," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2024.

[30] Z. Liang, Y. Mu, Y. Wang, T. Chen, W. Shao, W. Zhan, M. Tomizuka, P. Luo, and M. Ding, "Dexhanddiff: Interaction-aware diffusion planning for adaptive dexterous manipulation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[31] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, 2021.

[32] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis," in *International Conference on Representation Learning*, 2024.

[33] Y. Mu, T. Chen, Z. Chen, S. Peng, Z. Lan, Z. Gao, Z. Liang, Q. Yu, Y. Zou, M. Xu, *et al.*, "Robotwin: Dual-arm robot benchmark with generative digital twins," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

[34] T. Ren, S. Liu, A. Zeng, J. Lin, K. Li, H. Cao, J. Chen, X. Huang, Y. Chen, F. Yan, Z. Zeng, H. Zhang, F. Li, J. Yang, H. Li, Q. Jiang, and L. Zhang, "Grounded SAM: Assembling open-world models for diverse visual tasks," *arXiv preprint arXiv:2401.14159*, 2024.

[35] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6D pose estimation and tracking of novel objects," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2024.