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

SE(3)-PoseFlow: Estimating 6D Pose Distributions for Uncertainty-Aware Robotic Manipulation

Yufeng Jin[1,2], Niklas Funk[1], Vignesh Prasad[1], Zechu Li[1], Mathias Franzius[2], Jan Peters[1,3,4], Georgia Chalvatzaki[1,4]

***Abstract*— Object pose estimation is a fundamental problem in robotics and computer vision, yet it remains challenging due to *partial observability*, *occlusions*, and *object symmetries*, which inevitably lead to *pose ambiguity* and multiple hypotheses consistent with the same observation. While deterministic deep networks achieve impressive performance under well-constrained conditions, they are often overconfident and fail to capture the multi-modality of the underlying pose distribution. To address these challenges, we propose a probabilistic framework that leverages *flow matching on the SE(3) manifold* for estimating 6D object pose distributions. Unlike existing methods that regress a single deterministic output, our approach models the full pose distribution with a sample-based estimate and enables reasoning about *uncertainty* in ambiguous cases such as symmetric objects or severe occlusions. We achieve state-of-the-art results on REAL275, YCB-V and LM-O, and demonstrate how our sample-based pose estimates can be leveraged in downstream robotic manipulation tasks such as active perception for disambiguating uncertain viewpoints, or guiding grasp synthesis in an uncertainty-aware manner.**

***Index Terms*— Object Pose Uncertainty Estimation, Flow Matching, SE(3) Manifold**

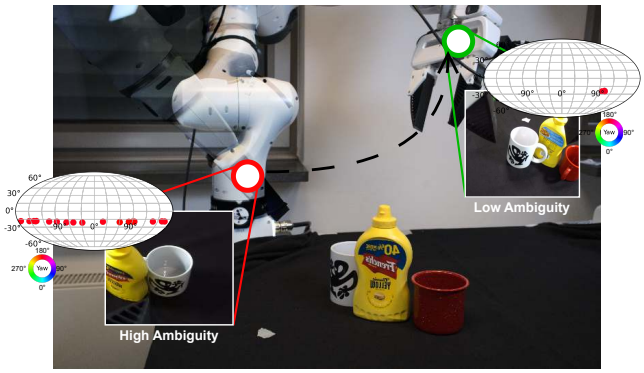

Fig. 1: We propose an uncertainty-aware 6D object pose estimation approach based on SE(3) flow matching. Our probabilistic framework predicts full 6D pose distributions to handle ambiguities, enabling reliable robotic manipulation under challenging real-world conditions (partial observability, occlusions, and symmetries). SO(3) distributions are visualized on a Mollweide projection, where latitude (pitch) and longitude (roll) map the orientation, and color encodes yaw.

I. Introduction

Estimating the 6D pose of objects is a fundamental problem in robotics, as it enables embodied agents to perceive, manipulate, and interact safely with their environment. In practical applications such as robotic grasping, assembly, and human–robot collaboration, it is not sufficient to output a single deterministic pose estimate. Instead, reasoning about *uncertainty* is critical for ensuring safe and reliable manipulation [1], [2]. Probabilistic models that capture the multi-modality of pose distributions provide richer information than deterministic pose estimates, especially in safety-critical manipulation scenarios where downstream decisions rely on calibrated confidence.

A central challenge in 6D object pose estimation arises from *pose ambiguity*. Symmetries in object geometry, partial observability, and occlusions often yield multiple feasible poses that are indistinguishable from sensor observations [3]–[7]. Deterministic deep learning approaches such as FoundationPose [8] have recently advanced the state of the art by leveraging large-scale synthetic training and transformer-based architectures, but they remain limited in their ability to represent multi-hypothesis pose distributions. As a result, they can be over-confident in ambiguous cases, which is undesirable for robotic planning and control.

To address these limitations, recent research has explored *probabilistic formulations of object pose estimation*. These methods aim to model the full distribution of feasible poses rather than commit to a single prediction. Early approaches leverage *directional probability distributions* such as the von Mises–Fisher or Bingham distribution to represent rotational uncertainty [9]–[11]. While theoretically principled, such parametric models are typically unimodal and require mixtures to capture multi-modal ambiguities, leading to computational inefficiency and numerical instability. More recently, *generative models* based on diffusion [3], [6], [7] and normalizing flows [12] have been proposed to directly model complex distributions on SE(3) through sample-based estimates. These approaches naturally capture multi-modality and uncertainty, but often rely on intermediate representations or remain constrained to synthetic benchmarks. This motivates the development of new methods that combine the scalability of modern architectures with principled probabilistic modeling on the SE(3) manifold for safe and robust robotics. Our main contributions are summarized as follows:

- We propose a probabilistic framework based on *flow matching on the* SE(3) *manifold* for 6D object pose es-

Corresponding author: Yufeng Jin (yufeng.jin@tu-darmstadt.de).
This work was supported by the German Research Foundation (DFG) Emmy Noether Programme (CH 2676/1-1), the EU's Horizon Europe project "ARISE" (Grant no.: 101135959), the German Federal Ministry of Education and Research (BMBF) Project "RiG" (Grant no.: 16ME1001)
[1]Department of Computer Science, TU Darmstadt, Germany.
[2]Honda Research Institute Europe GmbH, Offenbach, Germany.
[3]DFKI, Research Department SAIROL, Darmstadt, Germany.
[4]Hessian.AI, Darmstadt, Germany

timation. Our method provides a sample-based estimate of the pose distribution that naturally captures uncertainty in ambiguous cases such as object symmetries or severe occlusions.
- We introduce an adapted DiT module with masked cross-attention into the $\mathrm{SE}(3)$ flow model, which improves robustness under occlusions and cluttered real-world scenes, thereby achieving competitive state-of-the-art performance across multiple benchmarks.
- We demonstrate how the learned $\mathrm{SE}(3)$ distribution can be leveraged for downstream robotic tasks, such as guiding active perception to resolve viewpoint ambiguity, and enabling reliable and effective single-view grasp generation under partial observability.

II. Related Work

A. Probabilistic Object Pose Estimation

Recent work has sought to overcome the limits of deterministic regressors by explicitly modeling uncertainty in object pose estimation. One approach leverages *directional probability distributions* for rotational uncertainty: the von Mises–Fisher distribution for Euler angles [9], and the Bingham distribution for unit quaternions [10], [11], [13], [14]. These models handle symmetries well but have drawbacks: (i) computing normalization constants on non-Euclidean manifolds is costly, (ii) they are unimodal and require mixtures for multi-modality, risking mode collapse, and (iii) parameterizations can be unstable and scale poorly.

Beyond closed-form distributions, *generative probabilistic models* better capture complex, multi-modal pose distributions. DiffusionNOCS [6] uses image-to-image diffusion to predict NOCS maps aligned with depth, naturally handling symmetry but incurring inference overhead. GenPose [7] applies score-based diffusion on point clouds, sampling multiple hypotheses but perturbing $\mathrm{SO}(3)$ with Gaussian noise and requiring an auxiliary energy network for likelihood estimation. Möller et al. [15] adopt a particle-based diffusion formulation for point clouds, which discards texture cues important for fine-grained alignment. More theoretically, Hsiao et al. [3] and Liu et al. [12] study diffusion and normalizing flows directly on $\mathrm{SO}(3)$, showing promise for synthetic benchmarks but not extending to real-world robotics.

In summary, probabilistic methods capture pose ambiguity and uncertainty better than deterministic ones but often trade efficiency or generality for expressiveness, motivating scalable methods that connect theoretical advances with robotic deployment.

B. Flows on Manifolds

Flow matching [16] has emerged as an alternative to diffusion for generative learning. It trains continuous normalizing flows (CNFs) by regressing vector fields along probability paths, yielding simulation-free training, closed-form objectives, and faster inference. Chen and Lipman [17] extended this to Riemannian manifolds via *Riemannian Flow Matching*, which generalizes conditional flow matching using geodesic or spectral premetrics.

Several works apply manifold-aware flow matching across domains. In robotics, Braun et al. [18] introduced Riemannian Flow Matching Policies for efficient motion generation, while Funk et al. [19] and Zhang and Gienger [20] demonstrated SE(3)-equivariant flows for action and affordance learning. Beyond robotics, Miller et al. [21] used it for crystalline material discovery, and SE(3)-flow matching has been applied to protein backbone generation [22], [23].

These studies show the versatility of manifold-aware flows in robotics, materials science, and biology. However, applications on $\mathrm{SO}(3)$ or $\mathrm{SE}(3)$ remain largely confined to synthetic or simulation settings [3], [12]. To our knowledge, our work is the first to employ flow matching on $\mathrm{SE}(3)$ for *real-world 6D object pose estimation*, connecting manifold generative modeling with practical robotic deployment.

III. Method

Given an RGB-D input, our goal is to provide a sample-based estimate of the 6D object pose distribution rather than a single deterministic solution. Objects are localized using off-the-shelf detectors such as Mask R-CNN [24] or CNOS [25], from which object-centric RGB crops and partial point clouds are extracted. These observations are encoded by geometric and visual encoders and fused with DiT* blocks to drive conditional flow matching on the $\mathrm{SE}(3)$ manifold (Sec. III-A). This formulation enables efficient training, probabilistic sampling of multi-modal pose hypotheses, and naturally extends to pose tracking (Sec. III-B). For pose selection (Sec. III-C), we introduce two complementary strategies: a *model-free clustering approach* that aggregates the sample-based hypotheses into consensus modes, and a *model-based geometric scoring* that ranks hypotheses by their agreement with the 3D object model. Finally, we show in Sec. III-D how the learned $\mathrm{SE}(3)$ distributions enable active perception and uncertainty-aware downstream tasks, in particular *grasp planning under ambiguity*, leading to safer and more reliable robotic manipulation.

A. Overall Pipeline

Our framework integrates a dual-stream encoder for visual and geometric features, DiT* blocks with masked cross-attention, and an $\mathrm{SE}(3)$ flow matching module (see Fig. 2). The image stream employs a pretrained DINOv2 ViT [26] to extract semantic patch embeddings, with the backbone kept frozen during training. In parallel, the point cloud stream uses a PointNet++ [27] encoder, trained from scratch to preserve fine-grained spatial structure. To stabilize training, we first normalize partial point clouds by shifting them to zero-mean, which removes dependence on absolute camera-frame translations and instead emphasizes relative offsets in the object frame, thereby avoiding training collapse. Both modalities are projected into a shared 256-dimensional feature space. The continuous flow timestamp t is encoded via Fourier features and injected through adaptive layer normalization to condition the attention blocks.

In place of the original DiT blocks [28], which employ self-attention for image-to-image generation, we intro-

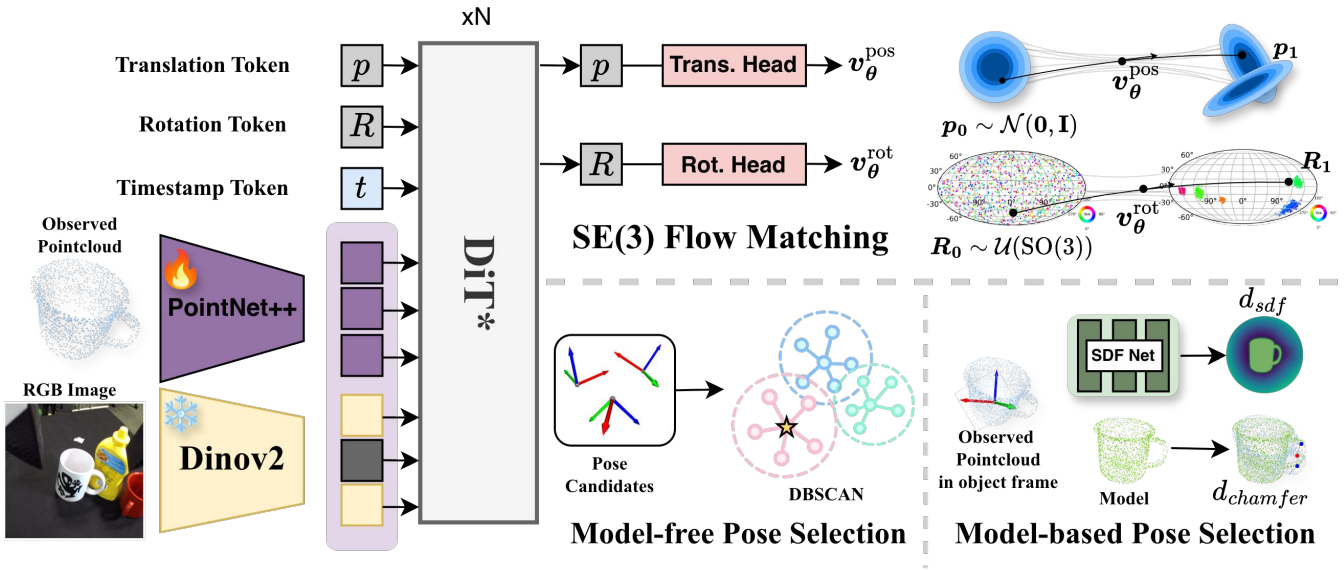

Fig. 2: Overview of SE(3)-PoseFlow. Given an RGB-D input, we extract object-centric RGB crops and partial point clouds using off-the-shelf detectors. The visual and geometric features, together with timestep and sampled poses, are encoded and fused via DiT* blocks with masked cross-attention to predict conditional velocity fields for SE(3) Flow Matching. The framework enables probabilistic sampling of multi-modal pose hypotheses and supports two complementary pose selection strategies: a model-free clustering approach and a model-based geometric scoring.

duce *masked cross-attention blocks*. Separate pose tokens are assigned for translation in $\mathbb{R}^3$ and rotation in SO(3), thereby disentangling the two distributions. The cross-attention mechanism allows pose tokens to learn relations with image and point cloud features. A segmentation-derived binary mask specifies the set of *active tokens* that participate in attention, filtering out background noise. Moreover, cross-attention reduces the computational complexity to linear $O(n)$ rather than quadratic $O(n^2)$.

The refined pose tokens are decoded by translation and rotation heads to yield velocity fields v_θ^{pos} and v_θ^{rot}. These velocity fields parameterize the conditional SE(3) flow for distribution sampling and integration.

B. Flow Matching on SE(3)

We model the conditional pose distribution $p(R,p \mid O,I)$, where $R \in \text{SO}(3)$ and $p \in \mathbb{R}^3$. Here, O denotes the observed point cloud and I the corresponding image. Following the flow-matching framework [16], [17], we employ the *Rectified Linear Flow (RLF)*, which defines a probability path between a random initialization (R_0, p_0) and a target pose (R_1, p_1). The translation is interpolated linearly in $\mathbb{R}^3$, while the rotation follows the geodesic on SO(3):

$$p_t = (1-t)\,p_0 + t\,p_1, \qquad R_t = R_0 \exp\big(t \cdot \log(R_0^\top R_1)\big), \quad (1)$$

for $t \in [0,1]$.

Differentiating the path yields the ground-truth conditional velocity fields for the translation $\dot{p}_t \in \mathbb{R}^3$ and rotation $\dot{r}_t \in \mathbb{R}^3$

$$\dot{p}_t = \tfrac{p_1 - p_t}{1-t}, \qquad \dot{r}_t = \tfrac{1}{1-t}\log(R_t^\top R_1). \quad (2)$$

Here, $\log(\cdot)$ and $\exp(\cdot)$ denote the Lie algebra logarithm and exponential maps on SO(3).

During training, the initial pose (R_0, p_0) is sampled uniformly at random from SE(3), and the network predicts the conditional velocity fields

$$\ell_{\text{pos}} = \big\| v_\theta^{\text{pos}}(p_t \mid O,I,t) - \dot{p}_t \big\|^2, \;\; \ell_{\text{rot}} = \big\| v_\theta^{\text{rot}}(R_t \mid O,I,t) - \dot{r}_t \big\|^2.$$

The overall flow-matching objective is

$$\mathcal{L}_{\text{FM}}(\theta) = \mathbb{E}_{t,(R_0,p_0),(R_1,p_1)}\big[\lambda\,\ell_{\text{pos}} + \ell_{\text{rot}}\big], \quad (3)$$

where λ is a weighting factor between translation and rotation losses, set to $\lambda = 10$ in our experiments to account for their different sensitivity scales.

At inference time, pose samples are generated by starting with a set of randomly sampled pose candidates and integrating the learned vector field from (R_0, p_0) to $t = 1$ for each candidate. This process produces a set of pose hypotheses, which, depending on the degree of pose ambiguity in the observation, may either cluster around a single mode or exhibit greater diversity. For pose tracking, we naturally extend the framework by initializing (R_0, p_0) from the previous estimate instead of starting from random samples, thereby enforcing temporal coherence across frames.

C. Pose Selection

We study two complementary strategies for selecting representative poses: a *model-free* clustering approach in SE(3) and a *model-based* geometric evaluation.

Model-free Pose Selection Given a hypothesis set $\{T_i\}$, we apply DBSCAN [29] with a distance that combines rotational and translational differences:

$$d(T_1,T_2) = \sqrt{\left(\tfrac{\theta(R_1,R_2)}{\varepsilon_R}\right)^2 + \left(\tfrac{\|t_1 - t_2\|}{\varepsilon_t}\right)^2}, \quad (4)$$

where $\theta(\cdot,\cdot)$ denotes the geodesic angle between two rotations. The largest cluster is selected and its representative is computed as the Karcher mean [30] on $\mathfrak{se}(3)$. This approach requires no object model and adapts naturally to multi-modal distributions. In practice we use $\varepsilon_R = 10^\circ$ and $\varepsilon_t = 0.03$cm.

Model-based Pose Selection To further resolve ambiguity, we evaluate pose candidates against the object geometry using two objectives. The *Chamfer loss* [31] measures the one-sided distance from observed points P to the transformed model points M:

$$d_{\text{chamfer}}(P,M) = \frac{1}{|P|}\sum_{p\in P}\min_{m\in T(M)}\|p-m\|_2^2. \tag{5}$$

It is simple and requires no pre-training, but is sensitive to point density and observation noise. The *SDF loss* [32]–[35] employs a learned signed distance function f, with coordinates normalized to $[-1,1]^3$, to compute

$$d_{\text{sdf}}(P,T) = \frac{1}{|P|}\sum_{p\in P} f\big(\text{Norm}(T^{-1}p)\big)^2. \tag{6}$$

This provides continuous and global surface feedback, but requires high-quality meshes or dense point clouds for training, and may fail with sparse supervision.

For both objectives, we convert residuals into log-likelihood scores and retain the top 20% of poses. Chamfer loss thus serves as a lightweight, training-free baseline, while SDF offers stronger geometric consistency when reliable shape supervision is available.

D. Exploiting the Sample-based Pose Estimation for Downstream Tasks

Active Perception In realistic scenarios, the robot's sensors often provide only partial observations due to occlusions or restricted viewpoints. To reduce pose uncertainty, we leverage the covariance of pose hypotheses. Given a set of sampled transformations $\{T_i\}$, we estimate the mean and covariance of translations and rotations, where rotational uncertainty is quantified in the tangent space of $\text{SO}(3)$. Based on this uncertainty, the robot actively selects the next-best viewpoint on an object-centric viewing sphere at a fixed distance. Formally, the next viewpoint v^* is chosen to maximally reduce the predicted rotational covariance:

$$v^* = \arg\min_{v\in\mathcal{V}} \mathbb{E}\big[\text{tr}\big(\Sigma_R \mid v\big)\big], \tag{7}$$

where $\mathcal{V}$ denotes the discrete set of admissible viewpoints, and Σ_R is the covariance of rotations estimated from $\{T_i\}$. Since this objective has no closed-form solution, we approximate it by particle sampling: candidate viewpoints are uniformly sampled on the viewing sphere, their induced pose covariances are evaluated, and the viewpoint with minimal rotational uncertainty is selected as the next best view. This strategy allows the robot to actively move its sensor around the object to disambiguate symmetric configurations and acquire confident pose estimates with minimal exploration cost.

Robotic Grasping In cluttered scenes, grasp planning must explicitly account for uncertainty in object orientation.

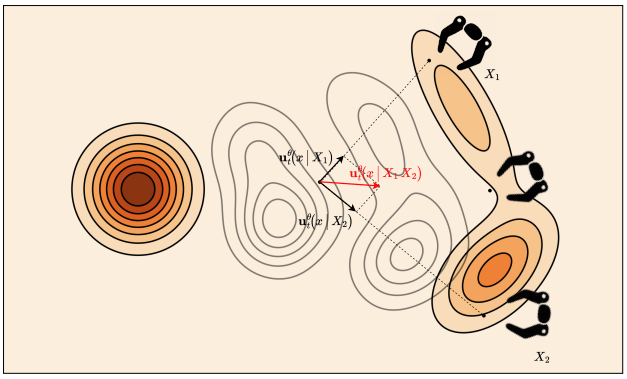

Fig. 3: **Illustrating the mean grasp pose velocity under pose uncertainty.** EquiGraspFlow velocities are averaged per pose hypothesis to form a mean field, which is integrated to sample grasps that are robust to pose ambiguity (e.g., favouring top grasps for a mug with an occluded handle).

We synthesize grasps by marginalizing over the pose candidates at the *velocity* level. Let $M = \{x_i\}_{i=1}^N \subset \mathbb{R}^3$ denote the object point cloud and $\{T_k\}_{k=1}^K \subset \text{SE}(3)$ be pose hypotheses drawn (e.g. by sampling or clustering) from the posterior $p(T \mid O, I)$. For each hypothesis, we transform the model into the world frame and then normalize the transformed cloud into a canonical space:

$$M_k^{(w)} = \{T_k x_i\}_{i=1}^N, \qquad \tilde{M}_k = M_k^{(w)} - \frac{1}{N}\sum_{i=1}^N M_{k,i}^{(w)}.$$

Given the models representing the pose hypothesis, we then leverage a flow-matching-based generative model, i.e., EquiGraspFlow [36], for grasp synthesis. In particular, starting from randomly initialized grasp pose candidates, we obtain the velocity update vectors for every combination of grasp pose and pose hypothesis (represented by the different models) and form the pose (hypothesis) marginal (mean) velocity field:

$$\bar{v}(M,t) \approx \frac{1}{K}\sum_{k=1}^K v_\theta\big(\tilde{M}_k, t\big). \tag{8}$$

Integrating $\bar{v}$ yields grasp samples that are consistent with the multiple pose hypotheses and thereby clustered across high-probability pose modes. For instance, for a mug with an occluded handle (or ambiguous azimuth), averaging velocities on the canonical space suppresses side-grasp modes tied to uncertain yaw and concentrates mass on top grasps that remain valid across the plausible orientations.

IV. Experiments

Dataset We conduct experiments on three widely used benchmarks for 6D object pose estimation. REAL275 [37] contains real RGB-D sequences of 6 object categories with large intra-class variations, which serves as the standard benchmark for evaluating category-level methods. YCB-Video (YCB-V) [38] consists of 92 video sequences of 21 YCB objects in cluttered scenes, widely adopted for instance-level evaluation. LINEMOD-Occlusion (LM-O) [39] con-

tains 8 textureless household objects under heavy occlusion, and is commonly used to benchmark instance-level methods in challenging real-world scenes. Together, these three datasets provide a comprehensive evaluation setting, covering both category-level generalization and challenging instance-level scenarios.

Evaluation Metrics Following GenPose [7] and DiffusionNOCS [6], we evaluate pose accuracy using rotation and translation thresholds. A prediction is considered correct if its rotation error is below α degrees and its translation error is below β centimeters. We report results under the commonly used 5°2cm, 5°5cm, and 10°5cm criteria, averaged across all objects and scenes. For symmetric objects, the minimum geodesic rotation error over the discrete symmetry set is adopted.

Baseline We compare against both deterministic and probabilistic approaches. Deterministic baselines include NOCS [37], DualPoseNet [40] and SPD [41], which are representative methods reported on the NOCS benchmark. For these methods, we report the results listed on the official leaderboard of [7]. Probabilistic baselines include GenPose [7] and DiffusionNOCS [6]. We directly evaluate them using their publicly available implementations and pre-trained checkpoints. Since DiffusionNOCS does not release code for recovering poses from NOCS maps, we follow the protocol described in [37] to estimate object poses from the predicted NOCS maps. For fair comparison, all methods are evaluated on the same test splits under the unified protocol without per-object tuning.

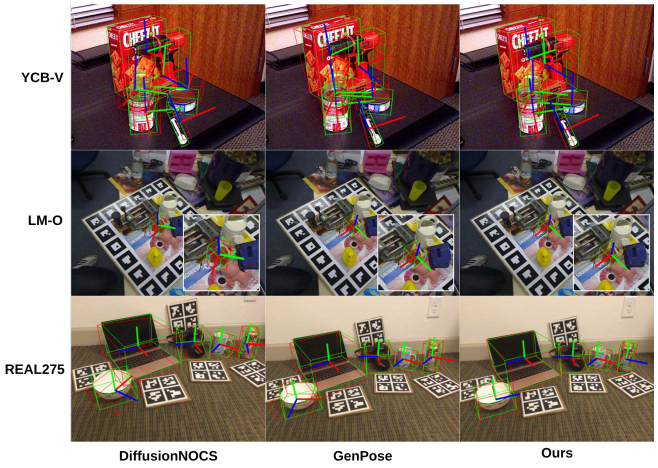

Fig. 4: Qualitative comparison of pose estimation on YCB-V, LM-O and REAL275 datasets.

A. Quantitative Analysis

Results on REAL275 Following the evaluation protocol of GenPose [7], we sample 50 pose hypotheses per object and retain the top 40% according to the pose selection strategy. Since REAL275 does not provide meshes or dense point clouds, SDF-based scoring cannot be applied and Chamfer distance is used for re-ranking. Table I shows that probabilistic methods consistently outperform regression-based

TABLE I: Quantitative comparison of category-level object pose estimation on REAL275 dataset.

	Method	5°2cm↑	5°5cm↑	10°2cm↑	10°5cm↑
Deterministic	NOCS [37]	-	9.5	13.8	26.7
	DualPoseNet [40]	29.3	35.9	50.0	66.8
	SPD [41]	19.3	21.4	43.2	54.1
Probabilistic	DiffusionNOCS [6]	-	35.0	66.6	77.1
	GenPose [7]	**52.1**	**60.9**	72.4	84.0
	Ours	48.8	56.3	**76.3**	**89.1**

TABLE II: Comparison of generative model-based 6D object pose estimation methods on the BOP dataset.

Method	YCB-V		LM-O	
	5°5cm↑	10°5cm↑	5°5cm↑	10°5cm↑
DiffusionNOCS [6]	23.4	54.8	15.5	42.5
GenPose [7]	**46.2**	63.8	32.2	48.2
Ours	45.4	**68.2**	**35.2**	**53.7**

baselines such as NOCS, DualPoseNet, and SPD, underlining the benefit of explicitly modeling multiple pose hypotheses in ambiguous settings. Relative to DiffusionNOCS, our approach achieves higher accuracy across most thresholds. This improvement stems from differences in inference design: DiffusionNOCS predicts normalized object coordinate (NOC) maps from masked RGB input and subsequently computes poses in SE(3), a process that is sensitive to depth–NOC misalignment and mask errors. Our method directly samples and evaluates hypotheses in SE(3), ensuring geometric consistency and yielding more reliable estimates under sparse or noisy depth.

At the strictest thresholds (5°2cm and 5°5cm), our accuracy is slightly lower than GenPose. This is mainly due to the scoring stage: Chamfer distance is sensitive to the sparse and noisy point clouds provided in REAL275, reducing its effectiveness. GenPose instead employs a learned energy network to approximate dataset-specific likelihoods, which is effective on REAL275 but less transferable across datasets.

Results on BOP Datasets The BOP benchmarks (YCB-V and LM-O) provide denser RGB-D input and high-quality mesh supervision, which allows our probabilistic pipeline to fully exploit geometry-aware scoring (Table II). SDF-based re-ranking provides a smoother and more global error landscape than Chamfer distance, improving the separation between valid poses and structurally inconsistent hypotheses. With this strategy, our method achieves the highest accuracy on LM-O and competitive results on YCB-V, outperforming prior probabilistic approaches. These findings support our hypothesis that probabilistic sampling combined with geometry-aware re-ranking improves the handling of pose ambiguity, particularly when strong geometric supervision is available. Qualitative results visualizing pose estimates on REAL275, YCB-V and LM-O are shown in Fig. 4.

TABLE III: Ablation on input modalities and kv masking under two thresholds.

	REAL275		YCB-V		LM-O	
	5°5cm ↑	10°5cm ↑	5°5cm ↑	10°5cm ↑	5°5cm ↑	10°5cm ↑
w/o RGB	**52.3**	77.9	20.9	43.6	**30.5**	46.7
w/o kv mask	47.8	73.8	36.4	51.2	22.8	40.2
ours	51.3	**79.3**	**40.2**	**54.7**	30.1	**48.3**

B. Ablation Study

Input modalities and attention mask We conduct an ablation study to examine the influence of input modalities and the proposed masking mechanism in Sec. III-A. To remove stochasticity, all experiments are performed without pose selection. Table III compares three settings: *W/o RGB*, which employs only point clouds as input; *W/o kv mask*, which removes the masking applied to visual tokens; and *Ours*, which integrates both visual and geometric features with the mask-attention design.

On REAL275, using only point clouds slightly outperforms the RGB-augmented variants. This suggests that visual features are less useful when occlusions are limited and objects are mostly symmetric and texture-less, as is the case in REAL275 where only the camera and laptop categories contain significant textures. On LM-O, performance differences remain marginal for similar reasons. In contrast, on YCB-V, which contains many textured objects, the addition of RGB improves accuracy by more than 10%, highlighting the importance of visual cues in textured scenes. Across all datasets, models with kv mask consistently perform better than those without, indicating that the mask helps suppress background clutter and extract cleaner, more generalizable visual features.

	REAL275		YCB-V	
	5°5cm ↑	10°5cm ↑	5°5cm ↑	10°5cm ↑
None	51.3	79.3	40.2	54.7
Model-free	52.5	88.2	42.5	65.6
Model-based (Chamfer)	56.3	89.1	43.1	67.5
Model-based (SDF)	-	-	**45.4**	**68.2**

TABLE IV: Ablation study on pose selection strategies under two accuracy thresholds on REAL275 and YCB-V.

Pose Selection Probabilistic sampling inevitably generates outlier hypotheses, making robust pose selection essential for reliable performance. We compare four strategies under a unified setting of 50 samples in Table IV: (i) *None*, where poses are taken without selection, (ii) *Model-free*, which evaluates all candidates in the best cluster, (iii) *Model-based (Chamfer)*, ranking poses by Chamfer distance to the object model, and (iv) *Model-based (SDF)*, which scores poses using a neural signed distance function. For model-based methods, we retain the top 40% hypotheses, while the model-free variant evaluates all cluster members.

Model-free selection improves over None by filtering spurious outliers, but remains weaker than model-based approaches due to the lack of geometric priors. On REAL275, no mesh or dense point cloud is available, causing SDF training to collapse; hence results are missing. On YCB-V, SDF-based scoring achieves the best accuracy, confirming its advantage over Chamfer distance in handling small alignment errors. Overall, these results indicate that model-based scoring, particularly with SDF supervision, is key to fully exploiting probabilistic sampling for resolving pose ambiguity.

	Pose Estimation		Pose Tracking		Runtime
Steps	5°5cm ↑	10°5cm ↑	5°5cm ↑	10°5cm ↑	Speed(FPS) ↑
1	7.6	20.6	51.2	83.6	**35.2**
2	44.4	75.7	**57.4**	88.2	21.3
3	52.5	81.3	56.2	86.3	16.7
5	56.3	**89.1**	56.9	87.1	10.1
10	**57.8**	88.6	57.2	**88.8**	4.3

TABLE V: Ablation study on the effect of varying the number of available inference steps for pose estimation and tracking on REAL275.

Inference Steps across Pose Estimation and Tracking We study the effect of varying the number of ODE integration steps on the pose estimation and tracking quality on REAL275 in Table V. For pose estimation, we follow the model-based protocol with 50 pose samples per frame. For tracking, the first frame is initialized from a perturbed ground truth pose (rotation up to 20°, translation up to 5 cm), and subsequent frames use the previous prediction.

Increasing the number of inference steps improves the pose estimation results for up to 5 steps, after which the accuracy saturates while the runtime increases. Tracking, however, achieves strong performance even with a single inference step, as the model only needs to refine a near-correct initialization. These results demonstrate the efficiency of the rectified flow matching objective: accurate results can be obtained within very few steps, owing to its continuous and constant velocity field formulation.

C. Uncertainty-aware Robotic Tasks

Active Perception We validate our active perception strategy on a Franka Panda arm equipped with a ZED Mini RGB-D camera mounted on the wrist. To obtain consistent object masks under varying viewpoints, we employ SAM2 [42] for mask tracking and extend it with dynamic prompts in multi-object scenes. Viewpoints are sampled on an object-centric sphere at a fixed radius, orienting the camera toward the object center. Each candidate viewpoint is scored by the induced pose covariance, and the next-best view is selected as the one minimizing rotational uncertainty, as illustrated in Fig. 1. The robot subsequently executes the chosen motion, actively moving the camera around the object to reduce ambiguity and improve the pose estimate. This setup enables quantitative evaluation of uncertainty reduction and pose accuracy in real-world conditions. In our experiments, we sample 20 candidate viewpoints uniformly on the upper hemisphere at a fixed radius of 0.4 m from the object center,

and evaluate up to 5 successive next-best views per trial. At each step, we draw 50 pose hypotheses from SE(3)-PoseFlow and estimate the rotational covariance in the tangent space of SO(3). We evaluate the strategy on symmetric objects, specifically bowls and bottles from the REAL275 categories, where a single frontal view typically yields high rotational ambiguity. Across these trials, the mean rotational covariance consistently decreases with each additional view, confirming that the uncertainty-driven viewpoint selection actively resolves pose ambiguity. Furthermore, the improvement in pose accuracy saturates after 2–3 views in most cases, suggesting that a small number of targeted observations is sufficient to disambiguate even highly symmetric objects.

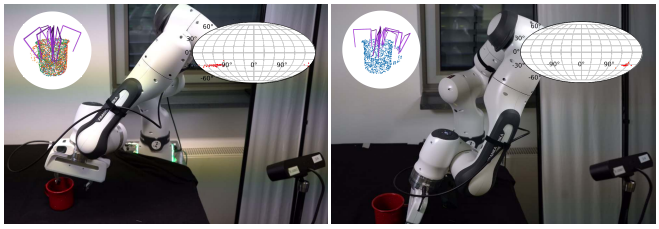

Fig. 5: Uncertainty-aware grasping on a mug. **Left:** Occluded case with a multi-modal sample-based distribution of pose hypotheses; Sampling grasps using EquiGraspFlow while marginalizing over the multiple pose hypotheses generates top-down grasps that remain valid across all pose hypotheses. **Right:** Non-occluded case with a unimodal distribution, i.e., all the samples agree on a single pose hypothesis; Sampling grasps using EquiGraspFlow while marginalizing over the multiple pose hypotheses (which now coincide to one pose) also produces side grasps targeting the handle.

Uncertainty-Aware Grasping Task. The grasping experiments are conducted using an external ZED2 RGB-D camera and a Franka Panda arm controlled through ROS/MoveIt. Grasp poses are generated with EquiGraspFlow [36] as described in Sec. III-D: the observed point cloud is first transformed into the canonical frame, the mean velocity field is then computed over the point cloud, and 10 candidate grasps are sampled from the resulting distribution. The nearest feasible grasp among the candidates is subsequently executed on the robot.

We evaluate two distinct scenarios using a mug as the target object: one where the handle is clearly visible (non-occluded) and another where the handle is self-occluded by the mug body. In the non-occluded case, the pose distribution is unimodal, and the grasp pose generation therefore also yields side grasps that reach for the handle in a natural and consistent manner. In contrast, under self-occlusion the pose distribution becomes multi-modal due to the ambiguous yaw angle of the object, and the proposed grasp sampling strategy automatically shifts to top-down grasps that remain valid across all pose hypotheses regardless of the specific yaw configuration.

As a baseline, we consider grasp generation based on only a single pose hypothesis. While this simplified approach works adequately in the non-occluded case, it fails under occlusion as it continues to propose handle-reaching side grasps that are only valid for a specific assumed pose. Our uncertainty-aware approach, by contrast, consistently selects top-down grasps that succeed across all plausible poses, effectively hedging against pose ambiguity.

Quantitatively, we executed 10 trials per visibility condition on the same mug to ensure a fair and controlled comparison. The results summarized in Table VI show that the baseline variant, *EquiGraspFlow (single)*, achieves a success rate of **75.0%** overall (9/10 in the non-occluded case and 6/10 under occlusion), whereas our uncertainty-aware *EquiGraspFlow (multi)* reaches **95.0%**, performing perfectly when the handle is visible and remaining robust under self-occlusion. This confirms that marginalizing grasp sampling over multiple pose hypotheses substantially improves grasp reliability in the presence of occlusion and pose ambiguity. Qualitative examples of both scenarios are shown in Fig. 5.

	Non-occluded	Occluded	Total (#/20)
EquiGraspFlow (single)	9/10	6/10	15/20 (75.0%)
EquiGraspFlow (multi)	10/10	9/10	**19/20 (95.0%)**

TABLE VI: Real-robot evaluation of uncertainty-aware grasping on a mug. Each cell reports 10 grasp attempts under two visibility conditions: **Non-occluded** (handle visible, unimodal distribution) and **Occluded** (handle self-occluded, multi-modal yaw ambiguity). The baseline uses a single pose hypothesis, while ours marginalizes over multiple pose hypotheses.

V. Conclusion

We presented a probabilistic framework for 6D object pose estimation based on SE(3) flow matching. Unlike deterministic regressors, our method generates sample-based hypotheses that capture multi-modality and calibrated uncertainty, which is crucial for handling symmetries, occlusions, and partial observability. The integration of visual and geometric cues through DiT blocks with masked cross-attention enables robust performance across challenging benchmarks. We further showed that the resulting pose hypotheses can be directly exploited in downstream robotics tasks such as active perception and uncertainty-aware grasp generation.

Our approach still has limitations. It does not yet generalize seamlessly to all object categories, and the modality gap between images and point clouds makes a unified representation difficult—point maps may provide a promising alternative. Moreover, the framework is sample-based, and principled Bayesian utilization of these samples remains an open question. While our real-robot experiments demonstrate promising results, the current evaluation is limited to tabletop scenarios with a small set of object categories, and extending to more diverse geometries, deformable objects, and dynamic environments remains important future work. Future work will also address multi-object scenes and long-horizon manipulation, and explore sequential Bayesian methods such as particle filtering for robust online tracking.

REFERENCES

[1] J. Lee, M. Lee, and D. Lee, "Uncertain pose estimation during contact tasks using differentiable contact features," *arXiv preprint arXiv:2305.16778*, 2023.

[2] J. Michaux, P. Holmes, B. Zhang, C. Chen, B. Wang, S. Sahgal, T. Zhang, S. Dey, S. Kousik, and R. Vasudevan, "Can't touch this: Real-time, safe motion planning and control for manipulators under uncertainty," *IEEE Transactions on Robotics*, 2025.

[3] T.-C. Hsiao, H.-W. Chen, H.-K. Yang, and C.-Y. Lee, "Confronting ambiguity in 6d object pose estimation via score-based diffusion on se (3)," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 352–362.

[4] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold *et al.*, "Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3364–3372.

[5] F. Manhardt, D. M. Arroyo, C. Rupprecht, B. Busam, T. Birdal, N. Navab, and F. Tombari, "Explaining the ambiguity of object detection and 6d pose from visual data," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6841–6850.

[6] T. Ikeda, S. Zakharov, T. Ko, M. Z. Irshad, R. Lee, K. Liu, R. Ambrus, and K. Nishiwaki, "Diffusionnocs: Managing symmetry and uncertainty in sim2real multi-modal category-level pose estimation," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7406–7413.

[7] J. Zhang, M. Wu, and H. Dong, "Generative category-level object pose estimation via diffusion models," *Advances in Neural Information Processing Systems*, vol. 36, pp. 54 627–54 644, 2023.

[8] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 17 868–17 879.

[9] S. Prokudin, P. Gehler, and S. Nowozin, "Deep directional statistics: Pose estimation with uncertainty quantification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 534–551.

[10] H. Deng, M. Bui, N. Navab, L. Guibas, S. Ilic, and T. Birdal, "Deep bingham networks: Dealing with uncertainty and ambiguity in pose estimation," *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1627–1654, 2022.

[11] I. Gilitschenski, R. Sahoo, W. Schwarting, A. Amini, S. Karaman, and D. Rus, "Deep orientation uncertainty learning based on a bingham loss," in *International conference on learning representations*, 2019.

[12] Y. Liu, H. Liu, Y. Yin, Y. Wang, B. Chen, and H. Wang, "Delving into discrete normalizing flows on so (3) manifold for probabilistic rotation modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 21 264–21 273.

[13] J. Glover, G. Bradski, and R. B. Rusu, "Monte carlo pose estimation with quaternion kernels and the bingham distribution," in *Robotics: science and systems*, vol. 7, 2012, p. 97.

[14] B. Okorn, M. Xu, M. Hebert, and D. Held, "Learning orientation distributions for object pose estimation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 580–10 587.

[15] C. Möller, N. Funk, and J. Peters, "Particle-based 6d object pose estimation from point clouds using diffusion models," *arXiv preprint arXiv:2412.00835*, 2024.

[16] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," *arXiv preprint arXiv:2210.02747*, 2022.

[17] R. T. Chen and Y. Lipman, "Flow matching on general geometries," *arXiv preprint arXiv:2302.03660*, 2023.

[18] M. Braun, N. Jaquier, L. Rozo, and T. Asfour, "Riemannian flow matching policy for robot motion learning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 5144–5151.

[19] N. Funk, J. Urain, J. Carvalho, V. Prasad, G. Chalvatzaki, and J. Peters, "Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching," *arXiv preprint arXiv:2409.04576*, 2024.

[20] F. Zhang and M. Gienger, "Affordance-based robot manipulation with flow matching," *arXiv preprint arXiv:2409.01083*, 2024.

[21] B. K. Miller, R. T. Chen, A. Sriram, and B. M. Wood, "Flowmm: Generating materials with riemannian flow matching," *arXiv preprint arXiv:2406.04713*, 2024.

[22] A. J. Bose, T. Akhound-Sadegh, G. Huguet, K. Fatras, J. Rector-Brooks, C.-H. Liu, A. C. Nica, M. Korablyov, M. Bronstein, and A. Tong, "Se (3)-stochastic flow matching for protein backbone generation," *arXiv preprint arXiv:2310.02391*, 2023.

[23] G. Huguet, J. Vuckovic, K. Fatras, E. Thibodeau-Laufer, P. Lemos, R. Islam, C. Liu, J. Rector-Brooks, T. Akhound-Sadegh, M. Bronstein *et al.*, "Sequence-augmented se (3)-flow matching for conditional protein generation," *Advances in neural information processing systems*, vol. 37, pp. 33 007–33 036, 2024.

[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[25] V. N. Nguyen, T. Groueix, G. Ponimatkin, V. Lepetit, and T. Hodan, "Cnos: A strong baseline for cad-based novel object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2134–2140.

[26] 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," *arXiv preprint arXiv:2304.07193*, 2023.

[27] 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*, vol. 30, 2017.

[28] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 4195–4205.

[29] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining (KDD)*, 1996, pp. 226–231.

[30] M. Moakher, "Means and averaging in the group of rotations," *SIAM Journal on Matrix Analysis and Applications*, vol. 24, no. 1, pp. 1–16, 2002.

[31] H. Fan, H. Su, and L. Guibas, "A point set generation network for 3d object reconstruction from a single image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 605–613.

[32] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *CVPR*, 2019.

[33] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *ICML*, 2020.

[34] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, R. Basri, and Y. Lipman, "Volume rendering of neural implicit surfaces," in *NeurIPS*, 2021.

[35] P. Wang, L. Liu *et al.*, "Neus: Learning neural implicit surfaces by volume rendering," in *NeurIPS*, 2021.

[36] B. Lim, J. Kim, J. Kim, Y. Lee, and F. C. Park, "Equigraspflow: Se (3)-equivariant 6-dof grasp pose generative flows," in *8th Annual Conference on Robot Learning*, 2024.

[37] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2642–2651.

[38] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," in *Robotics: Science and Systems (RSS)*, 2017.

[39] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Asian Conference on Computer Vision (ACCV)*. Springer, 2012, pp. 548–562.

[40] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. Li, "Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3560–3569.

[41] M. Tian, M. H. Ang Jr, and G. H. Lee, "Shape prior deformation for categorical 6d object pose and size estimation," in *European Conference on Computer Vision*. Springer, 2020, pp. 530–546.

[42] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson *et al.*, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.