

HOGraspFlow: Exploring Vision-based Generative Grasp Synthesis with Hand-Object Priors and Taxonomy Awareness

Yitian Shi, Zicheng Guo, Rosa Wolf, Edgar Welte, Rania Rayyes



Fig. 1: Grasp demonstrations for parallel jaw grippers via HOGraspFlow

Abstract—We propose Hand-Object(*HO*)*GraspFlow*, an affordance-centric approach that retargets a single RGB with hand-object interaction (HOI) into multi-modal executable parallel jaw grasps without explicit geometric priors on target objects. Building on foundation models for hand reconstruction and vision, we synthesize $SE(3)$ grasp poses with denoising flow matching (FM), conditioned on the following three complementary cues: RGB foundation features as visual semantics, HOI contact reconstruction, and taxonomy-aware prior on grasp types. Our approach demonstrates high fidelity in grasp synthesis without explicit HOI contact input or object geometry, while maintaining strong contact and taxonomy recognition. Another controlled comparison shows that *HOGraspFlow* consistently outperforms diffusion-based variants (*HOGraspDiff*), achieving high distributional fidelity and more stable optimization in $SE(3)$. We demonstrate a reliable, object-agnostic grasp synthesis from human demonstrations in real-world experiments, where an average success rate of over 83% is achieved.

I. INTRODUCTION

Recent years have witnessed a growing body of approaches in learning robotic manipulation behaviors directly from human demonstrations, ranging from teleoperation to internet-scale video corpora [1], [2], [3]. Specifically, learning from human-object interaction (HOI) demonstrations

is increasingly framed as a retargeting problem, where a robotic end-effector (EE) is mapped from anthropometric hand motion discovered in the wild to the EE kinematics.

Towards the challenge of kinematic mismatch between parallel jaw (PJ) EE and anthropometric hand in the video observations, a popular abstraction in vision-based imitation learning aligns the EE with the human thumb-index pair [1], [4], [5], enabling simple pinch retargeting. While appealingly concise, this proxy collapses the diversity of human grasp taxonomy [6] and neglects contact-dependent antipodal force closure [7]. Hence, this approach is limited to pinch-like grasps and cannot be robustly applied to dynamic demonstrations that exhibit diverse grasp types.

Moreover, recent research on multi-embodiment grasp generations [8], [9] has demonstrated that object-conditioned grasp priors can be learned and modulated by geometric embeddings (e.g., Signed Distance Function (SDF), point graph features) that capture EE morphology. While such approaches enable transfer across different EEs, they are fundamentally object-centric and typically assume access to reliable 3D geometries and pose estimation at test time. Critically, they do not parse human intent or contact semantics from HOI, and thus cannot directly exploit in-the-wild video data, where hand pose, contact detection, and occlusions introduce significant noise.

Targeting these limitations, in this work, we develop a vision-based generative hand-to-EE retargeting framework that supports adaptable transfer to dynamic, diverse anthropometric grasp demonstrations, while explicitly recognizing grasp taxonomy and contact with a single RGB crop of the hand. Our design is motivated by the asymmetry between intent and object variability, where objects vary widely while the set of underlying grasp intents remains relatively limited. On the object side, geometric diversity (shape, scale, etc.) and sensing brittleness (missing depth, specular failures) induce large, unstructured test-time shifts. In contrast, conditioned on a demonstrated grasp, the space of physically valid human hand poses with contact is constrained by anatomy, kinematics, compact grasp taxonomy, and affordance-driven intent, thus lying on a low-dimensional, structured manifold.

Moreover, building on recent advances in multi-modal grasp synthesis with diffusion models [10], [11], [12], we adopt denoising generative modeling on $SE(3)$ for intent-conditioned retargeting given the following advantages: (i) Due to the kinematic mismatch between the human hand and PJ grippers, a single human grasp type often corresponds to multiple valid gripper approaches realizing the same affordance. Denoising-based models naturally preserve this multi-modality by sampling diverse modes in $SE(3)$; (ii) The iterative denoising process admits controllable guidance [13], allowing generated poses to be steered by differentiable constraints (e.g., collision avoidance [14] or curated affordance priors [12]); (iii) In contrast to end-to-end methods that rely on post hoc filtering [15], [16], [17], where feasibility check and affordance matching are performed after grasp generation and invalid candidates are discarded, our in-the-loop guidance enforces constraints during generation, thereby reducing rejections and improving sample efficiency—a critical factor given the scarcity and sampling cost of force-closure, affordance-consistent ground-truth datasets [18].

In summary, our contributions are: (i) We designed a vision-based affordance-centric HOI retargeting framework that produces multi-modal 6-DoF PJ grasps from a single RGB frame. This is achieved by conditioning on foundational features on HOI, without requiring explicit object models or pose estimation, while accounting for the diversity of human grasp taxonomy; (ii) We introduce two generative retargeting frameworks, *HOGraspDiff* and *HOGraspFlow*, inspired by SOTA $SE(3)$ generative approaches to a vision-based setting. By integrating state-of-the-art visual foundation models’ features as contact priors, our method operates without explicit 3D geometric conditions or pose estimation of objects; (iii) Through extensive ablations and real-world deployment, we demonstrate consistent improvements in grasp-type prediction, contact accuracy, and distributional fidelity relative to the baselines, including a *contact-oracle* variant. With a minor translational correction with depths information at deployment, we achieved over 83% success in grasp transfer in our real-world robot experiments.

II. RELATED WORKS

A. Generative 6-DoF grasp synthesis

Learning-based grasp generators model a distribution over executable gripper poses conditioned on scene observations, enabling sampling-based exploration rather than hand-crafted proposal scoring. For instance, GraspNet-1B [19] and Contact-GraspNet [15] have established the effectiveness of learning from local contact geometry for grasp generation. In contrast, denoising-based grasp generators learn a latent density field through an iterative denoising process [20], representing a recent trend in robotic manipulation learning. Among these, $SE(3)$ diffusion fields [10] adapt denoising diffusion to the Lie group [21], which learn grasp densities and refine samples directly on the $SE(3)$ manifold, coupling pose synthesis with motion optimization for grasp execution. To handle symmetries and improve consistency under object rigid motions, EquiGraspFlow [11] enforces $SE(3)$ equivariance while adopting flow-based denoising to handle symmetries and rigid object motions. To incorporate task constraints and human intent, HGDdiffuser [12] augments the diffusion process with hand-intent cues, producing task-oriented 6-DoF grasps via guidance [22]. While effective, these generators typically assume accurate object geometry at inference, and thus usually fail under sensing artifacts and large out-of-distribution (OOD) variability in object shape and appearance.

B. Hand-object interaction (HOI) and reconstruction

Recent progress in HOI recovery has been driven by the MANO parametric hand model [23] and modern monocular reconstructions that deliver accurate 3D hand pose and shape from RGB [24], [25]. As the foundation, rich HOI datasets provide contact supervision and cross-object variability such as DexYCB [26] and OakInk [27], which comprises affordances and human interactions over diverse household objects. Meanwhile, HOGraspNet [28] contributes dense HOI annotations with grasp taxonomy in dynamic sequences, facilitating systematic analysis of everyday human grasps. Aligned with these trends, we condition grasp generation on reconstructed hand poses and dense contact maps, while deliberately avoiding direct dependence on object geometry, which is prone to reconstruction artifacts and time latency.

C. Learning PJ manipulation from human demonstrations

Learning from human demonstrations has progressed from teleoperation to in-the-wild human videos, which aim to narrow down the gap between robotic imitation learning and internet-scale demonstrations. As specific instances for learning manipulation with PJ EE, R+x [1] mines large video corpora to retrieve and adapt relevant HOI behaviors, while Point Policy [2] learns visuomotor policies by extracting EE-hand keypoint correspondences from demonstrations. Focusing on grasping, GAT-Grasp [4] conditions on gestures and affordances to translate human hand signals into task-aware robotic grasps. However, these approaches typically assume a thumb-index pair template as the basis for grasp

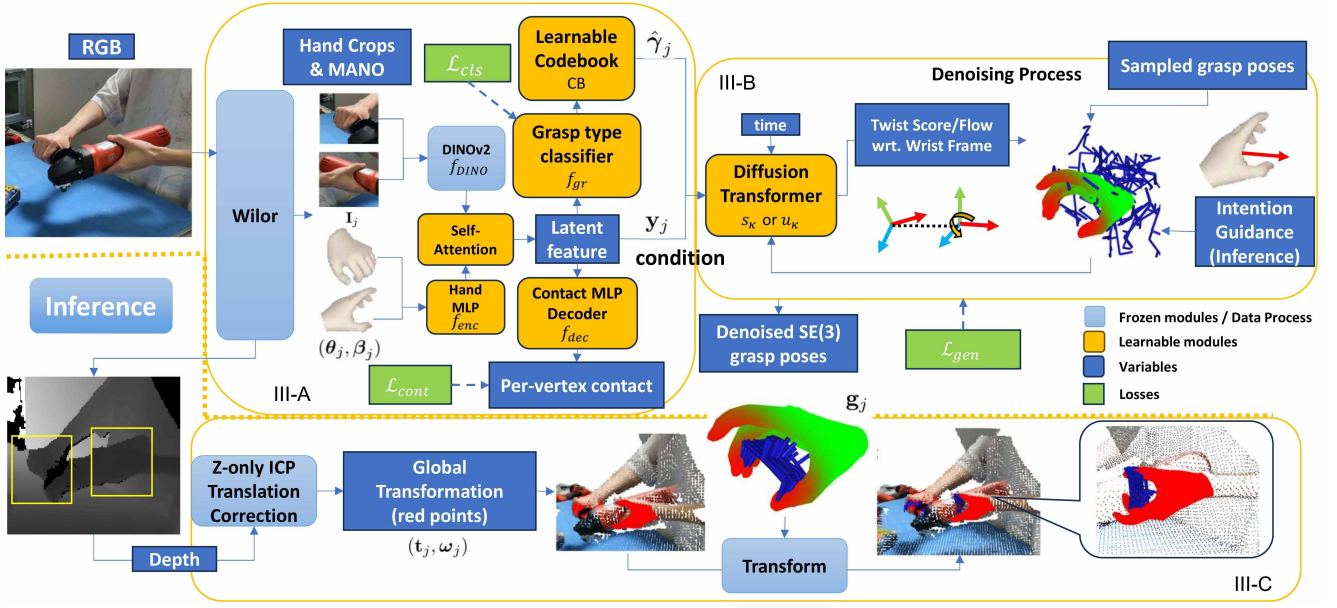


Fig. 2: Pipeline for *HOGraspFlow*

retargeting, which severely limits their generalizability to diverse and unconstrained in-the-wild demonstrations.

III. METHODOLOGY

We aim to recover affordance-centric grasp intent from a human demonstration and retarget it to a PJ, which answers two fundamental questions for affordance-centric grasping: *where to grasp* and *how to grasp*. Fig. 2 summarizes the full pipeline of our approach. Unlike prior pipelines that require object meshes or partial point clouds at both training and inference [8], [12], we condition grasp generation solely on the outputs of a foundational hand reconstructor WiLoR [25], represented by MANO [23] parameters and hand detections with semantics from DINOv2 [29].

To achieve this, our system converts a single RGB observation and the extracted hand poses from WiLoR into a compact grasp-intent embedding (Sec. III-A), and then synthesizes multi-modal PJ poses via a generative denoising process in the SE(3) manifold (Sec. III-B). At inference (Sec. III-C), depth information is only used to refine the absolute translation of the reconstructed hand pose via *Z-only ICP*.

A. Hand-object Perception and Feature Extraction

We first extract HOI semantics from foundational features (Sec. III-A.0.a). Then, these representations get refined and guided with two complementary substreams (Sec. III-A.0.b): (i) *hand contact estimation*, which principally encodes the localization of feasible grasps by predicting contact maps on the hand surfaces (i.e., answering *where to grasp*); and (ii) *grasp taxonomy recognition*, which shapes categorical prior on the distribution of PJ grasps, used jointly with a trainable codebook (CB) as a reference embedding (i.e., answering *how to grasp*).

a) *HOI feature extraction*: Given an RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ with a HOI demonstration, WiLoR [25] detects n hand boxes $\mathcal{B} = \{\mathbf{b}_j\}_{j=1}^n$, and regresses the global hand translation $\mathbf{t}_j \in \mathbb{R}^3$ and MANO parameters $(\omega_j, \theta_j, \beta_j)$ with global orientation $\omega_j \in \mathbb{R}^3$, joint angles $\theta_j \in \mathbb{R}^{45}$, and shape $\beta_j \in \mathbb{R}^{10}$. We then crop the image around each detection, $\mathbf{I}_j = \mathcal{C}(\mathbf{I}; \mathbf{b}_j)$, and apply background augmentation using ground-truth HOI masks from the dataset [28].

To capture the underlying semantics in HOI, explicit CAD or reconstructed meshes can fail under occlusion, specularity, and category diversity. In contrast, foundation models like DINOv2 [29] are broadly invariant across instances and appearances, making them well-suited for grasp representation and transfer without requiring explicit object models or poses. Therefore, we process each augmented crop \mathbf{I}_j with DINOv2, obtaining patch tokens $\mathbf{Z}_j = f_{\text{DINO}}(\mathbf{I}_j) \in \mathbb{R}^{P \times D}$ that serve as compact, semantics-aware descriptors of the local HOI context, where P and D denote the number of patches and the feature dimensionality, respectively.

We map the hand parameters $[\theta_j, \beta_j] \in \mathbb{R}^{55}$ to the same dimension D using a small MLP, yielding a hand feature $\mathbf{h}_j = f_{\text{enc}}(\theta_j, \beta_j) \in \mathbb{R}^D$. Notably, to avoid ambiguity in grasp localization under global hand orientation ω , grasp synthesis is grounded in the hand wrist frame and the feature extractor omits ω , using only articulated pose θ and shape β . Thus, the estimated grasp poses can be transformed to the world coordinate via (\mathbf{t}_j, ω_j) . The feature fusion is performed with a self-attention operator after concatenation between hand and image features¹:

$$\mathbf{y}_j = \text{SelfAttn}([\mathbf{h}_j, \text{ReLU}(f^*(\mathbf{Z}_j))]) \in \mathbb{R}^D, \quad (1)$$

the output \mathbf{y}_j is a single HOI-aware descriptor and f^* is the feature adaptation layers.

¹Only equations referenced in the text are numbered.

b) *Hand contact and grasp taxonomy recognition*: To enrich the hand representation, we learn a contact-aware HOI embedding with a lightweight MLP decoder adapted from Prakash et al. [30], without requiring explicit contact input.

Principally, MANO parametric model [23] map the hand pose parameters (θ, β) to compact hand meshes via: $M = \mathcal{M}(\theta, \beta) \in \mathbb{R}^{N_v \times 3}$, with $N_v = 778$ vertices in structured order. Given the fused descriptor \mathbf{y}_j , the decoder $f_{\text{dec}} : \mathbb{R}^D \rightarrow \mathbb{R}^{N_v}$ predicts per-vertex contact logits $\hat{\mathbf{c}}_j$ over M :

$$\hat{\mathbf{c}}_j = \text{Sigmoid}(f_{\text{dec}}(\mathbf{y}_j)) \in [0, 1]^{N_v}.$$

Given the per-vertex contact labels $\mathbf{c}_j \in [0, 1]^{N_v}$ that are retrieved following [28], we optimize a weighted binary cross entropy:

$$\mathcal{L}_{\text{cont}} = -\frac{1}{N_v} \sum_{v=1}^{N_v} \left[w_1 \mathbf{c}_{jv} \log \hat{\mathbf{c}}_{jv} + (1 - \mathbf{c}_{jv}) \log(1 - \hat{\mathbf{c}}_{jv}) \right],$$

$w_1 = 5$ is chosen to address class imbalance empirically.

While the *hand contact estimation* captures instance-specific HOI semantics over contact localizations, grasp retargeting between the anthropometric hand and PJ is underconstrained due to their morphological differences and kinematic mismatch. To mitigate this, we leverage the *grasp taxonomy recognition* as the morphological prior to complement the retargeted PJ grasp distributions.

Specifically, a grasp type classifier MLP: $f_{\text{gr}}(\mathbf{y}_j)$ categorizes $K = 33$ grasp types, defined by the GRASP Taxonomy [6], and trained via cross entropy: $\mathcal{L}_{\text{cls}} = \text{CE}(f_{\text{gr}}(\mathbf{y}_j), \text{cls}_j)$ with ground-truths class labels cls_j . We maintain a learnable codebook $\text{CB} = \{\gamma_k \in \mathbb{R}^D\}_{k=1}^K$ of size K , and obtain a semantics-aware prior as a softmax-weighted mixture to mitigate the classification errors:

$$\pi_j = \text{Softmax}(f_{\text{gr}}(\mathbf{y}_j)), \quad \hat{\gamma}_j = \sum_{k=1}^K \pi_{j,k} \gamma_k.$$

This taxonomy-conditioned prior complements the contact embedding by regularizing both contact topological and grip orientational distributions.

B. $SE(3)$ Pose Synthesis via Generative Denoising

Given the HOI-aware descriptor \mathbf{y}_j and the induced CB embedding $\hat{\gamma}_j$ from Sec. III-A, we aim to generate diverse, retargeted $SE(3)$ grasp poses. We deliberately adopt denoising-based generative models since they preserve multimodality by sampling diverse modes in $SE(3)$. Moreover, iterative samplers admit controllable, differentiable guidance [13], enforcing feasibility during generation and improving sample efficiency over post hoc filtering [15], [16].

We introduce two frameworks *HOGraspDiff* and *HOGraspFlow* that follow the leading families of recent denoising generative models: score-based diffusion [31], [10] and flow matching [32] in $SE(3)$, respectively, both based on the Diffusion Transformer (DiT) architecture [33].

We formulate the hand grasp retargeting as a conditional sampling process from a $SE(3)$ pose distribution of PJ:

$$\mathbf{g}_j := (\mathbf{p}_j, \mathbf{q}_j) \sim p(\mathbf{g} | \mathbf{y}_i, \hat{\gamma}_j), \quad \mathbf{g} \in SE(3),$$

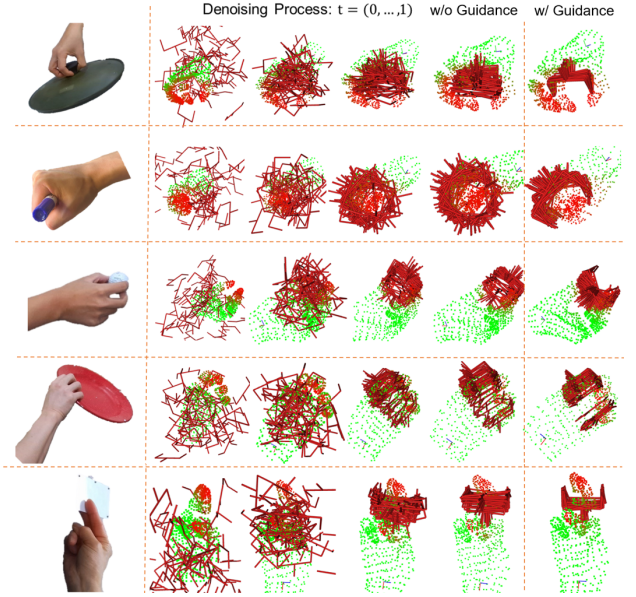


Fig. 3: Denoising process and generation results. Vertices in contact are in red. Parameters for guidance: $\theta_{\text{thr}} = 0.8$, $\lambda^{gd} = 1e - 3$.

where $\mathbf{p}_j \in \mathbb{R}^3$ denotes the Euclidean position, and $\mathbf{q}_j \in S^3 \subset \mathbb{R}^4$ is the unit quaternion of orientation. Since the reference frame of the denoising process is constructed with respect to the hand wrist frame, the framework naturally inherits equivariance to the global transformation (\mathbf{t}_j, ω_j) .

a) *HOGraspDiff with score matching (SM)*: We model the forward diffusion as left-invariant Brownian motion in score matching (SM), by pushing forward isotropic Wiener noise on the Lie algebra $\mathfrak{se}(3)$ analogous to [34].

Specifically, let $\mathbf{g}_t \in SE(3)$ denote the grasp pose at denoising time $t \in [0, 1]$ in the forward diffusion process. We learn a time-dependent left-invariant score function ² $s_\kappa(\mathbf{g}_t, t) \approx \nabla_{\mathbf{g}_t} \log p_t(\mathbf{g}_t) \in \mathfrak{se}(3)$ parametrized by κ . The learned score then drives the reverse-time diffusion:

$$d\mathbf{x}_t = u_\kappa(\mathbf{g}_t, t) dt + \sqrt{2\beta(t)} \cdot dW_t, \quad (2)$$

where $\beta(t) \geq 0$ is the diffusion schedule and W_t is left-invariant Wiener process on $\mathfrak{se}(3)$.

In training, a displacement $\Delta\mathbf{g}_t = (\Delta\mathbf{p}_t, \Delta\mathbf{q}_t) \in SE(3)$ is sampled at a random noise level $t \sim \mathcal{U}(0, 1]$, with decomposed translational and rotational elements [34]:

$$p_t(\Delta\mathbf{p}_t) = \mathcal{N}(\mathbf{p}; \mathbf{0}, \sigma_t^2 \mathbf{I}), p_t(\Delta\mathbf{q}_t) = \mathcal{IG}_{SO(3)}(R_{\mathbf{q}}; \epsilon_t). \quad (3)$$

Here, \mathcal{N} and $\mathcal{IG}_{SO(3)}$ denote the isotropic Gaussian distributions in \mathbb{R}^3 and $SO(3)$, parameterized by their respective concentration $\sigma_t^2 = \alpha_{\mathbf{p}} t$ and $\epsilon_t = \frac{\alpha_{\mathbf{q}} t}{2}$. $R_{\mathbf{q}} \in SO(3)$ is the rotation matrix of \mathbf{q} . The grasp poses of each time step \mathbf{g}_t is then diffused by the twist in body frame using group

²We omit notations $\mathbf{y}_j, \hat{\gamma}_j$ as conditions in $s_\kappa(\cdot), u_\kappa(\cdot)$ for simplicity

product³:

$$\mathbf{g}_{t+\Delta t} = \mathbf{g}_t \circ \Delta \mathbf{g}_t =: \begin{bmatrix} \mathbf{p}_t + R_{\mathbf{q}_t} J(\text{Log}(R_{\Delta \mathbf{q}_t})) \Delta \mathbf{p}_t \\ \mathbf{q}_t \otimes \Delta \mathbf{q}_t \end{bmatrix}, \quad (4)$$

where \otimes denotes quaternion multiplication and $J(\cdot)$ is the left Jacobian. Hence, the objective is to learn a score head that estimates the translational and rotational scores, namely $s_{\kappa}^{\mathbf{p}}(\mathbf{g}_t, t)$ and $s_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t)$, by minimizing the mean squared error to the ground-truth scores:

$$\mathcal{L}_{\text{score}} = \mathbb{E}_{\mathbf{g}_0, t, \Delta \mathbf{g}_t} \left[\left\| \nabla_{\Delta \mathbf{p}_t} \log p_t(\Delta \mathbf{p}_t) - s_{\kappa}^{\mathbf{p}}(\mathbf{g}_t, t) \right\|_2^2 + \left\| \nabla_{\Delta \mathbf{q}_t}^{\mathbb{L}} \log p_t(\Delta \mathbf{q}_t) - s_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t) \right\|_2^2 \right].$$

Closed-form solutions are calculated via [34]:

$$\nabla_{\Delta \mathbf{p}_t} \log p_t(\Delta \mathbf{p}_t) = -\Delta \mathbf{p}_t \sigma_t^{-2},$$

$$\nabla_{\Delta \mathbf{q}_t}^{\mathbb{L}} \log p_t(\Delta \mathbf{q}_t) = \sum_{i=1}^3 \mathbb{L}_i \log \mathcal{IG}_{SO(3)}(R_{\Delta \mathbf{q}_t}; \epsilon_t) \mathbf{e}_i,$$

\mathbb{L}_i is the left-trivialized Lie derivative of $\mathcal{IG}_{SO(3)}(R_{\mathbf{q}}; \epsilon)$ along the i -th orthogonal basis $\{\mathbf{e}_i\}_{i=1,2,3}$ on $\mathfrak{so}(3)$.

In sampling, for each translational or rotational element $\mathbf{x} \in \{\mathbf{p}, \mathbf{q}\}$, the update increment follows Eq. (2):

$$\Delta \mathbf{x}_t = \beta_{\mathbf{x}}(t) s_{\kappa}^{\mathbf{x}}(\mathbf{g}_t, t) \Delta t + \sqrt{2\beta_{\mathbf{x}}(t)} \mathbf{z}_{\mathbf{x}, t},$$

where $\beta_{\mathbf{x}}(t) = \frac{1}{2} \alpha_{\mathbf{x}}^2 t^{\alpha_t}$ with α_t as the time exponent and the sampling stochasticity of Wiener process: $\mathbf{z}_{\mathbf{x}, t} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}^3)$. We then update the pose by left multiplication iteratively via: $\mathbf{g}_{t-\Delta t} = \mathbf{g}_t \circ \Delta \mathbf{g}_t^{-1}$ with $SE(3)$ product in Eq. (4).

b) HOGraspFlow with flow matching (FM): We adapted the SM's formulations to construct the flow-based alternative, which learns a deterministic flow and transports the mass from a base distribution P_0 to $P_1 = P_{\text{data}}$.

Principally, flow matching (FM) parametrizes the drift $\beta_{\mathbf{x}}(t) s_{\kappa}^{\mathbf{x}}(\mathbf{g}_t, t)$ by learning left-trivialized velocity components $u_{\kappa}^{\mathbf{p}}(\mathbf{g}_t, t)$ and $u_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t)$ along the geodesics in $\mathfrak{se}(3)$, bypassing noise schedules and score normalization.

The smooth time-dependent linear and angular velocity fields $u_{\kappa}^{\mathbf{p}}(\mathbf{g}_t, t), u_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t) : [0, 1] \times SE(3) \rightarrow \mathbb{R}^3$, mapping from $\mathbf{g}_t, t \in [0, 1]$ to $\mathbf{g}_1 \sim P_1$ are learned, such that poses \mathbf{g}_1 follow the distribution of the ground-truth data. Therefore, the deterministic flow is constructed and applied via:

$$\mathbf{g}_t = \mathbf{g}_0 \circ [u_{\kappa}^{\mathbf{p}}(\mathbf{g}_t, t), u_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t)]^{\top}. \quad (5)$$

While the body twist naturally unifies translation and rotation, the fully coupled formulation in Eq. (5) creates strong correlations between them, often impeding stable optimization and denoising process. We therefore adopt a decoupled product manifold flow on $\mathbb{R}^3 \times SO(3)$ with independent priors \mathbf{p}_0 and \mathbf{q}_0 that are sampled same as Eq. (3). Given a ground-truth pose $\mathbf{g}_1 = (\mathbf{p}_1, \mathbf{q}_1)$, the geodesics can then be calculated as (in body frame):

$$\Delta \mathbf{p} = R_{\mathbf{q}_0}^{\top} (\mathbf{p}_1 - \mathbf{p}_0), \Delta \phi = \text{Log}(R_{\mathbf{q}_0^{-1} \mathbf{q}_1}). \quad (6)$$

³Log : $SO(3) \rightarrow \mathfrak{so}(3)$ and Exp : $\mathfrak{so}(3) \rightarrow SO(3)$ denote the logarithm and exponential maps, whereas "log" is the standard scalar logarithm.

Subsequently, the interpolation between the initial grasp pose \mathbf{g}_0 and the ground-truth grasp pose $\mathbf{g}_1 \sim P_1$ is performed to get the linear and angular transformation approaching the ground-truth value:

$$\mathbf{p}_{t+\Delta t} = \mathbf{p}_t + \Delta t R_{\mathbf{q}_0} \Delta \mathbf{p}, \mathbf{q}_{t+\Delta t} = \mathbf{q}_t \otimes \text{Exp}(\Delta t \Delta \phi). \quad (7)$$

In this way, the translational field is independent of the initial rotation \mathbf{q}_0 . The objective is to minimize the mean squared error between the ground-truth velocities and the predictions:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{\mathbf{g}_0, t, \Delta \mathbf{g}_t} \left[\left\| \Delta \mathbf{p} - u_{\kappa}^{\mathbf{p}}(\mathbf{g}_t, t) \right\|_2^2 + \left\| \Delta \phi - u_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t) \right\|_2^2 \right].$$

At inference, a grasp pose is calculated by sampling an initial pose and solving an ordinary differential equation (ODE) over $t \in [0, 1]$. We use either 4th-order Runge-Kutta [35] or the generic Euler sampling as Eq. (7) iteratively over t , trading off between sample quality and efficiency.

c) Guidance-based sampling: To better align the synthesized PJ grasps with the hand intention, guidance is applied in both SM and FM. In the hand frame, $\mathbf{e}_{\text{app}} = [0, 1, 0]^{\top}$ is empirically taken as the palm's axis in the local wrist frame. For a noisy pose \mathbf{g}_t , we apply soft guidance only when angular alignment (in cosine similarity) exceeds a threshold:

$$\xi_t = \nabla_{\mathbf{q}_t}^{\mathbb{L}} c_t \mathbb{1}[c_t < \theta_{\text{thr}}], \quad c_t = \langle \mathbf{R}_{\mathbf{q}_t} [y], \mathbf{e}_{\text{app}} \rangle.$$

With guidance weight $\lambda^{\text{gd}} > 0$, during sampling the rotational score/flow field is superposed via:

$$s_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t) \leftarrow s_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t) + \lambda^{\text{gd}} \xi_t, u_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t) \leftarrow u_{\kappa}^{\mathbf{q}}(\mathbf{g}_t, t) + \lambda^{\text{gd}} \xi_t.$$

We additionally apply classifier-free guidance [36] by sampling a weighted sum of the conditional and unconditional flow inspired by [11] for both approaches. The generation processes are illustrated in Fig. 3 for *HOGraspFlow*.

C. Deployment

Given grasp candidates \mathbf{g}_j sampled in the wrist frame, we transform them to the world frame via the hand's world pose $(\mathbf{t}_j, \boldsymbol{\omega}_j)$. In particular, to compensate for the translation bias of WiLoR in \mathbf{t}_j [4], [5], we apply and constrain the *ICP* algorithm [37] that only allows correction along the ray from the camera origin to the hand center ("*Z-only ICP*"), which enables accurate 3D hand pose with the monocular recognition, in contrast to multi-view setups in [2], [4]. In addition, this retargeting strategy improves generalization by decoupling grasp generation from the object's absolute pose.

IV. EXPERIMENTS

In the experiments, we aim to evaluate our framework along the following factors regarding the framework design: (i) the advantage of foundation vision features for synthesized grasp quality and representation; (ii) the relative performance of *HOGraspFlow* versus *HOGraspDiff* under matched settings; (iii) the real-world performance and generalization to unseen objects. We report distributional fidelity of synthesized grasps, per-vertex contact errors, and grasp-type classification scores in Sec. IV-B. To further assess

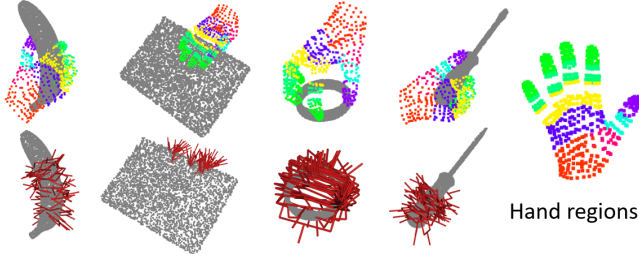


Fig. 4: Generated PJ grasps via *region-conditioned contact matching* with HOGraspNet annotations [28] via hand regions defined by [38].

generalization, we perform real-world evaluations to examine our approach on various objects and grasp types in Sec. IV-C.

A. Data preparation

We evaluate our method on the HOGraspNet dataset [28], which provides 1.5M HOI demonstrations in RGB-D frames, annotated with MANO hand parameters, object meshes and poses, contact maps, and grasp labels defined by GRASP taxonomy [6]. To generate supervision of PJ grasp distributions systematically, we synthesize grasp annotations offline in the following two steps: First, PJ grasps are sampled using the MetaGraspNet workflow [18] on each object mesh from the HOGraspNet. This produces collision-free grasps that fulfill antipodal constraints [7] and are agnostic to the affordance. Second, given the ground-truth hand vertices and their poses relative to the objects, the sampled PJ grasps are filtered via *region-conditioned contact matching* (Fig. 4): Specifically, we adopt the A-MANO [38] semantic partition of the MANO surface into 16 disjoint regions (Fig. 4, right) as a prior for filtering. A pre-generated force-closure candidate is accepted only if its implied contacts can be assigned to two distinct hand regions, reflecting the hypothesis that a realizable PJ grasp engages at least 2 opposing anatomical areas rather than a single region.

Baseline	TA(%) \uparrow	CA(%) \uparrow	EMD \downarrow
<i>HOGraspDiff</i> , w/o DINOv2	69.4	71.8	1.02 ± 0.22
w/ DINOv2	90.3	88.8	0.82 ± 0.23
w/ DINOv2 CB	91.4	89.5	0.80 ± 0.23
w/ Contact	78.8	90.5	0.88 ± 0.22
<i>HOGraspFlow</i> , w/o DINOv2	68.9	69.2	0.81 ± 0.24
w/ DINOv2	91.4	87.6	0.69 ± 0.21
w/ DINOv2 CB	92.5	88.1	0.67 ± 0.21
w/ Contact	78.5	90.5	0.76 ± 0.22

TABLE I: Comparison of Taxonomy Accuracy (TA(%)), Contact Accuracy (CA(%)), and earth mover’s distance (EMD).

B. Grasp generation performance with ablations

a) *Objective*: We train perception and generation end-to-end with $\mathcal{L} = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}} + \lambda_{\text{gen}}\mathcal{L}_{\text{gen}}$

Depending on the chosen generator, either the score or flow-based objective is enabled $\mathcal{L}_{\text{gen}} \in \{\mathcal{L}_{\text{flow}}, \mathcal{L}_{\text{score}}\}$. We choose $\lambda_{\text{cls}} = 0.1$, $\lambda_{\text{cont}} = 0.1$ and $\lambda_{\text{gen}} = 1$ in all baselines.

b) *Baselines*: Our first study probes two design choices in the framework: 2D vision features and the taxonomy-aware codebook. For this purpose, we evaluate four variants both on *HOGraspDiff* and *HOGraspFlow*: (i) *w/o DINOv2*: remove DINO-based self-attention and condition only on the hand MLP feature (i.e. $\mathbf{y}_j = \mathbf{h}_j$ in Eq. (1)), while holding identical architecture; (ii) *w/ DINOv2*: restore self-attention to DINOv2 patch tokens on the RGB hand crops; (iii) *w/ DINOv2 CB*: further augment with the taxonomy-aware trainable codebook (CB); and iv) *w/ Contact* (or *contact-oracle*): replace visual conditioning with the ground-truth per-vertex contact signal (with N_v dimensions) concatenated to the MANO input (i.e., $\mathbf{y}_i = \mathbf{h}_j = f_{\text{enc}}(\boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \mathbf{c}_j)$). Notably, the *contact-oracle* serves as a strong upper bound for conditioning and contact reconstruction quality, which is not available for deployment. In total, 8 baselines are considered in the ablation studies. For the sampling steps, to balance between the quality convergence and the evaluation time, *HOGraspFlow* samples with 40 steps, while *HOGraspDiff* requires 100 steps. The Euler sampler is used for both.

c) *Metrics*: Our evaluation metrics involve: (i) earth mover’s distance (EMD), which quantifies the mismatch between predicted and ground-truth grasp poses via $SE(3)$ geodesic (lower means better), where we generated 100 grasps for each data sample for the measurement (following [11]); (ii) Contact Accuracy (CA(%)), measuring the accuracy of reconstructed contact. In addition, the classification accuracy with respect to the grasp taxonomy (TA(%)) serves as a complementary metric. We evaluate each baseline on 26k validation samples from HOGraspNet at two granularities: (i) overall performance on the full validation split (Tab. I) in parallel with the EMD distribution for each baseline (Fig. 5); (ii) per grasp type performance (Tab. II) across eight representative grasp classes selected to evenly cover the power, intermediate, and precision categories [6].

d) *Results and analysis*: Tab. I reports the overall performance across all 8 baseline approaches. In the *HOGraspDiff* branch, the EMD drops from 1.02 to 0.82 when semantic cues from DINOv2 are integrated. This is further improved by 0.02 with the taxonomy-aware codebook. Moreover, A significant increase in the contact accuracy is identified given the semantic feature by over 15% compared to embedding from MANO only input, and has only 1 – 2% loss compared to the *contact-oracle* model. In contrast, the *HOGraspFlow* branch shows simultaneous gains: the mean drops from 0.81 to 0.69, and finally reached 0.67 with the integrated codebook. Regarding the classification performance, starting from MANO-only, taxonomy accuracy climbs up by over 25% when semantics are embedded in both denoising approaches. In comparison, the *contact-oracle* attains high CA but around 78% in TA, suggesting that per-vertex contact alone is less discriminative of grasp category than visual semantic cues.

Besides, Fig. 5 presents the concrete EMD distributions of

	Small Diameter Parallel Extension				Medium Wrap Sphere 4 Finger				Lateral		Stick		Palmar Pinch		Tripod	
	CA(%) \uparrow	EMD \downarrow	CA(%) \uparrow	EMD \downarrow	CA(%) \uparrow	EMD \downarrow	CA(%) \uparrow	EMD \downarrow	CA(%) \uparrow	EMD \downarrow	CA(%) \uparrow	EMD \downarrow	CA(%) \uparrow	EMD \downarrow	CA(%) \uparrow	EMD \downarrow
<i>HOGraspDiff</i> , w/o DINOv2	82.1	0.93	84.9	1.24	77.7	0.88	84.1	0.86	60.9	1.22	74.0	0.94	69.4	0.91	71.6	0.92
w/ DINOv2	88.1	0.85	88.9	0.71	85.3	0.87	88.2	0.75	89.5	0.72	84.1	0.86	91.3	0.85	84.8	0.80
w/ DINOv2 CB	90.6	0.71	88.5	0.67	87.0	0.86	88.2	0.70	89.3	0.72	87.2	0.83	86.7	0.84	87.6	0.77
w/ Contact	90.1	0.79	90.1	0.79	90.4	0.88	92.2	0.71	89.9	0.78	89.3	0.92	95.9	0.89	91.5	0.84
<i>HOGraspFlow</i> , w/o DINOv2	84.5	0.75	86.9	0.82	78.9	0.75	84.6	0.69	72.2	0.77	69.6	0.80	88.5	0.73	83.0	0.75
w/ DINOv2	87.2	0.63	86.5	0.66	84.6	0.63	84.8	0.66	88.6	0.69	87.5	0.69	92.5	0.70	88.9	0.67
w/ DINOv2 CB	87.0	0.61	88.1	0.64	84.6	0.62	87	0.67	89.6	0.67	86.1	0.66	95.0	0.67	89.9	0.61
w/ Contact	89.1	0.66	91.7	0.71	88.0	0.66	91.1	0.67	89.9	0.72	90.1	0.74	91.7	0.71	91.8	0.73

TABLE II: 8 typical grasp types and their corresponding contact accuracy and EMD.

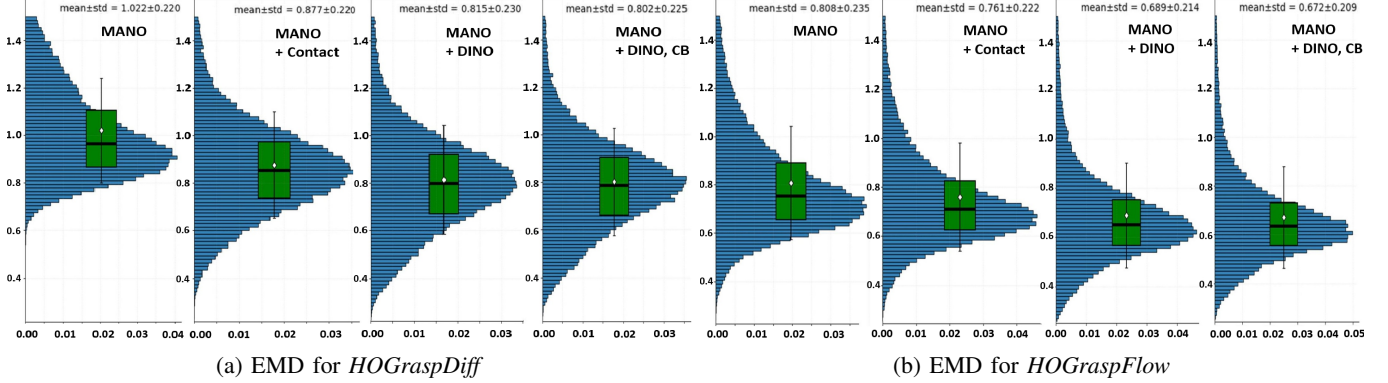


Fig. 5: Comparison of EMD histograms between *HOGraspDiff* and *HOGraspFlow* on HOGraspNet (in frequency).

each baseline. Across the board, adding RGB semantics produces a clear down-shift of the EMD. The taxonomy-aware codebook yields a consistent additional shift upon the others, represented by higher frequencies in the central area of distributions. *HOGraspFlow* consistently dominates *HOGraspDiff* across baselines, with lower central tendency and tighter spread in terms of interquartile range. Furthermore, Tab. II further reports per grasp type results and shows that the trends above hold across the grasp taxonomy. Consistent with Table I, adding RGB-derived semantic features and CB yields performance comparable to the *contact-oracle* model even under strong occlusions, while preserving low EMD. This justified that contact-only supervision is inadequate for semantic understanding compared with integrating object-aware visual priors.

In line with the findings from [11], we observe a significant EMD gap of over 0.15 between flow- and score-based models under identical encoding and denoising architectures. Our key insight is that, although both approaches aim to approximate the same data distribution through a time-dependent vector field, they differ fundamentally in the nature of their denoising targets. The SM models learns the score of perturbed marginals under a variance-exploding (VE) perturbation on the Lie algebra. This score is intrinsically heteroskedastic, since its typical norm scales as $1/\sigma(t) \propto t^{-1/2}$ for both translation and rotation, and its variance increases with t as $p_t(\cdot|x_0)$ spreads along the score directions of $SE(3)$ space. In contrast, FM directly



Fig. 6: Experiment setup (left) and objects in real-world experiments (right). The hammer, marker, mug and bowl are from the HOGraspNet (*In dist.*), others are OOD (*Out dist.*).

regresses the instantaneous velocity induced by a chosen bridge between the same marginals, parameterized by decoupled translational and rotational elements (Eq. 6) instead of explicit $SE(3)$ geodesics. Hence, the training reduces to unscaled standard regression, where every time step exposes the network to the same target magnitude and direction. This eliminates the scale drift present in score-based convergence and sampling, yielding a well-conditioned optimization on the Lie algebra.

C. Real-world Performance

a) *Setups*: We further conducted real-world experiments to evaluate the practical quality of our generated grasps. The experimental setup consists of a UR10e manipulator equipped with a Robotiq 2F-85 EE. A static Orbec

Femto Mega is used for perception. The object set in the experiments (shown in Fig. 6) consists of a representative set of daily objects, common tools and mechanical parts, which demand a wide range of grasping strategies in practice, spanning from firm power grasps to precise and fine manipulations. Each object was evaluated using 2 selected grasp types, with 4 trials per grasp type, chosen to match its functional utility.

b) Performance: Tab. III summarizes the real-world performance under the baselines of *HOGraSPFlow*, where our approach achieved 83.8% (67/80) in grasp success rate with 200ms latency on grasp synthesis, outperforming other aforementioned baselines. We also incorporate: (i) Thumb-index template, adopted from [4], [5]; (ii) GraspNet-1B [19], a point cloud-based grasp generation baseline, filtering its candidate grasps to match the observed hand approach direction and contact regions. While GraspNet-1B yields dense and robust grasps on clean, fully visible point clouds, it struggles to balance high grasp confidence with hand-alignment constraints after filtering and is sensitive to point artifacts (e.g., shiny angle grinder motor, the thin body of a pen), reaching 66.2% (53/80) overall success.

Fig. 1 further demonstrates qualitative grasp retargeting results, including power, tool-oriented, and fine manipulation grasps. Notably, given reliable *ICP* alignment, our method remains robust due to ignorance of sensory artifacts, whereas geometry-centric approaches often fail due to perception errors. In general, we show the flexibility and generalization of our approach in retargeting diverse human grasps to a PJ EE without accurate geometry or pose estimation on objects.

Nevertheless, three main failure types are identified: (i) imperfect hand-pose estimation from WiLoR, which can propagate through the entire pipeline; (ii) *ICP* registration errors of hands; (iii) motion planning failures during grasp executions.

Baseline	Overall	In dist.	Out dist.
Thumb-index template [4], [5]	33/80	13/32	20/48
GraspNet-1B [19]	53/80	21/32	32/48
<i>HOGraSPFlow</i> , w/o DINOv2	55/80	18/32	37/48
w/ DINOv2	63/80	21/32	42/48
w/ DINOv2 CB	67/80	26/32	41/48

TABLE III: Success in real-world grasp retargeting.

V. CONCLUSIONS

We proposed *HOGraSPFlow*, a vision-based, hand pose-centric retargeting framework that converts a single RGB HOI frame into multi-modal *SE(3)* parallel jaw grasps. By combining foundational RGB features with a learned contact decoder and a taxonomy-aware codebook, our method injects intent priors that improve distributional fidelity and semantic correctness, with FM consistently outperforming the score-based variant while maintaining high contact and taxonomy accuracy. Real-world experiments confirm robust transfer across diverse objects and grasp types with over 83% success

rate, which outperforms the existing template-based proxies and point-based grasp learning approach.

REFERENCES

- [1] G. Papagiannis *et al.*, “R+x: Retrieval and execution from everyday human videos,” in *ICRA*, 2025.
- [2] S. Haldar and L. Pinto, “Point policy: Unifying observations and actions with key points for robot manipulation,” *arXiv preprint arXiv:2502.20391*, 2025.
- [3] E. Welte and R. Rayyes, “Interactive imitation learning for dexterous robotic manipulation: Challenges and perspectives—a survey,” *arXiv preprint arXiv:2506.00098*, 2025.
- [4] R. Wang *et al.*, “Gat-grasp: Gesture-driven affordance transfer for task-aware robotic grasping,” *arXiv preprint arXiv:2503.06227*, 2025.
- [5] M. Lepert, J. Fang, and J. Bohg, “Phantom: Training robots without robots using only human videos,” *arXiv preprint arXiv:2503.00779*, 2025.
- [6] T. Feix *et al.*, “The grasp taxonomy of human grasp types,” *IEEE Transactions on human-machine systems*, vol. 46, no. 1, 2015.
- [7] J. Bohg *et al.*, “Data-driven grasp synthesis—a survey,” *IEEE TRO*, vol. 30, no. 2, 2013.
- [8] N. Khargonkar *et al.*, “Neuralgrasps: Learning implicit representations for grasps of multiple robotic hands,” in *CoRL*. PMLR, 2023.
- [9] M. Attarian *et al.*, “Geometry matching for multi-embodiment grasping,” in *CoRL*. PMLR, 2023.
- [10] J. Urain *et al.*, “Se(3)-diffusionfields: Learning smooth cost functions for joint grasp and motion optimization through diffusion,” in *ICRA*, 2023.
- [11] B. Lim *et al.*, “Equigrasppflow: Se (3)-equivariant 6-dof grasp pose generative flows,” in *8th CoRL*, 2024.
- [12] D. Huang *et al.*, “Hgdiffr: Efficient task-oriented grasp generation via human-guided grasp diffusion models,” *arXiv preprint arXiv:2503.00508*, 2025.
- [13] K. Frans *et al.*, “Diffusion guidance is a controllable policy improvement operator,” *arXiv preprint arXiv:2505.23458*, 2025.
- [14] H. Li *et al.*, “Language-guided object-centric diffusion policy for generalizable and collision-aware manipulation,” in *ICRA*. IEEE, 2025.
- [15] M. Sundermeyer *et al.*, “Contact-graspingnet: Efficient 6-dof grasp generation in cluttered scenes,” in *ICRA*. IEEE, 2021.
- [16] M. Breyer *et al.*, “Volumetric grasping network: Real-time 6 dof grasp detection in clutter,” in *CoRL*. PMLR, 2021.
- [17] Y. Shi *et al.*, “vmf-contact: Uncertainty-aware evidential learning for probabilistic contact-grasp in noisy clutter,” in *ICRA*, 2025.
- [18] M. Gilles *et al.*, “Metagrasspnetv2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping,” *TASE*, vol. 21, no. 3, 2024.
- [19] H.-S. Fang *et al.*, “Graspnet-1billion: A large-scale benchmark for general object grasping,” in *CVPR*, 2020.
- [20] R. Wolf *et al.*, “Diffusion models for robotic manipulation: a survey,” *Frontiers in Robotics and AI*, vol. Volume 12 - 2025, 2025.
- [21] J. Sola, J. Deray, and D. Atchuthan, “A micro lie theory for state estimation in robotics,” *arXiv preprint arXiv:1812.01537*, 2018.
- [22] P. Dhariwal and A. Nichol, “Diffusion models beat gans on image synthesis,” *NeurIPS*, vol. 34, 2021.
- [23] J. Romero, D. Tzionas, and M. J. Black, “Embodied hands: Modeling and capturing hands and bodies together,” *arXiv preprint arXiv:2201.02610*, 2022.
- [24] G. Pavlakos *et al.*, “Reconstructing hands in 3d with transformers,” in *CVPR*, 2024.
- [25] R. A. Potamias *et al.*, “Wilor: End-to-end 3d hand localization and reconstruction in-the-wild,” in *CVPR*, 2025.
- [26] Y.-W. Chao *et al.*, “Dexycb: A benchmark for capturing hand grasping of objects,” in *CVPR*, 2021.
- [27] L. Yang *et al.*, “Oakink: A large-scale knowledge repository for understanding hand-object interaction,” in *CVPR*, 2022.
- [28] W. Cho *et al.*, “Dense hand-object (ho) graspingnet with full grasping taxonomy and dynamics,” in *ECCV*. Springer, 2024.
- [29] M. Oquab *et al.*, “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023.
- [30] A. Prakash *et al.*, “How do i do that? synthesizing 3d hand motion and contacts for everyday interactions,” in *CVPR*, 2025.
- [31] Y. Song and S. Ermon, “Generative modeling by estimating gradients of the data distribution,” in *NeurIPS*, vol. 32, 2019.

- [32] Y. Lipman *et al.*, “Flow matching for generative modeling,” *arXiv preprint arXiv:2210.02747*, 2022.
- [33] W. Peebles and S. Xie, “Scalable diffusion models with transformers,” in *CVPR*, 2023.
- [34] H. Ryu *et al.*, “Diffusion-edfs: Bi-equivariant denoising generative modeling on $se(3)$ for visual robotic manipulation,” in *CVPR*, 2024.
- [35] M. Schober, D. Duvenaud, and P. Hennig, “Probabilistic ode solvers with runge-kutta means,” in *NeurIPS*, vol. 27, 2014.
- [36] J. Ho and T. Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [37] A. Segal, D. Haehnel, and S. Thrun, “Generalized-icp,” in *Robotics: science and systems*, vol. 2, no. 4. Seattle, WA, 2009.
- [38] L. Yang *et al.*, “Cpf: Learning a contact potential field to model the hand-object interaction,” in *ICCV*, 2021.