Affordance-Guided Diffusion Prior for 3D Hand Reconstruction

Naru Suzuki[1⋆], Takehiko Ohkawa[1⋆], Tatsuro Banno[1], Jihyun Lee[2], Ryosuke Furuta[1], and Yoichi Sato[1]

[1]The University of Tokyo [2]KAIST

Abstract. How can we infer a 3D hand pose when large portions of the hand are heavily occluded by itself or by objects? Humans often resolve such ambiguities by leveraging contextual knowledge—such as *affordances*, where an object's shape and function suggest how the object is typically grasped. Inspired by this observation, we propose a generative prior for 3D hand pose modeling guided by affordance-aware textual descriptions of hand-object interactions (HOI). Our method employs a diffusion-based generative model that learns the distribution of plausible hand poses conditioned on contextual signals, such as affordance descriptions and image features. The affordance descriptions are designed to represent the semantic intent and geometric structure of HOI, using the reasoning from a vision-language model (VLM) and grasp classification. We leverage the diffusion prior to refine the 3D pose predictions in hand reconstruction into more accurate and functionally coherent estimation. Our experiments demonstrate that our affordance-guided refinement significantly improves 3D hand pose estimation performance on 3D hand affordance datasets, HOGraspNet and HO3D, over state-of-the-art methods such as foundation models for hand reconstruction and the latest diffusion priors for 3D hands.

1 Introduction

Hand-object interactions (HOI) are ubiquitous in daily life, where human actions are inseparable from the surrounding context, including objects, environments, and the person's intent. Among them, reconstructing 3D hands (*e.g.*, estimating pose and shape parameters of the MANO model [47]) from a single RGB image has become a crucial task, enabling applications in virtual/augmented reality and robotic dexterous manipulation. However, this task remains highly challenging due to severe occlusions and inherent 3D ambiguities commonly present in real-life interactions. Despite these difficulties, humans can still infer plausible hand poses by exploiting *contextual knowledge* beyond the hand itself. In particular, the shape and function of interacting objects can provide strong cues for reasoning about feasible hand configurations during interaction.

Most previous works on 3D hand reconstruction [10,39,41,42] have paid little attention to leveraging such contextual knowledge. Only a few studies have explored contextual signals in the following forms: (a) temporal contact states (*e.g.*, pre-grasp, in-grasp, post-grasp) [65], (b) short interaction captions (*e.g.*, "a hand holding a spoon") [61], (c)

⋆ Equal contribution.

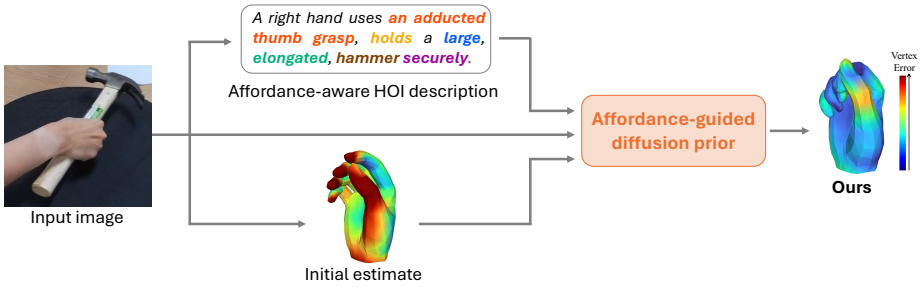

Fig. 1: Can affordance-aware semantics benefit 3D hand reconstruction? By leveraging contextual cues about how an object is grasped, our affordance-guided diffusion prior refines coarse 3D hand predictions into functionally coherent and physically grounded poses, even under strong occlusion. Vertex error is color-coded on the hand mesh.

action labels (*e.g.*, "screw", "open") [30,37], (d) intention labels (*e.g.*, intend to "use" or "receive" an object) [52, 60], etc. These semantic labels have been used as conditional input in 3D modeling [56, 60, 61], for additional semantic classification [44, 59, 65], or to create specific evaluation categories [4,30,37]. However, these existing labels remain coarse and fail to capture the rich details of HOI needed to resolve ambiguities.

To enrich contextual signals and better reflect the subtle nuances of HOI scenes, we emphasize the critical role of *affordances* [1,6,14] as semantic guidance, which directly encodes how object properties constrain and guide possible hand interactions. Existing affordance-related studies are typically categorized into recognizing (i) motor actions or (ii) grasp types, defined by discrete labels [63]. Motor actions, often expressed by verb labels (*e.g.*, "pick up"), represent the high-level semantic intent of manipulation. In contrast, grasp types, defined through rigorous taxonomies (*e.g.*, "lateral grasp", "precision grasp") [4, 12], provide the geometric structure of fine-grained hand configurations.

Based on this perspective, we integrate these two distinct aspects of affordances, *i.e.*, both semantic intent and geometric structure of HOI, into a unified contextual representation for hand reconstruction. However, the way to combine these heterogeneous elements of affordances in a unified form is not obvious. Unlike following existing categories (*e.g.*, "pick up" or "lateral grasp"), we tackle this issue by representing affordances through detailed textual descriptions that encode both semantic and geometric cues. This descriptive form offers greater flexibility to capture variations in object attributes, hand grasps, and interaction intent beyond predefined categories. In particular, we incorporate grasp types and object attributes (category, shape, and size), along with the underlying intent, such as interaction (action) and intention captions. To generate such a descriptive text reflecting multiple affordance aspects, we propose a step-by-step generation process by leveraging visual commonsense reasoning of a large vision-language model (VLM) [2] with grasp taxonomy knowledge from the hand affordance dataset [4], resulting in complete and interpretable descriptions of affordances.

With the affordance descriptions, we propose generative prior modeling that can integrate textual descriptions into 3D pose modeling (Fig. 1). We train a denoising diffusion model [16] on hand pose data, while newly adopting affordance descriptions, encoded with DistilBERT [48], and ViT-based image features [9] as conditional input. We further utilize this prior to refine 3D pose estimates, inspired by [38]. Specifically,

starting from an initial estimate by WiLoR [42] or Hamba [8], we iteratively denoise the pose so that it remains consistent with the affordance-aware diffusion prior and visible 2D evidence. To improve the refinement, we further detect occluded joints with ray casting from joints and hand segmentation to guide only uncertain joints.

Our experiments are conducted on 3D hand affordance datasets, HOGraspNet [4] and HO3D [15] where the grasp type labels are assigned by the classification model trained on HOGraspNet. Our affordance-guided diffusion prior consistently outperforms recent hand reconstruction models, including foundational transformers such as WiLoR [42] and HaMeR [41], the Mamba-based Hamba [8], and hand diffusion priors based on InterHandGen [24], with or without contextual reasoning. Furthermore, our qualitative study finds our method producing grasp poses that not only appear more natural but also align closely with the affordance descriptions.

Our main contributions are summarized as follows:

- We are the first to introduce affordance-aware textual descriptions into 3D hand reconstruction using VLMs and grasp classification.
- We design a diffusion-based generative prior that learns the hand pose distribution given the contextual signals, enabling semantically consistent generation.
- We propose an occlusion-aware refinement approach that leverages the affordance-aware diffusion prior for 3D hand reconstruction, yielding substantial improvements in pose accuracy specifically when estimated joints are uncertain.

2 Related Work

3D hand reconstruction: This task involves estimating hand pose and shape to reconstruct hand mesh, typically from a single RGB image. The earlier efforts [10, 23, 36, 39, 64] encode image features with CNN-based backbones and regress pose and shape parameters of the MANO model [47]. Particularly, HandOccNet [39] proposes robust feature extraction for occluded regions. Recent foundation models [41, 42] leverage a vision transformer [9] with large-scale pretraining to enhance generalization ability across different domains. While these recent methods typically provide strong alignment to 2D observation, ambiguities still remain in the 3D space, *e.g.*, depth ambiguity and uncertainty for occluded joints.

To address this, integrating with contextual cues is crucial as humans can infer feasible 3D poses given the context. Yet, this direction remains relatively underexplored. For instance, recent methods [44, 56, 59, 65, 67] attempt to integrate semantic or physical cues (*e.g.*, grasp type and state, body proportion, and action labels), but focus on coarse-level reasoning rather than fine-grained details of HOI scenes.

In contrast, our work is motivated by leveraging detailed textual descriptions regarding affordances (see Fig. 2) and we are the first to introduce such affordance descriptions into 3D hand reconstruction.

Affordance recognition and reasoning: The concept of *affordances*, introduced by Gibson [13], describes the functional relationship between perception and potential action (*e.g.*, a handle *affords* grasping or a button *affords* pressing). In computer vision, affordance recognition and reasoning have been studied through diverse formulations, including functional region segmentation [11, 18, 25, 34], grasp and contact prediction [4,6,14], affordance-based HOI understanding [7,22] and generation [3,43,62,66],

and adaptation to robot manipulation [1, 26, 58]. These approaches often operate at the object or scene level, capturing where and how an agent may interact. Yet, they typically rely on coarse labels [63] such as verb categories (*e.g.*, "pick up", "place") or predefined grasp taxonomies [12]. Recent 3D hand affordance datasets [4, 18] further encourage affordance understanding with precise hand poses, object geometry, and vertex-wise contact.

In this work, we propose to represent affordances as rich, detailed, and multi-level semantic descriptions that unify motor intent, grasp configuration, and object attributes (see Fig. 2). Furthermore, unlike the existing formulations, we demonstrate that such affordance descriptions resolve the fine-grained geometric ambiguities central to 3D hand reconstruction.

Diffusion models: Denoising diffusion models [16] have emerged as a powerful generative framework, owing to their strong capability to capture complex data distribution and their flexibility during inference. Diffusion models are trained to gradually transform Gaussian noise into realistic samples through an iterative denoising process.

Recent works adopt diffusion-based modeling in 3D hand and body modeling, outperforming the conventional VAEs-based modeling [31, 33]. The diffusion models can leverage various conditional inputs for 3D modeling [5, 24, 29, 38, 50, 54, 61]. Human motion generation commonly leverages text prompts to guide motion patterns [32, 53, 54], including hand-object motion generation [5, 27]. InterHandGen [24] synthesizes two-hand interactions conditioned on a single hand, while G-HOP [61] generates 3D hand-object interactions from object geometry. The trained diffusion priors can be further adapted to 3D reconstruction tasks with score-distillation sampling [24, 61] and refinement with keypoint guidance [38, 50].

Our work instead proposes a pose diffusion model that flexibly takes multi-modal input for detailed text and visual features, unlike existing frameworks, *e.g.*, text-to-motion [5, 27], 3D two-hand [24], and 3D hand-object generation [61]. It further enhances the diffusion-based refinement with guidance from occlusion detection unlike [38].

3 Method

Our work is motivated to leverage contextual signals from textual descriptions of affordances and visual evidence for 3D hand reconstruction. To incorporate these cues, we propose a generative prior model based on a diffusion process that learns the 3D hand pose distribution given the contextual signals. Our approach begins with an affordance description generation, which extracts key elements of affordances from input images using a vision-language model and grasp classification, as described in Sec. 3.1. We then construct a diffusion prior conditioned on the affordance descriptions and image features, as detailed in Sec. 3.2. Subsequently, the learned prior is adapted to correct occluded or uncertain joints with the refinement scheme in Sec. 3.3, yielding more accurate and functionally coherent hand poses.

3.1 Affordance description generation

Affordances represent how an object can be interacted with hands by capturing its actionable properties; thus we utilize them as contextual representation and guidance for

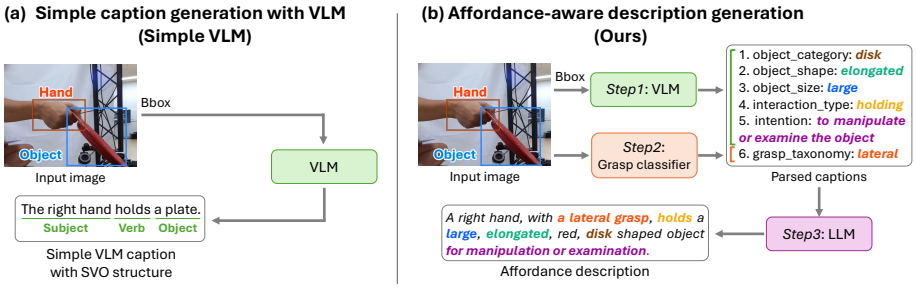

Fig. 2: Affordance description generation. (a) We show a naive baseline of caption generation (Simple VLM) from off-the-shelf VLMs (*e.g.*, Qwen2.5-VL [2]) given hand-object bounding boxes, lacking fine-grained affordance cues. (b) Inspired by chain-of-thought reasoning, our proposed generation first obtains parsed captions (items 1-5) by prompting the VLM to extract intended affordance elements. It further infers a grasp taxonomy (item 6) with a tailored classifier due to VLM's limited capability to predict hand grasps in a zero-shot manner. These generated affordance elements are summarized using LLM [19] , resulting in a single detailed description.

3D hand reconstruction. Specifically, beyond standard categorical labels (*e.g.*, grasp or action labels) of affordance recognition [4, 22, 44, 63], we enrich its form with textual descriptions that reflect both semantic and precise geometric properties, such as object category, shape, size, interaction (action), intention, and grasp type.

Our preliminary study finds that direct (one-step) captioning, which prompts VLMs to generate a description from an image, often produces a simplified sentence lacking fine-grained HOI properties. An example of this simply generated caption ("Simple VLM") is found in Fig. 2(a). Additionally, while being able to reasonably describe HOI scenes, recent VLMs still struggle to capture precise grasp taxonomy as this annotation is not common in general QA tasks or VLMs' pretraining data. To provide finer affordance details, we generate the affordance descriptions step-by-step in Fig. 2(b). Inspired by the chain-of-thought reasoning [57] of LLMs, we break down the complex description task into small executable and interpretable steps. Our pipeline consists of three steps: (i) extracting key elements of affordances with zero-shot VLM reasoning, (ii) classifying the grasp with separate supervised training, and (iii) integrating all elements into a final description with LLM. This procedure ensures the completeness and interpretability of the generated affordance descriptions with improved caption quality. We detail the individual steps below.

Step 1 – Extract parsed captions with VLM: We leverage a vision-language model (Qwen2.5-VL [2]) to extract structured captions describing different aspects of affordances from an image. Specifically, given the image together with the bounding boxes of the hand and the object, we design a prompt that queries the model to describe five key elements that characterize HOI properties. These include object category (*e.g.*, "disk", "sphere"), object shape (*e.g.*, "elongated", "flat"), object size (*e.g.*, "small", "large"), interaction (action) type (*e.g.*, "holding", "reaching"), and intention caption (*e.g.*, "to lift", "to manipulate"). These elements are chosen because they capture the physical properties of the object and the semantic intent of the action. Further details are found in the supplement.

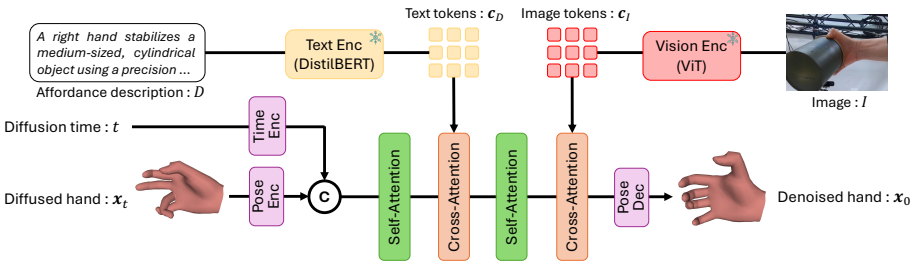

Fig. 3: Denoising framework of affordance-guided diffusion prior (AffHandGen). The input diffused hand pose is denoised using a transformer decoder network conditioned on the affordance description and RGB image, each encoded by DistilBERT [48] and the ViT backbone [41]. These modality interactions across pose, text, and image are learned through cross-attention.

Step 2 – Grasp classification: Accessing grasp information is essential in determining hand configurations. Yet, we observe that classifying grasp taxonomies using existing VLMs is highly challenging, as rigorous grasp labels [12] may not be widely represented in their pretraining data. Therefore, we train a dedicated grasp classification model independently on a large-scale hand affordance dataset, HOGraspNet [4]. Specifically, we take the RGB image as input and extract visual features encoded by the ViT and MANO pose parameters using HaMeR [41]. The visual and pose features are concatenated and passed through a self-attention layer, followed by an MLP layer for final classification. Our model achieves 87.6% accuracy in the evaluation dataset, and further implementation details and results are provided in the supplement.

Step 3 – Summarize with LLM: Given the short captions for each element related to affordances, we transform these items into a single detailed description. We use a large language model (LLM), Mistral-7B [19], with a summarization prompt that generates a final affordance description. This LLM-based summarization allows us to combine nuanced variations in HOI properties, producing a coherent sentence that flexibly represents semantic and geometric cues beyond what predefined categories can express.

Overall, this divide-and-conquer approach comprised of the three steps ensures that the resulting descriptions are (i) complete and specific by eliciting all key elements, (ii) consistent with the predicted grasp type via supervised classification, and (iii) coherent and interpretable as the conditional input (used for the subsequent diffusion training).

3.2 Affordance-guided diffusion prior

Using the affordance descriptions, we present a diffusion prior model conditioned on text and image features, dubbed **AffHandGen**. This model is trained to generate and refine 3D hand poses aligned with the contextual signals. We choose the diffusion modeling to learn data distribution across different modalities, such as image, text, and pose, inspired by its successful results in 3D human synthesis [24, 38, 40, 54]. It also benefits from being directly applicable to various downstream tasks without additional training [50], making it flexible and broadly reusable.

Overview: With the goal of refining 3D hand pose in the downstream task, we construct a diffusion model in the hand pose space, where the shape estimation remains fixed. For

the input, we use the pose parameters of MANO [47] as the target variable of denoising, *i.e.*, $\mathbf{x}_0 = \boldsymbol{\theta} \in \mathbb{R}^{15\times 3}$. We also use a diffusion timestep t, an affordance description D (generated in Sec. 3.1), and the corresponding RGB image I as the conditional input. We encode the affordance description using a text encoder, yielding textual features $\mathbf{c}_D$, and encode the RGB image with an image encoder, yielding image features $\mathbf{c}_I$. The resulting diffusion model parameterized by ϕ is formulated as $\hat{\mathbf{x}}_0 = f_\phi(\mathbf{x}_t, t, \mathbf{c})$, where $\mathbf{x}_t$ is the diffused pose at the timestep t and $\mathbf{c} = \{\mathbf{c}_D, \mathbf{c}_I\}$ denotes the text and image condition.

The overall architecture of $f_\phi(\cdot)$ is illustrated in Fig. 3. Our **AffHandGen** is a transformer network trained to denoise noisy hand pose input $\mathbf{x}_t$ at the timestep t. The hand pose and timestep are first encoded through linear layers and concatenated, followed by processing with transformer decoder layers and a final pose decoder with a linear layer.
Cross-attention to modality tokens: For conditional data encoding, InterHandGen [24] treats the condition $\mathbf{c}$ as a global feature vector and learns the relationships across the target pose, conditions, and timestep using self-attention. However, representing each conditional modality with a single vector loses spatial and semantic granularity. For instance, when two hand images involve the same object but different grasp types, the global visual or textual embeddings (*e.g.*, from CNN or CLIP [45]) tend to be similar, making it difficult for the diffusion model to distinguish subtle interaction differences.

To capture such fine-grained contextual distinctions, we represent each modality as a set of *tokens* and employ a *cross-attention* mechanism [55] that learns localized correspondences between the denoising pose features and conditional modality tokens. This design enables the model to selectively attend to informative modality tokens, including those implicitly encoding grasp semantics, thereby achieving context-sensitive conditioning. Specifically, we use modality-specific tokenizers: DistilBERT [48] for the text embedding and the ViT network trained in [41] for the image embedding.
Diffusion process: We follow the DDPM formulation [16], where the diffusion process consists of forward and reverse paths spanned with diffusion steps $t \in [0, T]$. The forward process gradually adds standard Gaussian noise to the input data $\mathbf{x}_0$, which is formulated as

$$\mathbf{x}_t := \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \tag{1}$$

where $\epsilon \sim \mathcal{N}(0, I)$ is a noise variable and $\alpha_{1:T} \in (0, 1]^T$ is a sequence that controls the degree of noise at each timestep t. In contrast, the reverse process learns to iteratively transform noise $\mathbf{x}_T \sim \mathcal{N}(0, I)$ toward the target data distribution $\mathbf{x}_0 \sim q(x_0)$, defined as

$$p_\phi(\mathbf{x}_{0:T}) := p(\mathbf{x}_T) \prod_{t=1}^{T} p_\phi(\mathbf{x}_{t-1} \mid \mathbf{x}_t), \tag{2}$$

where p_ϕ is a model distribution defined by the network $f_\phi(\cdot)$. We train the denoising network to approximate the original data $\mathbf{x}_0$ following [24, 38, 54].

The training loss is computed between the original data $\mathbf{x}_0$ and the generated data $\hat{\mathbf{x}}_0$. We take the L2 loss between $\mathbf{x}_0$ and $\hat{\mathbf{x}}_0$ for the pose space, denoted as L_θ. After passing to the MANO layer, we compute the L2 loss on the mesh vertices and 3D joint space, denoted as L_v and L_j. The overall loss is formulated as

$$L_{total} = \lambda_\theta L_\theta + \lambda_v L_v + \lambda_j L_j. \tag{3}$$

Algorithm 1: Diffusion-based pose refinement.

Diffuse 3D pose and visibility labels $\mathbf{v}$ *of* $\tilde{\mathbf{x}}_0$
$\tilde{\mathbf{x}}_n \leftarrow \sqrt{\bar{\alpha}_n}\tilde{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_n}\,\epsilon$
$\mathbf{v} = V(\tilde{\mathbf{x}}_0)$ *// 1: visible, 0: occluded*
for $t = n$ to 1 **do**
  $\hat{\mathbf{x}}_0 \leftarrow f(\tilde{\mathbf{x}}_t, t, \mathbf{c})$
  ▷ *2D keypoint fitting only for visible joints*
  $\hat{\mathbf{x}}_0 \leftarrow \hat{\mathbf{x}}_0 - \lambda_{2d}\nabla_{\hat{\mathbf{x}}_0}\mathcal{L}_2(M(\hat{\mathbf{x}}_0), M(\tilde{\mathbf{x}}_0)) \odot \mathbf{v}$
  $\epsilon_t \leftarrow \frac{1}{\sqrt{1-\bar{\alpha}_t}}(\tilde{\mathbf{x}}_t - \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0)$
  $\tilde{\mathbf{x}}_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}}\hat{\mathbf{x}}_0 + \sqrt{1-\bar{\alpha}_{t-1}}\epsilon_t$
end for
return $\hat{\mathbf{x}}_0$

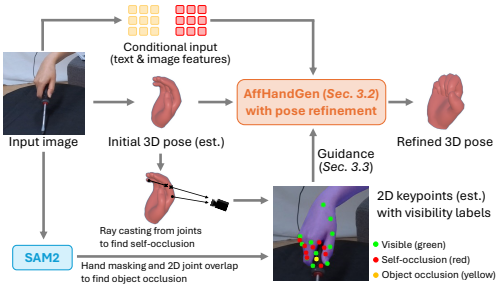

Fig. 4: Diffusion-based pose refinement algorithm. Our refinement takes the initial 3D pose estimates $\tilde{\mathbf{x}}_0$ from recent hand reconstruction methods (*e.g.*, [42]), conditional text and image features $\mathbf{c}$, and detected visibility $V(\cdot)$ in Sec. 3.3, and then corrects occluded joints (red and yellow in the right figure) while visible joints (green) remain unchanged. The visibility labels are obtained from two detection criteria: self-occlusion with ray casting on the MANO mesh and object occlusion with SAM2 [46] mask for hands. This refinement is resumed at the middle of diffusion steps n, reducing the total inference cost, and optimized with the 2D keypoint loss based on camera projection function $M(\cdot)$ and its loss weight λ_{2d}.

λ_θ, λ_v, and λ_j are the weights for each loss term.

To improve the robustness of the model, we randomly dropout the conditional input $\mathbf{c_D}$ and $\mathbf{c}_I$, inspired by [17, 24, 33, 50] . Specifically, we randomly mask the affordance description and the RGB image with a probability of 0.1. This emulates conditional and unconditional generation within a single model. This strategy brings several benefits: (i) preventing the model from overfitting to a specific conditioning modality, leading to a strong prior learning over 3D hand poses, and (ii) improving flexibility at inference time, allowing seamless switching between conditional and unconditional sampling without retraining.

3.3 Occlusion-aware prior adaptation to 3D hand reconstruction

Once the diffusion prior is trained, we adapt the prior to guide and improve 3D hand reconstruction, specifically aiming to refine the 3D pose parameters. While foundation models for hand reconstruction (*e.g.*, HaMeR [41]) effectively align hand meshes with the visible 2D evidence in RGB images, they still struggle to recover accurate 3D poses under severe occlusion, where visual cues are limited. SCGen [38] addresses this by applying a 3D human diffusion prior to refine 3D pose estimates with 2D keypoint constraints. Given an initial 3D pose $\tilde{\mathbf{x}}_0$ (the target of refinement) and reliable 2D keypoints $\mathbf{y}_0$ detected by human foundation models like Sapiens [20], it first sets a diffused pose $\tilde{\mathbf{x}}_n$ at the intermediate step n ($0 < n < T$) and iteratively denoises the pose until $t = 0$, while minimizing the 2D projection loss between $\hat{\mathbf{x}}_0$ and $\mathbf{y}_0$.

Inspired by the diffusion-based refinement for human pose [38], we introduce an *affordance-guided diffusion refinement* tailored to 3D hand reconstruction (Fig. 4). Our approach newly incorporates contextual cues from textual descriptions and image features, enabling the model to reconstruct hand poses that are physically plausible and semantically consistent with the intended interaction. Unlike body poses, hands exhibit high articulation and frequent object contacts, leading to more complex occlusion pat-

terns. To address this, we detect the visibility of each joint and apply selective 2D guidance only to visible joints, allowing the diffusion prior to correct uncertain (occluded) joints through contextual conditioning.

Our refinement framework requires the initial 3D pose $\tilde{\mathbf{x}}_0$, its camera projection (*i.e.*, 2D keypoints) $M(\tilde{\mathbf{x}}_0)$, joint-wise visibility labels $\mathbf{v} \in \{0,1\}^J$ (J is the number of joints), and contextual conditions $\mathbf{c} = \{\mathbf{c}_D, \mathbf{c}_I\}$ from text and image features. The visibility labels $\mathbf{v}$ are obtained through the detection scheme below and are used to mask the 2D keypoint loss. Since reliable 2D hand keypoints are more difficult to obtain than those in the body domain [38], we instead use the projected 2D keypoints of the initial 3D pose, $M(\tilde{\mathbf{x}}_0)$, along with the visibility labels as supervision signals. This allows the model to align the pose to confident 2D evidence and denoise uncertain occluded joints based on the contextual diffusion prior. We further apply the 2D keypoint fitting to the denoised pose $\hat{\mathbf{x}}_0$ as post-processing. This ensures that the resulting pose after the diffusion-based refinement is well-aligned to the visible 2D keypoints.

Occlusion detection: Estimating joint visibility is useful for downstream HOI tasks, yet challenging. We employ two criteria to determine the occlusion of hand joints regarding *self-occlusion* and *object occlusion*, resulting in the binary visibility labels $\mathbf{v}$. For self-occlusion, we determine occluded joints by ray casting given the hand mesh derived from $\tilde{\mathbf{x}}_0$. A ray is cast from each hand joint toward the camera, and its intersections with the hand surface are computed as [49]. If the number of intersections is greater than or equal to two, the joint is occluded by the part of hand mesh, regarded as self-occlusion. For instance, a ray from the pinky fingertip of Fig. 4 penetrates into the index finger's mesh, judged as the self-occluded joint. For object occlusion, we identify object-occluded joints by comparing the projected 2D joints from the initial pose predictions, $M(\tilde{\mathbf{x}}_0)$, and the hand mask obtained from SAM2 [46]. If the 2D joint position is outside the mask, the joint is regarded as object occlusion.

4 Experiments

We present quantitative results on pose refinement, comparing against foundational models for 3D hand reconstruction and diffusion prior baselines with or without using contextual signals. We then present ablations and qualitative analyses of pose refinement and controllable generations. Additional results are found in the supplement.

Experimental settings: We use the HOGraspNet [4] dataset for complex and diverse hand-object interactions, which is the only dataset offering RGB images, 3D assets, and grasp taxonomy labels needed for our task setup. We follow 28 of the 33 grasp types defined by [12] to train a grasp classification network. We also use the HO3D [15] dataset including RGB images and 3D hand-object annotations, as an additional validation and the HInt dataset [41] for in-the-wild evaluation. As these two datasets lack grasp annotations, we apply the grasp classifier trained on [4] to assign pseudo grasp labels. We subsample all datasets and use sparse frames as training and evaluation, because consecutive frames sharing similar visuals often contain the same grasp labels and almost identical VLM descriptions. To ensure diverse contextual signals without heavy semantic overlaps, we resample every 10 frames, resulting in 119K training and 3K evaluation images in [4] and 65K training and 1K evaluation images in [15]. HOGrasp-

HOGraspNet										
Method	Avg.↓	Low	Med.	High	PA-MPVPE↓	PCK@5↑	PCK@10↑	PCK@15↑	F@5↑	F@15↑
HandOccNet [39]	15.09	14.34	15.55	17.65	12.93	12.0	42.8	67.4	0.482	0.935
MeshGraphormer [28]	10.02	9.39	10.27	13.55	8.99	24.7	68.7	87.5	0.625	0.977
HaMeR [41]	8.38	7.37	8.92	12.68	7.66	38.6	78.6	91.7	0.696	0.988
Hamba [8]	8.46	7.44	9.00	12.72	7.57	38.3	78.9	91.8	0.698	0.987
+ **Ours**	7.85	7.49	10.21	9.46	7.11	41.2	81.6	93.2	0.718	0.991
WiLoR [42]	8.22	7.42	8.56	12.41	7.57	39.7	79.4	92.2	0.702	0.989
+ InterHandGen [24]	8.99	7.64	9.72	14.61	7.91	34.8	75.4	90.1	0.681	0.986
+ w/ Full Cond. (D+I)	8.06	7.50	8.29	11.12	7.29	40.3	79.6	92.3	0.713	0.990
+ **Ours**	**7.53**	**7.29**	**7.57**	**9.37**	**6.90**	**44.1**	**82.5**	**93.6**	**0.731**	**0.992**
HO3D										
Method	Avg.↓	Low	Med.	High	PA-MPVPE↓	PCK@5↑	PCK@10↑	PCK@15↑	F@5↑	F@15↑
HandOccNet [39]	8.83	8.46	9.44	9.86	8.80	31.4	69.3	86.7	0.572	0.965
MeshGraphormer [28]	12.55	11.71	13.15	18.79	12.83	16.6	52.0	73.3	0.436	0.908
HaMeR [41]	8.15	7.34	9.59	9.54	8.33	37.7	74.5	89.0	0.632	0.973
Hamba [8]	7.42	**7.10**	7.88	8.71	7.66	40.3	77.9	91.3	0.655	0.982
+ **Ours**	**7.29**	7.25	**7.14**	8.39	**7.37**	**41.1**	**78.9**	91.9	**0.657**	**0.983**
WiLoR [42]	7.43	7.19	7.85	7.87	7.63	39.0	77.9	91.9	0.650	**0.983**
+ InterHandGen [24]	7.55	7.30	7.99	7.99	7.64	38.1	77.6	91.4	0.638	0.980
+ w/ Full Cond. (D+I)	7.65	7.35	8.15	8.51	7.72	37.6	77.1	91.0	0.633	0.979
+ **Ours**	7.37	7.11	7.83	**7.80**	7.46	39.1	78.7	**92.0**	0.645	**0.983**

Table 1: Results of 3D hand pose estimation on HOGraspNet and HO3D. We report PA-MPJPE in millimeters for varying occlusion levels: Low (0–5 occluded joints), Medium (6–10), and High (11–21), together with PA-MPVPE in mm, PCK@5/10/15, and F@5/15 mm. The diffusion baselines (InterHandGen, w/ Full Cond.) are applied on top of WiLoR, while our refinement is applied on top of both WiLoR and Hamba. **Bold** and underline denote the best and second-best.

Net's evaluation includes unseen subjects (S1 split [4]) and HO3D comprises unseen views, subjects, and objects.

For baselines of 3D hand reconstruction, we test a graph-based transformer, **MeshGraphormer** [28], an occlusion-robust network, **HandOccNet** [39], two foundational transformers, **HaMeR** [41] and **WiLoR** [42], and a Mamba-based model, **Hamba** [8]. For the diffusion prior, we train a 3D hand diffusion model, **InterHandGen** [24], by adapting its architecture from two-hand input to a single-hand input for this task, which lacks any contextual conditions. For a fair comparison with our method, we further implement **InterHandGen w/ Full Cond.** modified to take the text and image conditions as global feature vectors. These diffusion baselines are trained from scratch and applied to refine 3D poses from WiLoR with the refinement method in Sec. 3.3.

We evaluate reconstruction performance using both 3D and 2D metrics. For 3D joints and mesh vertices, we report Procrustes-aligned mean errors (PA-MPJPE and PA-MPVPE). We also report mesh F-scores under 5 mm and 15 mm thresholds (F@5 and F@15), which evaluate surface reconstruction accuracy within the corresponding distance thresholds. For 2D joint accuracy, we report PCK@5, PCK@10, and PCK@15, which measure the percentage of keypoints whose 2D errors fall within 5, 10, and 15 pixels, respectively. To measure the model's robustness under increasing ambiguities, we further provide evaluation subsets according to different occlusion levels (Low, Medium, High), grouped by the number of visible joints detected in Sec. 3.3.

Implementation details: For generation, our attention layers follow a latent size of 512 and four heads. The loss weights are set to $\lambda_\theta = 1$, $\lambda_v = 10000$, and $\lambda_j =$

(a) Text type	RGB?	Avg.	Low	Med.	High
(1) N/A	-	9.28	8.16	9.82	14.6
(2) N/A	✓	7.90	7.69	7.89	9.91
(3) Simple VLM	-	8.67	8.00	9.03	11.4
(4) Simple VLM	✓	7.76	7.64	7.70	9.50
(5) Aff.	-	8.43	7.62	8.90	11.5
(6) Aff.	✓	7.53	7.29	7.57	9.37
(7) Aff. w/ GT grasp	✓	**7.51**	**7.28**	**7.54**	**9.35**

(b) Text input	Avg.↓	FPS↑
(1) Qwen + Mistral	**7.53**	0.31
(2) Qwen keywords	7.59	0.39
(3) Grasp labels	7.69	1242
(4) Qwen (one-shot)	7.68	1.19
(5) Aff. classifier *	7.57	96
Seg. mask	**Avg.↓**	**FPS↑**
(6) SAM2 [46]	**7.53**	8.1
(7) InterFormer [51]	7.54	16.3
(8) None	7.55	—

(c) 2D guid.	Avg.↓
λ_{2d}=0.005	**7.53**
λ_{2d}=0	7.65
V+L enc.	**Avg.↓**
HaMeR+DistilBERT	**7.53**
CLIP	11.36

(d) Aff. field	Avg.↓
Full	**7.53**
w/o obj. category	7.53
w/o interaction	7.54
w/o intention	7.54
w/o obj. shape	7.56
w/o obj. size	7.59
w/o grasp tax.	7.71

Table 2: Ablation study on HOGraspNet (PA-MPJPE). **(a)** Conditional modality compares text inputs and RGB image features. **(b)** Text input and visibility masks, **(c)** 2D keypoint guidance and V+L encoder, and **(d)** affordance fields; * denotes a lightweight classifier trained on GT-based affordance elements. Shaded rows mark our setting. **Bold** and underline denote the best and second-best.

10000. We set the number of tokens for text and image embeddings to 64 and 192, respectively. We train diffusion models for 80 epochs with a batch size of 64 using an Adam optimizer [21] with a learning rate of 1e-3. We set the maximum diffusion steps to $T = 1000$ and use a cosine scheduler [35] for the forward process. For refinement, we implement faster sampling techniques. We adaptively set the start refinement step $n \in [n_{min}, n_{max}]$ to the value chosen linearly according to the number of occluded joints. Larger occlusion counts result in larger values of n that require more refinement steps, conversely setting fewer steps in less occluded samples. We also sparsely sample every m step in the reverse process. On HOGraspNet, we set $\lambda_{2d} = 0.005$, $n = [100, 1000]$, $m = 20$. HO3D requires fewer refinement steps and increased 2D fitting as the initial predictions are more reliable: we set $\lambda_{2d} = 0.5$, $n = [10, 20]$, $m = 3$. Additional analysis of hyperparameter sensitivity is found in the supplement.

4.1 Results

3D hand pose estimation: Tab. 1 reports the 3D pose evaluation on HOGraspNet [4] and HO3D [15]. Recent foundation models outperform the CNN-based HandOccNet [39], yet the high-occlusion regime remains challenging (∼12 mm error on the "High" subset). The diffusion baseline InterHandGen [24] barely corrects poses without contextual signals, and even "w/ Full Cond." (image and text as global vectors) only partially improves the "Med."/"High" splits over WiLoR [42]. In contrast, our prior consistently improves WiLoR across *all* occlusion levels, reaching an average error of 7.53 mm, with a substantial 3.04 mm (24%) reduction on the "High" occlusion subset. This gain comes from fine-grained cross-attention to image and text tokens and occlusion-aware refinement, rather than attending conditions as coarse global vectors like "IHGen w/ Full Cond." The same trend holds on HO3D: although WiLoR and Hamba are already highly competitive, our prior refines both without degradation on average, achieving 7.37 mm with WiLoR and the best 7.29 mm with Hamba. This confirms that the gains are not specific to a single base estimator, but generalize across different estimators.

Ablation studies: All ablations are conducted on HOGraspNet and reported in Tab. 2. We ablate the main components of our pipeline, including conditional modalities, affordance text generation, visibility masking, 2D keypoint guidance, encoder design, and individual affordance fields.

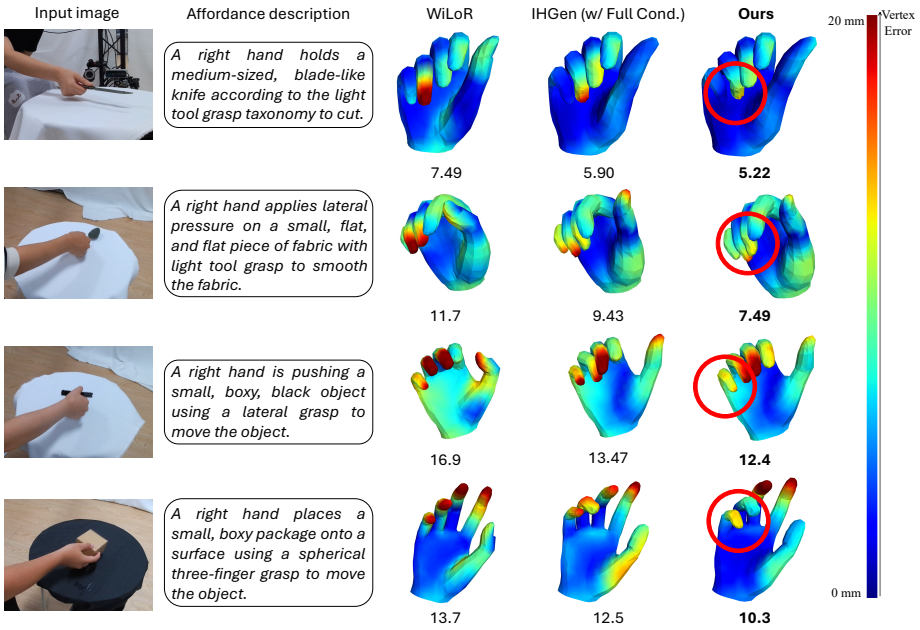

Fig. 5: Qualitative results of diffusion-based refinement on the HOGraspNet dataset [4]. We show refinement results with our diffusion prior and a comparison diffusion baseline, InterHand-Gen (IHGen) [24] with Full Cond. The vertex error compared to the ground-truth mesh is color-coded, which is clipped to $[0, 20]$ mm. While the initial hand pose from WiLoR [42] struggles to estimate joints in occluded regions, our method reasonably refines the estimates to plausibly grasp the object even under occlusion. The values under the samples denote PA-MPJPE.

(a) Conditional modality: We evaluate the impact of image and text conditioning, comparing the guidance effect from affordance descriptions with simple VLM captions. The Simple VLM text is generated by prompting Qwen2.5-VL [2] to describe the given HOI scene (*e.g.*, "The right hand holds a screwdriver"). Compared with results of $(1, 3, 5)$, the simple caption method does not offer significant improvement over the baseline without text conditioning (N/A), while the affordance descriptions (Aff.) help reduce the ambiguities. Conditioning image features with ViT encoding $(2, 4, 6)$ further proves to be effective, serving to align the predicted pose with visual evidence. We observe that our configuration (6) performs the best 7.53 mm by utilizing the affordance descriptions and the image features jointly. We also assess the effect of predicted grasp types (see "Step 2" of Sec. 3.1) against using ground-truth grasp labels at test time. We find the results $(6, 7)$ almost on par, indicating that the predicted grasps in (6) include few failures in classification and do not negatively affect the downstream task.

(b) Text input & segmentation mask: For the text input, our two-stage Qwen+Mistral description (1), where the VLM extracts affordance keywords that Mistral summarizes into a coherent sentence, outperforms three simpler alternatives: raw VLM keywords without summarization (2), grasp taxonomy labels alone (3), and a single-shot VLM prompt (4). This confirms that the structured summarization provides a stronger condi-

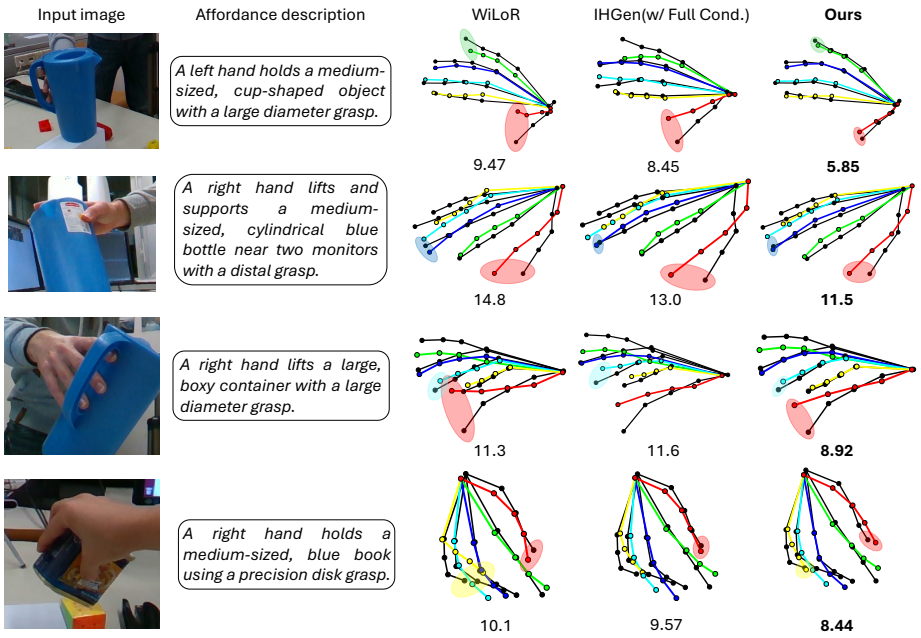

Fig. 6: Qualitative results of diffusion-based refinement on the HO3D dataset [15]. We compare the refinement results of our diffusion prior against a diffusion baseline, InterHandGen (IHGen) [24] with Full Cond., both applied to the initial hand poses predicted by WiLoR [42]. Predicted hand joints are colored by finger, with ground truth displayed in black. Some joint errors are highlighted using colored ellipses. The values under the samples denote PA-MPJPE.

tioning signal than flat labels or unsummarized text. A lightweight affordance classifier trained with GT-based affordance elements is then used to generate the affordance descriptions (5), recovering most of the accuracy (7.57 mm) while running at 96 fps. For the visibility mask, replacing SAM2 [46] (6) with the faster InterFormer [51] (7) doubles throughput at a negligible accuracy cost, and removing the mask entirely (8) only mildly degrades performance.

(c) 2D guidance & V+L encoder: The 2D keypoint guidance (λ_{2d}=0.005) grounds the refinement to visual evidence during inference, and removing this guidance by setting λ_{2d}=0 degrades the accuracy. We further ablate the V+L encoder, which denotes the vision and language encoding modules used to condition our diffusion prior. Our modality-specific encoder, implemented with HaMeR for visual features and DistilBERT for textual affordance descriptions, is also essential to performance. In contrast, replacing these task-specific encoders with a CLIP encoder substantially degrades the error to 11.36 mm. This result suggests that CLIP's global embeddings lack the hand-specific spatial granularity required to distinguish subtle hand-object interactions.

(d) Affordance fields: Ablating each field of the affordance description shows that the grasp taxonomy contributes the most to the refinement. The object category, shape, size, interaction, and intention also provide complementary gains, highlighting the benefit of combining multiple elements for accurate refinement.

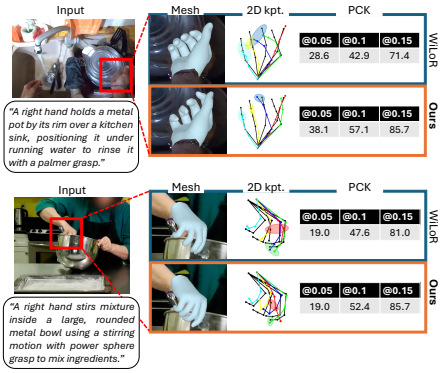

Fig. 7: In-the-wild sample evaluation on the HInt dataset [41]. We present PCK (%) to the 2D keypoint GT (black skeleton). The highlighted circle and color in 2D kpt indicate the 2D joint error and its finger type (red circle means thumb error).

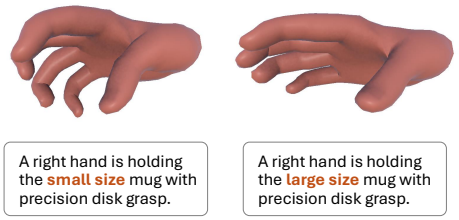

Fig. 8: Qualitative results of our diffusion prior. We show samples generated from partially different descriptions. We find that the diffusion-based generation responds to the change of object size, while remaining consistent with the given description.

Qualitative results: Fig. 5 presents qualitative comparisons on HOGraspNet between the WiLoR predictions, the refined poses by InterHandGen (IHGen) with Full Cond., and our diffusion prior. The initial poses exhibit large errors (red vertices) in the occluded parts (*e.g.*, the ring and pinky fingers are often hidden in the examples). Our method corrects these errors and improves over the IHGen (w/ Full Cond.) baseline.

Similarly, Fig. 6 visualizes qualitative comparisons on the HO3D dataset. As highlighted by the ellipses, the initial WiLoR predictions suffer from high errors at severely occluded joints. While the diffusion baseline, IHGen (w/ Full Cond.), attempts to correct them, it struggles to effectively reflect the textual affordance descriptions into the refined poses. In contrast, our diffusion prior successfully leverages these text cues, effectively correcting the hidden articulations to form plausible grasps.

Fig. 7 shows in-the-wild evaluation on the HInt dataset [41], including proxy 2D keypoint evaluation (PCK). We find stable captioning with generalized VLM capability and successful diffusion-based refinement over WiLoR with improved PCK metrics. This suggests the applicability of our method in such unconstrained cooking environments where the affordance signals are abundant.

Controllability of diffusion-based generation: Fig. 8 presents qualitative examples of our diffusion model given different textual descriptions. We validate how pose generation responds to different affordance elements with the rest of the description fixed. Specifically, we alter only the object size (small *vs*. large), demonstrating that our model can produce hand poses that adapt to the size change, such as wider grasps for large objects and tighter grasps for small ones. This suggests that our diffusion prior can faithfully incorporate actionable semantics embedded in the affordance descriptions.

Object-property-wise evaluation: We provide further granular evaluations on [4] with different object types and sizes. As shown in Fig. 9, while WiLoR suffers large errors on medium cylinders, large disks, and articulated shapes, our method consistently achieves robust refinement across these challenging categories.

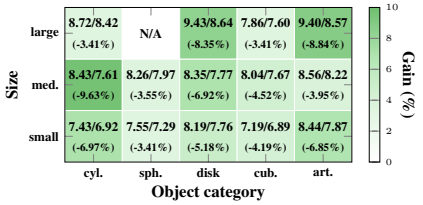

Fig. 9: Object property-wise performance evaluation on HOGraspNet. The cell value "WiLoR / Ours (gain%)" shows the WiLoR baseline, our method, and the relative gain in PA-MPJPE. The x-axis represents cylinder, sphere, disk, cuboid, and articulated shapes.

Method	Unit	GFLOPs↓	Throughput↑ (frame/sec)
WiLoR	per sample	138.06	21.0
AffHandGen	per diff. step	0.99	389.1
AffHandGen	per sample (≈8 steps)	**7.94**	**48.6**

Table 3: Computational cost evaluation on HOGraspNet. We measure the inference burden of foundational transformers and our diffusion model. As it sparsely samples diffusion steps in inference, our method only requires on average approximately 8 diffusion steps per sample, resulting in significantly lower GFLOPs and higher throughput compared to WiLoR.

Inference cost: Owing to our adaptive start steps in refinement (n in Fig. 4) and sparse time sampling, our method does not worsen the inference burden in GFLOPs and Throughput compared to the foundation model WiLoR in Tab. 3.

5 Conclusion

We present a novel diffusion prior that leverages affordance descriptions generated with VLM's reasoning and grasp classification. Our model learns the distribution of plausible and functionally coherent 3D hand poses, and then is adapted to refine 3D pose estimates, enabling robust reasoning about occluded joints and more accurate reconstructions in challenging hand-object interactions. Our experiments demonstrate that the guidance by affordance descriptions substantially improves the refinement quality over both foundational transformer methods and diffusion priors, with strong gains under severe occlusions. Our method also provides controllable and interpretable refinements, aligning pose generation with affordance descriptions.

We note some limitations that open interesting avenues for future work. Our method currently focuses on static single-hand reconstruction and does not yet explicitly account for 3D object geometry, relying instead on textual descriptions as a proxy to include the object information. Extending the approach to explicitly model 3D object interactions is promising for future work. Furthermore, extending our affordance-conditioned prior from static images to continuous temporal domains, such as egocentric video analysis and human motion synthesis, is a highly promising and exciting direction for modeling dynamic hand-object interactions.

Acknowledgements

This work was supported by JST K Program Grant Number JPMJKP25V1 and JST ASPIRE Grant Number JPMJAP2303.

References

1. Bahl, S., Mendonca, R., Chen, L., Jain, U., Pathak, D.: Affordances from human videos as a versatile representation for robotics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13778–13790 (2023)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Cha, J., Kim, J., Yoon, J.S., Baek, S.: Text2hoi: Text-guided 3d motion generation for hand-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1577–1585 (2024)
4. Cho, W., Lee, J., Yi, M., Kim, M., Woo, T., Kim, D., Ha, T., Lee, H., Ryu, J.H., Woo, W., et al.: Dense hand-object (ho) graspnet with full grasping taxonomy and dynamics. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 284–303 (2024)
5. Christen, S., Hampali, S., Sener, F., Remelli, E., Hodan, T., Sauer, E., Ma, S., Tekin, B.: Diffh2o: Diffusion-based synthesis of hand-object interactions from textual descriptions. In: Proceedings of the ACM SIGGRAPH Asia Conference. pp. 1–11 (2024)
6. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., Rogez, G.: Ganhand: Predicting human grasp affordances in multi-object scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5031–5041 (2020)
7. Do, T.T., Nguyen, A., Reid, I.: Affordancenet: An end-to-end deep learning approach for object affordance detection. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). pp. 5882–5889 (2018)
8. Dong, H., Chharia, A., Gou, W., Carrasco, F.V., De la Torre, F.: Hamba: Single-view 3d hand reconstruction with graph-guided bi-scanning mamba. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 2127–2160 (2024)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: Proceedings of the International Conference on Learning Representations (ICLR). (2021)
10. Fan, Z., Ohkawa, T., Yang, L., Lin, N., Zhou, Z., Zhou, S., Liang, J., Gao, Z., Zhang, X., Zhang, X., et al.: Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 428–448 (2024)
11. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2vec: Reasoning object affordances from online videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2139–2147 (2018)
12. Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The grasp taxonomy of human grasp types. IEEE Transactions on human-machine systems **46**(1), 66–77 (2015)
13. Gibson, J.J.: The theory of affordances:(1979). In: The people, place, and space reader, pp. 56–60. Routledge (2014)
14. Goyal, M., Modi, S., Goyal, R., Gupta, S.: Human hands as probes for interactive object understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3293–3303 (2022)
15. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3196–3206 (2020)
16. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 6840–6851 (2020)

17. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
18. Jian, J., Liu, X., Li, M., Hu, R., Liu, J.: Affordpose: A large-scale dataset of hand-object interactions with affordance-driven hand pose. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14713–14724 (2023)
19. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de Las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2024)
20. Khirodkar, R., Bagautdinov, T.M., Martinez, J., Zhaoen, S., James, A., Selednik, P., Anderson, S., Saito, S.: Sapiens: Foundation for human vision models. In: Proceedings of the European Conference on Computer Vision (ECCV). vol. 15062, pp. 206–228 (2024)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR). (2015)
22. Koppula, H.S., Gupta, R., Saxena, A.: Learning human activities and object affordances from rgb-d videos. The International journal of robotics research **32**(8), 951–970 (2013)
23. Kulon, D., Guler, R.A., Kokkinos, I., Bronstein, M.M., Zafeiriou, S.: Weakly-supervised mesh-convolutional hand reconstruction in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4990–5000 (2020)
24. Lee, J., Saito, S., Nam, G., Sung, M., Kim, T.K.: Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 527–537 (2024)
25. Li, G., Sun, D., Sevilla-Lara, L., Jampani, V.: One-shot open affordance learning with foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3086–3096 (2024)
26. Li, G., Tsagkas, N., Song, J., Mon-Williams, R., Vijayakumar, S., Shao, K., Sevilla-Lara, L.: Learning precise affordances from egocentric videos for robotic manipulation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10581–10591 (2025)
27. Li, M., Christen, S., Wan, C., Cai, Y., Liao, R., Sigal, L., Ma, S.: Latenthoi: On the generalizable hand object motion generation with latent hand diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17416–17425 (2025)
28. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12939–12948 (2021)
29. Liu, A.L., Chao, Y.W., Chen, Y.T.: Task-oriented human grasp synthesis via context- and task-aware diffusers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10375–10385 (2025)
30. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21013–21022 (2022)
31. Lu, J., Lin, J., Dou, H., Zeng, A., Deng, Y., Liu, X., Cai, Z., Yang, L., Zhang, Y., Wang, H., et al.: Dposer-x: Diffusion model as robust 3d whole-body human pose prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9988–9997 (2025)
32. Lv, X., Xu, L., Yan, Y., Jin, X., Xu, C., Wu, S., Liu, Y., Li, L., Bi, M., Zeng, W., et al.: Himo: A new benchmark for full-body human interacting with multiple objects. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 300–318 (2024)

33. Müller, L., Ye, V., Pavlakos, G., Black, M., Kanazawa, A.: Generative proxemics: A prior for 3d social interaction from images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9687–9697 (2024)
34. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8688–8697 (2019)
35. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Proceedings of the International Conference on Machine Learning (ICML). vol. 139, pp. 8162–8171 (2021)
36. Ohkawa, T., Furuta, R., Sato, Y.: Efficient annotation and learning for 3d hand pose estimation: A survey. International Journal of Computer Vision (IJCV) **131**, 3193–3206 (2023)
37. Ohkawa, T., He, K., Sener, F., Hodan, T., Tran, L., Keskin, C.: Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12999–13008 (2023)
38. Ohkawa, T., Lee, J., Saito, S., Saragih, J., Prado, F., Xu, Y., Yu, S.I., Furuta, R., Sato, Y., Shiratori, T.: Generative modeling of shape-dependent self-contact human poses. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 906–915 (2025)
39. Park, J., Oh, Y., Moon, G., Choi, H., Lee, K.M.: Handoccnet: Occlusion-robust 3d hand mesh estimation network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1496–1505 (2022)
40. Patel, C., Nakamura, H., Kyuragi, Y., Kozuka, K., Niebles, J.C., Adeli, E.: Uniegomotion: A unified model for egocentric motion reconstruction, forecasting, and generation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 10318–10329 (2025)
41. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9826–9836 (2024)
42. Potamias, R.A., Zhang, J., Deng, J., Zafeiriou, S.: Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12242–12254 (2025)
43. Prakash, A., Lundell, B., Andreychuk, D., Forsyth, D., Gupta, S., Sawhney, H.: How do i do that? synthesizing 3d hand motion and contacts for everyday interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7026–7036 (2025)
44. Prakash, A., Tu, R., Chang, M., Gupta, S.: 3d hand pose estimation in everyday egocentric images. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 183–202 (2024)
45. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
46. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
47. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: modeling and capturing hands and bodies together. ACM Transactions on Graphics (ToG) **36**(6), 245:1–245:17 (2017)
48. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)

49. Smith, B., Wu, C., Wen, H., Peluse, P., Sheikh, Y., Hodgins, J.K., Shiratori, T.: Constraining dense hand surface tracking with elasticity. ACM Transactions on Graphics (ToG) **39**(6), 219:1–219:14 (2020)
50. Stathopoulos, A., Han, L., Metaxas, D.: Score-guided diffusion for 3d human recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 906–915 (2024)
51. Su, Y., Wang, Y., Yao, L., Cui, Y., Chau, L.P.: Interaction-aware representation modeling with co-occurrence consistency for egocentric hand-object parsing. In: Proceedings of the International Conference on Learning Representations (ICLR). (2026)
52. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: GRAB: A dataset of whole-body human grasping of objects. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 581–600 (2020)
53. Tevet, G., Gordon, B., Hertz, A., Bermano, A.H., Cohen-Or, D.: Motionclip: Exposing human motion generation to clip space. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 358–374 (2022)
54. Tevet, G., Raab, S., Gordon, B., Shafir, Y., Cohen-Or, D., Bermano, A.H.: Human motion diffusion model. In: Proceedings of the International Conference on Learning Representations (ICLR). (2023)
55. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). vol. 30 (2017)
56. Wang, Y., Sun, Y., Patel, P., Daniilidis, K., Black, M.J., Kocabas, M.: Prompthmr: Promptable human mesh recovery. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1148–1159 (2025)
57. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). vol. 35, pp. 24824–24837 (2022)
58. Wei, Y.L., Jiang, J.J., Xing, C., Tan, X.T., Wu, X.M., Li, H., Cutkosky, M., Zheng, W.S.: Grasp as you say: Language-guided dexterous grasp generation. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). vol. 37, pp. 46881–46907 (2024)
59. Wen, Y., Pan, H., Yang, L., Pan, J., Komura, T., Wang, W.: Hierarchical temporal transformer for 3d hand pose estimation and action recognition from egocentric RGB videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21243–21253 (2023)
60. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: Oakink: A large-scale knowledge repository for understanding hand-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20953–20962 (2022)
61. Ye, Y., Gupta, A., Kitani, K., Tulsiani, S.: G-HOP: generative hand-object prior for interaction reconstruction and grasp synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1911–1920 (2024)
62. Ye, Y., Li, X., Gupta, A., De Mello, S., Birchfield, S., Song, J., Tulsiani, S., Liu, S.: Affordance diffusion: Synthesizing hand-object interactions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 22479–22489 (2023)
63. Yu, Z., Huang, Y., Furuta, R., Yagi, T., Goutsu, Y., Sato, Y.: Fine-grained affordance annotation for egocentric hand-object interaction videos. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 2155–2163 (2023)
64. Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular rgb image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2354–2364 (2019)

65. Zhang, Y., Cui, Z., Kephart, J.O., Ji, Q.: Diffusion-based 3d hand motion recovery with intuitive physics. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7306–7317 (2025)
66. Zhou, B., Zhan, Y., Zhang, Z., Lu, Z.: MEgoHand: Multimodal egocentric hand-object interaction motion generation. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS). vol. 38, pp. 49464–49490 (2025)
67. Zhou, Y., Ohkawa, T., Zhou, G., Goto, K., Hirose, T., Sekikawa, Y., Inoue, N.: Df-mamba: Deformable state space modeling for 3d hand pose estimation in interactions. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 5352–5363 (2026)

Affordance-Guided Diffusion Prior for 3D Hand Reconstruction

— Supplementary Material —

A Grasp Classification

As mentioned in Sec. 3.1, we adopt a grasp taxonomy classifier to classify grasp types from a single image. The overall architecture for grasp taxonomy classification is illustrated in Fig. 10. The input image is first encoded by a ViT to obtain visual tokens. In parallel, a MANO regressor predicts the hand pose parameters from the encoded visual tokens. Since grasp taxonomy depends not only on the hand pose but also on the visual context surrounding the hand-object interaction, we combine the visual tokens and the MANO pose parameters as a joint input to a self-attention module. We use the ViT encoder and the MANO regressor pretrained on [41]. The output of the self-attention module is then fed into an MLP layer to produce the final grasp taxonomy classification. For training, we use the HOGraspNet [4] dataset, adopting the same data size and train/validation split as in the AffHandGen training setup.

The overall classification accuracy is 87.6% on the evaluation split. We show the confusion matrix for grasp taxonomy classification in Fig. 13, where the original 33 grasp types are grouped into 6 categories based on grasp power and thumb posture, following the grouping introduced in [63] and Fig. 14. While predicted taxonomies are mostly accurate, we find some failure examples of our classification model in Fig. 11. These failure cases correspond to grasp types that the model frequently confuses in the confusion matrix. Although the predictions are incorrect, the misclassified grasp types are visually and functionally similar. As shown in Tab. 2 rows (6) and (7), the performance gain when replacing predicted grasps with the ground-truth is marginal, indicating that these failures in grasp classification have limited practical impact.

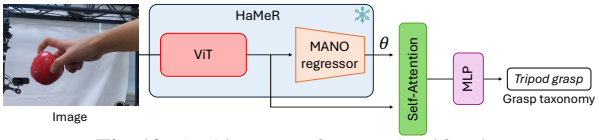

Fig. 10: Architecture of grasp classification.

Fig. 11: Failure cases of grasp classification.

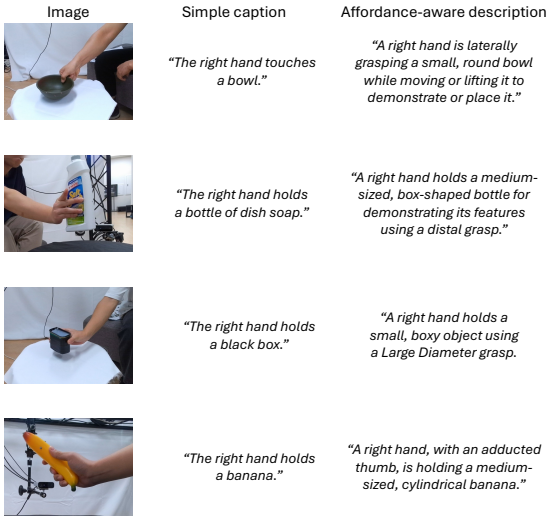

Fig. 12: Examples of generated simple captions and affordance-aware descriptions.

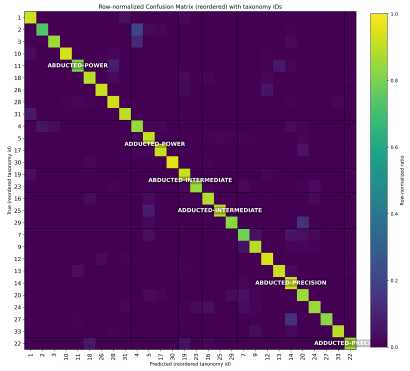

Fig. 13: Confusion matrix of grasp classification.

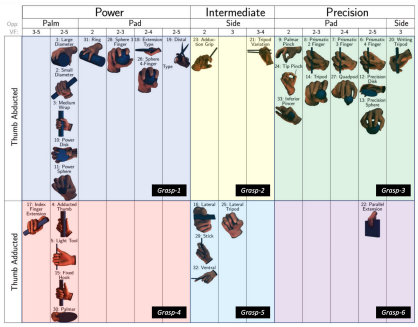

Fig. 14: The 33 grasp labels [12] are grouped into 6 categories. The figure is adapted from [63].

B Description Generation

We design two types of descriptions in Sec. 3.1: a simple caption and an affordance-aware description. These descriptions are generated using a VLM (Qwen2.5-VL [2]). Examples of both types of generated descriptions are provided in Fig. 12. Compared to the simple captions, our affordance-aware descriptions offer more detailed explanations of hand-object interaction, including grasp taxonomy, object size and shape, and other interaction-relevant attributes.

For simple caption generation, we feed the image together with the hand-object bounding boxes into the VLM and directly prompt it using the template shown in Fig. 15. In contrast, the affordance-aware description integrates additional structured cues. Specifically, we first feed the image and the hand-object bounding boxes into the VLM with the prompt shown in Fig. 16. In parallel, we perform grasp classifica-

tion on the input image and insert the predicted grasp taxonomy into a parsed caption template. The combined parsed caption is then provided to an LLM (Mistral-7B [19]), which summarizes it into an affordance-aware description. We prompt the LLM with "Summarize the given parsed captions in a single sentence starting with 'A right/left hand…'", and use the resulting summary as our caption.

```
You are an assistant that outputs a single short English phrase describing a hand-object
interaction.

Use ONLY this format:
The {right/left} hand [verb phrase] [object phrase]

Rules:
- Use a concise verb phrase (e.g., holding, touching, grasping, pointing).
- Use a concise noun phrase for the object (e.g., a cup, a smartphone).
- If no object is visible, use "nothing".
- Output exactly one short line, no punctuation, no explanations.
- Do NOT use parentheses, colons, or quotation marks.

Context:
- Hand bbox: (x1, y1, x2, y2) or none
- Object bbox: (x1, y1, x2, y2) or none

Now produce the output line starting with "The {right/left} hand".
```

Fig. 15: Prompt used to generate a simple caption.

```
Instructions:
1. Use the provided detected hand and object bounding boxes (x1, y1, x2, y2) as reference to
describe the interaction. If no bounding box is detected, use "none" for that field. The
bounding boxes are given below:
- Right hand bbox: (241, 172, 398, 307)
- Object bbox: (228, 0, 622, 363)

2. Focus on the following key aspects of the hand-object interaction:
- Shape and size: Describe the geometric shape of the object and its relative size to the hand.
- Contact and interaction: What physical relationship is happening between the hand and object?
- Pose and action: Describe how the hand is positioned and what action is being performed.

3. Use the following output keys only. Do not include any other fields, explanations, or
comments. Each value should be brief (<= 1-2 sentences) and formatted exactly as specified.

Write your answers in the following format:

1. hand_caption: ...
2. object_category: ...
3. object_shape: ...
4. object_size: ...
5. interaction_type: ...
6. intention: ...

Do not use JSON, curly brackets, or quotation marks. Only use the format above.
Output Keys and Descriptions:

1. hand_caption
- What to include: A concise description of the primary action the hand is performing.
- Write in the form: A right hand [does something]. Use present tense and active voice. Limit to
<= 25 words.

2. object_category
- What to include: The semantic category of the object the hand is interacting with.
- Format: Use a single noun (singular). If the object cannot be identified, use "unknown".

3. object_shape
- Describe the geometric shape in simple words like "boxy", "spherical", "flat", or others.
- Prefer short phrases. If unclear, use "unknown".

4. object_size
- What to include: The object's size relative to the size of the hand. Use: "tiny", "small",
"medium", "large", "huge", or "unknown".

5. interaction_type
- What to include: Describe the type of interaction the hand is performing with the object in a
concise verb phrase.
- Focus on the action itself (e.g., grabbing, tapping, twisting).
- You may use your own wording. If uncertain, use "unknown".

Examples: "grabbing", "tapping", "pushing", "supporting", "pointing", "twisting", "pressing a
button", "using as a tool", "resting on", "unknown"

6. intention
- What to include: The inferred high-level goal of the hand. Start with "to". Use <= 10 words.
If unclear, write "undetermined".
```

Fig. 16: Prompt used to generate an affordance-aware description.

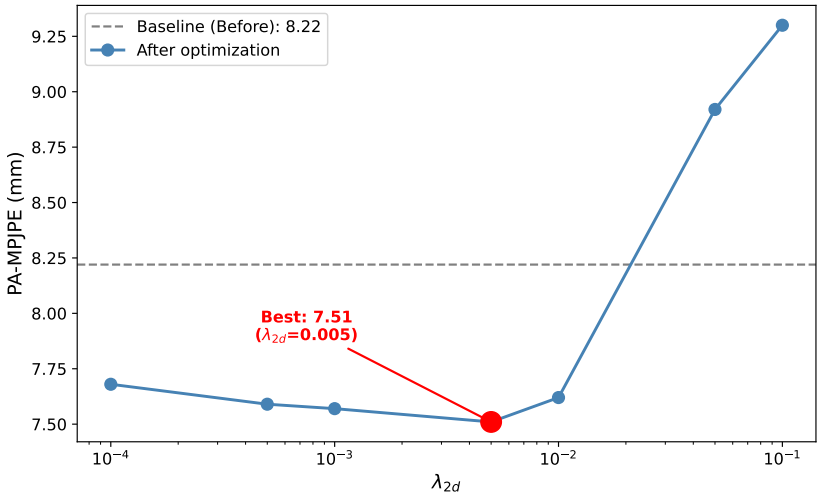

Fig. 17: Sensitivity analysis of the 2D reprojection loss weight λ_{2d} on the HOGraspNet dataset.

C Additional Results

Hyperparameter sensitivity: In the refinement stage (Sec. 3.3), we incorporate a 2D reprojection loss weighted by λ_{2d} to encourage consistency between the optimized 3D hand pose and 2D keypoint estimates from an off-the-shelf 3D hand reconstruction method. We analyze the model's sensitivity to λ_{2d} on the HOGraspNet dataset, keeping the maximum diffusion steps T=1000, start step $n \in [100, 1000]$, sampling interval m=20, and 100 post-processing steps, with image conditioning enabled. As shown in Fig. 17, sweeping λ_{2d} across four orders of magnitude (10^{-4}–10^{-1}) reveals a broad plateau of near-optimal performance in the range of 10^{-4} to 10^{-2}. We achieve the best PA-MPJPE of 7.51 mm at λ_{2d}=0.005 (compared to 8.22 mm before refinement). This broad stability region demonstrates that our refinement pipeline is highly robust and obviates the need for exhaustive hyperparameter tuning in practice.

In-the-wild results: To further assess generalization beyond the training datasets, we evaluate our method on HInt [41], which includes challenging in-the-wild images from the New Days, VISOR, and Ego4D splits. As shown in Tab. 4, our prior improves WiLoR on occluded joints across all splits and HaMeR on most splits, indicating its effectiveness under unconstrained real-world ambiguities. We further present qualitative results in Fig. 19, where our method produces plausible hand poses that remain consistent with the interaction context across diverse real-world scenarios.

Failure cases: We finally inspect cases where our prior offers limited benefit (Fig. 18). For interactions far from the training distribution, such as unusual grasps or rare object affordances in in-the-wild HInt images, affordance descriptions can become less reliable, leading to weaker refinement over the initial WiLoR estimate. Broadening the affordance vocabulary and grounding it in explicit 3D object geometry may help close this gap, as discussed in the limitations below.

Text-conditional generation: In Fig. 20, we provide the qualitative results of our diffusion-based pose sampling from the text input. Our diffusion prior generates hand

	Method	New Days			VISOR			Ego4D		
		@0.05↑	@0.1↑	@0.15↑	@0.05↑	@0.1↑	@0.15↑	@0.05↑	@0.1↑	@0.15↑
All	HandOccNet [39]	7.1	22.0	37.8	5.5	18.6	33.2	2.9	10.2	19.6
	MeshGraphormer [28]	15.0	34.7	49.1	15.9	39.0	55.2	7.2	19.8	31.8
	Hamba [8]	25.1	54.3	70.7	19.7	49.8	69.1	12.0	33.0	49.0
	HaMeR [41]	**27.7**	54.3	68.9	**25.7**	55.2	71.3	**18.3**	41.6	56.4
	+ **Ours**	24.6	52.6	68.5	24.3	**55.8**	72.8	17.5	**42.3**	**58.0**
	WiLoR [42]	24.9	**54.9**	71.6	21.9	53.5	72.6	12.5	34.2	50.3
	+ **Ours**	23.6	54.5	**72.8**	21.7	54.9	**74.8**	12.5	35.6	52.5
Visible	HandOccNet [39]	8.4	25.6	43.0	6.2	19.8	34.7	3.1	10.7	20.5
	MeshGraphormer [28]	18.0	40.3	55.0	21.9	49.5	64.6	9.4	24.1	36.1
	Hamba [8]	30.7	63.5	78.8	25.7	60.9	79.0	14.7	39.5	56.3
	HaMeR [41]	**33.2**	62.3	76.1	**33.0**	**65.5**	79.1	**23.6**	49.3	62.6
	+ **Ours**	29.3	60.5	75.8	29.9	64.6	79.8	21.2	**49.4**	**63.7**
	WiLoR [42]	30.0	**63.6**	79.9	27.4	63.9	81.8	15.4	40.7	57.4
	+ **Ours**	28.0	62.6	**80.5**	26.0	63.9	**83.0**	14.8	41.2	58.4
Occluded	HandOccNet [39]	4.6	16.1	30.1	4.7	17.0	31.1	2.6	9.3	18.0
	MeshGraphormer [28]	8.4	24.1	39.1	8.3	26.7	45.1	4.5	14.6	26.7
	Hamba [8]	14.6	41.0	60.6	12.6	36.9	58.6	8.6	25.3	41.1
	HaMeR [41]	**17.2**	**42.1**	59.6	16.0	43.0	62.7	11.6	32.5	49.4
	+ **Ours**	15.9	40.4	59.2	**16.6**	**45.4**	65.1	**12.6**	**33.7**	**51.5**
	WiLoR [42]	15.4	**42.1**	61.2	15.4	41.3	62.9	9.0	26.6	42.5
	+ **Ours**	15.1	**42.1**	**63.0**	16.3	44.3	**66.1**	9.5	28.6	45.5

Table 4: Quantitative evaluation on the HInt benchmark [41] (PCK@0.05/0.1/0.15). On occluded joints, our affordance-guided prior improves WiLoR across all three splits (New Days, VISOR, Ego4D) and HaMeR on most of them.

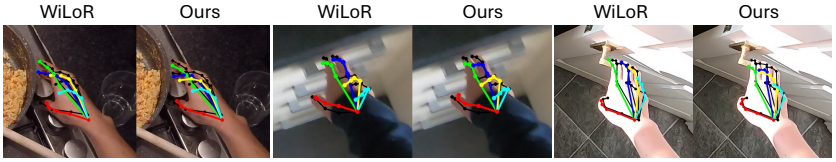

Fig. 18: Failure cases on HInt. When the interaction deviates from the training distribution, the affordance prior provides weaker refinement over the initial WiLoR estimate.

poses semantically aligned with affordance-aware descriptions, with high sample diversity.

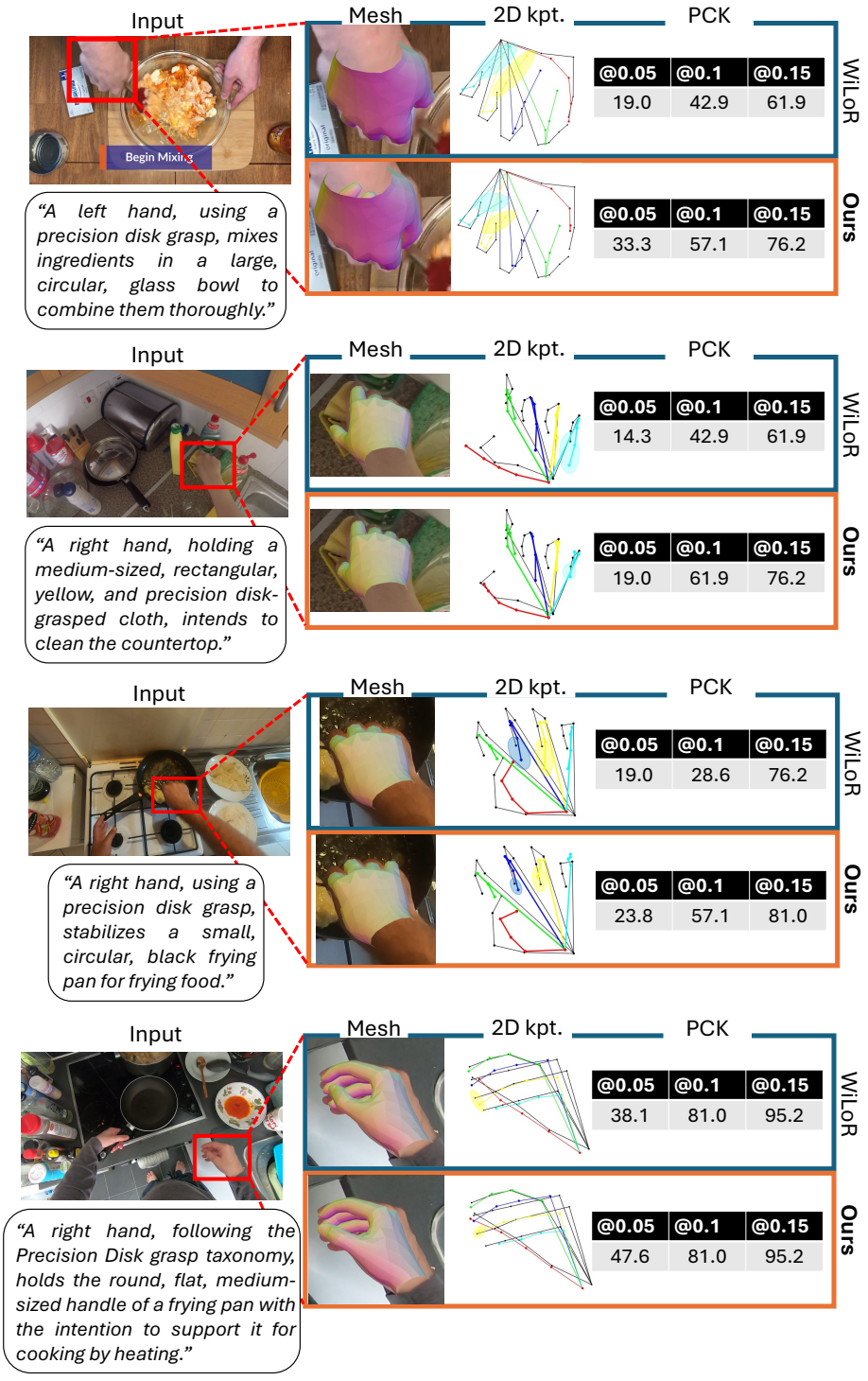

Fig. 19: Additional qualitative results on in-the-wild images from the HInt dataset [41].

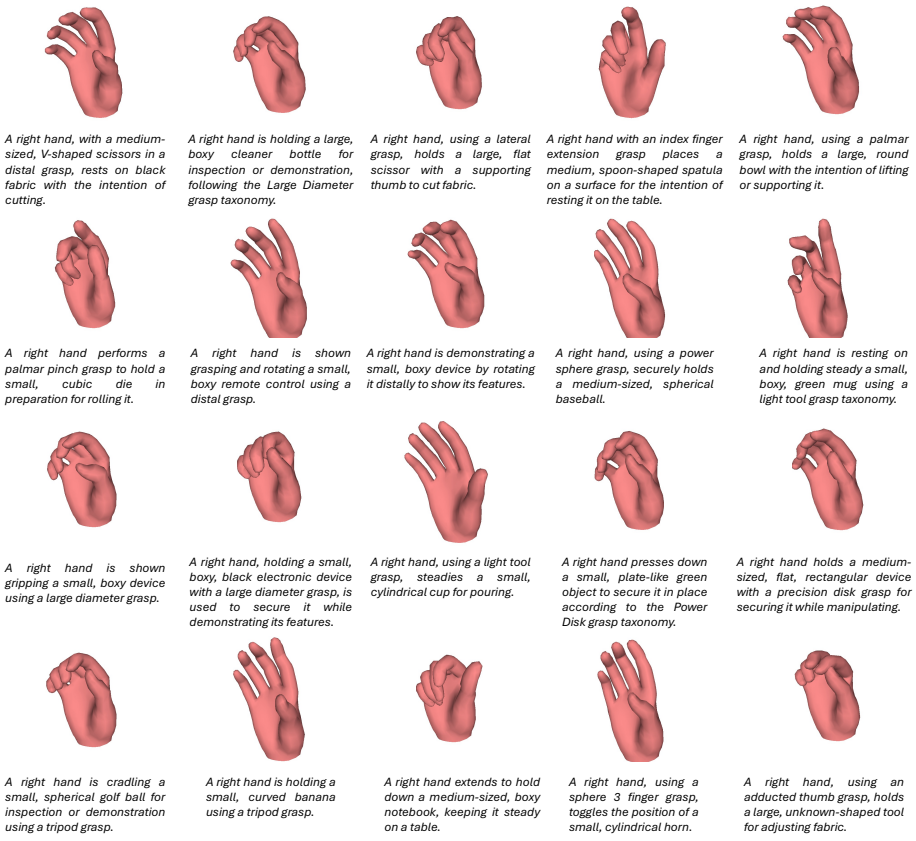

Fig. 20: Sampled by our affordance-description conditioned prior.