Affordance-Guided Diffusion Prior for 3D Hand Reconstruction

Naru Suzuki[1] Takehiko Ohkawa[1] Tatsuro Banno[1] Jihyun Lee[2] Ryosuke Furuta[1] Yoichi Sato[1]
[1]The University of Tokyo [2]KAIST

Abstract

How can we reconstruct 3D hand poses 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 hand pose refinement 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 affordance descriptions, which are inferred from a large vision-language model (VLM). This enables the refinement of occluded regions into more accurate and functionally coherent hand poses. Extensive experiments on HOGraspNet, a 3D hand-affordance dataset with severe occlusions, demonstrate that our affordance-guided refinement significantly improves hand pose estimation over both recent regression methods and diffusion-based refinement lacking contextual reasoning.*

1. Introduction

Reconstructing 3D hand pose from a single RGB image is crucial for applications in virtual/augmented reality and robotic dexterous manipulation. In everyday life, hand-object interactions (HOI) are ubiquitous, yet they often involve severe occlusions that make accurate hand pose estimation particularly challenging. Despite these difficulties, humans can still infer plausible hand poses by exploiting *contextual knowledge* beyond the hand itself, such as cues from objects, environments, and the person's intent. In particular, the geometry and function of interacting objects can provide strong cues for determining plausible hand configurations during interaction.

Most previous works on 3D hand reconstruction [5, 17, 18] have paid little attention to 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) [30], (b) short interaction captions (*e.g.*, "a hand holding a spoon") [28], (c) action labels (*e.g.*, "screw", "open") [13, 15], (d) intention labels (*e.g.*, intend to "use" or "receive" an object) [27], etc. These semantic cues are used as conditional input in training [25, 27, 28, 30], auxiliary learning target [30], or test set design for evaluation [3, 13, 15]. However, such contextual signals remain coarse and fail to capture the rich details needed to resolve ambiguities in 3D hand reconstruction.

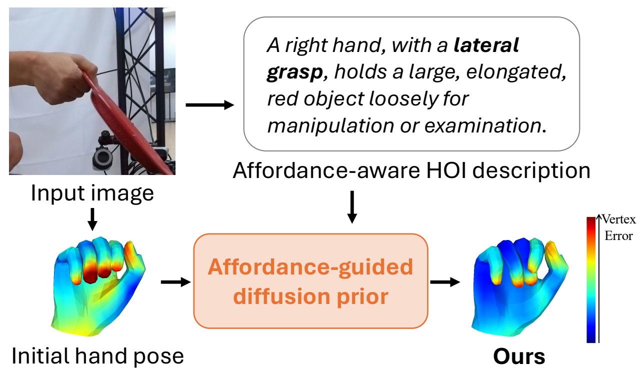

Figure 1. **Can affordance-aware textual descriptions benefit 3D hand reconstruction?** We develop an affordance-guided diffusion prior model that refines 3D hand pose into more accurate and functionally coherent poses. Given the initial pose estimates from HaMeR [17], our method achieves robust hand pose refinement under self-occlusion and object-occlusion on the HOGraspNet dataset [3]. The vertex error is color-coded on the hand mesh.

Among these contextual signals, *affordances* stand out as a promising representation for interaction understanding [1, 4, 8], which directly encode how object properties constrain and guide possible hand actions. Existing formulations of affordances fall into (i) motor actions and (ii) grasp types [29]. The motor actions are typically represented by verb labels (*e.g.*, "pick-up") [6], but such labels are coarse and unable to differentiate non-identical object functionalities (*e.g.*, "pick-up bowl" *vs.* "pick-up knife"). In contrast, the grasp types have been studied with rigorous taxonomies [3, 7], which provide precise definitions of hand poses with fine-grained structure.

Based on this observation, we adopt affordances as our contextual representation for 3D hand reconstruction, while further extending beyond categorical labels into rich tex-

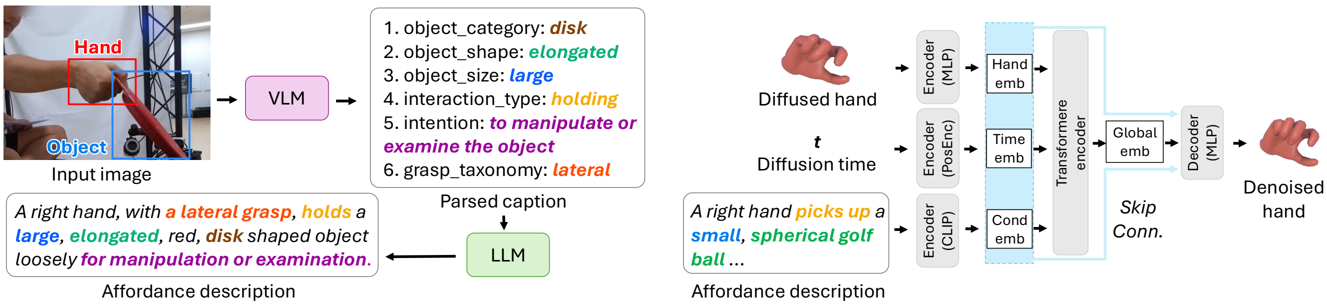

Figure 2. **Affordance description generation with VLM (left) and affordance-guided diffusion prior for hand poses (right).** Our proposed description generation scheme consists of two steps. The first is to obtain a parsed caption from the image and hand-object bounding boxes with VLM (QwenVL2.5 [2]). The second is to summarize the parsed caption using an LLM (Mistral-7B [10]) to obtain detailed descriptions of affordances. Our proposed model is trained to generate hand poses that align with the generated descriptions. The overall architecture is inspired by InterHandGen [12], by replacing the counter hand conditioning with affordance descriptions.

tual descriptions. In addition to incorporating both motor actions and grasp types, we propose to capture object attributes such as category, shape, and size, as well as the intent underlying the activity. These enriched descriptions can serve as more fine-grained contextual cues that directly inform hand interaction patterns in 3D. To achieve this, we employ a large vision-language model (VLM), QwenVL2.5 [2], to encode these affordances properties and aggregate them into a single description (Fig. 1).

With these affordance descriptions, we propose generative prior modeling that can integrate textual descriptions into the learning of 3D pose. Inspired by text-motion generation [14, 23, 24], we design a diffusion-based generative prior guided by affordance descriptions. We train a denoising diffusion model [9] on hand pose data, while we newly adopt affordance descriptions encoded with CLIP [19] as conditional features. We further utilize this prior to refine 3D pose estimates, inspired by [16]. Specifically, starting from an initial hand pose from HaMeR [17], we iteratively denoise the pose so that it remains consistent with the learned prior, while aligning its 2D projection with the visible keypoints. To improve the refinement, we further detect occluded joints to guide only uncertain joints with the prior, while leveraging textual descriptions for test images.

In our experiments, we use HOGraspNet [3], which provides affordance annotations such as grasp taxonomy. Our affordance-guided diffusion prior consistently outperforms recent regression methods such as HaMeR [17]. It also surpasses a simplified variant of InterHandGen [12], where the one-hand conditioning is removed to adapt to our task. Furthermore, we show through qualitative results that our method produces grasp poses that not only appear more natural and physically plausible 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, where a VLM provides rich and fine-grained contextual knowledge.
- We design a diffusion-based framework that jointly learns affordance descriptions and hand poses, enabling controllability through affordance-guided conditioning.
- We propose a refinement approach that leverages the diffusion prior, yielding substantial improvements in pose accuracy, particularly under heavy occlusion.

2. Method

We propose leveraging textual descriptions of affordances for 3D hand reconstruction and diffusion models for learning across text and pose. Our approach begins with an affordance description generation, which extracts key elements of affordances from input images using a vision-language model, as described in Section 2.1. We then employ a diffusion-based generative prior conditioned on the affordance descriptions to model the distribution of plausible 3D hand poses, as detailed in Section 2.2. Subsequently, the learned prior is used for the refinement to correct occluded or uncertain joints in initial estimates, yielding more accurate and functionally coherent hand poses.

2.1. Affordance description generation with VLM

Affordances inherently encode how an object can be interacted with hands by capturing its actionable properties; thus we utilize them as an effective representation and guidance for 3D hand reconstruction. Particularly, beyond categorical labels in existing studies for visual affordances [3, 11], we enrich their form to textual descriptions that reflect both physical and functional properties, such as object category, shape, size, interaction (action), intention, and grasp type.

We generate such affordance descriptions step-by-step in two phases. Our preliminary study finds that direct (one-step) captioning, by prompting to generate a description from an image, often produces incomplete or ambiguous outputs due to the complexity of the task. To address this, we are inspired by the chain-of-thought reasoning [26] of LLMs to break down the task into two steps: (i) extracting each key element independently and (ii) integrating all elements into a final description. This procedure enables 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 (QwenVL2.5 [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 six key elements that characterize HOI properties: object category (*e.g.*, "disk", "sphere"), object shape (*e.g.*, "elongated", "flat"), object size (*e.g.*, "small", "large"), interaction type (*e.g.*, "holding", "reaching"), which is equal to motor action as in [29], intention (*e.g.*, "to lift", "to manipulate") as [27], and grasp taxonomy (28-class [7]). These elements are chosen because they comprehensively capture both the physical properties of the object and the functional aspects of the interaction, forming a rich affordance description.

Step 2 – Summarize with LLM: Given the short captions for each element, we use LLM (Mistral-7B [10]) with a summarization prompt to combine these elements into a final affordance description. We find that classifying grasp taxonomy with existing vision-language models is highly challenging. Consequently, we rely on the ground truth annotations provided by HOGraspNet [3] to supply accurate taxonomy information for our framework.

2.2. Affordance-guided diffusion prior

We present our diffusion prior conditioned on affordance descriptions, which learns to generate and refine 3D hand poses aligned with the text condition. We choose diffusion models to learn data distributions across different modalities, such as text and pose, inspired by their successful results in text-motion synthesis [12, 14, 24]. Diffusion-based modeling also benefits from being directly applicable to various downstream tasks without additional training [22], thus making it both flexible and broadly reusable.

Overview: The overall architecture is illustrated in Fig. 2 (right). For the input, we use the pose parameters of MANO [21] as the target variables for our diffusion model, *i.e.*, $\mathbf{x}_0 = \boldsymbol{\theta} \in \mathbb{R}^{15\times 3}$. For conditioning input, we encode affordance descriptions using CLIP [19], which yields a textual feature $\mathbf{c}$. The resulting diffusion model is then formulated as $\hat{\mathbf{x}}_0 = f_\phi(\mathbf{x}_t, t, \mathbf{c})$, where $\mathbf{x}_t$ is the diffused pose at diffusion time step t.

Single-view pose refinement: Our refinement process is illustrated in Fig. 3. We adopt a refinement strategy similar to [16]. Given the hand poses inferred from a single-view estimator HaMeR [17], the occluded parts of the hand joints

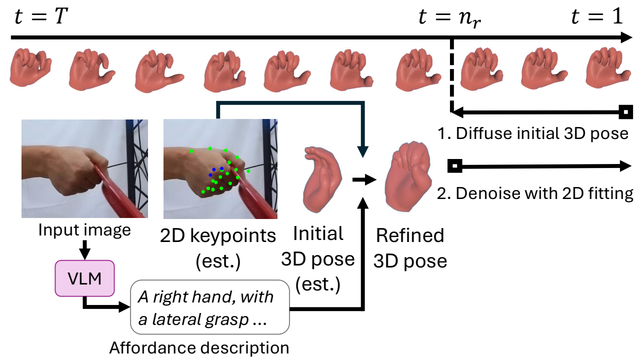

Figure 3. **Single-view pose refinement.** We diffuse the initial 3D hand poses for n_r steps and denoise it with 2D keypoint fitting, inspired by [16]. Our refinement utilizes affordance descriptions as the condition and corrects occluded joints (blue), while visible joints (green) remain fixed. Occlusion labels are obtained from two criteria: Self-occlusion (ray casting on the MANO mesh) and object-occlusion (SAM2 mask).

are often inaccurately estimated due to the lack of visual evidence. To address this, we develop a single-view pose refinement framework by adapting our diffusion prior to the visible 2D keypoints, where only the occluded joints are refined with this prior.

We employ two criteria to determine the occlusion of hand joints. (1) **Self-occlusion judged by ray-surface intersection:** We determine self-occluded joints by casting a ray toward the camera and computing its intersection over the hand surface. (2) **Object-occlusion judged by mask overlap:** We find object-occluded joints by comparing the projected 2D joints from the initial pose predictions and the hand mask obtained from SAM2 [20]. If the point is outside the mask, the joint is regarded as occluded by the object.

3. Experiments

We present quantitative results on single-view pose refinement, followed by qualitative analyses of pose correction and controlled generation.

Experimental settings: We evaluate on HOGraspNet [3], a large-scale benchmark with severe occlusions, including 1.5M RGB images, 3D hand poses, and grasp taxonomy labels of HOI scenes. The dataset spans 28 grasp taxonomies [7], covering a wide range of hand poses, object categories, and interaction types. We sub-sample one-tenth of the dataset, which is sufficient to preserve the diversity of grasp types and occlusion levels for reliable evaluation. The dataset is split by subjects following the S1 protocol [3]. We set the maximum diffusion steps to $T = 1000$, and the added noise level n_r is chosen linearly in the range $[100, 1000]$ according to the number of occluded joints, with larger occlusion counts corresponding to larger n_r.

Single-view pose refinement: Table 1 reports PA-MPJPE

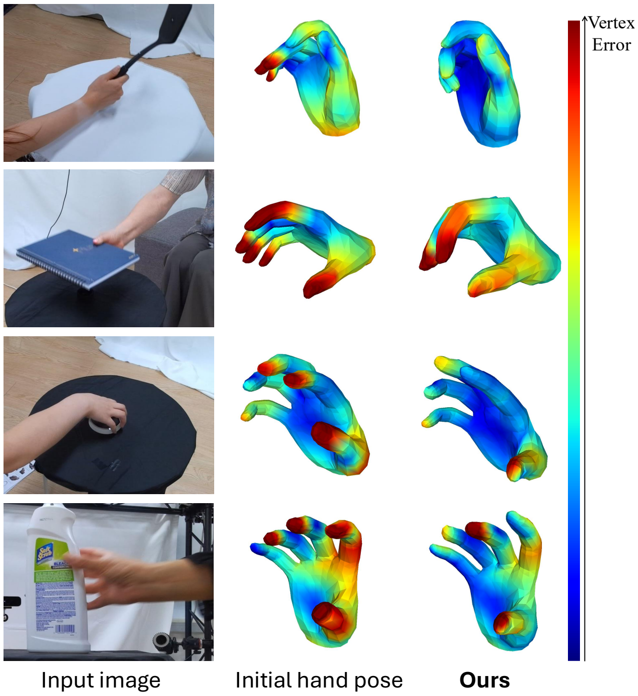

Figure 4. **Qualitative results of our diffusion-based refinement.** We show examples of refinement result with our diffusion prior with the vertex error is color-coded on the hand mesh. While the initial hand pose from HaMeR [17] struggles to estimate joints in occluded regions, our method reasonably refines them so that the hand can plausibly grasp the object even under occlusion.

Method	Avg.	Low	Med.	High
HaMeR [17]	8.51	7.69	11.8	21.3
+ IHGen [12] (w/o text)	8.59	7.78	12.3	19.0
+ Ours (w/ text)	**8.28**	**7.67**	**10.9**	**16.4**

Table 1. **Results of hand pose refinement with description.** We evaluate our diffusion-based refinement on the test set, given initial estimates from HaMeR [17]. We report PA-MPJPE (mm) for the overall average and different occlusion levels: Low (0–5 occluded joints), Medium (6–10), and High (11+) out of 21 joints. We sample 300 instances for this evaluation.

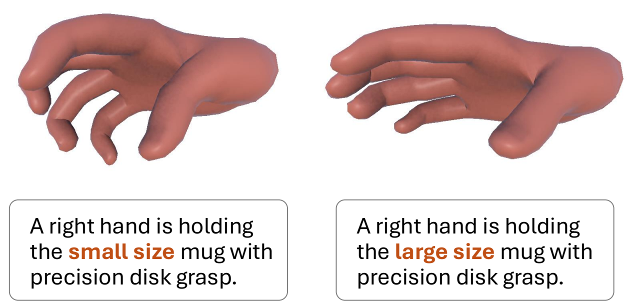

Figure 5. **Qualitative results of our diffusion prior.** We show samples generated from partially different descriptions. We find that the diffusion-based generation responds to object size, while remaining consistent with the grasp taxonomy.

on the test set. We implement an unconditional baseline using InterHandGen (IHGen) [12]. This baseline shows limited refinement, yielding a 2.3 mm reduction under high occlusion but even degrading performance at lower occlusion levels. In contrast, our diffusion prior consistently improves performance across all occlusion levels, achieving an average reduction of 0.23 mm over HaMeR [17], and a substantial 4.9 mm (25%) reduction under high occlusion. These results highlight the effectiveness of leveraging affordance descriptions, particularly for reasoning about highly occluded joints.

Fig. 4 presents qualitative comparisons between the initial hand poses estimated by HaMeR and the refined poses obtained with our diffusion prior, along with vertex error heatmaps compared to the ground-truth mesh. While the initial hand poses exhibit large errors in the occluded parts (*e.g.*, the middle to pinky fingers in the top row), our method significantly reduces these errors, yielding more plausible and accurate hand poses.

Generation controllability of our diffusion prior: Fig. 5 presents qualitative examples of our diffusion model given different textual descriptions. We validate how pose generation (without any reference images) responds to different affordance elements while keeping 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 not only yields kinematically plausible poses but also faithfully incorporates actionable semantics embedded in the affordance descriptions.

4. Conclusion

We presented a novel diffusion-based hand pose refinement framework that leverages affordance descriptions generated by a vision-language model. By conditioning on affordance descriptions, our model learns a distribution of plausible and functionally coherent hand poses, which enables robust reasoning about occluded joints and yields more accurate reconstructions in challenging hand-object interaction scenarios. Our experiments on HOGraspNet demonstrate that the guidance by affordance descriptions substantially improves refinement quality over both regression-based methods and diffusion priors without contextual conditioning, with strong gains under severe occlusions. Our method also provides controllable and interpretable refinements, aligning pose generation with affordance descriptions.

References

[1] Shikhar Bahl, Russell Mendonca, Lili Chen, Unnat Jain, and Deepak Pathak. Affordances from human videos as a versatile representation for robotics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13778–13790, 2023. 1

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3

[3] Woojin Cho, Jihyun Lee, Minjae Yi, Minje Kim, Taeyun Woo, Donghwan Kim, Taewook Ha, Hyokeun Lee, Je-Hwan Ryu, Woontack Woo, et al. Dense hand-object (ho) graspnet with full grasping taxonomy and dynamics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 284–303, 2024. 1, 2, 3

[4] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5031–5041, 2020. 1

[5] Zicong Fan, Takehiko Ohkawa, Linlin Yang, Nie Lin, Zhishan Zhou, Shihao Zhou, Jiajun Liang, Zhong Gao, Xuanyang Zhang, Xue Zhang, et al. Benchmarks and challenges in pose estimation for egocentric hand interactions with objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 428–448. Springer, 2024. 1

[6] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2139–2147, 2018. 1

[7] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on human-machine systems*, 46(1):66–77, 2015. 1, 3

[8] Mohit Goyal, Sahil Modi, Rishabh Goyal, and Saurabh Gupta. Human hands as probes for interactive object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3293–3303, 2022. 1

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 2

[10] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023. 2, 3

[11] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. Learning human activities and object affordances from rgb-d videos. *The International journal of robotics research*, 32(8):951–970, 2013. 2

[12] Jihyun Lee, Shunsuke Saito, Giljoo Nam, Minhyuk Sung, and Tae-Kyun Kim. Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 527–537, 2024. 2, 3, 4

[13] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. 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)*, pages 21013–21022, 2022. 1

[14] Xintao Lv, Liang Xu, Yichao Yan, Xin Jin, Congsheng Xu, Shuwen Wu, Yifan Liu, Lincheng Li, Mengxiao Bi, Wenjun Zeng, et al. Himo: A new benchmark for full-body human interacting with multiple objects. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–318, 2024. 2, 3

[15] Takehiko Ohkawa, Kun He, Fadime Sener, Tomas Hodan, Luan Tran, and Cem Keskin. Assemblyhands: Towards egocentric activity understanding via 3d hand pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12999–13008, 2023. 1

[16] Takehiko Ohkawa, Jihyun Lee, Shunsuke Saito, Jason Saragih, Fabian Prado, Yichen Xu, Shoou-I Yu, Ryosuke Furuta, Yoichi Sato, and Takaaki Shiratori. Generative modeling of shape-dependent self-contact human poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 906–915, 2025. 2, 3

[17] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9826–9836, 2024. 1, 2, 3, 4

[18] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. 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)*, pages 12242–12254, 2025. 1

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763. PmLR, 2021. 2, 3

[20] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[21] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (ToG)*, 36(6): 245:1–245:17, 2017. 3

[22] Anastasis Stathopoulos, Ligong Han, and Dimitris Metaxas. Score-guided diffusion for 3d human recovery. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 906–915, 2024. 3

[23] Guy Tevet, Brian Gordon, Amir Hertz, Amit H Bermano, and Daniel Cohen-Or. Motionclip: Exposing human motion generation to clip space. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 358–374, 2022. 2

[24] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano. Human motion diffusion model. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 2, 3

[25] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1148–1159, 2025. 1

[26] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 24824–24837, 2022. 3

[27] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. 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)*, pages 20953–20962, 2022. 1, 3

[28] Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. 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)*, pages 1911–1920, 2024. 1

[29] Zecheng Yu, Yifei Huang, Ryosuke Furuta, Takuma Yagi, Yusuke Goutsu, and Yoichi Sato. Fine-grained affordance annotation for egocentric hand-object interaction videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2155–2163, 2023. 1, 3

[30] Yufei Zhang, Zijun Cui, Jeffrey O Kephart, and Qiang Ji. Diffusion-based 3d hand motion recovery with intuitive physics. *arXiv preprint arXiv:2508.01835*, 2025. 1