HANDCRAFTER: MONOCULAR 3D HAND MOTION RECONSTRUCTION WITH MULTI-VIEW VIDEO DIFFUSION PRIORS

Anonymous authors
Paper under double-blind review

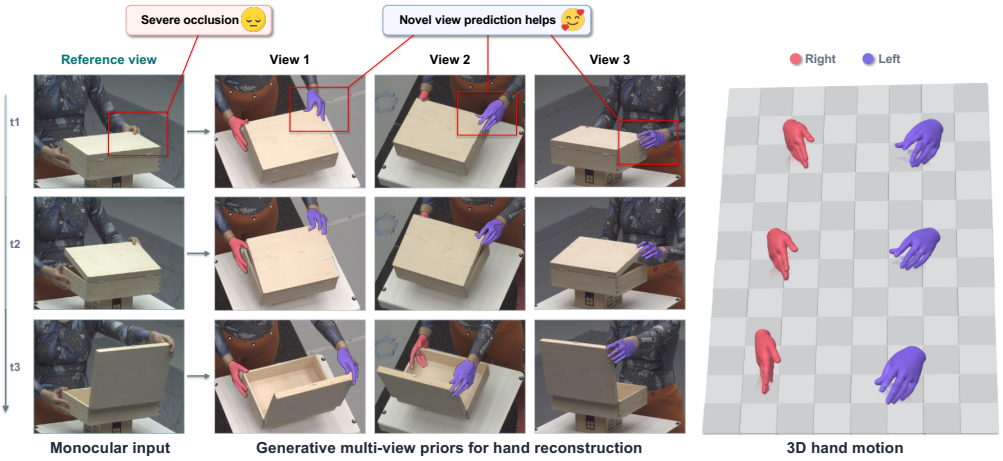

Figure 1: We introduce **HandCrafter**, a generative multi-view framework for monocular 3D hand motion reconstruction under severe hand-object occlusions. HandCrafter synthesizes complementary viewpoints to provide spatial cues for recovering hidden hand articulations and reconstructs temporally coherent 3D hand motion. The generated views are shown for visualization, while our hand estimator directly exploits the underlying multi-view diffusion representations.

ABSTRACT

Monocular 3D hand motion reconstruction remains challenging in hand-object interaction videos due to persistent occlusions and the resulting spatial ambiguity. We present **HandCrafter**, a generative multi-view framework that reconstructs 3D hand motion by synthesizing complementary viewpoints from a single monocular video. We first leverage an off-the-shelf geometry foundation model to warp the input video into novel viewpoints, producing incomplete but spatially aligned conditions that are subsequently completed by an adapted video diffusion model. Hand motion is then recovered from the resulting multi-view diffusion representations. Instead of relying on hand detection and cropping, our estimator uses sparse 2D joint anchors for spatial localization and decodes unified MANO parameters from cross-view diffusion representations, enabling crop-free 3D hand motion reconstruction. To mitigate the scarcity of multi-view HOI data, we further mix limited multi-view data with scalable monocular supervision, enabling cross-view and temporal modeling to benefit from a broader training distribution. Extensive experiments demonstrate superior performance under severe occlusions and strong generalization across diverse in-the-wild HOI scenarios.

1 INTRODUCTION

3D hand motion reconstruction from monocular videos has long attracted significant attention due to its broad applications in virtual reality, augmented reality, and embodied interaction. However, recovering accurate hand motion remains particularly challenging in monocular hand-object interaction (HOI) videos, where fingers and palms are frequently occluded by manipulated objects. Such

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

occlusions can persist over multiple frames, leaving critical hand articulations unobserved throughout the input sequence. While temporal cues help maintain motion continuity, persistent occlusions limit their ability to resolve the spatial ambiguity of hidden hand structures.

Existing methods have sought to alleviate this ambiguity by introducing increasingly powerful visual and temporal priors. Large-scale image-based models such as HaMeR Pavlakos et al. (2024) and WiLoR Potamias et al. (2025) substantially improve in-the-wild hand reconstruction through data and model scaling. Video-based approaches further exploit temporal information to recover coherent hand motion and trajectories; for example, HaWoR Zhang et al. (2025b) reconstructs world-space hand motion from egocentric videos, while OmniHands Lin et al. (2026) models interactive hand motion through 4D temporal reasoning. More recently, ViDiHand Wang et al. (2026) demonstrates that pretrained video diffusion models provide strong representations for hand motion reconstruction. Despite these advances, existing methods primarily infer hand motion from visual and temporal evidence within the observed monocular sequence, without explicitly constructing complementary viewpoints to provide additional spatial cues.

Motivated by this observation, we explore whether synthesized complementary views can provide useful spatial cues for monocular hand motion reconstruction under persistent occlusion. Recent advances in camera-controlled video diffusion Ren et al. (2025); Bai et al. (2025b;a); Yu et al. (2025a) offer a promising solution for generating such views while preserving temporal coherence. Building on this capability, we introduce a multi-view video diffusion prior that leverages 3D geometry estimated by an off-the-shelf foundation model to construct spatially aligned conditions for complementary view synthesis. Rather than decoding the generated multi-view latents into auxiliary images, we directly exploit the spatial and visual cues encoded in these representations as priors to resolve ambiguities in the original viewpoint. Joint reasoning across these multi-view representations enables the reconstruction to leverage both temporal context and complementary spatial information for unified 3D hand motion recovery.

Specifically, we present **HandCrafter**, a framework that couples multi-view generation with structured hand motion estimation. Given a monocular interaction video, we use an off-the-shelf geometry foundation model to warp the input video to novel viewpoints, producing incomplete but spatially aligned conditions that are completed by an adapted video diffusion model. During the multi-view denoising process, we extract hand features in a feed-forward manner from the multi-view diffusion latent representations, which are subsequently aggregated by our proposed MANO decoder to predict a unified set of bimanual MANO parameters. This design allows generative view completion to provide complementary spatial priors for monocular reconstruction without requiring additional real views at inference time. Quantitative and qualitative results demonstrate that our model is robust to severe occlusions and generalizes well to in-the-wild scenarios.

Our main contributions are as follows:

- We introduce, to our knowledge, the first generative multi-view framework for monocular 3D hand motion reconstruction that synthesizes complementary viewpoints through video diffusion and integrates their spatial and visual cues for unified 3D motion recovery.
- Instead of relying on detection-based hand crops for MANO estimation, we introduce sparse 2D joint anchors for hand localization and decode unified MANO parameters from cross-view diffusion representations, enabling crop-free 3D hand motion reconstruction.
- To alleviate the scarcity of multi-view HOI data, we devise a hybrid training strategy that combines multi-view supervision with scalable monocular degraded-to-clean video pairs, enabling the generative prior and hand estimator to generalize beyond limited multi-view training data.

2 RELATED WORK

Monocular 3D Hand Reconstruction. Monocular 3D hand reconstruction commonly employs parametric models such as MANO Romero et al. (2017) to recover hand pose and shape from visual observations. Early efforts improve generalization to in-the-wild interactions by reducing the domain gap between motion capture data and natural images Boukhayma et al. (2019); Moon (2023). HaMeR Pavlakos et al. (2024) further improves reconstruction accuracy and robustness by scaling transformer models and training data, while WiLoR Potamias et al. (2025) integrates hand localization and mesh recovery into an end-to-end framework. Beyond per-frame estimation, HMP Duran

et al. (2024) introduces a learned generative motion prior to recover temporally consistent hand pose and shape under occlusion. Subsequent approaches combine camera motion estimation with hand reconstruction to recover global trajectories from monocular videos Yu et al. (2025b); Zhang et al. (2025b). Other methods exploit attention mechanisms to aggregate temporal context and model hand relationships for coherent motion recovery Ye et al. (2026); Lin et al. (2026). Our method complements these advances by directly incorporating spatial cues from multi-view video diffusion latents to improve hand motion reconstruction under persistent occlusion.

Generative Priors for Hand Reconstruction. Generative modeling enables learning priors over hand geometry and motion for reconstruction under diverse conditioning signals. Early diffusion-based approaches estimate hand poses and meshes through image-conditioned denoising Li et al. (2023); Ivashechkin et al. (2023). The flexibility of conditional generation also supports unified formulations: HHMR Li et al. (2024) employs a multi-modally controllable graph diffusion model for mesh generation, inpainting, reconstruction, and fitting, while UniHand Sun et al. (2026) brings motion estimation and generation into a shared latent diffusion framework. Beyond visual conditioning, physical motion constraints and affordance-aware interaction descriptions guide diffusion-based refinement of hand motion and poses Zhang et al. (2025c); Suzuki et al. (2026). These methods primarily model generative distributions of hand geometry or motion. More recently, ViDiHand Wang et al. (2026) explores a different source of prior knowledge by adapting pretrained video diffusion representations for hand motion reconstruction.

Diffusion Priors for 3D/4D Reconstruction. Diffusion models have increasingly served as powerful priors for recovering 3D structure from limited observations, from view-conditioned object reconstruction Zhou & Tulsiani (2023); Liu et al. (2024a); Wu et al. (2025b); Chang et al. (2026) to dynamic scene and human modeling with synthesized multi-view videos Xie et al. (2025); Wu et al. (2025a); Jin et al. (2025); Zhi et al. (2025). Recent works further strengthen the coupling between generation and geometry. FantasyWorld Dai et al. (2026) jointly models video and 3D structure, while NeoVerse Yang et al. (2026) combines 4D reconstruction with diffusion-based novel-view generation. Diffusion-based modeling has also been explored for fine-grained 3D hand interaction generation Zhang et al. (2025a; 2026) and HOI reconstruction Chen et al. (2026); Dang et al. (2026b;a). These advances motivate us to leverage multi-view video diffusion as a spatial prior for 3D hand motion reconstruction.

3 METHOD

3.1 OVERVIEW

Given a monocular hand-object interaction video $\boldsymbol{V}_0 \in \mathbb{R}^{T\times H\times W\times 3}$ and its camera intrinsic matrix $\boldsymbol{I}_0 \in \mathbb{R}^{3\times 3}$, our goal is to recover the 3D motion of both hands in the camera coordinate system. For each hand, we represent its motion using MANO parameters $\mathcal{M}$, consisting of the hand pose $\boldsymbol{\theta} \in \mathbb{R}^{T\times 45}$, global translation $\boldsymbol{t} \in \mathbb{R}^{T\times 3}$, global orientation $\boldsymbol{o} \in \mathbb{R}^{T\times 3}$, shape parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$, and a visibility confidence score $c \in [0, 1]$ indicating whether the hand is present in the frame.

To address severe occlusions in HOI videos, we present **HandCrafter**, to our knowledge, the first framework that leverages generative multi-view video priors for 3D hand motion reconstruction. Given n target viewpoints specified by camera intrinsics $\{\boldsymbol{I}_i\}_{i=1}^n$ and extrinsics $\{\boldsymbol{E}_i\}_{i=1}^n$, where $\boldsymbol{I}_i \in \mathbb{R}^{3\times 3}$ and $\boldsymbol{E}_i \in \mathbb{R}^{4\times 4}$, HandCrafter synthesizes the complementary videos $\{\boldsymbol{V}_i\}_{i=1}^n$ and aggregates visual evidence across the generated viewpoints. The resulting multi-view representation is then decoded in a feed-forward manner to predict the MANO parameters of the left and right hands, $\mathcal{M}_L, \mathcal{M}_R$. The reconstructed MANO meshes are geometrically consistent across views, allowing their projections to align with the observed hands in both the original and generated viewpoints.

3.2 MULTI-VIEW HOI VIDEO GENERATION

Our multi-view video generator is built upon Wan2.1-T2V-1.3B Wan et al. (2025), a latent video generative model based on a Diffusion Transformer (DiT) Peebles & Xie (2023) and flow matching Lipman et al. (2023). Given a monocular source-view video $\boldsymbol{V}_0$, we adapt the model to jointly synthesize videos $\boldsymbol{V}_1, \ldots, \boldsymbol{V}_n$ from n novel camera viewpoints. The resulting multi-view latent representations are subsequently used as spatial priors for 3D hand motion reconstruction.

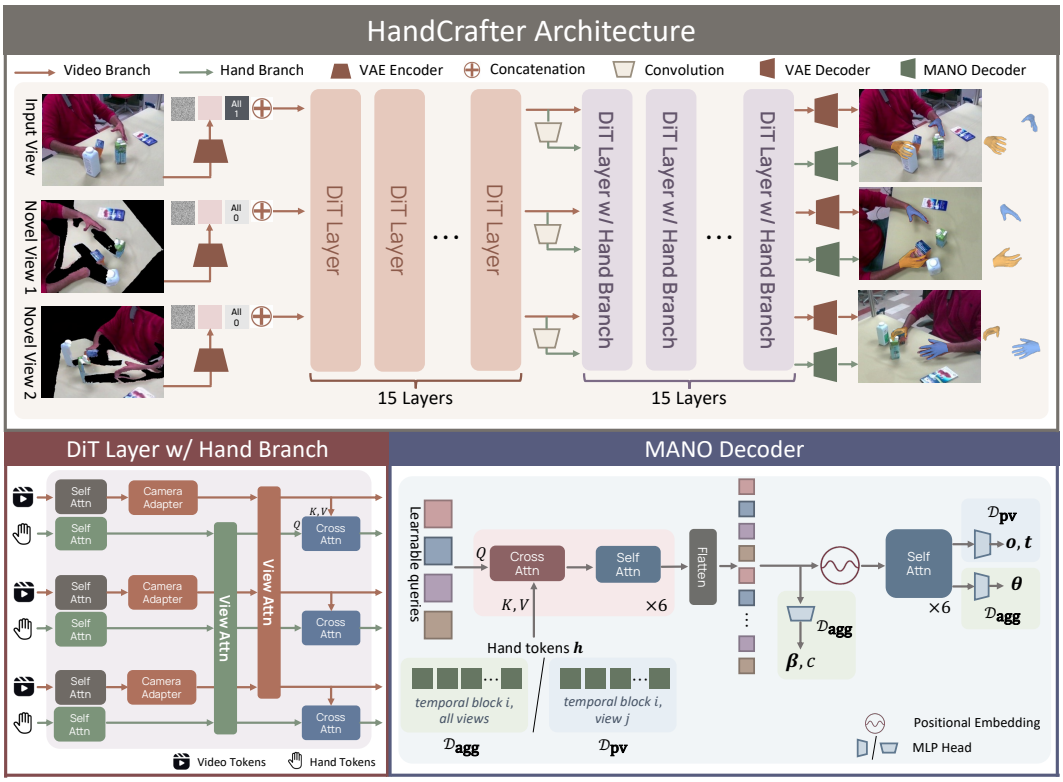

Figure 2: **Overview of HandCrafter architecture.** A multi-view diffusion backbone jointly models the input and complementary views, while a parallel hand branch extracts cross-view representations for 3D hand motion reconstruction. The hand branch integrates intra-view, cross-view, and video-to-hand interactions, and the MANO decoder combines cross-view aggregation $\mathcal{D}_{\text{agg}}$ with per-view decoding $\mathcal{D}_{\text{pv}}$ to reconstruct unified 3D hand motion. For clarity, only two novel views are shown.

Warp Video Condition and Embedding Design. We first estimate the depth of the source video using a monocular depth estimation model. Together with the camera intrinsic matrix $\boldsymbol{I}_0$, the estimated depth is back-projected into 3D space to construct a dynamic point cloud $\boldsymbol{P} \in \mathbb{R}^{T\times N\times 3}$, where N denotes the number of points at each timestamp. We then project $\boldsymbol{P}$ onto each target camera to obtain the corresponding warped videos $\boldsymbol{W}_1, \ldots, \boldsymbol{W}_n$, which provide geometrically aligned but incomplete conditions for novel-view generation. The source and warped videos are encoded by the pretrained Wan VAE:

$$\boldsymbol{v} = \mathcal{E}(\boldsymbol{V}_0), \qquad \boldsymbol{w}_i = \mathcal{E}(\boldsymbol{W}_i), \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{v}, \boldsymbol{w}_i \in \mathbb{R}^{t\times h\times w\times d}$, with t, h, and w denoting the compressed temporal and spatial dimensions, and d the latent channel dimension.

As illustrated in Figure 2, we construct view-specific conditioning for the noisy video latents. For each novel view, the corresponding warp latent $\boldsymbol{w}_i$ is concatenated with its noisy latent along the channel dimension. For the source view, we instead concatenate the clean source latent $\boldsymbol{v}$ with its noisy counterpart, using the clean latent as an explicit appearance condition. To explicitly distinguish the source view from the generated novel views, four additional binary indicator channels are appended to the latent representation. The resulting source- and novel-view embeddings are jointly fed into the video-generation DiT backbone.

Modification of Wan DiT. As shown in Figure 2, we augment the original Wan DiT with two additional modules, a *Camera Adapter* and *View Attention*. The Camera Adapter first encodes the Plücker embeddings of each camera and injects the resulting camera features into the video generation process through cross-attention. This provides explicit geometric conditioning for synthesizing the corresponding target viewpoints. To further encourage spatial consistency across the generated views, we introduce View Attention, which allows tokens from different viewpoints at the same timestamp to exchange information. Such cross-view interaction enables the model to jointly reason

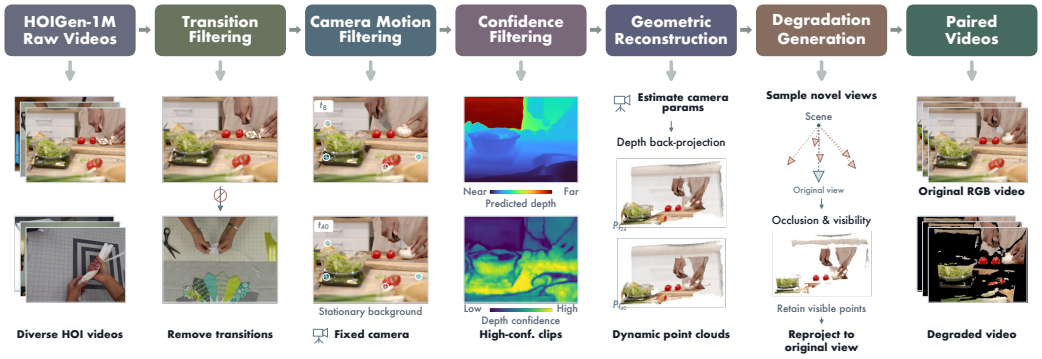

Figure 3: **HOIGen-1M data processing pipeline.** We first filter monocular HOI videos based on temporal continuity, camera motion, and geometric confidence to obtain clips suitable for reconstruction. The retained videos are reconstructed into dynamic point clouds, from which we sample novel viewpoints and perform visibility filtering. Visible points are then reprojected to the source view to synthesize degraded videos, which are paired with the original RGB sequences for degraded-to-clean supervision. Further details are provided in Sec. A.3.

about complementary observations and maintain coherent spatial structures across views. Except for the newly introduced Camera Adapter and View Attention modules, all remaining components are initialized from the pretrained Wan model.

3.3 Cross-view Hand Latent Representation

To extract hand-centric representations from the multi-view video features, we introduce a parallel hand branch shared across viewpoints. These hand branches progressively construct cross-view hand latent representations for subsequent 3D hand motion decoding. Inspired by FantasyWorld Dai et al. (2026), we initialize the hand branch from the video tokens at the 15th layer of the DiT backbone. Specifically, the intermediate video features $\boldsymbol{f} \in \mathbb{R}^{t\times h\times w\times d}$ are copied and spatially compressed through convolutional layers into $\boldsymbol{f'} \in \mathbb{R}^{t\times m\times d}$, which initializes the corresponding hand branch, with m denoting the number of hand tokens within each temporal block.

Within each hand-branch layer, the hand tokens first interact within each view via self-attention. We then apply a View Attention block, analogous to that introduced in Sec. 3.2, to facilitate information exchange across viewpoints. Finally, the hand tokens cross-attend to the corresponding video tokens at the end of each layer, continuously extracting and refining hand-relevant cues from the video representations. Through these intra-view, cross-view, and video-to-hand interactions, the hand branch progressively builds a compact cross-view latent representation for subsequent 3D hand motion decoding.

3.4 Structured Hand Motion Decoding

Our structured hand motion decoder employs two decoders, $\mathcal{D}_{\text{agg}}$ and $\mathcal{D}_{\text{pv}}$, to recover MANO parameters from the hand latent representations (Figure 2). $\mathcal{D}_{\text{agg}}$ aggregates hand tokens across all viewpoints to predict the hand pose $\boldsymbol{\theta}$, shape $\boldsymbol{\beta}$, and visibility confidence c, while $\mathcal{D}_{\text{pv}}$ operates on the hand tokens of each viewpoint to estimate the global orientation $\boldsymbol{o}$ and translation $\boldsymbol{t}$.

Since each temporal latent block in Wan corresponds to four video frames, we assign two learnable query tokens to each frame, one for each hand side, resulting in eight queries per temporal block for frame-level motion decoding. Let $\boldsymbol{h} \in \mathbb{R}^{n\times t\times m\times d}$ denotes the hand tokens produced by the final hand-branch layer. Within each decoder layer, the frame-level hand queries first cross-attend to the corresponding hand-token context and then interact through self-attention. The resulting queries are flattened along the temporal dimension, augmented with temporal positional embeddings, and further refined through temporal self-attention to capture motion dependencies across frames. Finally, separate MLP heads map the refined query features to the corresponding hand parameters, while an additional head predicts sparse 2D joint anchors to provide explicit image-space localization cues without relying on detection-based hand crops.

Dataset	Clips	Views	Resolution	Hand Anno.
Hand-object interaction datasets				
DexYCB	1,000	8	640×480	✓
TACO	2,317	8	4096×3000	✓
OakInk v2	2,754	3	848×480	✓
ARCTIC	915	8	2800×2000	✓
GigaHands	1,000	8	1280×720	✓
Video datasets w/o hand annotation				
SynCamVideo	3,401	10	1280×1280	✗
HOIGen-1M	15,736	Mono.	720p+	✗

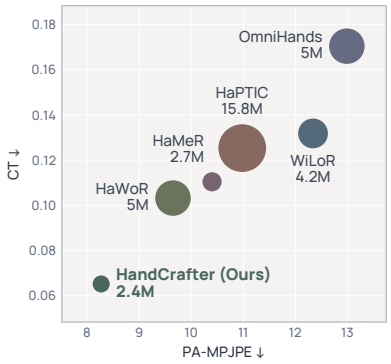

Table 1: **Training data and data-efficiency comparison. Left:** Statistics of our training datasets, including multi-view HOI datasets and video datasets without hand annotation. **Right:** Comparison of training data scale and performance on HO-Cap. HandCrafter achieves lower reconstruction error with substantially less annotated hand data. The size of the circle represents the number of frames used for training.[†]

3.5 Two-Stage Hybrid Training Strategy

Training HandCrafter poses two distinct challenges. First, learning reliable multi-view video generation requires spatial supervision across viewpoints, while existing multi-view HOI datasets are relatively limited in visual diversity. Second, hand motion estimation requires accurate geometric annotations while preserving the generative prior learned by the video backbone. We therefore adopt a two-stage training strategy: the first stage trains the multi-view video generator using hybrid multi-view and monocular supervision, while the second stage freezes the video branch and optimizes the hand representation and decoding modules.

Stage I: Hybrid Multi-View Video Generation Training. We collect and preprocess five multi-view HOI datasets, including ARCTIC Fan et al. (2023), DexYCB Chao et al. (2021), TACO Liu et al. (2024b), OakInk-v2 Zhan et al. (2024), GigaHands Fu et al. (2025) and one synthetic multi-view video dataset, SynCamVideo Bai et al. (2025b). All sequences are converted into clips of 49 frames at a resolution of 640×480. These datasets provide direct cross-view supervision for learning geometrically consistent HOI video generation.

However, multi-view HOI datasets are typically captured in controlled environments, resulting in relatively limited diversity in backgrounds, camera configurations, and interaction patterns. Training solely on such data therefore tends to limit generalization to unconstrained scenarios. In contrast, large-scale monocular HOI videos provide substantially richer visual and motion diversity, but lack paired observations from different viewpoints. To complement the limited diversity of multi-view data, we apply degradation-based supervision to these monocular videos, allowing the model to benefit from a broader training distribution beyond controlled multi-view captures (Sec. A.5), following a similar philosophy to NeoVerse Yang et al. (2026).

Monocular Video Degradation. As described in Sec. 3.2, the multi-view generation model is conditioned on the geometrically warped video that inevitably contains missing regions and projection artifacts. Robust recovery from such imperfect conditions is therefore essential for high-quality novel-view synthesis. We adopt HOIGen-1M Liu et al. (2025) as the source of scalable monocular HOI videos. As illustrated in Figure 3, we apply a dedicated data cleaning and preprocessing pipeline to construct degraded-to-clean video pairs, with details provided in Sec. A.3. The final mixture of multi-view and monocular training data is summarized in Table 1. During this stage, all parameters of the multi-view video generation model are jointly optimized.

Stage II: Hand Motion Estimation Training. After adapting the generative prior, we freeze the video-generation branch and optimize only the hand representation and decoding modules. Training is performed on the same annotated multi-view HOI datasets using their available hand annotations. Following common practice in 3D hand reconstruction , we supervise the predicted MANO param-

[†]We include WiLoR's hand-reconstruction training data in HaWoR's count, as HaWoR uses WiLoR's frozen pretrained backbone.

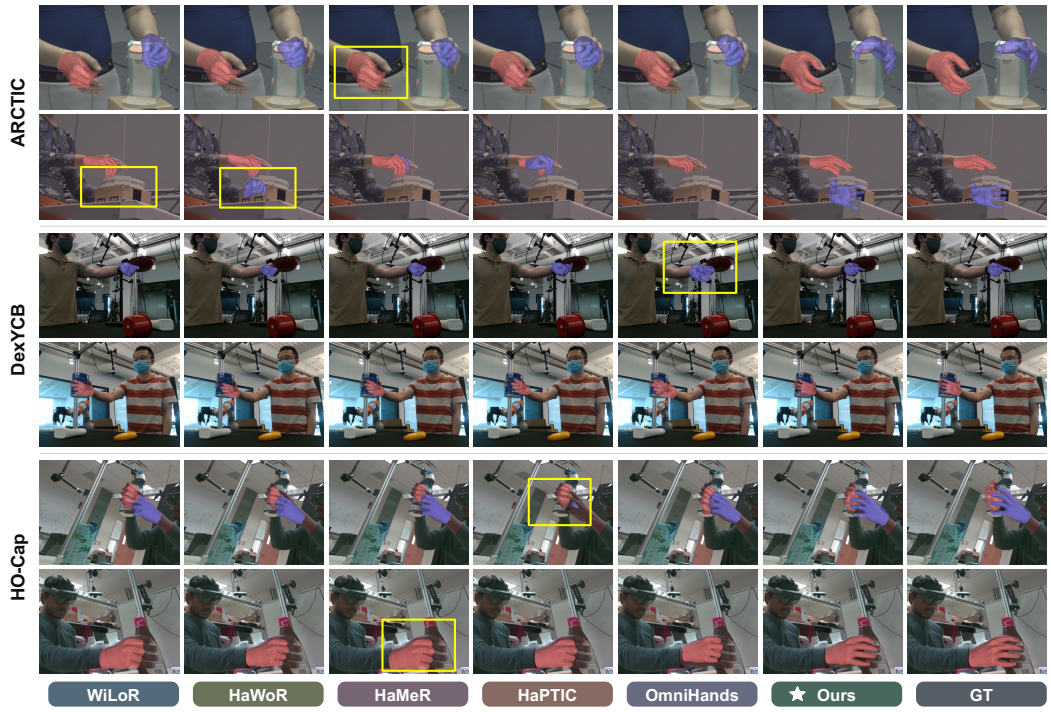

Figure 4: **Qualitative comparison.** HandCrafter produces more accurate 3D hand reconstruction results under challenging occlusions. Yellow boxes highlight failure cases of other baselines.

eters, hand geometry, image-space localization, and visibility using

$$\begin{aligned}\mathcal{L}_{\text{hand}} = \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{orient}}\mathcal{L}_{\text{orient}} + \lambda_{\text{shape}}\mathcal{L}_{\text{shape}} + \lambda_{\text{joint}}\mathcal{L}_{\text{joint}} \\ + \lambda_{\text{proj}}\mathcal{L}_{\text{proj}} + \lambda_{\text{uv}}\mathcal{L}_{\text{uv}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{conf}}\mathcal{L}_{\text{conf}}.\end{aligned} \tag{2}$$

Detailed definitions of these loss terms and their weighting coefficients are provided in Sec. A.1, with implementation details in Sec. A.2.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate HandCrafter on three hand-object interaction benchmarks: ARCTIC Fan et al. (2023), DexYCB Chao et al. (2021), and HO-Cap Wang et al. (2025). For ARCTIC and DexYCB, we follow their standard train/test splits. HO-Cap remains unseen during training for all evaluated methods, including HandCrafter, providing a common benchmark for assessing their generalization to unseen capture settings and interaction distributions. The training data used in our two-stage framework are detailed in Sec. 3.5.

Evaluation Protocols and Metrics. We follow the standard evaluation protocols of each benchmark. We report FAcc and F1 for hand detection performance. For 3D hand reconstruction, we use MPJPE-p, PA-MPJPE-p, MPVPE-p and PA-MPVPE-p to measure joint-position accuracy, EPE-p for image-space localization, and GO-p and CT-p for global orientation and camera-space translation errors, respectively.

4.2 Comparison with State-of-the-Art Methods

Baselines. We compare HandCrafter with representative state-of-the-art image- and video-based hand reconstruction methods, including HaMeR Pavlakos et al. (2024), WiLoR Potamias et al. (2025), HaWoR Zhang et al. (2025b), HaPTIC Ye et al. (2026), and OmniHands Lin et al. (2026). We use the official implementations and pretrained checkpoints whenever available, and evaluate all

Table 2: **Quantitative comparison.** We report metrics for hand detection, pose reconstruction, and global localization. HO-Cap is excluded from training and used solely for cross-dataset generalization evaluation. Bold and underlining mark the **Best** and **Second-best** final results, respectively.

	Method	Detection		Pose Reconstruction				Localization		
		FAcc ↑	F1 ↑	MPJPE-p ↓ (mm)	PA-MPJPE-p ↓ (mm)	MPVPE-p ↓ (mm)	PA-MPVPE-p ↓ (mm)	EPE-p ↓ (px)	GO-p ↓ (°)	CT-p ↓ (m)
ARCTIC	WiLoR	0.7223	0.9204	32.66	23.42	31.52	22.72	122.45	28.35	0.7515
	HaMeR	0.7966	0.9413	29.30	19.00	28.24	18.43	91.80	25.63	0.6569
	HaWoR	0.8352	0.9168	35.00	19.18	34.86	18.78	86.11	30.08	0.6137
	HaPTIC	0.7895	0.9398	20.74	17.03	22.21	16.88	96.12	18.94	0.5184
	OmniHands	0.7581	0.9226	32.85	22.85	31.92	22.17	113.90	28.81	0.7430
	HandCrafter	**0.9949**	**0.9960**	**14.02**	**6.57**	**13.60**	**6.38**	**18.16**	**11.20**	**0.1593**
DexYCB	WiLoR	0.9417	**0.9557**	**14.02**	10.95	**14.02**	10.61	32.54	11.79	0.0952
	HaMeR	0.5556	0.7416	17.47	14.86	17.29	14.37	51.20	15.88	0.1188
	HaWoR	0.8495	0.8382	15.01	8.38	14.75	8.11	28.52	**11.28**	0.0670
	HaPTIC	0.6069	0.7628	23.89	21.28	24.37	21.35	84.80	22.51	0.2396
	OmniHands	0.5727	0.7572	34.99	23.75	33.02	22.82	84.87	30.12	0.2138
	HandCrafter	**0.9754**	0.9151	14.61	**6.74**	14.06	**6.42**	**13.56**	11.41	**0.0430**
HO-Cap	WiLoR	0.9237	0.9695	21.72	12.34	20.79	11.91	38.22	19.31	0.1318
	HaMeR	0.8395	0.9408	21.93	10.40	20.97	10.02	26.21	17.22	0.1104
	HaWoR	0.9063	0.9497	22.98	9.65	21.47	8.90	**21.98**	16.57	0.1032
	HaPTIC	0.8480	0.9439	**20.20**	10.98	**19.26**	10.32	31.16	16.57	0.1253
	OmniHands	0.8243	0.9311	26.10	12.99	24.48	12.69	35.46	19.67	0.1704
	HandCrafter	**0.9670**	**0.9801**	21.09	**8.27**	20.25	**7.97**	24.44	**15.79**	**0.0652**

methods under the same protocols described above. Among the baselines, WiLoR, HaPTIC, and HaWoR use ARCTIC in their training data, while all baselines use DexYCB.

Quantitative Results. As shown in Table 2, HandCrafter consistently outperforms existing methods on ARCTIC across both hand detection and 3D hand estimation metrics. For *hand detection*, HandCrafter improves FAcc and F1 over existing methods, demonstrating robust detection under frequent hand-object occlusions. For *3D hand estimation*, HandCrafter reduces PA-MPJPE-p from the previous best of 17.03 mm to 6.57 mm, while also substantially improving MPJPE-p. It further achieves consistent improvements in EPE-p, GO-p, and CT-p, indicating more accurate image-space localization, global orientation, and camera-space translation.

We attribute these improvements to the complementary spatial cues provided by the generated viewpoints. By jointly reasoning over cross-view diffusion representations, HandCrafter can better resolve hand configurations that are ambiguous or occluded in the original monocular view.

Robustness to Occlusion. To assess robustness across occlusion levels, we select ARCTIC clips with light, moderate, and severe occlusion using rule-based methods (Sec. A.6) from its hand and object annotations. We compare against WiLoR and HaPTIC, both of which also use ARCTIC for training. Table 3 shows that their performance degrades markedly as occlusion increases, while HandCrafter remains robust across all three levels.

4.3 Generalization to In-the-Wild Scenarios

To evaluate generalization beyond the training datasets, we directly test HandCrafter on HO-Cap dataset, which is excluded from training. As shown in Table 2, HandCrafter maintains strong quantitative performance under this cross-dataset setting, particularly improving PA-MPJPE-p from the previous best of 9.65 mm to 8.27 mm. The qualitative results in Figure 4 further demonstrate robust hand reconstruction under changes in capture setup, object appearance, and interaction distribution.

In-the-Wild Generalization. We further evaluate HandCrafter on diverse out-of-distribution Internet HOI videos. As shown in Figure 5, our method recovers plausible and temporally coherent hand motion across varied objects, backgrounds, viewpoints, and interaction patterns. These results demonstrate that the learned multi-view diffusion prior can provide useful spatial cues beyond the controlled datasets used during training.

4.4 Ablation Studies

Benefits of Multi-View Modeling. To assess the value of multi-view modeling, we vary the number of synthesized novel views from zero (monocular input) to four on ARCTIC. As shown in Table 4,

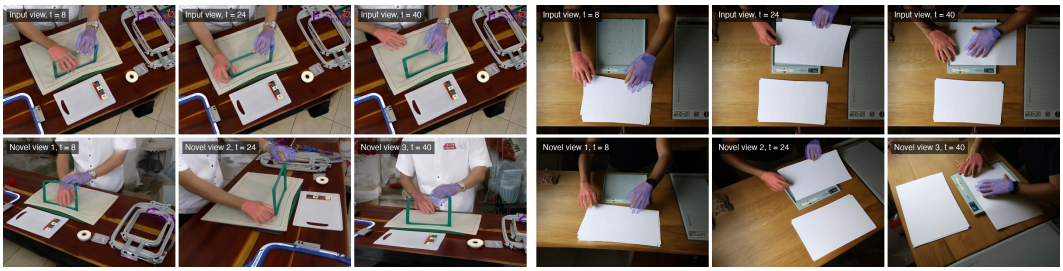

Figure 5: **Qualitative results on out-of-distribution Internet HOI videos.** We visualize selected source frames and decoded complementary views with reconstructed MANO meshes. HandCrafter generalizes to diverse in-the-wild objects, backgrounds, viewpoints, and interaction patterns while maintaining coherent multi-view generation.

metrics generally improve with more views, indicating that complementary viewpoints help resolve monocular ambiguities. The gains diminish as the view count increases, with performance largely saturating at three to four novel views.

Effect of View Attention. We next examine the contribution of View Attention to hand motion estimation by selectively removing the cross-view interaction modules from the video and hand branches. As shown in Table 5, using View Attention in both branches achieves the best performance across all metrics, demonstrating the benefit of cross-view information exchange.

Number of MANO Query Tokens. We study the influence of the number of learnable query tokens used for frame-level MANO decoding. Specifically, we vary the number of query tokens among 1, 4, and 8 while keeping all other settings fixed. We retrain all three configurations using the same 20% subset of the mixed training dataset. As reported in Table 6, 8 queries achieves the best performance.

Table 3: **Performance under different occlusion levels on ARCTIC.**

Method	Light		Moderate		Severe	
	F1 ↑	MPJPE-p ↓	F1 ↑	MPJPE-p ↓	F1 ↑	MPJPE-p ↓
WiLoR	0.9800	20.85	0.8750	40.85	0.7157	68.89
HaPTIC	0.9914	9.72	0.9235	24.18	0.7698	55.49
HandCrafter	**0.9998**	**9.57**	**0.9999**	**11.65**	**0.9976**	**11.21**

Table 4: **Ablation on the number of novel views on ARCTIC.**

# Novel Views	MPJPE-p ↓	PA-MPJPE-p ↓	EPE-p ↓	GO-p ↓	CT-p ↓
Mono	14.70	6.59	15.36	11.14	0.2383
1	12.49	5.82	13.88	9.62	0.1762
2	11.80	5.62	13.43	9.08	0.1650
3	11.45	5.53	**13.13**	**8.78**	0.1554
4	**11.45**	**5.47**	13.32	8.90	**0.1505**

Table 5: **Ablation study on View Attention on ARCTIC.**

Method	MPJPE-p ↓	PA-MPJPE-p ↓	EPE-p ↓	GO-p ↓	CT-p ↓
w/o ViewAttn	27.76	10.45	34.88	25.00	0.2582
Hand ViewAttn	22.25	8.67	30.88	18.54	0.2832
Video ViewAttn	20.70	7.32	22.30	18.79	0.1853
Video + Hand ViewAttn	**14.02**	**6.57**	**18.16**	**11.20**	**0.1593**

Table 6: **Ablation on the number of MANO query tokens on ARCTIC.**

# Query Tokens	MPJPE-p ↓	PA-MPJPE-p ↓	EPE-p ↓	GO-p ↓	CT-p ↓
1	45.02	18.44	71.31	38.92	0.7124
4	38.98	15.93	58.21	31.96	0.4972
8	**35.40**	**15.80**	**50.58**	**27.11**	**0.3943**

5 CONCLUSION

We presented HandCrafter, a generative multi-view framework for monocular 3D hand motion reconstruction under severe hand-object occlusions. By synthesizing complementary viewpoints and directly exploiting cross-view diffusion representations, HandCrafter provides spatial cues for resolving monocular ambiguities. We further introduce a structured decoder for crop-free MANO estimation and a two-stage hybrid training strategy combining multi-view supervision with scalable monocular HOI videos. Extensive experiments demonstrate accurate reconstruction under challenging occlusions and strong generalization across diverse datasets and in-the-wild HOI scenarios.

Limitations and future work. Our current framework is designed for videos with limited camera motion and is therefore less effective for videos with substantial viewpoint changes, such as egocentric sequences, where reliable multi-view synthesis becomes challenging. Handling large camera motion and extending the framework to egocentric settings are promising directions for future work.

AI USE STATEMENT

In this work, we used generative AI tools to improve the grammar, clarity, and readability of the manuscript. All AI-assisted revisions were reviewed by the authors to ensure accuracy and preserve the intended scientific meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Jianhong Bai, Menghan Xia, Xiao Fu, Xintao Wang, Lianrui Mu, Jinwen Cao, Zuozhu Liu, Haoji Hu, Xiang Bai, Pengfei Wan, et al. Recammaster: Camera-controlled generative rendering from a single video. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14834–14844. IEEE, 2025a.

Jianhong Bai, Menghan Xia, Xintao Wang, Ziyang Yuan, Zuozhu Liu, Haoji Hu, Pengfei Wan, and Di Zhang. Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints. In *International Conference on Learning Representations*, volume 2025, pp. 58038–58060, 2025b.

Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10843–10852, 2019.

Jiahao Chang, Chongjie Ye, Yushuang Wu, Yuantao Chen, Yidan Zhang, Zhongjin Luo, Chenghong Li, Yihao Zhi, and Xiaoguang Han. Reconviagen: Towards accurate multi-view 3d object reconstruction via generation. In *International Conference on Learning Representations*, volume 2026, pp. 90387–90401, 2026.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9040–9049. IEEE, 2021.

Yuantao Chen, Jiahao Chang, Chongjie Ye, Chaoran Zhang, Zhaojie Fang, Chenghong Li, and Xiaoguang Han. Forehoi: Feed-forward 3d object reconstruction from daily hand-object interaction videos. *arXiv preprint arXiv:2602.06226*, 2026.

Yixiang Dai, Fan Jiang, Chiyu Wang, Mu Xu, and Yonggang Qi. Fantasyworld: Geometry-consistent world modeling via unified video and 3d prediction. In *International Conference on Learning Representations*, volume 2026, pp. 103603–103622, 2026.

Lingwei Dang, Juntong Li, Zonghan Li, Hongwen Zhang, Liang An, Wei Min, Yebin Liu, and Qingyao Wu. Harmohoi: Harmonizing appearance and 3d motion for multi-view hand-object interaction synthesis. *arXiv preprint arXiv:2607.17097*, 2026a.

Lingwei Dang, Ruizhi Shao, Hongwen Zhang, Wei Min, Yebin Liu, and Qingyao Wu. Svimo: Synchronized diffusion for video and motion generation in hand-object interaction scenarios. *Advances in Neural Information Processing Systems*, 38:76689–76718, 2026b.

Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, and Michael J Black. Hmp: Hand motion priors for pose and shape estimation from video. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6341–6351. IEEE, 2024.

Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12943–12954. IEEE, 2023.

Rao Fu, Dingxi Zhang, Alex Jiang, Wanjia Fu, Austin Funk, Daniel Ritchie, and Srinath Sridhar. Gigahands: A massive annotated dataset of bimanual hand activities. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17461–17474. IEEE, 2025.

Maksym Ivashechkin, Oscar Mendez, and Richard Bowden. Denoising diffusion for 3d hand pose estimation from images. In *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3128–3137. IEEE, 2023.

Yudong Jin, Sida Peng, Xuan Wang, Tao Xie, Zhen Xu, Yifan Yang, Yujun Shen, Hujun Bao, and Xiaowei Zhou. Diffuman4d: 4d consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11047–11057. IEEE, 2025.

Lijun Li, Li'an Zhuo, Bang Zhang, Liefeng Bo, and Chen Chen. Diffhand: End-to-end hand mesh reconstruction via diffusion models. *arXiv preprint arXiv:2305.13705*, 2023.

Mengcheng Li, Hongwen Zhang, Yuxiang Zhang, Ruizhi Shao, Tao Yu, and Yebin Liu. Hhmr: Holistic hand mesh recovery by enhancing the multimodal controllability of graph diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 645–654. IEEE, 2024.

Dixuan Lin, Yuxiang Zhang, Mengcheng Li, Wei Jing, Qi Yan, Qianying Wang, Yebin Liu, and Hongwen Zhang. Omnihands: Robust motion capture of interactive hands via a versatile transformer. *ACM Transactions on Graphics*, 45(5):1–17, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Kun Liu, Qi Liu, Xinchen Liu, Jie Li, Yongdong Zhang, Jiebo Luo, Xiaodong He, and Wu Liu. Hoigen-1m: A large-scale dataset for human-object interaction video generation. In *CVPR*, 2025.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *International conference on learning representations*, volume 2024, pp. 27676–27697, 2024a.

Yun Liu, Haolin Yang, Xu Si, Ling Liu, Zipeng Li, Yuxiang Zhang, Yebin Liu, and Li Yi. Taco: Benchmarking generalizable bimanual tool-action-object understanding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21740–21751. IEEE, 2024b.

Gyeongsik Moon. Bringing inputs to shared domains for 3d interacting hands recovery in the wild. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17028–17037. IEEE, 2023.

Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3d with transformers. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9826–9836. IEEE, 2024.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182. IEEE, 2023.

Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. Wilor: End-to-end 3d hand localization and reconstruction in-the-wild. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12242–12254. IEEE, 2025.

Xuanchi Ren, Tianchang Shen, Jiahui Huang, Huan Ling, Yifan Lu, Merlin Nimier-David, Thomas Müller, Alexander Keller, Sanja Fidler, and Jun Gao. Gen3c: 3d-informed world-consistent video generation with precise camera control. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6121–6132. IEEE, 2025.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), November 2017.

Zhihao Sun, Tong Wu, Ruirui Tu, Daoguo Dong, and Zuxuan Wu. Unihand: A unified model for diverse controlled 4d hand motion modeling. *arXiv preprint arXiv:2602.21631*, 2026.

Naru Suzuki, Takehiko Ohkawa, Tatsuro Banno, Jihyun Lee, Ryosuke Furuta, and Yoichi Sato. Affordance-guided diffusion prior for 3d hand reconstruction. In *European Conference on Computer Vision (ECCV)*, 2026.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jikai Wang, Qifan Zhang, Yu-Wei Chao, Bowen Wen, Xiaohu Guo, and Yu Xiang. HO-cap: A capture system and dataset for 3d reconstruction and pose tracking of hand-object interaction. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

Yuxi Wang, Chengkai Jin, Yufei Liu, Wenqi Ouyang, Tianyi Wei, Zhiwei Zeng, Siyuan Huang, Zhiqi Shen, and Xingang Pan. The surprising effectiveness of video diffusion models for hand motion reconstruction. *arXiv preprint arXiv:2606.30308*, 2026.

Rundi Wu, Ruiqi Gao, Ben Poole, Alex Trevithick, Changxi Zheng, Jonathan T Barron, and Aleksander Holynski. Cat4d: Create anything in 4d with multi-view video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26057–26068, 2025a.

Sibo Wu, Congrong Xu, Binbin Huang, Andreas Geiger, and Anpei Chen. Genfusion: Closing the loop between reconstruction and generation via videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6078–6088. IEEE, 2025b.

Yiming Xie, Chun-Han Yao, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d: Dynamic 3d content generation with multi-frame and multi-view consistency. In *International Conference on Learning Representations*, volume 2025, pp. 33421–33441, 2025.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Yufei Ye, Yao Feng, Omid Taheri, Haiwen Feng, Shubham Tulsiani, and Michael J Black. Predicting 4d hand trajectory from monocular videos. In *2026 International Conference on 3D Vision (3DV)*, pp. 860–870. IEEE, 2026.

Mark Yu, Wenbo Hu, Jinbo Xing, and Ying Shan. Trajectorycrafter: Redirecting camera trajectory for monocular videos via diffusion models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 100–111. IEEE, 2025a.

Zhengdi Yu, Stefanos Zafeiriou, and Tolga Birdal. Dyn-hamr: Recovering 4d interacting hand motion from a dynamic camera. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27716–27726. IEEE, 2025b.

Xinyu Zhan, Lixin Yang, Yifei Zhao, Kangrui Mao, Hanlin Xu, Zenan Lin, Kailin Li, and Cewu Lu. Oakink2: A dataset of bimanual hands-object manipulation in complex task completion. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 445–456. IEEE, 2024.

Jiajun Zhang, Yuxiang Zhang, Liang An, Mengcheng Li, Hongwen Zhang, Zonghai Hu, and Yebin Liu. Manidext: Hand-object manipulation synthesis via continuous correspondence embeddings and residual-guided diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025a.

Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1805–1815. IEEE, 2025b.

Yufei Zhang, Zijun Cui, Jeffrey O Kephart, and Qiang Ji. Diffusion-based 3d hand motion recovery with intuitive physics. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7306–7317. IEEE, 2025c.

Zimu Zhang, Yucheng Zhang, Xiyan Xu, Ziyin Wang, Sirui Xu, Kai Zhou, Bing Zhou, Chuan Guo, Jian Wang, Yu-Xiong Wang, and Liang-Yan Gui. Handx: Scaling bimanual motion and interaction generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2274–2284, June 2026.

Yihao Zhi, Chenghong Li, Hongjie Liao, Xihe Yang, Zhengwentai Sun, Jiahao Chang, Xiaodong Cun, Wensen Feng, and Xiaoguang Han. Mv-performer: Taming video diffusion model for faithful and synchronized multi-view performer synthesis. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–14, 2025.

Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *2023 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 12588–12597. IEEE, 2023.

A Appendix

A.1 Training Objectives

We provide detailed definitions of the individual loss terms in the overall hand reconstruction objective introduced in Eq. 2. For simplicity, we describe the regression losses for a single hand in one video and omit the sample and hand-side indices. Let T and N denote the number of frames and viewpoints, respectively. We use $\widehat{\cdot}$ and $(\cdot)^\star$ to denote predicted and ground-truth quantities. As described in Sec. 3.4, all viewpoints share the MANO shape parameters $\boldsymbol{\beta}$ and the frame-wise local hand pose $\boldsymbol{\theta}$, whereas the global orientation $\boldsymbol{o}$ and translation $\boldsymbol{t}$ are estimated for each viewpoint and frame.

A.1.1 MANO Parameter and Geometry Supervision

Local pose loss. We supervise the local rotations of the 15 non-root MANO joints. Let $R_{tk} \in \mathrm{SO}(3)$ denote the local rotation matrix of joint k at frame t. After converting the predicted 6D rotation representation to a rotation matrix, we compute the mean squared error over matrix elements:

$$\mathcal{L}_{\text{pose}} = \frac{1}{9T \cdot 15} \sum_{t=1}^{T} \sum_{k=1}^{15} \left\| \widehat{R}_{tk} - R^{\star}_{tk} \right\|_F^2, \tag{3}$$

where $\|\cdot\|_F$ denotes the Frobenius norm and the factor 9 accounts for the nine elements of a rotation matrix.

Global orientation loss. Let $G_{vt} \in \mathrm{SO}(3)$ denote the global hand orientation at frame t in viewpoint v. After transforming the ground-truth orientation into the corresponding camera coordinate system, we apply the same rotation-matrix loss:

$$\mathcal{L}_{\text{orient}} = \frac{1}{9NT} \sum_{v=1}^{N} \sum_{t=1}^{T} \left\| \widehat{G}_{vt} - G^{\star}_{vt} \right\|_F^2 \tag{4}$$

Shape loss. We supervise the MANO shape parameters $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{10}$ using a Euclidean loss:

$$\mathcal{L}_{\text{shape}} = \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\star} \right\|_2 \tag{5}$$

3D joint loss. We further supervise the 3D hand geometry. Let $\widehat{\mathbf{J}}^0_{vtj}$ and $\mathbf{J}^{0,\star}_{vtj}$ denote the MANO joint locations generated from the predicted and ground-truth parameters, respectively, where the global orientation in the corresponding camera coordinate system is retained while the global translation is set to zero. We compute the average Euclidean distance over all $J = 16$ native MANO joints, including the wrist:

$$\mathcal{L}_{\text{joint}} = \frac{1}{NTJ} \sum_{v=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| \widehat{\mathbf{J}}^0_{vtj} - \mathbf{J}^{0,\star}_{vtj} \right\|_2 \tag{6}$$

This loss is independent of the predicted global translation.

A.1.2 IMAGE-SPACE LOCALIZATION SUPERVISION

To improve hand localization and exploit visual evidence from different hand regions, we select the wrist and five metacarpophalangeal joints as six sparse localization anchors. For each viewpoint and frame, the model predicts the image coordinates and depth of these anchors. Their recovered 3D locations are subsequently used to estimate the global hand translation.

Let $\mathcal{S}$ denote the anchor set, with $|\mathcal{S}| = 6$. We first project camera-space 3D joints onto the image plane using the camera intrinsic matrix K_v. The horizontal and vertical coordinates are normalized by the image width $W = 640$ and height $H = 480$, respectively, followed by subtracting 0.5. The resulting normalized image domain is therefore $[-0.5, 0.5]^2$, with the image center at $(0, 0)$. We denote this projection and normalization operation by $\overline{\pi}_v$.

2D joint anchor loss. Let $\widehat{\mathbf{u}}_{vta} \in \mathbb{R}^2$ denote the predicted normalized image coordinate of anchor a. Its ground-truth target is obtained by projecting the corresponding camera-space 3D joint:

$$\mathbf{u}^{\star}_{vta} = \overline{\pi}_v \left(\mathbf{J}^{\star}_{vta} \right) \tag{7}$$

We supervise the anchor locations using

$$\mathcal{L}_{\mathrm{uv}} = \frac{1}{6NT} \sum_{v=1}^{N} \sum_{t=1}^{T} \sum_{a \in \mathcal{S}} \| \widehat{\mathbf{u}}_{vta} - \mathbf{u}^{\star}_{vta} \|_2 \tag{8}$$

Depth loss. The localization branch additionally predicts a focal-normalized log-depth for each anchor. Let $z^{\star}_{vta}$ denote its ground-truth depth in camera coordinates, and let $f_{x,v}$ and $f_{y,v}$ denote the horizontal and vertical focal lengths. The depth target is defined as

$$d^{\star}_{vta} = \log \left(\frac{z^{\star}_{vta}}{f_v} \right) \qquad f_v = \sqrt{\frac{f_{x,v} f_{y,v}}{WH}} \tag{9}$$

We supervise the predicted depth $\widehat{d}_{vta}$ using the Smooth L1 loss:

$$\mathcal{L}_{\mathrm{depth}} = \frac{1}{6NT} \sum_{v=1}^{N} \sum_{t=1}^{T} \sum_{a \in \mathcal{S}} \rho \left(\widehat{d}_{vta} - d^{\star}_{vta} \right) \tag{10}$$

where

$$\rho(e) = \begin{cases} \frac{1}{2} e^2, & |e| < 1, \\ |e| - \frac{1}{2}, & |e| \geq 1. \end{cases} \tag{11}$$

Anchor-based localization. To recover the global translation, we first back-project the predicted image coordinates and depths of the six anchors into camera space, yielding $\widehat{\mathbf{Q}}_{vta}$. We then average the offsets between the recovered anchor positions and the corresponding zero-translation MANO joints:

$$\widehat{\mathbf{t}}_{vt} = \frac{1}{6} \sum_{a \in \mathcal{S}} \left(\widehat{\mathbf{Q}}_{vta} - \widehat{\mathbf{J}}^{0}_{vta} \right) \tag{12}$$

During back-projection, we recover metric depth using $f_v \exp(\mathrm{clip}(\widehat{d}_{vta}, -3, 2))$ for numerical stability. The depth loss in Eq. 10, however, is applied to the unclipped prediction.

2D projection loss. We additionally constrain the projection of the reconstructed 3D hand. Let

$$\widehat{\mathbf{J}}_{vtj} = \widehat{\mathbf{J}}^{0}_{vtj} + \widehat{\mathbf{t}}_{vt} \tag{13}$$

denote the predicted camera-space joint after applying the recovered translation, and let $\mathbf{J}^{\star}_{vtj}$ denote its ground-truth counterpart. We minimize the average image-space projection error:

$$\mathcal{L}_{\mathrm{proj}} = \frac{1}{NTJ} \sum_{v=1}^{N} \sum_{t=1}^{T} \sum_{j=1}^{J} \left\| \overline{\pi}_v \left(\widehat{\mathbf{J}}_{vtj} \right) - \overline{\pi}_v \left(\mathbf{J}^{\star}_{vtj} \right) \right\|_2 . \tag{14}$$

A.1.3 HAND DETECTION SUPERVISION

Hand detection loss. The confidence branch predicts whether the left or right hand is present in the input video. Let $p_s = \sigma(c_s)$ denote the predicted probability for hand side $s \in \{\text{left}, \text{right}\}$, and let $h_s \in \{0, 1\}$ indicate whether the corresponding hand has a valid MANO annotation. We use the binary cross-entropy loss averaged over the two hand sides:

$$\mathcal{L}_{\text{conf}} = -\frac{1}{2} \sum_{s \in \{\text{left},\text{right}\}} \left[h_s \log p_s + (1 - h_s) \log(1 - p_s) \right]. \tag{15}$$

This term provides video-level supervision for hand detection.

A.1.4 LOSS WEIGHTS

We use the following loss weights in all experiments:

$$\begin{aligned} \lambda_{\text{pose}} &= 1, & \lambda_{\text{orient}} &= 0.1, \\ \lambda_{\text{shape}} &= 0.05, & \lambda_{\text{joint}} &= 5, \\ \lambda_{\text{proj}} &= 1, & \lambda_{\text{uv}} &= 1, \\ \lambda_{\text{depth}} &= 0.1, & \lambda_{\text{conf}} &= 1. \end{aligned} \tag{16}$$

A.2 IMPLEMENTATION DETAILS

We train the multi-view video generation model for three days in Stage I. In Stage II, we adopt a progressive training schedule to improve computational efficiency under the available GPU memory budget. The first phase uses only the input view, with all View Attention modules disabled, to learn hand pose estimation from monocular features. The second phase enables all View Attention modules to incorporate complementary information across multiple views. The two phases last two and three days, respectively. All training is conducted on eight NVIDIA A100 80 GB GPUs with a learning rate of 1×10^{-4}.

During training, we sample at most four novel viewpoints, ensuring that the angular difference between each sampled viewpoint and the input viewpoint does not exceed 85°. The angular constraint is intended to introduce useful viewpoint variation without requiring excessive extrapolation into regions unobserved in the input.

A.3 HOIGEN-1M DATA PROCESSING

We describe the data processing pipeline used to construct degraded-to-clean video pairs from Internet hand-object interaction (HOI) videos. As illustrated in Figure 3, the pipeline consists of two stages. We first apply transition, camera-motion, and geometric-confidence filtering to identify video clips suitable for geometric reconstruction. We then reconstruct frame-wise 3D point clouds and synthesize degraded videos by sampling novel viewpoints and simulating view-dependent occlusions. The resulting degraded sequences are paired with the original RGB videos for training.

A.3.1 DATA FILTERING AND PREPROCESSING

Data source and filtering strategy. We use HOIGen-1M Liu et al. (2025) as the source of scalable monocular HOI videos. The dataset covers diverse object categories, interaction motions, and capture conditions. For our geometric processing, we retain clips with temporally continuous content, sufficiently reliable geometry, and limited camera motion.

To reduce the computational cost of subsequent reconstruction and synthetic degradation, we adopt a conservative cascading filtering strategy. Videos are sequentially processed by transition detection, static-camera filtering, preliminary quality checks, and geometric-confidence filtering. Only clips with sufficient evidence of satisfying the quality criteria are passed to the next stage; ambiguous or inconsistent samples are discarded. This strategy prioritizes geometric reliability at the cost of a lower data retention rate.

Transition filtering. We detect shot boundaries and gradual transitions by measuring color and structural differences between adjacent frames together with short-term brightness variations. Only

temporally continuous clips are retained, preventing content from different shots from being mixed during geometric reconstruction and temporal modeling.

Static-camera filtering. We use optical-flow-based motion analysis to identify clips with approximately static backgrounds. This filtering step suppresses videos with substantial camera motion so that the observed changes are dominated by human and object motion.

Quality and geometric-confidence filtering. We first remove videos with undesirable visual properties, such as black borders, portrait layouts, or abnormal content changes. We then reconstruct the candidate clips using VGGT-Ω and evaluate their geometric reliability using the confidence predicted for each reconstructed point. Clips with consistently low geometric confidence are discarded to reduce the influence of unreliable geometry on subsequent degradation generation.

Clip standardization. The retained videos are divided into fixed-length clips and resized to a consistent spatial resolution. Each sample contains $T = 49$ frames at 640×480 resolution, providing frame-wise correspondence between the original and degraded videos used during training.

A.3.2 Synthetic Degraded-to-Clean Pair Construction

Given a retained monocular video, we synthesize incomplete observations that mimic the artifacts encountered during cross-view warping. Specifically, we sample virtual novel viewpoints, determine the scene points visible from each viewpoint, and warp these visible points back to the original camera. This procedure introduces structured missing regions caused by occlusion and viewpoint changes while preserving temporal and spatial alignment with the original RGB sequence.

Frame-wise geometry reconstruction. For a standardized RGB sequence $\mathcal{I} = \{I_t\}_{t=1}^{T}$ with $T = 49$, we process the entire sequence using VGGT-Ω to obtain the per-frame depth D_t, confidence map S_t, camera intrinsics K_t, and world-to-camera extrinsics $[R_t \mid \mathbf{t}_t]$.

Reliable dynamic point-cloud construction. Before back-projection, we remove low-confidence points, invalid depths, and unstable points near depth discontinuities. Let $\mathcal{N}(\mathbf{u})$ denote the 3×3 neighborhood of pixel $\mathbf{u}$. We define the local relative depth variation as

$$e_t(\mathbf{u}) = \frac{\max_{\mathbf{v}\in\mathcal{N}(\mathbf{u})} D_t(\mathbf{v}) - \min_{\mathbf{v}\in\mathcal{N}(\mathbf{u})} D_t(\mathbf{v})}{\max\left(|D_t(\mathbf{u})|, 10^{-6}\right)}. \tag{17}$$

We retain a pixel only when its depth and confidence are finite, its depth is positive, $S_t(\mathbf{u}) > \tau_{\text{point}}$, and $e_t(\mathbf{u}) \leq 0.03$. The last criterion removes floating points that frequently occur around depth discontinuities. The remaining pixels are back-projected to construct a dynamic point cloud $\mathcal{P}_t$ for each frame.

Scene-center estimation and novel-view sampling. We estimate a reference depth from the reconstructed point clouds using the 33rd percentile of the valid point depths and place a reference center $\mathbf{s}$ at the corresponding position along the optical axis of the source camera.

We then sample virtual cameras on a sphere centered at $\mathbf{s}$, with all sampled cameras oriented toward the reference center. We sample 12 novel viewpoints for each clip. To introduce diverse occlusion patterns while avoiding excessively large viewpoint changes, the sampling region is restricted to the lateral and upper directions relative to the source camera. Let α and β denote the yaw and upward pitch offsets, respectively. We constrain the sampling range using an upper-half ellipse:

$$\frac{\alpha^2}{\alpha_{\max}^2} + \frac{\beta^2}{\beta_{\max}^2} \leq 1, \qquad 0 \leq \beta \leq \beta_{\max}, \tag{18}$$

where $\alpha_{\max}$ and $\beta_{\max}$ denote the maximum yaw and pitch offsets. This coupled constraint prevents both directions from simultaneously taking extreme values.

Each virtual camera inherits the intrinsics of the fixed source camera, while its orientation is adjusted such that its optical axis passes through $\mathbf{s}$. The cameras remain fixed throughout each video clip.

Novel-view visibility filtering and source-view reprojection. For each frame, we first project $\mathcal{P}_t$ into a sampled virtual camera and retain only points lying inside its field of view. We then apply a depth-based visibility test to remove points occluded by closer surfaces along similar viewing directions. The retained points therefore approximate the portion of the reconstructed scene visible from the sampled novel viewpoint.

These visible points are subsequently projected back to the source camera to produce a degraded frame $\widetilde{I}_t^{(v)}$. Pixels not covered by any projected point are filled with black. This procedure mimics the structured missing regions caused by occlusion and field-of-view differences when observations from a novel viewpoint are warped back to the source view. Applying the procedure to all frames produces a degraded sequence that is temporally aligned with the original RGB video.

The synthesized degraded video and its corresponding original RGB sequence form a degraded-to-clean training pair. During training, the latent representation of the degraded video is used as the conditioning input, while the latent representation of the original video serves as the clean reconstruction target. Since HOIGen-1M contains diverse scenes, objects, interaction motions, and capture conditions, and our novel-view sampling further introduces varying occlusion patterns, the resulting training pairs provide substantially broader visual and geometric diversity than controlled multi-view HOI datasets. This enables the model to learn recovery from incomplete warp conditions across diverse HOI content rather than relying on specific scenes, objects, or interaction patterns.

A.4 Evaluation Metrics

Following ViDiHand Wang et al. (2026), we evaluate hand reconstruction from two aspects: hand detection and pose reconstruction, using nine metrics in total. All 3D quantities are evaluated in the camera coordinate system. We use $\widehat{\cdot}$ and $(\cdot)^\star$ to denote predicted and ground-truth quantities, respectively.

A.4.1 Hand Detection

Evaluation targets and matching. A ground-truth hand is considered to be inside the image if at least one of its joints has depth $z > 0.01\,\mathrm{m}$ and projects inside the image region $[0, W) \times [0, H)$. This criterion only determines whether the hand lies inside the camera field of view and does not exclude joints occluded by objects.

We retain predictions with hand confidence greater than 0.5. Predicted and ground-truth hands are matched using the 2D bounding boxes obtained by projecting their MANO meshes onto the image plane. For ground-truth hands, we enlarge the bounding-box width and height by a factor of 1.1 around the box center. Candidate matches are required to have the same hand side and an IoU greater than 0.1. We then perform one-to-one greedy matching in descending order of IoU. Matched pairs are counted as true positives (TP), unmatched ground-truth hands as false negatives (FN), and unmatched valid predictions as false positives (FP). Let n_{TP}, n_{FN}, and n_{FP} denote their counts over the entire test set.

Frame accuracy (FAcc). FAcc measures the fraction of frames containing neither missed detections nor false positives. A frame is counted as correct only when all in-frame ground-truth hands are detected and no extra valid prediction is produced.

F_1 score. We report the F_1 score to jointly account for false positives and false negatives:

$$\mathrm{F}_1 = \frac{2n_{\mathrm{TP}}}{2n_{\mathrm{TP}} + n_{\mathrm{FN}} + n_{\mathrm{FP}}}. \tag{19}$$

A.4.2 Pose Reconstruction

For a single hand, let $\widehat{\mathbf{J}}_j$ and $\mathbf{J}_j^\star$ denote the predicted and ground-truth joint positions, and let $\widehat{\mathbf{V}}_v$ and $\mathbf{V}_v^\star$ denote the corresponding MANO mesh vertices. We evaluate $J = 21$ joints and $V = 778$ vertices. Joint $j = 1$ denotes the wrist.

MPJPE and PA-MPJPE. For MPJPE, the predicted and ground-truth joints are independently aligned to their corresponding wrist locations before computing the mean Euclidean distance:

$$e_{\mathrm{MPJPE}} = \frac{1}{J}\sum_{j=1}^{J} \left\| (\widehat{\mathbf{J}}_j - \widehat{\mathbf{J}}_1) - (\mathbf{J}_j^\star - \mathbf{J}_1^\star) \right\|_2. \tag{20}$$

Wrist alignment removes the effect of global translation while retaining differences in global orientation, scale, and hand articulation.

PA-MPJPE is computed after Procrustes alignment. Let $\widehat{\mathbf{J}}_j^{\mathrm{PA}}$ denote the aligned predicted joint:

$$e_{\text{PA-MPJPE}} = \frac{1}{J}\sum_{j=1}^{J}\left\|\widehat{\mathbf{J}}_j^{\mathrm{PA}} - \mathbf{J}_j^{\star}\right\|_2. \tag{21}$$

The alignment permits a proper rotation but excludes reflections. Both MPJPE and PA-MPJPE are reported in millimeters.

MPVPE and PA-MPVPE. MPVPE and PA-MPVPE are defined analogously to MPJPE and PA-MPJPE, respectively, with the 21 hand joints replaced by the 778 corresponding MANO mesh vertices. These metrics measure the reconstruction error of the full hand surface and are reported in millimeters.

End-point error (EPE). EPE measures the image-space distance between corresponding projected joints. Let π denote perspective projection using the camera intrinsics. For joint j, we define

$$e_{\mathrm{EPE},j} = \min\left(\left\|\pi(\widehat{\mathbf{J}}_j) - \pi(\mathbf{J}_j^{\star})\right\|_2, \sqrt{W^2+H^2}\right). \tag{22}$$

We evaluate only ground-truth joints with depth greater than $0.01\ \mathrm{m}$ whose projections lie inside the image. The final EPE is averaged over all such joints and reported in pixels.

Global orientation error (GO). GO measures the geodesic angular difference between the predicted and ground-truth MANO global rotations. Let $\widehat{G}, G^{\star} \in \mathrm{SO}(3)$ denote the corresponding rotation matrices in the camera coordinate system. We compute

$$e_{\mathrm{GO}} = \arccos\left(\frac{\mathrm{tr}(\widehat{G}^{\top}G^{\star}) - 1}{2}\right). \tag{23}$$

The error is reported in degrees and lies in $[0, 180]$.

Camera-space translation error (CT). CT measures the Euclidean distance between the predicted and ground-truth MANO global translations:

$$e_{\mathrm{CT}} = \left\|\widehat{\mathbf{t}} - \mathbf{t}^{\star}\right\|_2. \tag{24}$$

Both translations follow the same camera-coordinate convention. CT is reported in meters and measures the error of the global hand position, including depth. We use the MANO translation parameter rather than the decoded wrist-joint location.

A.4.3 Presence-Aware Pose Evaluation

Evaluating pose errors only on successfully detected hands excludes missed detections from the statistics. Since challenging cases, such as severe occlusion or complex hand-object interactions, are more likely to be missed, such an evaluation may bias the results toward easier examples and overestimate reconstruction performance. Following ViDiHand Wang et al. (2026), we therefore include every in-frame ground-truth hand in pose evaluation.

We use the suffix "-p" to denote presence-aware pose metrics, where p stands for *presence*. For a successfully matched prediction, the corresponding error is computed using the definitions above. For a missed ground-truth hand, we assign an error according to the following rules.

MPJPE-p and PA-MPJPE-p. For a missed hand, we construct a MANO hand of the same side using the mean shape and mean local pose, with identity global rotation and zero translation. We wrist-align this canonical hand and the ground-truth hand and compute the MPJPE. The resulting error is used for both MPJPE-p and PA-MPJPE-p. We do not apply Procrustes alignment for missed detections, preventing rotation and scale alignment from artificially reducing the miss penalty.

MPVPE-p and PA-MPVPE-p. We use the vertices of the same canonical MANO hand and compute MPVPE after wrist alignment with the ground-truth mesh. The resulting value is used for both MPVPE-p and PA-MPVPE-p, again without Procrustes alignment.

EPE-p. For a missed hand, each in-frame ground-truth joint is assigned the image diagonal $\sqrt{W^2+H^2}$ as its pixel error.

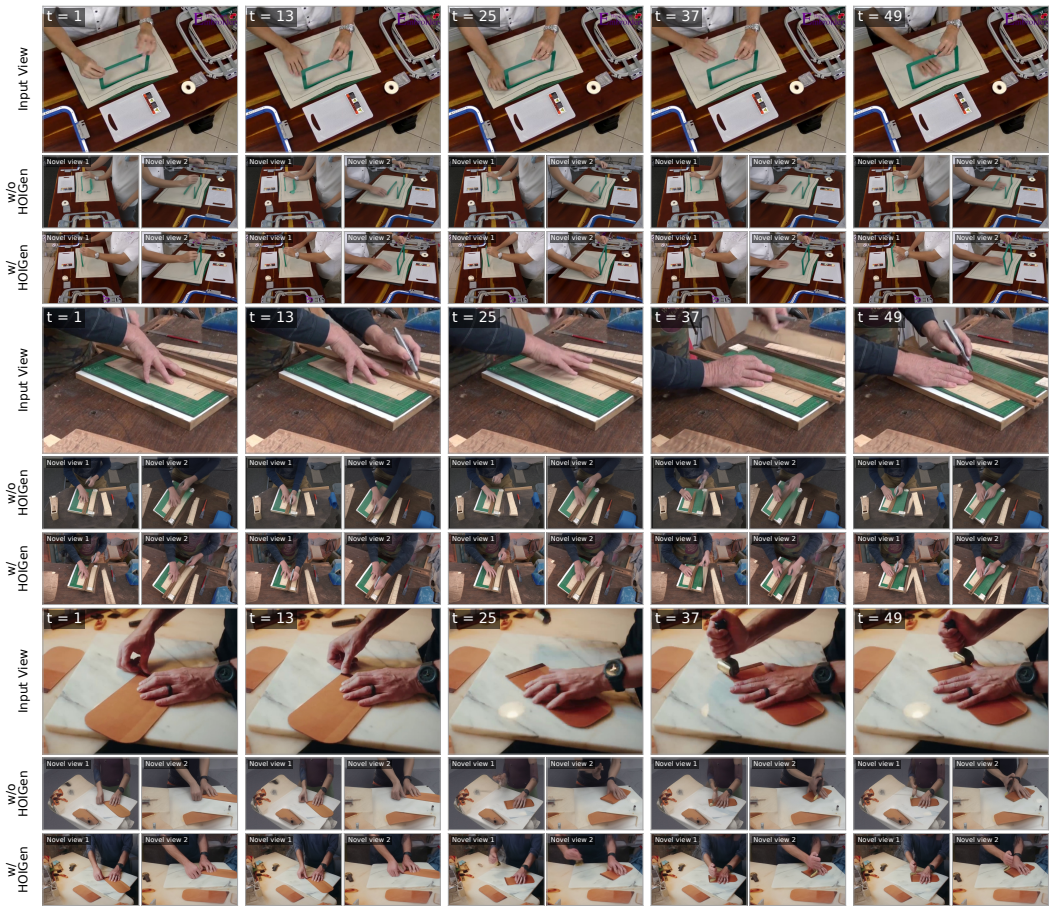

Figure 6: **Effect of HOIGen-1M training data on in-the-wild novel-view synthesis.**

GO-p. For a missed hand, we set the predicted global rotation to the identity, $\widehat{G} = I_3$, and compute the angular error to the ground-truth orientation using Eq. 23.

CT-p. For a missed hand, we set the predicted translation to zero, $\widehat{\mathbf{t}} = \mathbf{0}$. Its contribution is therefore

$$e_{\text{CT-p}} = \|\mathbf{t}^{\star}\|_2 \,. \tag{25}$$

The resulting presence-aware metrics jointly reflect detection coverage and reconstruction accuracy, preventing methods that miss difficult samples from being favored solely because pose errors are computed only on successfully detected hands.

A.5 HOIGen-1M Ablation

We compare novel-view synthesis results with and without HOIGen-1M in the training data in Figure 6. All input videos are collected from the Internet and depict hand–object interactions in unconstrained, real-world settings. For each example, we show selected input frames and two novel views generated by models trained without and with HOIGen-1M. Incorporating HOIGen-1M better preserves scene appearance and hand–object interactions across the generated views.

A.6 Occlusion Stratification

We construct occlusion subsets of ARCTIC using geometry-based visibility rules, considering only hand occlusions caused by the manipulated object or the other hand. For each of the 49 sampled frames in a clip, we reconstruct the MANO hand meshes and object mesh from the dataset annotations and render their depth maps separately using the camera parameters. We define each hand's

occlusion ratio as the number of occluded pixels divided by its projected area when rendered in isolation; a pixel is considered occluded only if the occluding surface is more than 1 mm closer to the camera than the hand surface. We classify clips according to the following rules:

- **Severe.** At least one hand has an occlusion ratio of at least 90% in at least 50% of the sampled frames. The occluded hand may alternate between the left and right hands across frames.
- **Moderate.** The clip does not qualify as severe, and the same hand has an occlusion ratio of at least 50% in at least 50% of the sampled frames.
- **Light.** Each hand's mean occlusion ratio is at most 10%; both hands have occlusion ratios of at most 20% in at least 90% of the sampled frames; and neither hand exceeds 30% occlusion in any sampled frame.

Clips that satisfy none of these criteria are excluded, yielding three mutually exclusive occlusion subsets.