arXiv:2603.20850v2 [cs.CV] 8 Jun 2026

Glove2Hand: Synthesizing Natural Hand-Object Interaction from Multi-Modal Sensing Gloves

Xinyu Zhang[†,‡] Ziyi Kou Chuan Qin Mia Huang Ergys Ristani
Ankit Kumar Lele Chen Kun He Abdeslam Boularias[‡] Li Guan

Meta Reality Labs [‡]Rutgers University mlzxy.github.io/glove2hand

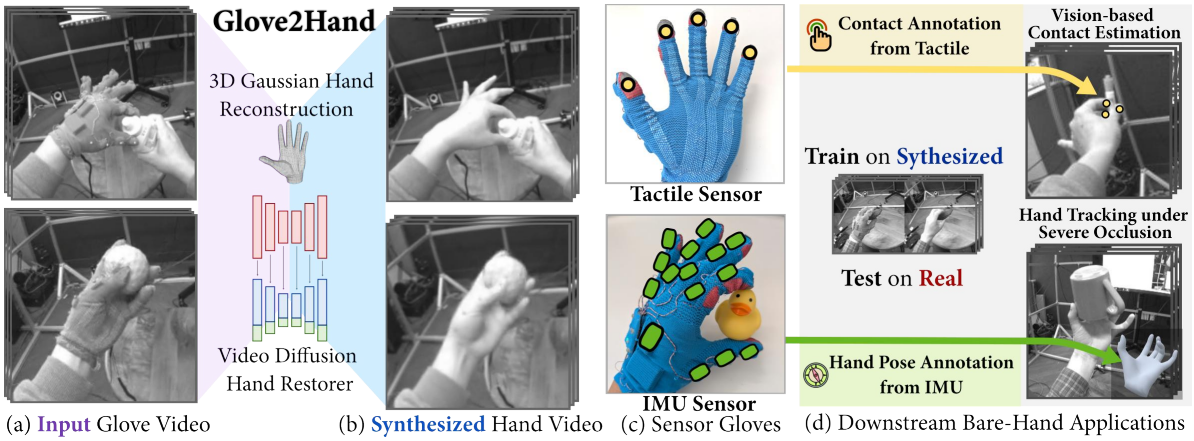

Figure 1. Glove2Hand translates egocentric glove videos (a) into photorealistic, temporally consistent, bare-hand videos (b) capturing complex interactions with non-rigid objects. The translated videos are accompanied with synchronized tactile and IMU signals (c) from sensing gloves, which significantly enhances downstream bare hand tasks (d) by providing contact events from tactile as ground-truth for vision-based contact estimation; and occlusion-free hand poses from IMU as ground-truth for hand tracking under severe occlusion.

Abstract

Understanding hand-object interaction (HOI) is fundamental to computer vision, robotics, and AR/VR. However, conventional hand videos often lack essential physical information such as contact forces and motion signals, and are prone to frequent occlusions. To address the challenges, we present Glove2Hand, a framework that translates multi-modal sensing glove HOI videos into photorealistic bare hands, while faithfully preserving the underlying physical interaction dynamics. We introduce a novel 3D Gaussian hand model that ensures temporal rendering consistency. The rendered hand is seamlessly integrated into the scene using a diffusion-based hand restorer, which effectively handles complex hand-object interactions and non-rigid deformations. Leveraging Glove2Hand, we create HandSense, the first multi-modal HOI dataset featuring glove-to-hand videos with synchronized tactile and IMU signals. We demonstrate that HandSense significantly enhances downstream bare-hand applications, including video-based contact estimation and hand tracking under severe occlusion.

1. Introduction

Understanding human hand-object interaction (HOI) is a critical problem in computer vision, robotics, augmented reality (AR), virtual reality (VR), and human-computer interaction [6, 19, 23, 24, 30, 44, 47, 60]. One common approach is to capture egocentric videos that contain various human hands, objects, and their interactions for developing data-driven algorithms [16, 26, 46, 68]. However, most HOI capturing systems nowadays are fundamentally limited to vision-only modality. As a result, they i) lack the rich physical grounding of force and contact information of hands; and ii) suffer from hand occlusion due to limited camera views, which makes robust hand tracking difficult [21].

Several vision-based solutions have been proposed to address the above limitations of HOI videos. For exam-

[†]Work done during an internship at Meta Reality Labs.

ple, ContactPose [4] estimates binary fingertip contact from reconstructed mesh. But it only works with pre-scanned rigid objects and is not capable of providing continuous force measurements. Multi-camera studio [49] is designed to mitigate the hand occlusion issue, but is impractical for in-the-wild capture due to complex setup and calibration. Recently, sensing gloves have emerged as a new wearable device that captures various sensor modalities (e.g., IMU [40, 65], tactile [45, 82]) and can be accompanied with egocentric camera devices for multi-modal data collection. Fig. 1.a shows an example of a sensing glove with full-hand IMU and fingertip tactile sensors. Nonetheless, significant appearance gaps between such sensing gloves and human bare hands lead to low generalizability of vision models trained on glove data to bare hand tasks.

To address these limitations, we propose Glove2Hand, a generative 3D reconstruction-based video approach that visually translates the sensing gloves in captured egocentric videos to photorealistic bare hands while faithfully preserving valuable IMU and tactile signals obtained from the sensor gloves. There are two major gaps in existing work that need to be addressed in order to achieve photo-realistic quality videos: (1) maintaining temporal and multi-view consistency across frames instead of focusing on only static images [8, 43]; and (2) handling interactions with complex objects (including non-rigid objects with unknown shapes) instead of those with known and rigid shapes [66, 76].

Glove2Hand combines the strengths of 3D reconstruction for consistency and generative modeling for flexibility and photo-realism. We first build a minimalist 3D Gaussian hand model, which defines statistical surface distributions over a learnable canonical hand mesh. The design provides a strong geometrical prior, naturally enables relighting, and renders spatiotemporally consistent object-free hand videos. Subsequently, unlike previous methods that explicitly model object geometry [2, 76], we represent objects and background in the pixel domain, allowing for greater flexibility in handling objects with unknown or deformable shapes. We then train a diffusion-based hand restorer to seamlessly integrate the rendered hands into the scene, refining hand-object details and overall coherence.

To demonstrate the effectiveness of Glove2Hand on enhancing bare-hand learning tasks, we further create a multi-modal HOI video dataset *HandSense* that contains accurate glove poses, realistically synthesized bare-hand videos, and time-synchronized sensors including IMU and tactile signals. In particular, we design two bare-hand tasks on HandSense (Fig. 1d): *Vision-based Fingertip Contact Estimation* and *Hand Pose Tracking under Heavy Occlusion*. We conduct extensive experiments to show that Glove2Hand effectively enhances the model performance on both tasks for bare hands. Our contributions are three folds:

- We propose the Glove2Hand framework for glove-to-hand video translation, which leverages a novel 3D Gaussian hand for consistent rendering, and a diffusion hand restorer for generative refinement of hand-object details. To our best knowledge, this is the first work to generate photorealistic and authentic hand-object interactions from sensor glove captures, bridging the large appearance gap between the sensor glove and human hand.
- We evaluate the generative quality of our Glove2Hand on our new dataset *HandSense*, the first HOI dataset with synchronized bare-hand and sensor glove videos, and directly measured continuous tactile and IMU signals.
- We demonstrate the value of synthesized hand videos on two novel applications: Vision-based contact estimation and robust hand tracking under severe occlusion.

2. Related Work

Generative Models for Hands. Research on generating hand imagery falls into three categories. **Hand avatars** [9, 11, 32, 36, 78, 80] use rendering methods like raytracing [5], NeRF [48] or Gaussian Splatting [5, 37] to offer precise control, but typically generate only the hand without backgrounds or interaction. **Hand diffusion models** [8, 43, 51, 52, 54, 58, 74] generate plausible hands within general image synthesis pipelines, with backgrounds, but lack precise control and focus on single, low-detail, static images. Finally, **hand-object interaction synthesis** [66, 72, 76] specializes in generating close-up, static images of contact, for instance by estimating grasp poses from object images [72] or rendering from pre-scanned meshes [76].

Our framework distinguishes from these approaches in two key aspects. First, we generate dynamic **videos** of hand-object interaction, capturing actions rather than the static images produced by prior work. Second, the generated hand motions are **grounded by an input glove video**. This grounding provides precise controllability, enabling multi-modal sensor fusion and scalable data curation without requiring hard-to-obtain non-rigid object meshes.

Video-to-Video Translation. Early image-to-image translation learned mappings from aligned [31] or unpaired data [83], but cannot handle the large embodiment differences in our task. Recent video translation work [3, 10, 28, 69, 70] often relies on latent space interpolation in pre-trained image diffusion models. In contrast, our method learns a direct glove-to-hand video translation from unpaired data, addressing a significant domain gap without being constrained by a pre-trained model's capabilities.

Video Inpainting. Video inpainting traditionally focused on object removal and background completion, often using optical flow for temporal consistency [39, 81]. More recent methods leverage diffusion models for language-conditioned inpainting [18, 34, 73]. Our work is distinct in its focus on inpainting highly detailed hand-object contact regions, a particularly challenging sub-problem due to the

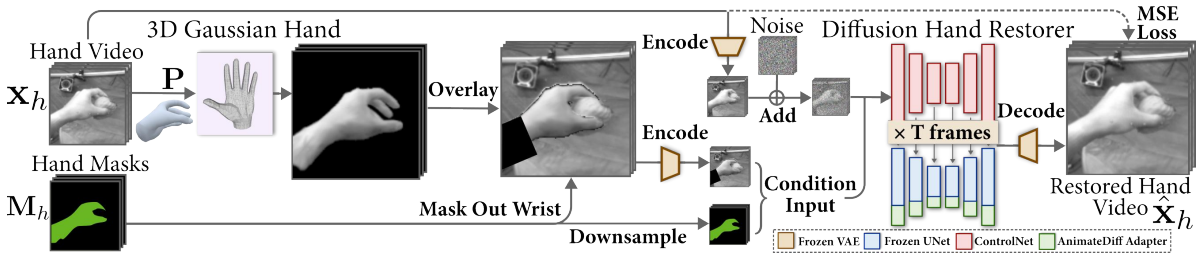

Figure 2. **Glove2Hand Training.** We extract hand poses from input videos and render corresponding hand only frames using our 3D Gaussian hand model. The rendered hands are then overlayed onto the original videos with wrist regions masked. The resulting video is encoded into a latent space and serves as a condition for our diffusion hand restorer, which is trained to restore the original hand video.

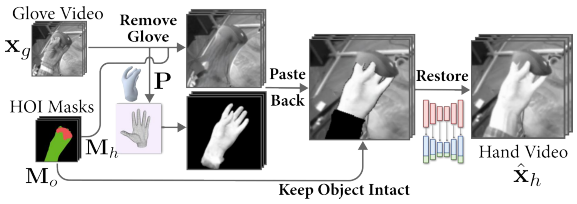

Figure 3. **Glove2Hand Inference.** We render hand only frames using hand poses from glove videos. Prior to pasting, we erase the glove region with an optical-flow based background inpainter [81], and ensures object pixels remain intact during pasting. The resulting video is encoded and fed as input to diffusion hand restorer.

dexterity of hands and the complexity of contact physics.

Hand Datasets with Tactile Sensing. Datasets for hand-object interaction [2, 42, 64] often estimate binary contact from meshes [4, 12, 38]. While some provide continuous force, they are often limited to flat tablets (restricting dexterity) [1, 14, 15, 79], are simulation-based [67], or are glove-only datasets lacking visual data [33]. To our knowledge, our work creates the first dataset combining sensor-measured tactile and IMU signals with synchronized, photorealistic bare-hand videos of dexterous manipulation.

3. Glove-to-Hand Framework

3.1. Problem Formulation

We define a video clip as $\mathbf{x} \in \mathbb{R}^{T\times H\times W\times C}$ in either glove domain $\mathbf{x}_g \in \mathcal{I}_{\text{glove}}$ or hand domain $\mathbf{x}_h \in \mathcal{I}_{\text{hand}}$. Each clip is associated with hand poses $\mathbf{P} \in \mathbb{R}^{T\times J\times 3}$, hand and object masks, $\mathbf{M}_h, \mathbf{M}_o \in \{0,1\}^{T\times H\times W}$, forming a sample tuple $s = (\mathbf{x}, \mathbf{P}, \mathbf{M}_h, \mathbf{M}_o)$. Let $\mathcal{D}_{\text{glove}} = \{s \mid \mathbf{x} \in \mathcal{I}_{\text{glove}}\}$, our objective is to learn a mapping $f : \mathcal{D}_{\text{glove}} \mapsto \mathcal{I}_{\text{hand}}$ that translates a glove sequence into a photorealistic hand video without manual edits. We assume $\mathcal{I}_{\text{glove}}$ and $\mathcal{I}_{\text{hand}}$ are unpaired, where no sample correspondence exists across domains.

3.2. Method

The key challenges in realistic glove-to-hand video translation are twofold: (1) achieving temporal and multi-view consistency, which is notoriously difficult for generative models, and (2) modeling complex interactions with non-rigid objects. This is further compounded by the significant visual domain gap between the sensor glove and a bare hand. Our core insight is that despite this appearance gap, both the glove and hand share the same underlying articulated structure, represented by the hand pose $\mathbf{P}$.

This realization allows us to decompose the problem into two more tractable subproblems: (1) transforming the glove video into a consistent, in-air hand sequence, and (2) integrating this hand sequence back into the scene while refining interaction details. Therefore, we combine the consistency of a reconstruction-based hand avatar (to solve the first subproblem) with the generative flexibility of a diffusion model (to solve the second). This approach allows us to refine complex hand-object boundary details without requiring explicit 3D representations of objects or the background. Our framework is visualized in Fig. 2 and Fig. 3.

3.2.1. 3D Gaussian Hand

Hand avatar methods deterministically reconstruct 3D hand representations and render novel views from a given hand pose and camera pose, typically using neural rendering techniques rather than generative sampling. However, applying existing avatar methods to our scenario presents several challenges: (1) state-of-the-art methods often require dense, multi-view camera setups (e.g., hundreds of views [32, 49]), whereas our egocentric headset provides only two closely positioned cameras; (2) existing avatar datasets are typically captured under controlled, calibrated lighting conditions. In contrast, our egocentric scenario features frequent and dynamic lighting changes inherent to real-world human movement. Ignoring lighting variations leads to ambiguous reconstructions; (3) established methods often rely on heavy, multi-stage pipelines involving NeRF or ray marching [9, 32, 36, 50]. While recent work has explored Gaussian Splatting for hand avatars [11, 61, 78], their implementations are complex and not publicly available. As this avatar serves as a component within our larger framework, we prioritize simplicity and efficiency. We therefore develop a minimalist 3D Gaussian hand model

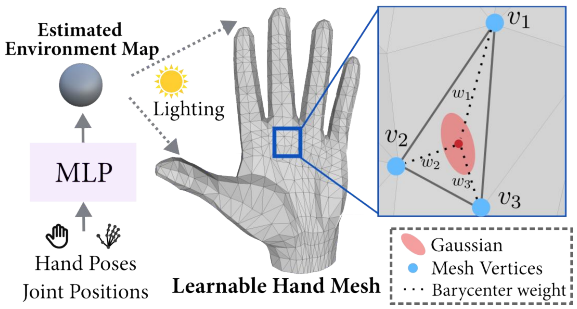

Figure 4. **Surface-grounded and Relightable Gaussian Hand.** We defines Gaussians over the surface of a hand mesh, which benefits from the mesh geometric priors and naturally enables lighting estimation based on the consistent mesh surface normals.

adapted for sparse camera inputs and variable lighting.

Our design is motivated by a key insight: Canonical hand meshes provide a strong geometric prior but lack learning flexibility, whereas Gaussian splatting is highly flexible but lacks inherent structure. We unify these advantages by defining 3D Gaussians directly on the canonical mesh surface. This simple yet effective modification also naturally addresses the challenge of variable lighting. The underlying mesh provides consistent surface normals which are essential for enabling robust lighting estimation. Our two proposed techniques are detailed below and illustrated in Fig 4.

Surface-Grounded Gaussians. We define Gaussians directly on mesh surfaces. For a triangle with vertices v_1, v_2, v_3, each Gaussian is characterized by barycentric weights $\mathbf{w} = [w_1, w_2, w_3]$ where $\sum_i w_i = 1$, 2D scale $\mathbf{s} = [s_x, s_y]$, and rotation ϕ. Unlike 2DGS [29] which defines Gaussians in 3D space with regularization to encourage surface formation, our method leverages a stronger prior and directly anchors Gaussians to mesh surfaces.

During hand deformation, instead of applying linear blend skinning to Gaussians (which can break surface integrity [77]), we transform only the mesh triangles and recompute Gaussians on the deformed surfaces. This leverages the hand mesh's pre-existing bone weights, avoiding the learning of skinning weights for individual Gaussians.

Formally, given triangle vertices v_1, v_2, v_3 in canonical frame and v'_1, v'_2, v'_3 in deformed frame, we compute edges $\mathbf{e}_1 = v_2 - v_1$, $\mathbf{e}_2 = v_3 - v_1$ and similarly $\mathbf{e}'_1 = v'_2 - v'_1$, $\mathbf{e}'_2 = v'_3 - v'_1$ for the deformed frame. These edges are projected to local 2D coordinates using an orthonormal basis $(\mathbf{u}, \mathbf{v})$ derived via Gram-Schmidt: $\mathbf{u} = \frac{\mathbf{e}_1}{\|\mathbf{e}_1\|}$, $\mathbf{w}_\perp = \mathbf{e}_2 - \text{proj}_\mathbf{u}(\mathbf{e}_2)$, $\mathbf{v} = \frac{\mathbf{w}_\perp}{\|\mathbf{w}_\perp\|}$, yielding $\mathbf{e}_{1,D} = (\mathbf{e}_1 \cdot \mathbf{u}, \mathbf{e}_1 \cdot \mathbf{v})$ and $\mathbf{e}_{2,D} = (\mathbf{e}_2 \cdot \mathbf{u}, \mathbf{e}_2 \cdot \mathbf{v})$. We point out that any point p on the triangle can be expressed as $p = c_1\mathbf{e}_{1,D} + c_2\mathbf{e}_{2,D}$, and the coefficients $\mathbf{c} = [c_1, c_2]^T$ shall remain constant under deformation. This gives us $p = \mathbf{M}_{\text{canon}}\mathbf{c}$ and $p' = \mathbf{M}_{\text{deform}}\mathbf{c}$ for the deformed frame, where $\mathbf{M}_{\text{canon}} = [\mathbf{e}_{1,D}; \mathbf{e}_{2,D}]$ and $\mathbf{M}_{\text{deform}} = [\mathbf{e}'_{1,D}; \mathbf{e}'_{2,D}]$. The deformation gradient $\mathbf{A} = \mathbf{M}_{\text{deform}}\mathbf{M}^{-1}_{\text{canon}}$ maps point from canonical frame to deformed frame by $p' = \mathbf{A}p$. A surface-grounded Gaussian can be represented by an ellipse in quadratic form $\mathbf{Q} = \mathbf{RSR}^T$. $\mathbf{R}$ is the 2D rotation matrix for ϕ and $\mathbf{S} = \text{diag}(1/s_x^2, 1/s_y^2)$. The deformed ellipse becomes $\mathbf{Q}' = \mathbf{A}^{-T}\mathbf{Q}\mathbf{A}^{-1}$, from which new scale s' and rotation ϕ' are obtained via eigen decomposition.

Each Gaussian is associated with a triangle. The adaptive control algorithm can be directly applied to flexibly allocate Gaussians—during each control cycle, we reassign Gaussians to their closest triangles. We learn a small per-Gaussian offset along the surface normal, ensuring overlapping Gaussians do not share the same depth, thereby avoiding ambiguity in point-based rendering [17]. We train these Gaussians using only image reconstruction loss through differentiable rendering. This hierarchical definition allows gradients to jointly finetune the canonical mesh vertices.

Relightable Hand Gaussians. To model variable illumination, we adapt the relighting framework proposed by LumiGauss [35]. This approach learns an environment map, represented as spherical harmonics (SH) coefficients $\mathbf{l} \in \mathbb{R}^{3\times(n^2+1)}$, where n is the SH order. The final color is computed as $\mathbf{c} \odot \text{SH}(\mathbf{l}, \mathbf{n})$, where $\mathbf{c}$ is the Gaussian's intrinsic color (albedo) and $\text{SH}(\mathbf{l}, \mathbf{n})$ evaluates the diffuse lighting component given the surface normal $\mathbf{n}$. However, this method presents two limitations for our dynamic, egocentric setting: (1) LumiGauss assumes a single static environment map, whereas our lighting conditions change dynamically; (2) it suffers from the well-known albedo-illumination ambiguity. As noted in LumiGauss [35], the learned Gaussian normals $\mathbf{n}$ and intrinsic colors $\mathbf{c}$ can co-adapt to incorrectly bake shadows into the albedo.

We address both limitations. To solve the first, we model lighting as a dynamic function of the hand's configuration. We employ a small MLP to predict the SH coefficients $\mathbf{l}$ from hand pose $\mathbf{P}$. This is effective because local lighting and self-shadowing are primarily determined by the hand's location and articulation. To solve the second, our core design of defining Gaussians on the mesh provides a natural solution. The surface normals $\mathbf{n}$ are provided by the mesh geometry but not Gaussian, which substantially mitigates the ambiguity. In practice, to better capture local effects, our MLP predicts two separate environment maps, $\mathbf{l}_{\text{palm}}$ and $\mathbf{l}_{\text{back}}$, for the palm and back of the hand, respectively.

3.2.2. Diffusion Hand Restorer

Our 3D Gaussian hand renders a consistent, hand-only sequence from the glove video. However, naively overlaying the render introduces several artifacts, such as implausible object interactions like penetration or floating, as the render is not interaction-aware; and an unnatural wrist connection, as the arm is not modeled. Some of these artifacts are il-

lustrated in Fig. 8. Furthermore, the bulky glove's larger shape, when replaced by the slimmer hand, leaves visible gaps around the hand. To address these limitations, we introduce a diffusion hand restorer based on ControlNet [75] and AnimateDiff [57], and trained on bare-hand videos. We generate the conditional input by overlaying the rendered hand, dilating its mask, and masking the wrist region. The network is trained to restore the original bare-hand video from such corrupted input, as shown in Fig. 2.

Glove-to-Hand Inference. Gloves have a different and larger shape than hands. Therefore, overlaying a rendered hand onto glove videos creates visible glove artifacts around the rendered hand's edges (illustrated in Fig. 8, columns 2-4), and can also incorrectly occlude foreground objects. To resolve this, we employ a glove-to-hand inference pipeline (Fig. 3). Specifically, we first extract glove and object masks using SAM-2 [55], prompted by bounding boxes detected with Grounding DINO [41]. Next, we use the glove mask to erase the glove region with Propainter [81], an optical flow-based background inpainter. We leverage the object mask to ensure that foreground object pixels remain intact. The entire pipeline is fully automatic.

4. HandSense Dataset

To validate the effectiveness of Glove2Hand, we create *HandSense*, a multi-modal HOI video dataset that contains egocentric videos with time-synchronized sensor signals. In particular, we recruited 5 subjects and performed two sessions of data collection from each. In this first session, the subject was asked to wear the sensing glove and complete 6 object manipulation tasks (e.g., in-hand rotation, pick-and-place). In the second session, the subject followed the same object manipulation protocol but didn't wear the gloves. We show the collected data in Figure 5 and more data collection details in the supplementary materials.

Data Modalities. For both sensing glove and bare hand sessions, we collect egocentric grayscale video recordings from headset cameras. To obtain accurate 3D hand poses, we adopt MoCap system [56, 63] with markers on the gloves/hands. For the sensing glove session, we collected IMU signals distributed on multiple hand phalanges, and tactile signals located at each fingertip. Both IMU and tactile signals are continuous values and time-synchronized with egocentric videos based on the closest timestamps. For the bare hand session, we manually annotate discrete binary contact/no-contact labels for each fingertip when a subject is manipulating an object. To mitigate the hand occlusion issue, we carefully designed the instructions for subjects to use specific fingers so that we have strong prior knowledge about contact events.

Measured Tactile Signals and Non-rigid Objects. We hightlight two key features that distinguish HandSense from existing datasets. First, to our best knowledge, HandSense

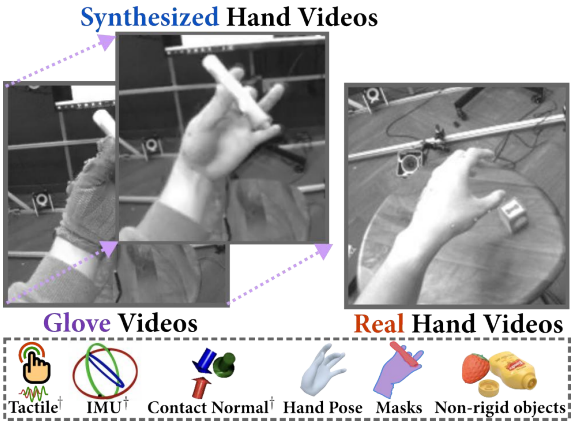

Figure 5. **HandSense Dataset.** HandSense provides egocentric HOI videos collected by both sensing gloves and human bare hands. † denotes multi-modal sensing modalities from gloves.

provides the first continuous, sensor-measured tactile signals synchronized with bare-hand HOI videos. In contrast, datasets like H2O [38] and ARCTIC [12] provide only binary, post-processed contact labels, which are estimated from hand and rigid object poses rather than measured. Second, HandSense features interactions with diverse object categories, including deformable (e.g., squishy toys) and articulated items (e.g., detachable lids), unlike prior methods that are often limited to pre-scanned, rigid objects.

With HandSense dataset, we demonstrate the value of Glove2Hand in two bare-hand applications: i) video-based contact estimation, benefited from tactile signals by sensing gloves as annotation-free ground truth, and ii) robust hand tracking under occlusion that augments estimated hand poses in occluded scenarios using IMU signals as pseudo ground-truth. The objective of both tasks is to facilitate deeper physical understanding of hand-object interaction based on sensor signals from gloves. We show more detailed experiment setups and results in Section 5.2.

5. Experiments

We design our experiments to answer three key questions:

1. How realistic are the hand-object interaction (HOI) videos generated by our Glove2Hand framework?
2. What is the contribution of each component of our proposed framework to the final generation quality?
3. What is the value of our generated HOI data with synchronized sensor signals, for bare hand applications?

To address the first, we evaluate our framework on HandSense using quantitative metrics and human evaluation. Second, we conduct ablations on each component. For the third, we demonstrate our data's value in video-based contact estimation and hand tracking under heavy occlusion.

Sensing Glove. Our sensing glove includes one tactile sen-

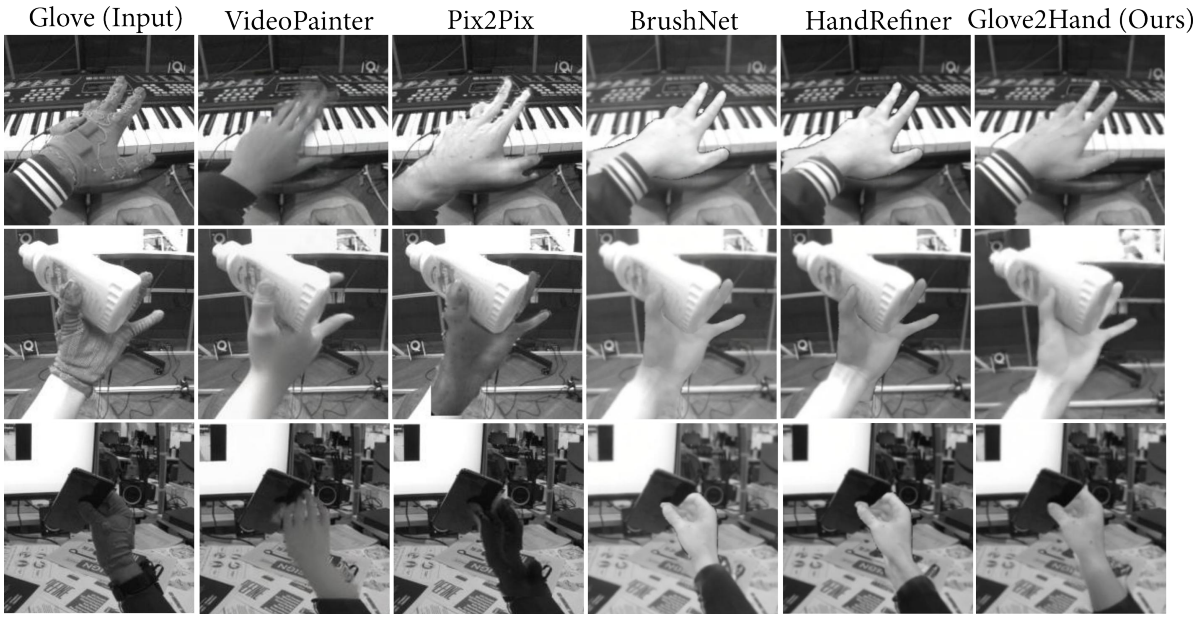

Figure 6. **Qualitative Comparison for Glove-to-Hand.** Our Glove2Hand produces photorealistic hands with clean object contact. In contrast, existing methods generate blurry results, or suffer from strong visual artifacts, unrealistic textures, and inconsistent lighting. The bottom row demonstrates the high generalization of Glove2Hand to unseen subjects, objects, and backgrounds. For a fair comparison, Pix2Pix, BrushNet, and HandRefiner were retrained on our dataset using the same inputs and HOI masks as our method (Fig. 2 and Fig. 3).

Method	**FID** ↓	**FVD** ↓	**FVD-long** ↓
Vanilla Glove	74.9	38.3	45.5
InpaintAnything* [73]	63.6	32.7	39.6
FoundHand* [8]	82.2	47.9	50.5
VideoPainter* [3]	72.2	35.7	42.2
CycleGAN [83]	63.6	30.8	37.4
Pix2Pix [31]	38.6	24.7	31.4
BrushNet [34]	37.9	34.5	40.4
HandRefiner [43]	35.5	24.2	29.7
Glove2Hand (Ours)	**30.1**	**19.5**	**24.5**

Table 1. **Evaluation for Glove-to-Hand Video Translation.** We evaluate Glove2Hand on HandSense dataset with FID, FVD (2-sec. clips), and FVD-long (on 10-30 sec. clips). (∗) denotes methods evaluated off-the-shelf without fine-tuning on HandSense.

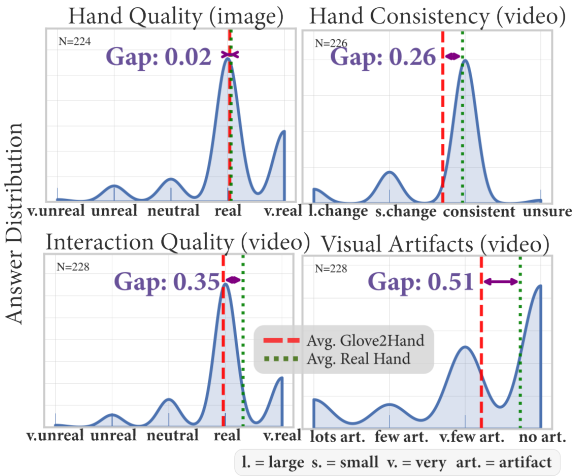

Figure 7. **Human Evaluation Results** . Our study compares the synthesized videos by Glove2Hand with real hand videos across four metrics (N=number of answers). The average of score distribution for our synthesized hands are close to real hands, with a max gap of 0.51. This gap (lower is better) is small relative to the 1.0 distance between adjacent categories, showing high realism.

sor per fingertip which produces one channel of continuous force values, and twelve 6-DoF IMU sensors distributed on the full hand for tracking hand pose. We provide more details of the glove configurations in the Supplementary.

5.1. Video Quality Evaluation

Setup. We train our framework on the HOT3D [2] and HandSense datasets, and evaluate on the testing split of HandSense. We firstly optimize a 3D Gaussian hand model for each subject. The subject-specific models are then frozen while training the single diffusion restorer. We perform same-subject glove-to-hand translation by using the corresponding 3D Gaussian hand. We measure realism against real hand videos using Fréchet Inception Distance (FID) [25] (clean-fid [53]) for image quality and Fréchet Video Distance (FVD) [62] (cd-fvd [13]) for video fidelity.

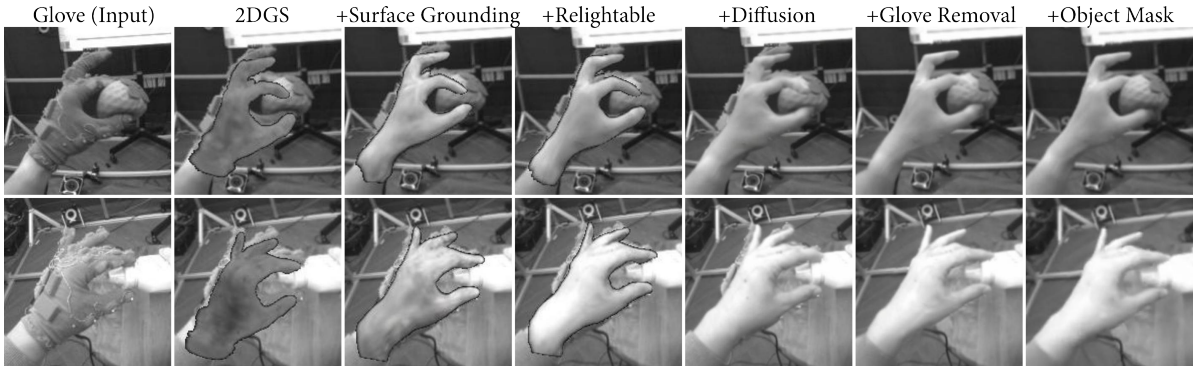

Figure 8. **Qualitative Ablations.** Our surface-grounded Gaussian reconstructs coherent hand geometry and appearance compared to 2DGS [29]. Adding the relightable Gaussian improves shading. The diffusion hand restorer further enhances skin texture, connects the wrist, and removes visual artifacts. Finally, glove removal and object masks eliminate glove artifacts and refine the hand-object boundary.

Results. We present quantitative and qualitative comparisons in Tab. 1 and Fig. 6. Glove2Hand significantly outperforms existing generative and vidoe inpainting baselines in both FID and FVD metrics. We omit some baselines from Fig. 6 as they fail on glove-to-hand translation by returning the input gloves largely unchanged. Compared to diffusion-only methods like HandRefiner [43], our 3D Gaussian representation synthesizes finer geometry and texture.

Human Evaluation. We conduct a user study to assess the realism of our generated hand-object interaction videos. We recruited five participants and presented them with a mix of our generated and real hand videos. For each sample, subjects answered questions regarding hand realism, the plausibility of the hand-object interaction, temporal consistency (e.g., identity preservation), the presence of visual artifacts, and hand motion stability. Each participant evaluated approximately over 40 images and 40 videos. We visualize the aggregated response distributions in Fig. 7. The results demonstrate that our generated videos achieve a high degree of realism, approaching the scores of real videos. Notably, for still images, participant responses indicate that our generated hands are nearly indistinguishable from real ones. We provide further details in the Supplementary.

Ablation Studies. We present quantitative and qualitative ablation studies in Tab. 2 and Fig. 8, respectively. The quantitative results show that each component incrementally improves image fidelity (FID). While we observe minor FVD fluctuations, the qualitative results in Fig. 8 clearly demonstrate the distinct and necessary contribution of each module. Specifically, our surface-grounded and relightable 3D Gaussian hand establishes a realistic, geometrically-accurate foundation. The Diffusion Hand Restorer builds upon this, fine-tuning skin textures, generating a natural wrist, and refining hand-object boundaries. The glove removal module is critical for eliminating visual artifacts from the input sensor glove, and the object mask is essential for preserving fine-grained contact details.

Configuration	FID ↓	FVD ↓	FVD-long ↓
2DGS [29]	91.1	50	62.9
+Surface Grounding	60.3	35.1	46.6
+Relightable	56.7	30.7	40.2
+Diffusion	32.3	19.8	**22.7**
+Glove Removal	31.2	20.9	25
+Object Mask	**30.1**	**19.5**	24.5

Table 2. **Ablation Results.** We perform an incremental ablation study to evaluate the contribution of each proposed component. The results demonstrate that our components work cohesively, with each addition progressively improving the final performance. Surface grounding and the diffusion yield the most significant gains. Qualitative ablations are presented in Fig. 8.

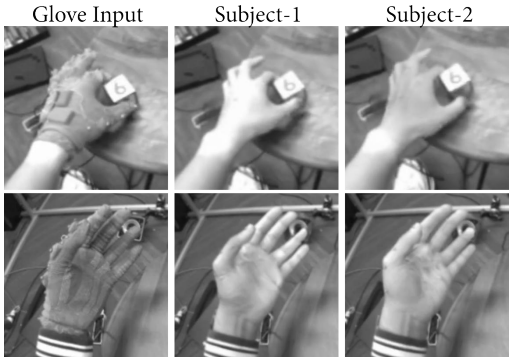

Figure 9. **Glove-to-Hand (different subjects).** We customize the output hand identity by training a 3D Gaussian hand per-subject, while sharing the diffusion hand restorer. The figure shows glove inputs (left) and the translated hand of different subjects (right).

5.2. Bare-Hand Applications

5.2.1. Vision-based Contact Estimation

Setup. Estimating hand-object contact from vision is challenging, and there is a lack of large-scale, accurately-

Method		Contact IoU (%)	Precision (%)	Recall (%)
PressureVision++ [14]		0.8	62.5	0.9
Ours	Glove only	71.5	82.8	83.9
	G2H only	75.6	90.6	82
	Hand only	85.3	90	94.2
	Hand+G2H	**88.2**	**92.6**	**94.9**

Table 3. **Results on Contact Estimation.** We train video contact estimator on glove-to-hand (G2H) videos, and evaluate on real, manually-annotated videos using Contact IoU [15], Precision, and Recall. Training on G2H videos significantly outperforms the glove ones. Combining real and G2H data yields the best result. G2H contact labels are automatically derived from tactile signals.

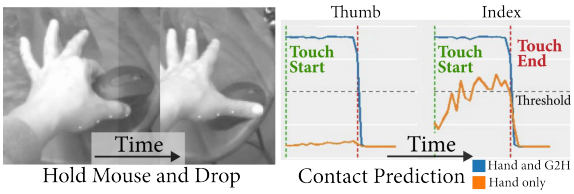

Figure 10. **Contact Estimation Example.** The model trained with only manually-annotated real hand data (orange) produces noisy and unstable predictions. In contrast, adding our glove-to-hand (G2H) data (blue) results in stable and accurate contact estimation.

labeled data. Existing methods like PressureVision [14, 15] capture contact on planar surfaces [1], but this setup is insufficient for dexterous hand-object interactions. Our Glove2Hand framework addresses this by generating realistic hand videos with synchronized tactile signals, which serve as dense ground-truth contact labels. We leverage this data to train a video-based contact estimator. The model takes a 32-frame clip as input and predicts binary contact states for each fingertip at each frame, yielding an output of size $\mathbb{R}^{32\times5}$. The architecture uses a frozen DINOv3 backbone [59] fine-tuned with LoRA [27]. We apply four temporal attention layers over the [CLS] tokens, followed by a linear classifier to predict contact logits. The model is trained on a combination of real hand videos and our generated glove-to-hand videos. Real hand videos are manually annotated, while ground-truth for our generated videos is obtained by thresholding the tactile signals.

Results. We evaluate all models on a held-out test set of manually annotated real hand videos in Tab. 3 with using Contact IoU [14], precision, and recall. A model trained on the translated videos significantly outperforms the one trained on the original glove videos, while combining our data with real videos yields the best performance. This validates our framework as a data curation engine and provides strong evidence for the generation quality.

5.2.2. Hand Tracking under Heavy Occlusion

Estimating hand pose during object interaction is critical, yet existing methods are frequently limited by self- and object-occlusion. A key barrier is the difficulty of acquiring ground-truth pose annotations in such occluded scenarios [7, 21]. We address this by leveraging on-glove IMU sensors. They provide accurate, vision-free joint measurements, enabling occluded pose capture without multi-camera studios which is suitable for in-the-wild recording. To evaluate, we finetune UmeTrack [22], a strong egocentric hand tracker, on our HandSense dataset using IMU-derived poses. We report the mean keypoint position error (MKPE) in mm, on a held-out test set of real videos with poses captured in a multi-camera MoCap system as ground truth. As shown in Tab. 4, finetuning with our glove-to-hand data significantly improves tracking accuracy, particularly in occluded cases (19.2 mm $\rightarrow$ 16.6 mm). Conversely, naive training on raw glove data degrades performance (19.5 mm $\rightarrow$ 26.5 mm), confirming a large domain gap. These results show that our framework can leverage IMU sensing to achieve robust hand pose tracking under severe occlusion.

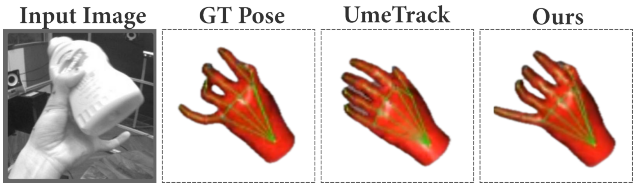

Figure 11. **Hand Tracking Visualization.** UmeTrack struggles with occlusions, our model, finetuned on our synthesized videos with IMU tracking annotation, robustly tracks the occluded hand.

Method	MKPE ↓		MKPE.T ↓	
	Occlusion	Overall	Occlusion	Overall
UmeTrack [22]	19.2	19.5	10.8	9.8
UmeTrack + Glove	27.2	26.5	11.5	11.2
UmeTrack + G2H (Ours)	**16.6**	**17.8**	**9.9**	**9.4**

Table 4. **Results on Hand Tracking.** We directly evaluate UmeTrack on HandSense and finetune it with IMU tracking results as ground-truth. Finetuning with Glove2Hand videos improves tracking accuracy over UmeTrack, especially for occluded cases as IMU sensors are robust to camera occlusion. In contrast, training on glove data degrades performance due to the domain gap.

6. Conclusion

We present the Glove2Hand framework for synthesizing photorealistic bare-hand videos from sensor-glove videos, capable of handling complex object interactions while preserving synchronization with multi-modal sensor signals. Leveraging Glove2Hand, we construct HandSense, the first multi-modal HOI dataset and demonstrate its effectiveness in significantly enhancing bare-hand applications. We believe our work enables more physically grounded and realistic analysis of hand-object interactions, benefiting broader applications in the computer vision community.

References

[1] Morph. sensel morph haptic sensing tablet. 3, 8

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, et al. Hot3d: Hand and object tracking in 3d from egocentric multi-view videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7061–7071. 2, 3, 6, 14

[3] Yuxuan Bian, Zhaoyang Zhang, Xuan Ju, Mingdeng Cao, Liangbin Xie, Ying Shan, and Qiang Xu. Videopainter: Any-length video inpainting and editing with plug-and-play context control. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 2, 6

[4] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, pages 361–378. Springer. 2, 3

[5] Johanna Bräunig, Christian Schüßler, Vanessa Wirth, Marc Stamminger, Ingrid Ullmann, and Martin Vossiek. A realistic radar ray tracing simulator for hand pose imaging. In *2023 20th European Radar Conference (EuRAD)*, pages 238–341. IEEE, 2023. 2

[6] Alessandro Carfì, Timothy Patten, Yingyi Kuang, Ali Hammoud, Mohamad Alameh, Elisa Maiettini, Abraham Itzhak Weinberg, Diego Faria, Fulvio Mastrogiovanni, Guillem Alenyà, et al. Hand-object interaction: From human demonstrations to robot manipulation. *Frontiers in Robotics and AI*, 8:714023, 2021. 1

[7] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9044–9053, 2021. 8

[8] Kefan Chen, Chaerin Min, Linguang Zhang, Shreyas Hampali, Cem Keskin, and Srinath Sridhar. Foundhand: Large-scale domain-specific learning for controllable hand image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17448–17460, . 2, 6

[9] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8683–8693, . 2, 3

[10] Ernie Chu, Tzuhsuan Huang, Shuo-Yen Lin, and Jun-Cheng Chen. Medm: Mediating image diffusion models for video-to-video translation with temporal correspondence guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1353–1361. 2

[11] Yilan Dong, Haohe Liu, Qing Wang, Jiahao Yang, Wenqing Wang, Gregory Slabaugh, and Shanxin Yuan. Handsplat: Embedding-driven gaussian splatting for high-fidelity hand rendering. *arXiv preprint arXiv:2503.14736*. 2, 3

[12] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12943–12954. 3, 5, 14

[13] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[14] Patrick Grady, Jeremy A Collins, Chengcheng Tang, Christopher D Twigg, Kunal Aneja, James Hays, and Charles C Kemp. Pressurevision++: Estimating fingertip pressure from diverse rgb images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8698–8708. 3, 8

[15] Patrick Grady, Chengcheng Tang, Samarth Brahmbhatt, Christopher D Twigg, Chengde Wan, James Hays, and Charles C Kemp. Pressurevision: estimating hand pressure from a single rgb image. In *European Conference on Computer Vision*, pages 328–345. Springer, 2022. 3, 8

[16] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 1

[17] Markus Gross and Hanspeter Pfister. *Point-based graphics*. Elsevier, 2011. 4

[18] Bohai Gu, Hao Luo, Song Guo, Peiran Dong, and Qihua Zhou. Coherent video inpainting using optical flow-guided efficient diffusion. *arXiv preprint arXiv:2412.00857*, 2024. 2

[19] Shuwei Guo, Cong Yang, Zhizhong Su, Wei Sui, Xun Liu, Minglu Zhu, and Tao Chen. From human-computer interaction to human-robot manipulation. *Engineering Proceedings*, 110(1):1, 2025. 1

[20] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. 15

[21] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3196–3206, 2020. 1, 8

[22] Shangchen Han, Po-chen Wu, Yubo Zhang, Beibei Liu, Linguang Zhang, Zheng Wang, Weiguang Si, Peizhao Zhang, Yujun Cai, Tomas Hodan, et al. Umetrack: Unified multi-view end-to-end hand tracking for vr. In *SIGGRAPH Asia 2022 conference papers*, pages 1–9, 2022. 8

[23] Shunran Hao, Dongxu Gao, Zhaojie Ju, and Qing Gao. Multimodal hand gesture recognition based on the fusion of surface electromyography and vision. In *2024 30th Interna-*

tional Conference on Mechatronics and Machine Vision in Practice (M2VIP), pages 1–6. IEEE, 2024. 1

[24] Aashni Haria, Archanasri Subramanian, Nivedhitha Asokkumar, Shristi Poddar, and Jyothi S Nayak. Hand gesture recognition for human computer interaction. *Procedia computer science*, 115:367–374, 2017. 1

[25] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[26] Ryan Hoque, Peide Huang, David J Yoon, Mouli Sivapurapu, and Jian Zhang. Egodex: Learning dexterous manipulation from large-scale egocentric video. *arXiv preprint arXiv:2505.11709*, 2025. 1

[27] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 8

[28] Zhihao Hu and Dong Xu. Videocontrolnet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*. 2

[29] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 4, 7

[30] Mingzhen Huang, Fu-Jen Chu, Bugra Tekin, Kevin J Liang, Haoyu Ma, Weiyao Wang, Xingyu Chen, Pierre Gleize, Hongfei Xue, Siwei Lyu, et al. Hoigpt: Learning long-sequence hand-object interaction with language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7136–7146, 2025. 1

[31] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134. 2, 6

[32] Shun Iwase, Shunsuke Saito, Tomas Simon, Stephen Lombardi, Timur Bagautdinov, Rohan Joshi, Fabian Prada, Takaaki Shiratori, Yaser Sheikh, and Jason Saragih. Relightablehands: Efficient neural relighting of articulated hand models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16663–16673. 2, 3

[33] Shuo Jiang, Haonan Li, Ruochen Ren, Yanmin Zhou, Zhipeng Wang, and Bin He. Kaiwu: A multimodal manipulation dataset and framework for robot learning and human-robot interaction. *arXiv preprint arXiv:2503.05231*, 2025. 3

[34] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. In *European Conference on Computer Vision*, pages 150–168. Springer, 2024. 2, 6

[35] Joanna Kaleta, Kacper Kania, Tomasz Trzciński, and Marek Kowalski. Lumigauss: Relightable gaussian splatting in the wild. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1–10. IEEE, 2025. 4

[36] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. Harp: Personalized hand reconstruction from a monocular rgb video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12802–12813. 2, 3

[37] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2

[38] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10138–10148. 3, 5, 14

[39] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting. *arXiv preprint arXiv:2501.10018*, 2025. 2, 15

[40] Yutong Li, Jieyi Zhang, Wenqiang Xu, Tutian Tang, and Cewu Lu. Fsglove: An inertial-based hand tracking system with shape-aware calibration. *arXiv preprint arXiv:2509.21242*, 2025. 2

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 5, 14

[42] 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*, pages 21013–21022. 3, 14

[43] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 7085–7093. 2, 6, 7

[44] Hao Luo, Yicheng Feng, Wanpeng Zhang, Sipeng Zheng, Ye Wang, Haoqi Yuan, Jiazheng Liu, Chaoyi Xu, Qin Jin, and Zongqing Lu. Being-h0: Vision-language-action pre-training from large-scale human videos. *arXiv preprint arXiv:2507.15597*, 2025. 1

[45] Yiyue Luo, Chao Liu, Young Joong Lee, Joseph DelPreto, Kui Wu, Michael Foshey, Daniela Rus, Tomás Palacios, Yunzhu Li, Antonio Torralba, et al. Adaptive tactile interaction transfer via digitally embroidered smart gloves. *Nature communications*, 15(1):868, 2024. 2

[46] Junyi Ma, Jingyi Xu, Xieyuanli Chen, and Hesheng Wang. Diff-ip2d: Diffusion-based hand-object interaction prediction on egocentric videos. *arXiv preprint arXiv:2405.04370*, 2024. 1

[47] Madhur Mangalam, Sanjay Oruganti, Gavin Buckingham, and Christoph W Borst. Enhancing hand-object interactions in virtual reality for precision manual tasks. *Virtual Reality*, 28(4):166, 2024. 1

[48] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf:

Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[49] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pages 548–564. Springer. 2, 3

[50] Akshay Mundra, Jiayi Wang, Marc Habermann, Christian Theobalt, Mohamed Elgharib, et al. Livehand: Real-time and photorealistic neural hand rendering. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 18035–18045. 3

[51] Supreeth Narasimhaswamy, Uttaran Bhattacharya, Xiang Chen, Ishita Dasgupta, Saayan Mitra, and Minh Hoai. Handiffuser: Text-to-image generation with realistic hand appearances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2468–2479. 2

[52] Junho Park, Kyeongbo Kong, and Suk-Ju Kang. Attention-hand: Text-driven controllable hand image generation for 3d hand reconstruction in the wild. In *European Conference on Computer Vision*, pages 329–345. Springer. 2

[53] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On aliased resizing and surprising subtleties in gan evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11410–11420, 2022. 6

[54] Anton Pelykh, Ozge Mercanoglu Sincan, and Richard Bowden. Giving a hand to diffusion models: a two-stage approach to improving conditional human image generation. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–10. IEEE. 2

[55] 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. 5, 14

[56] Lala Shakti Swarup Ray, Bo Zhou, Sungho Suh, and Paul Lukowicz. A comprehensive evaluation of marker-based, markerless methods for loose garment scenarios in varying camera configurations. *Frontiers in Computer Science*, 6: 1379925, 2024. 5

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 5

[58] 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), 2017. 2

[59] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 8

[60] Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3352–3360. IEEE, 2025. 1

[61] Zhoutao Sun, Xukun Shen, Yong Hu, Yuyou Zhong, and Xueyang Zhou. Jghand: Joint-driven animatable hand avater via 3d gaussian splatting. *arXiv preprint arXiv:2501.19088*. 3

[62] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 6

[63] Michael F Vallaro. Markerless motion capture for dual-handed teleportation of industrial robots. Master's thesis, University of Rhode Island, 2024. 5

[64] 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. *arXiv preprint arXiv:2406.06843*. 3, 14

[65] Xinyi Wang, Pengfei Ren, Haoyang Zhang, Xin Sheng, Da Li, Liang Xie, Yue Gao, and Erwei Yin. Vihand: Enhancing 3d hand pose estimation with visual-inertial benchmark. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12753–12760, 2025. 2

[66] Sirui Xu, Zhengyuan Li, Yu-Xiong Wang, and Liang-Yan Gui. Interdiff: Generating 3d human-object interactions with physics-informed diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14928–14940. 2

[67] Wenqiang Xu, Zhenjun Yu, Han Xue, Ruolin Ye, Siqiong Yao, and Cewu Lu. Visual-tactile sensing for in-hand object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8803–8812, 2023. 3

[68] Yue Xu, Yong-Lu Li, Zhemin Huang, Michael Xu Liu, Cewu Lu, Yu-Wing Tai, and Chi-Keung Tang. Egopca: A new framework for egocentric hand-object interaction understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5273–5284, 2023. 1

[69] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Rerender a video: Zero-shot text-guided video-to-video translation. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, . 2

[70] Shuai Yang, Yifan Zhou, Ziwei Liu, and Chen Change Loy. Fresco: Spatial-temporal correspondence for zero-shot video translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8703–8712, . 2

[71] Vickie Ye, Ruilong Li, Justin Kerr, Matias Turkulainen, Brent Yi, Zhuoyang Pan, Otto Seiskari, Jianbo Ye, Jeffrey Hu, Matthew Tancik, and Angjoo Kanazawa. gsplat: An open-source library for gaussian splatting. *Journal of Machine Learning Research*, 26(34):1–17, 2025. 14

[72] Yufei Ye, Xueting Li, Abhinav Gupta, Shalini De Mello, Stan Birchfield, Jiaming Song, Shubham Tulsiani, and Sifei Liu. Affordance diffusion: Synthesizing hand-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22479–22489. 2

[73] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything:

Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 2, 6

[74] Haozhuo Zhang, Bin Zhu, Yu Cao, and Yanbin Hao. Hand1000: Generating realistic hands from text with only 1,000 images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9905–9913, 2025. 2

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 5

[76] Mengqi Zhang, Yang Fu, Zheng Ding, Sifei Liu, Zhuowen Tu, and Xiaolong Wang. Hoidiffusion: Generating realistic 3d hand-object interaction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8521–8531. 2

[77] Xinyu Zhang, Haonan Chang, Yuhan Liu, and Abdeslam Boularias. Motion blender gaussian splatting for dynamic scene reconstruction. 2025. 4

[78] Lizhi Zhao, Xuequan Lu, Runze Fan, Sio Kei Im, and Lili Wang. Gaussianhand: Real-time 3d gaussian rendering for hand avatar animation. *IEEE Transactions on Visualization and Computer Graphics*. 2, 3

[79] Yiming Zhao, Taein Kwon, Paul Streli, Marc Pollefeys, and Christian Holz. Egopressure: A dataset for hand pressure and pose estimation in egocentric vision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27727–27738, 2025. 3

[80] Xiaozheng Zheng, Chao Wen, Zhuo Su, Zeran Xu, Zhaohu Li, Yang Zhao, and Zhou Xue. Ohta: One-shot hand avatar via data-driven implicit priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 799–810. 2

[81] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy. Propainter: Improving propagation and transformer for video inpainting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10477–10486, 2023. 2, 3, 5

[82] Jiahang Zhu, Aiguo Song, Ke Shang, and Zhikai Wei. A wearable smart glove for tactile interaction. In *Proceedings of the 2024 4th International Conference on Robotics and Control Engineering*, pages 160–164, 2024. 2

[83] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232. 2, 6

Appendix

In this appendix, we first introduce our sensing glove configuration. Next, we outline the data collection protocol for the *HandSense* dataset. We then detail the our user study process. Finally, we provide implementation details regarding the training and inference of the Glove2Hand framework. Fig. 15 provides additional qualitative comparison with existing work. Please refer to the supplementary video `glove2hand-visualization.mp4` for visualizations of synthesized hand-object interactions.

A. Sensing Glove Configuration

Although Glove2Hand is general to handle various sensor glove design, the glove chosen in this paper is our research platform designed for advanced hand sensing, tracking, and haptic feedback in AR/VR/MR applications. It features 12 IMUs for detailed hand pose estimation and 5 capacitive tactile sensors on the fingertips for touch and pressure detection, supporting microgestures and interactions with physical objects. The glove streams sensor data at high framerates suitable for real-time gesture recognition and hand tracking. More detailed information about the glove configuration will be shared upon acceptance of the paper to comply with the anonymity requirements of the double-blind review.

B. HandSense Data Collection

We collect data spanning six hand-object interaction tasks, summarized in Table 5.

The object set includes ten items: a mouse, phone, soda can, marker pen, piano key, squishy toy, cube, mug, table surface, and mustard bottle. For each subject, we record paired sessions—one wearing the tactile glove and one with a bare hand—using identical object configurations and task instructions. Data acquisition consists of 2–5 minute segments per task. Within each segment, subjects perform repeated trials while varying finger usage and grasp types. Prior to recording, retro-reflective markers are affixed to the dorsal surface of the subject's hand (or glove) to enable ground truth pose tracking via an optical motion capture system. We compare HandSense with other datasets with contact label in Tab. 6.

Motion Retarget from Gloves to Bare Hands. Although glove and human hands share the same kinematic structure, the glove is bulkier than the bare hand. Directly using the glove pose as a hand pose results in unrealistic glove2hand synthesis. For example, an OK sign collected from the glove, when synthesized to a bare hand, would show the index finger and thumb not touching due to the glove's surface thickness. We treat this as a per-joint displacement. To mitigate this, at the beginning of glove data collection, we ask each candidate to perform a set of standard hand poses, e.g., pinch, OK sign, while wearing the glove. We then optimize the joint displacements and apply them to each recorded glove hand pose. This process is analogous to motion retargeting in robotics where human motion is transferred to robot, but simpler because the gap between glove and bare hand is smaller.

C. Human Evaluation Details

We recruited five participants to evaluate the perceptual quality of our synthesized imagery. Each participant completed a 45-minute session, evaluating approximately 40 images and 40 videos. The evaluation protocol uses a 5-point Likert scale (with 1.0 intervals) to assess hand realism, hand-object interaction (HOI) realism, motion stability, identity consistency, and visual artifacts. The complete questionnaire is detailed in Fig. 14. Representative samples for different rating categories are shown in Fig. 12. Our user study was conducted under a protocol approved by the Institutional Review Board (IRB). All participants provided informed consent.

Evaluation Protocols. We evaluate Glove2Hand across three distinct scenarios to assess generalization capabilities:

1. **In-Domain:** Glove and bare hand identities match (Subject A's glove $\rightarrow$ Subject A's bare hand). The subject, background, and objects were seen during training. This setup validates the method's capacity for controlled data generation.
2. **Cross-Subject:** Glove and bare hand identities differ (Subject A's glove $\rightarrow$ Subject B's bare hand). The environment is seen, but the target hand morphology is synthesized from a different source subject.
3. **In-the-Wild:** A fully unseen setting where subjects, objects, and backgrounds were not present in the training set. This represents the most challenging scenario for scalable data collection.

During evaluation, samples from these three groups are randomly interleaved with real ground-truth data in a blind study design. This establishes a high-quality reference anchor (upper bound) for the ratings.

Results. Table 7 reports the Mean Opinion Score (MOS) and the perceptual gap (difference) between synthesized and real data. Higher MOS indicates better quality, while a lower gap indicates higher fidelity to the ground truth. We observe that Glove2Hand achieves high fidelity in controlled settings (In-Domain and Cross-Subject), making it suitable for large-scale data curation. While performance degrades in the challenging In-the-Wild setting, the results remain respectable. We hypothesize that scaling the training dataset size and improving HOI segmentation masks will further bridge the domain gap.

Table 5. List of interaction tasks and exemplar instructions used during data collection.

Task	Exemplar Instruction
1. Bottle Opening	"Open the mustard bottle using only your index finger and thumb."
2. Object Rotation	"Hold the object with index, thumb, and middle fingers; rotate it in front of you."
3. Piano Keystroke	"Press different piano keys sequentially using only your ring finger."
4. Pick-and-Place	"Pick up the object using a whole-hand grasp and place it on the table."
5. Surface Interaction	"Press and slide your index finger firmly against the table surface."
6. In-Hand Rotation	"Rotate the object within your hand, maximizing finger contact and occlusion."

Dataset	Images	Subjects/Objects	Pose	Contact
H2O [38]	572K	4 / 8	Optim.	Estim.
ARCTIC [12]	2.1M	10 / 11	MoCap	Estim.
HO-Cap [64]	699K	9 / 64	Optim.	-
HOI4D [42]	2.4M	4 / 800	Manual	-
HOT3D [2]	3.7M	19 / 33	MoCap	
HandSense (Ours)	200K	5 / 10	MoCap	Measured.

Table 6. **Datasets with Contact Labels.** HandSense is the first HOI video dataset to provide direct, sensor-measured contact.

D. Glove2Hand Details

HOI Segmentation Masks. To generate segmentation masks for hands and interacting objects, we implement a pipeline leveraging Grounding DINO [41] and SAM-2 [55]. First, we detect potential objects using Grounding DINO. To identify the specific object being manipulated, we compute the Intersection-over-Union (IoU) between the detected object bounding boxes and the rasterized projection of the fitted hand mesh. The object with the highest IoU (surpassing a valid threshold) is selected as the box prompt for SAM-2. For hand segmentation, we detect the full arm using Grounding DINO and use the resulting box to prompt SAM-2. To ensure temporal consistency, we initialize SAM-2 with prompts on a single reference frame and propagate the masks. Finally, the specific hand or glove mask is obtained by cropping the full arm mask using the bounding box of the projected hand mesh.

Pose Optimization. While we utilize an optical motion capture system, the ground truth hand pose $\mathbf{P}$ may exhibit inaccuracies due to marker occlusion or synchronization latency during rapid motion. To mitigate this, we introduce a learnable per-frame pose refinement term $\Delta\mathbf{P}$. This offset is optimized jointly with the Gaussian parameters during the reconstruction phase, following the camera pose optimization strategy in gsplat [71]. We empirically find this refinement significantly reduces artifacts in the reconstructed Gaussian hand. Note that $\Delta\mathbf{P}$ is not used during the subsequent training of the diffusion restorer.

Gaussian Parameterization. We anchor the 3D Gaussians to the mesh surface using barycentric coordinates. During optimization, we learn the unnormalized barycentric logits rather than the weights directly to ensure valid constraints. Additionally, we learn a scalar offset along the surface normal. This offset is parameterized via a sigmoid activation scaled by a hyper-parameter $z_{\max}$, ensuring the Gaussians remain tightly grounded to the underlying geometry. Each subject-specific Gaussian model is trained on approximately 10 minutes of egocentric hand-only videos.

Auxiliary Training Data (HOT3D). We incorporate the HOT3D dataset [2] to augment the training of the Diffusion Hand Restorer. Unlike our primary subjects, we do not train 3D Gaussian models for HOT3D sequences. Instead, we employ a 2D self-supervised strategy: we crop out the hand region with a dilated (larger) hand mask, and mask out the wrist region before overlaying the original hand pixels onto the background. This creates videos of missing wrist details and hand-object boundary. The diffusion model is then trained to restore these corrupted regions (i.e., inpainting the hand-object boundary and wrist), allowing us to leverage large-scale data without expensive 3D reconstruction.

Training and Inference Efficiency. All models are trained on NVIDIA A100 (80GB) GPUs. We crop hand regions from the raw headset footage at a resolution of 250×250. These crops are upsampled to 512×512 to satisfy the input constraints of the diffusion restorer's VAE. Furthermore, prior to training the 3D Gaussian hand model, we rectify the images and camera parameters to convert the raw fisheye distortion into a standard pinhole camera model. For the 3D Gaussian hand, we train each subject-specific model

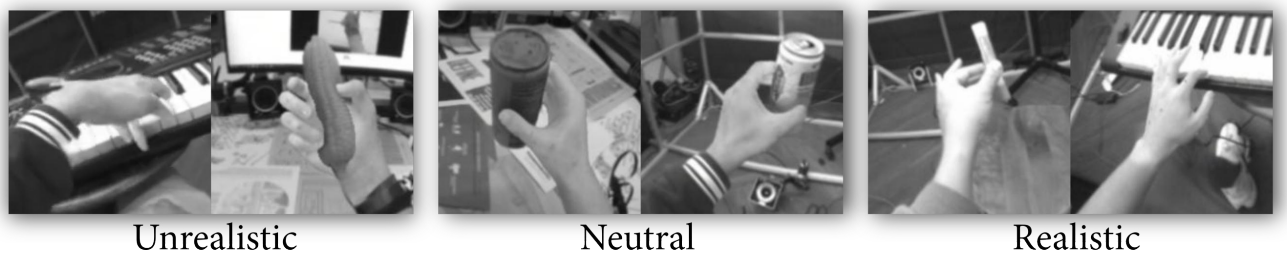

Figure 12. **Qualitative Examples of User Ratings.** Samples rated as "Realistic" or "Very Realistic" are perceptually indistinguishable from real hands. "Neutral" samples generally exhibit valid geometry and interaction plausibility but may lack skin details. "Unrealistic" samples typically display unnatural hand-object boundary or non-negligible visual artifacts.

Metric		**Mean Opinion Score ↑ / Gap to Real ↓**		
		In-Domain	**Cross-Subject**	**In-the-Wild**
Image	Hand Realism	**4.04 / 0.02**	3.67 / 0.39	2.68 / 1.37
	HOI Realism	**4.06 / 0.01**	3.51 / 0.56	2.61 / 1.46
Video	Hand Realism	**3.83 / 0.51**	3.65 / 0.69	2.69 / 1.65
	HOI Realism	**3.96 / 0.35**	3.78 / 0.53	2.84 / 1.47
	Motion Stability	**3.64 / 0.50**	3.37 / 0.77	2.54 / 1.60
	Identity Consistency	**2.70 / 0.26**	2.54 / 0.43	2.20 / 0.79
	Visual Artifacts	**3.22 / 0.51**	3.00 / 0.73	2.19 / 1.53

Table 7. **Human Evaluation Results.** We report the Mean Opinion Score (MOS) and the perceptual gap relative to real data (Gap). **In-Domain** synthesis achieves performance near ground truth (Gap < 0.05 for images). **Cross-Subject** synthesis maintains high efficacy with gaps consistently below 1.0. **In-the-Wild** performance reflects the expected challenge of unseen environments but remains within a reasonable qualitative range.

for 120k iterations ($\sim$12 hours), though varying the schedule shows convergence at $\sim$6 hours. Rendering speed is approximately 50 FPS without custom CUDA kernel optimization. For the diffusion hand restorer, training proceeds in two stages. First, the image-based restorer is trained for 60k iterations. Second, we insert AnimateDiff [20] motion adapters and fine-tune on 22-frame video clips for an additional 60k iterations. The total training time is approximately 72 hours. For long-video generation, we apply the temporal sliding window strategy from DiffuEraser [39] to ensure consistency. The inference speed is approximately 0.5 FPS.

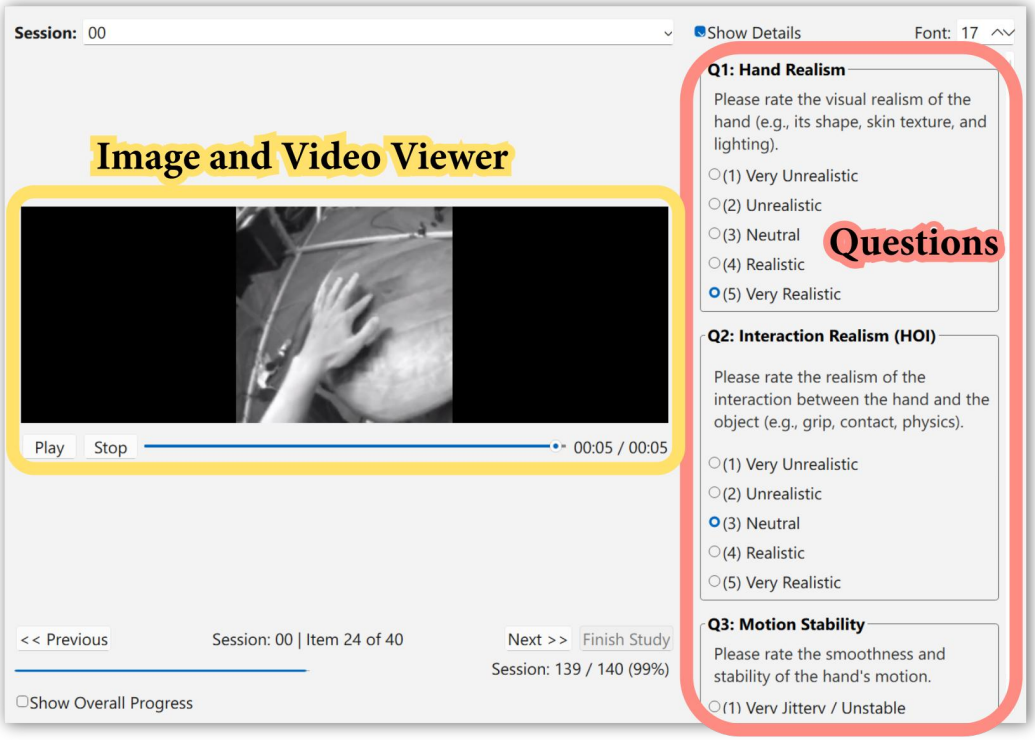

Figure 13. **Human Evaluation Interface.** Participants view a randomly sampled image or video and rate it against specific criteria before proceeding.

User Study Questionnaire

Per-Image Assessment

Q1: Hand Realism: Rate the visual realism of the hand (shape, skin texture, lighting). (1) Very Unrealistic (2) Unrealistic (3) Neutral (4) Realistic (5) Very Realistic

Q2: HOI Realism: Rate the plausibility of the interaction (grip, contact, physics). (1) Very Unrealistic (2) Unrealistic (3) Neutral (4) Realistic (5) Very Realistic

---

Per-Video Assessment

Q3: Hand Realism: Rate the visual realism of the hand. (1) Very Unrealistic (2) Unrealistic (3) Neutral (4) Realistic (5) Very Realistic

Q4: HOI Realism: Rate the plausibility of the interaction. (1) Very Unrealistic (2) Unrealistic (3) Neutral (4) Realistic (5) Very Realistic

Q5: Motion Stability: Rate the temporal smoothness of the hand motion. (1) Very Unstable (2) Unstable (3) Neutral (4) Stable (5) Very Stable

Q6: Identity Consistency: How consistent is the hand's appearance (shape/size) over time?
(1) Significant unnatural changes.
(2) Slight unnatural changes.
(3) Consistent appearance.
(4) (Unsure)

Q7: Visual Artifacts: Are there distracting visual artifacts (blur, flickering, texture issues)?
(1) Frequent/Severe artifacts.
(2) Noticeable artifacts.
(3) Minor artifacts.
(4) No artifacts observed.

Figure 14. **Human Evaluation Questionnaire.**

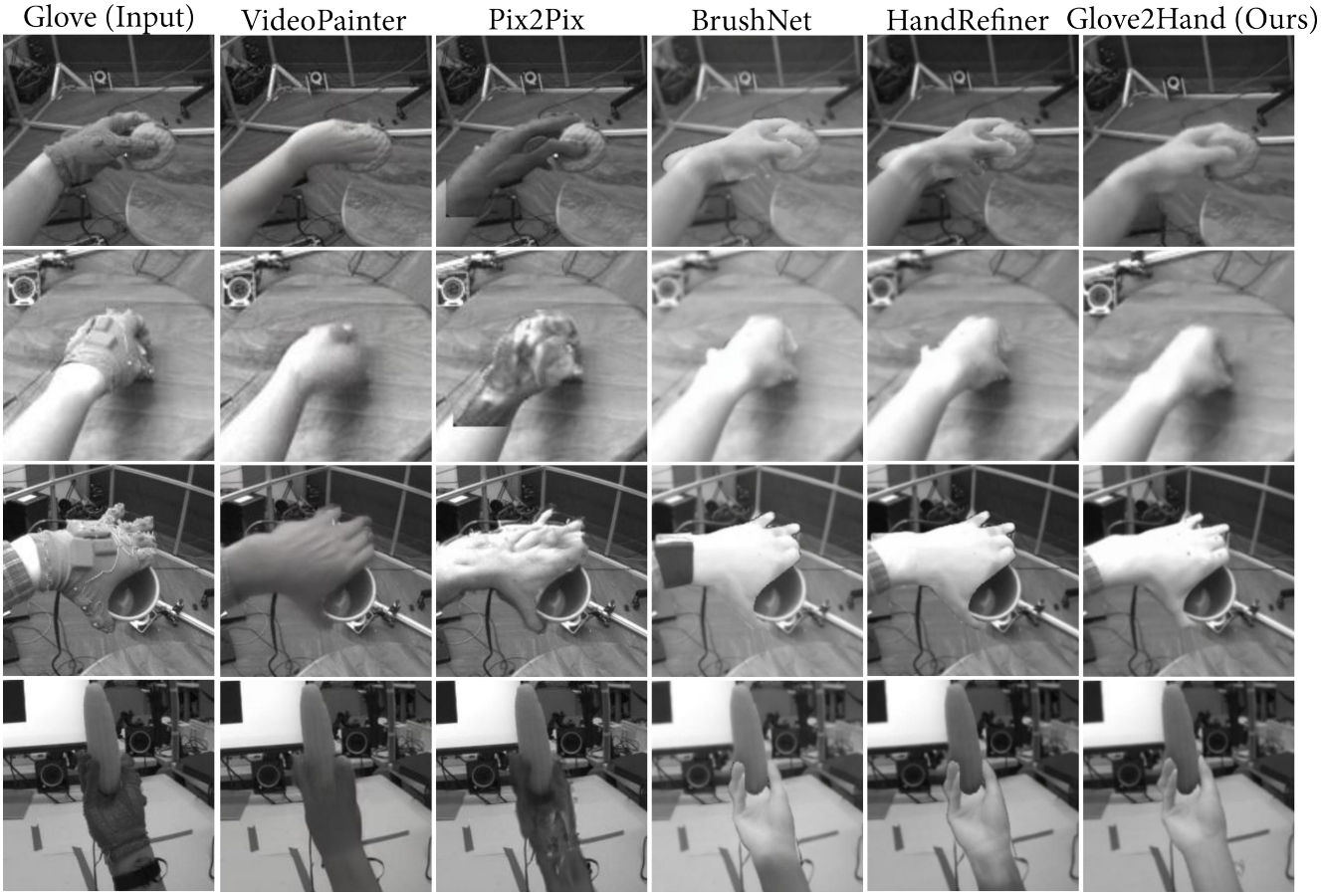

Figure 15. **Additional Qualitative Comparison for Glove-to-Hand.**