Supplementary Material for "Glove2Hand: Synthesizing Natural Hand-Object Interaction from Multi-Modal Sensing Gloves"

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. 4 provides additional qualitative comparison with existing work. Please refer to the supplementary video `glove2hand-visualization.mp4` for visualizations of synthesized hand-object interactions.

1. 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.

2. HandSense Data Collection

We collect data spanning six hand-object interaction tasks, summarized in Table 1.

Table 1. 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."

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. 2.

3. 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

Dataset	Images	Subjects/Objects	Pose	Contact
H2O [4]	572K	4 / 8	Optim.	Estim.
ARCTIC [2]	2.1M	10 / 11	MoCap	Estim.
HO-Cap [9]	699K	9 / 64	Optim.	-
HOI4D [7]	2.4M	4 / 800	Manual	-
HOT3D [1]	3.7M	19 / 33	MoCap	
HandSense (Ours)	200K	5 / 10	MoCap	Measured.

Table 2. **Datasets with Contact Labels.** HandSense is the first HOI video dataset to provide direct, sensor-measured contact.

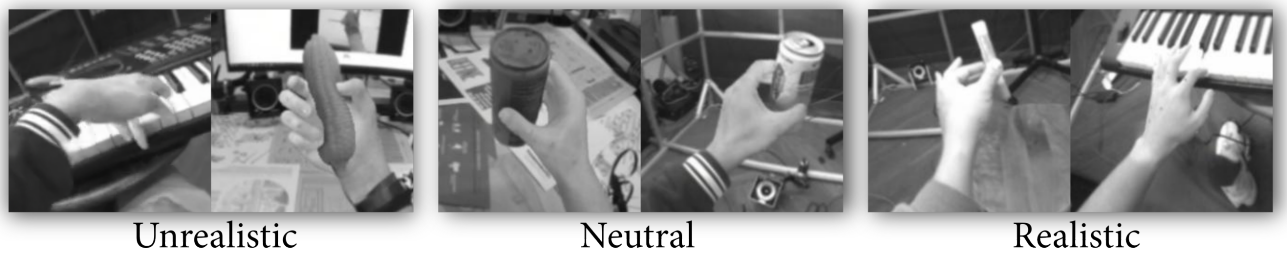

Figure 1. **Qualitative Examples of User Ratings.** Samples rated as "Realistic" or "Very Realistic" are perceptually indistinguishable from real hands. "Neutral" samples generally exhibit valid geometry and interaction plausibility but may lack skin details. "Unrealistic" samples typically display unnatural hand-object boundary or non-negligible visual artifacts.

(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. 3. Representative samples for different rating categories are shown in Fig. 1. 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 → 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 → 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 3 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.

4. Glove2Hand Details

HOI Segmentation Masks. To generate segmentation masks for hands and interacting objects, we implement a pipeline leveraging Grounding DINO [6] and SAM-2 [8]. 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.

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 3. **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.

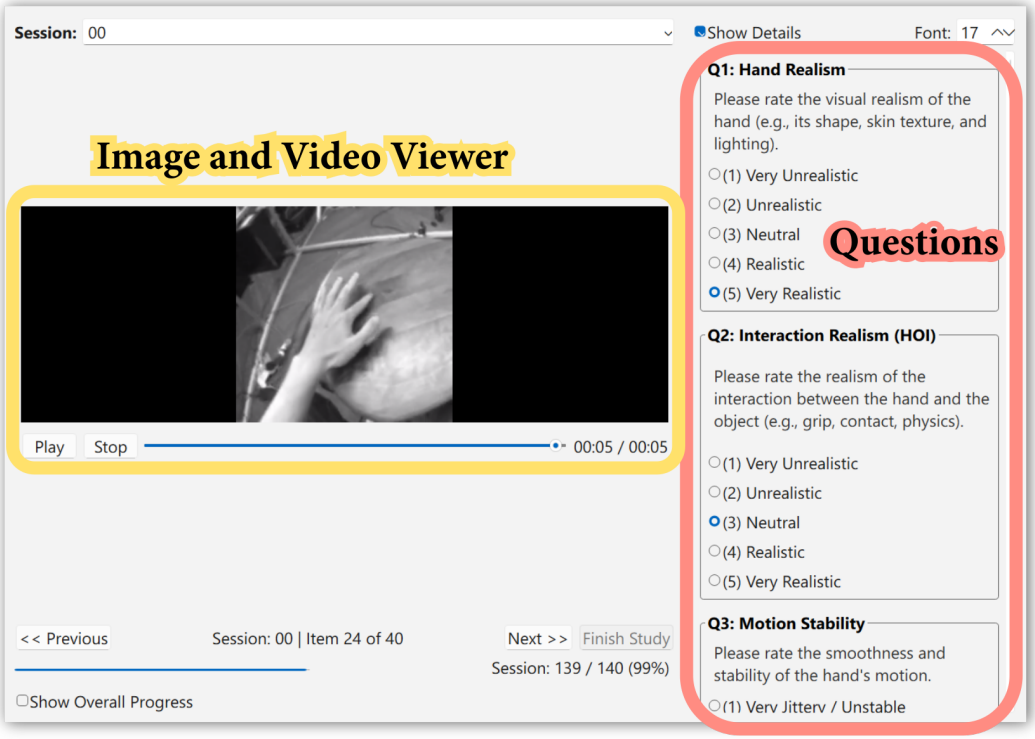

Figure 2. **Human Evaluation Interface.** Participants view a randomly sampled image or video and rate it against specific criteria before proceeding.

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 [10]. 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

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 3. **Human Evaluation Questionnaire.**

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 [1] 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 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 [3] 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 [5] to ensure consistency. The inference speed is approximately 0.5 FPS.

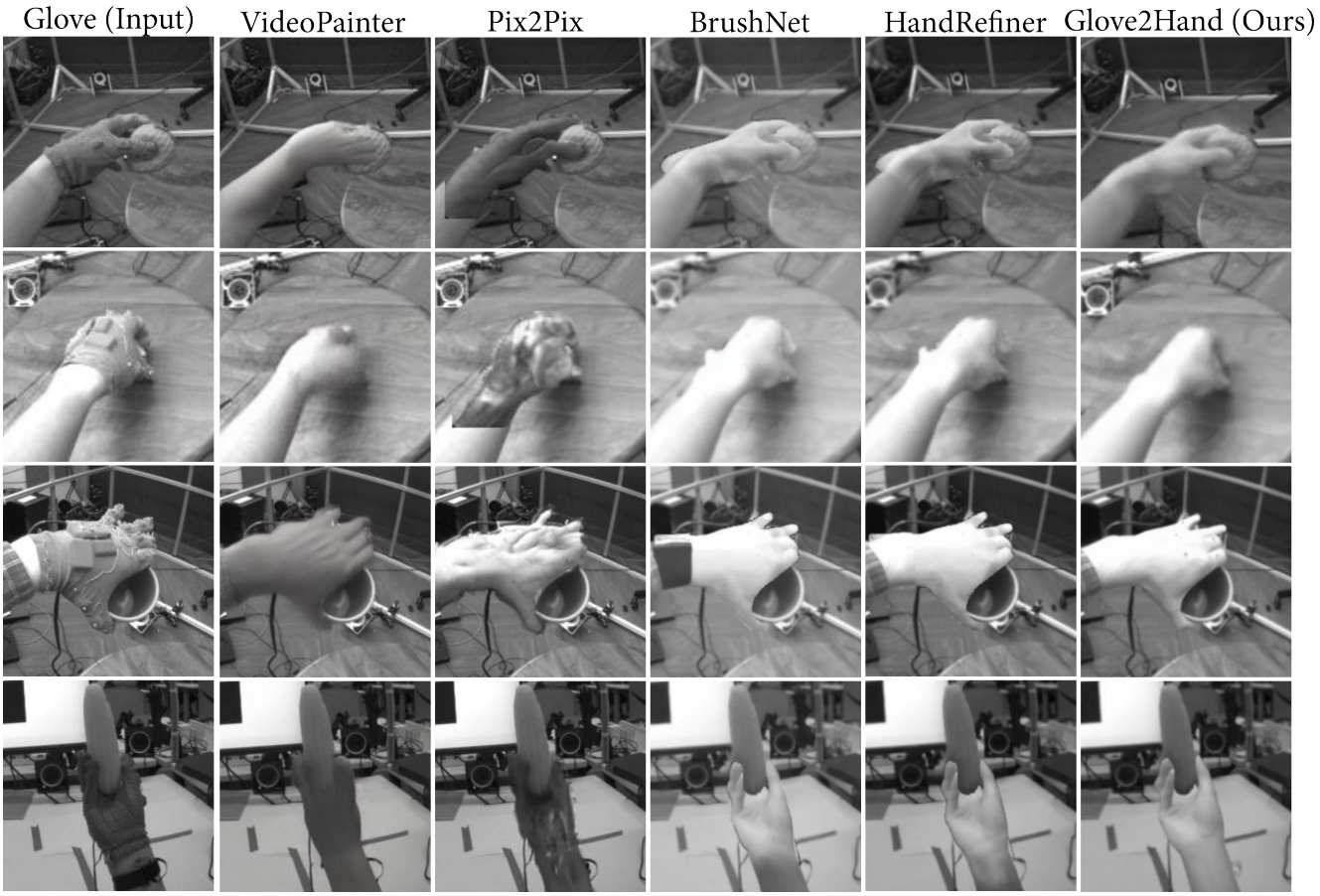

Figure 4. **Additional Qualitative Comparison for Glove-to-Hand.**

References

[1] 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, 4

[2] 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. 2

[3] 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. 4

[4] 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. 2

[5] Xiaowen Li, Haolan Xue, Peiran Ren, and Liefeng Bo. Diffueraser: A diffusion model for video inpainting. *arXiv preprint arXiv:2501.10018*, 2025. 4

[6] 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. 2

[7] 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. 2

[8] 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. 2

[9] 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*. 2

[10] 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. 3