Supplementary Material
RHINO: Reconstructing Human Interactions with Novel Objects from Monocular Videos

Lixin Xue[1] Chengwei Zheng[1,2] Georgios Paschalidis[3]
Chen Guo[1] Manuel Kaufmann[1] Juan Zarate[1] Dimitrios Tzionas[3,4]

[1]ETH Zürich [2]The University of Tokyo [3]University of Amsterdam [4]Aristotle University of Thessaloniki

We discuss data preprocessing (Sec. S.1) and implementation details (Sec. S.2), and present additional experimental results (Sec. S.3), including failure cases (Sec. S.3.4). Last, we discuss the possible negative societal impact (Sec. S.4).

S.1. Data Preprocessing

We adopt the data preprocessing framework of HSR [13] to register human bodies and scene cameras into a common world frame. We extend this by adding components for: (i) object pose estimation, (ii) camera-object motion disentanglement, and (iii) contact estimation. The motion disentanglement part has been detailed in the main paper (Sec. 3.2). Here we provide more details regarding the other two parts.

Implementation Details. For motion disentanglement (Sec. 3.2), RANSAC automatically identifies which frames have a stationary object, without assuming that any particular frames (*e.g.*, the first few ones) are static. For human initialization, the weak-perspective camera used for the initial body estimation via AiOS [11] applies *only* to that initialization step; we subsequently recover the body trajectory under a full perspective camera model by minimizing a 2D keypoint reprojection error.

S.1.1. 3D Object Pose Estimation

We detect object masks via Grounded-SAM2 [7, 8] using text prompts describing the object class. To extract features only on object pixels, we suppress the background by replacing non-object pixels with either black or white values, dynamically chosen to maximize contrast against the object's original color. These masked images are then processed by MASt3R [5] to extract feature matches. We compare our matching strategy against baselines in Fig. S.1. As shown in the first row, SuperPoint [1] keypoints exhibit poor repeatability, even on identical frames. Similarly, LoFTR [10] yields incorrect matches outside object boundaries due to its coarse-to-fine matching strategy. In contrast, MASt3R yields dense and precise matches, robustifying pose estimation, and by extension shape reconstruction.

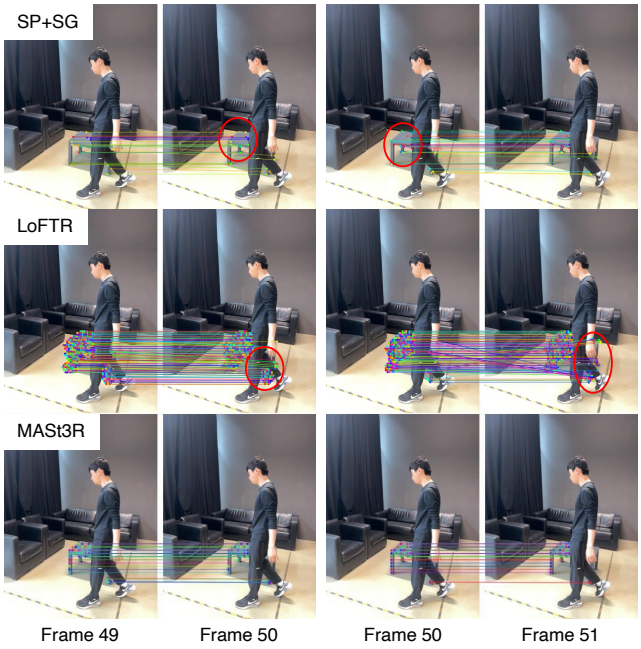

Figure S.1. **Comparison of feature matches** on object pixels across consecutive frames. We compare our `MASt3R`-based [5] strategy against established baselines. Red circles indicate failure modes in baselines: sparse, inconsistent keypoints (`SP+SG` [1, 9]) and incorrect background matches (`LoFTR` [10]).

S.1.2. 3D Contact Estimation

We detect 3D contact points on the human via InteractVLM [2], a SotA image-based model. However, InteractVLM often struggles. First, it operates on individual frames, so there is jitter in detected 3D contact points, while neighboring frames are inconsistently labeled as "contact" or "non-contact" ones. Second, there might be false positives; the model can detect contact when there is none in the image.

We mitigate these errors via a motion-based strategy. Our key insight is that an object typically moves only when manipulated by a human. Thus, we automatically label a frame as a "contact frame" if the temporal change in 3D

object pose exceeds a threshold, else we label this as "non-contact frame." However, noisy object pose estimates can cause these labels to fluctuate across frames. We tackle this by applying a temporal filter that removes short label spans by flipping them to match the surrounding label, and then extends each contact region by a margin of frames to account for the contact-but-stationary phase (*i.e.*, when the human is in contact with the object but not yet moving it). Then, we apply InteractVLM only for contact frames, avoiding false positives for non-contact frames.

Finally, we remove jitter in the predicted 3D contact points by averaging per-vertex contact predictions over a local temporal window and thresholding the result.

S.2. Implementation Details

S.2.1. Neural Representations

Human. The canonical human shape network, f^H_{sdf}, comprises 8 blocks, each consisting of a 256-unit fully-connected layer with weight normalization and Softplus activation. The human pose condition, $\boldsymbol{\theta}_b$, is a 69-dimensional vector formed by concatenating the axis-angle representations of all body joints. The appearance network, f^H_{rgb}, comprises 4 similar blocks but uses ReLU activation for the hidden layers and a Sigmoid for the output layer. To help convergence, we pretrain the shape network using a rough-shape loss using a SMPL-X [6] mesh in canonical pose.

Object & Scene. The shape and texture networks for the object (f^O_{sdf}, f^O_{rgb}) and the scene (f^S_{sdf}, f^S_{rgb}) share a similar architecture to the human networks. We employ geometric initialization following IGR [4]; the object canonical shape is initialized as a small sphere with outward-facing normals. Similarly, the scene geometry is initialized as a large sphere with inward-facing normals to enclose the capture volume.

S.2.2. Losses

Following HSR [13], we train our neural representations using a combination of RGB, normal, depth, and mask losses. However, we go beyond HSR by modeling dynamic human-object manipulation, which causes severe occlusions, while also requiring high-fidelity hand reconstruction. We tackle these challenges via two additional losses, L_{body} and L_{hand} described below, the effect of which is analyzed in Sec. S.3.

Body Prior Loss. Occlusions can cause the body geometry to appear truncated. To resolve this, we sample points x_{b} within the interior of the canonical SMPL-X [6] mesh, and penalize for predicted positive signed-distance values:

$$L_{\text{body}} = \gamma_1 \tanh\left(f^H_{\text{sdf}}(x_{\text{b}})/\gamma_2\right)^2 \quad \text{for } f^H_{\text{sdf}}(x_{\text{b}}) \geq 0. \quad \text{(S.1)}$$

Hand SDF Loss. The recovered hand shape can often be rough due to complex articulation and noisy pose estimation. We resolve these by supervising the human SDF in the hand region using the SMPL-X mesh as a geometric prior. Specifically, we sample 3D points x_{h} near the hands, compute their SDF values $\xi(x_{\text{h}})$ from the SMPL-X mesh, and penalize the $L1$ difference of the respective prediction:

$$L_{\text{hand}} = w(x_{\text{h}})|f^H_{\text{sdf}}(x_{\text{h}}) - \xi(x_{\text{h}})|. \quad \text{(S.2)}$$

The weight term $w(x_{\text{h}})$ attenuates supervision near the wrist region, where the SMPL-X mesh cannot model sleeves.

S.2.3. Training Details

Alternating Optimization. We employ a two-stage optimization strategy. In Stage I, we optimize for all network parameters with disabled physical losses (Eq. 12–13). In Stage II, we freeze the shape and appearance network parameters and optimize only for human and object poses with physical losses enabled. We follow Stage I exclusively for the first 25% of training iterations to reconstruct initial geometry and appearance. Subsequently, we alternate between the two stages every 10 epochs (6 epochs for Stage I and 4 epochs for Stage II). The model is trained for 100k steps, taking $\sim$18 hours on an NVIDIA RTX 4090 GPU.

Sampling Strategy. We use a weighted pixel sampling strategy, allocating 50% of samples to the human body, 30% to the object, 10% to the hand, and 10% to the scene. To help reconstruct the initial object shape, for the first 10 epochs we increase the object sampling rate to 70%, while reducing body samples to 20% and hand samples to 0%.

S.2.4. Evaluation Details

Different methods produce output in different formats (*e.g.*, meshes or point clouds, presence/absence of human or object) and operate within different coordinate frames. Thus, we tailor our evaluation protocols for fair comparisons.

Coordinate Frames. HOLD [3] and InterTrack [12] produce output in the camera frame, so post hoc we need to align it into the world frame. Instead, HSR [13] and RHINO operate in the world frame, but still require rigid alignment to account for differences in camera trajectory.

Aligning: Estimated Shapes to GT Shapes. To compute evaluation metrics, we first need to align the estimated shapes to the GT ones, as follows below for each method:

- HSR estimates clothed human meshes, and does not estimate moving objects. Therefore, we perform frame-wise alignment using the ground-truth (GT) human mesh.
- HOLD and its variant HOLD* (that uses RHINO's object pose estimates), estimate hands and a manipulated object, focusing on object reconstruction quality. Thus, we align the predicted object mesh with the GT one.
- InterTrack estimates unclothed human bodies as point clouds (SMPL vertices) rather than clothed meshes as in the GT, so direct surface alignment is infeasible. Instead, we perform Procrustes alignment on 3D joints derived from the InterTrack predictions and GT SMPL-X fits. We

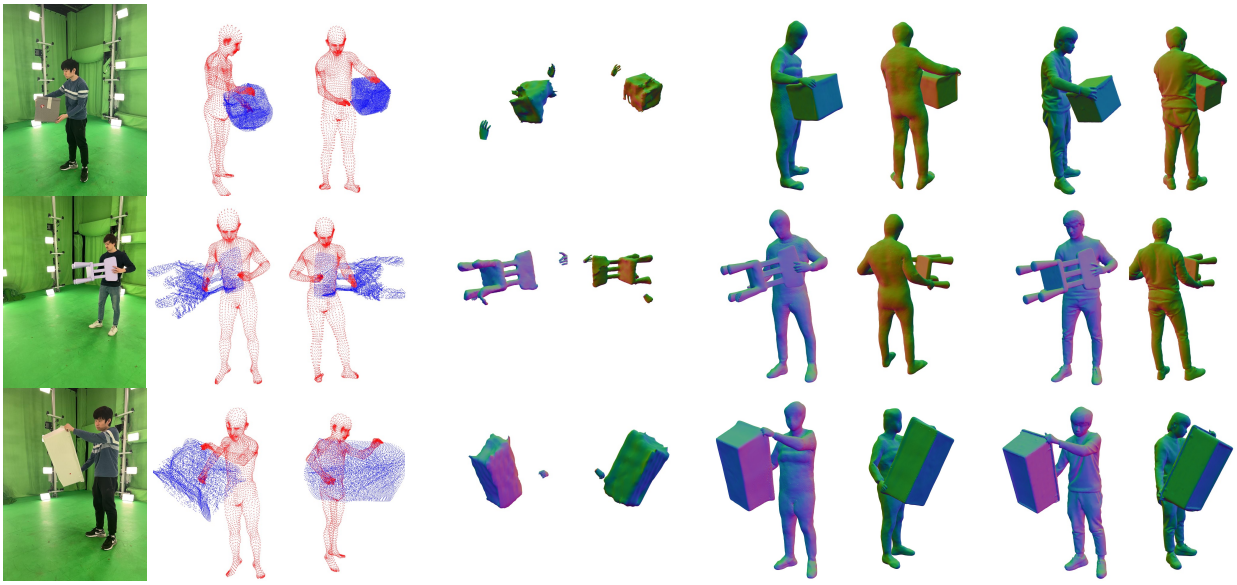

Figure S.2. **Evaluation on shape reconstruction (extending Fig. 6)** using our BenchRHINO dataset (Sec. 4.1). HOLD [3] struggles to reconstruct shapes under noisy object poses and fails to capture 3D hand-object interactions. InterTrack [12] often recovers reasonable object shape, but fails to accurately model the interaction due to errors in human and object pose or poor generalization to unseen objects. Our method (RHINO) faithfully reconstructs detailed interactions and object shapes that align closely with the ground truth (GT).

perform only human-based alignment, as InterTrack's object poses are often noisy, with wrong scale or orientation.
- RHINO estimates both the human and object, so we perform alignment based on the human, the object, or both.

Results. In Tab. 1 of the main manuscript, we report metrics after aligning the estimated shapes of each method to GT ones appropriately, as described above.

The BenchRHINO dataset contains 7 sequences: `S1_Take3`, `S1_Take7`, `S1_Take8`, `S2_Take5`, `S2_Take9`, `S3_Take6`, and `S4_Take2`. In Tab. 3, the "`SP+SG`" baseline yields plausible object shapes on only 5 of these, while "`LoFTR`" succeeds on 6. For fair comparison, we report aggregated metrics on the 5 sequences where all methods succeed: `S1_Take3`, `S1_Take7`, `S2_Take5`, `S2_Take9`, and `S3_Take6` (excluding `S1_Take8` and `S4_Take2`).

S.3. Additional Experiment Results

S.3.1. Shape Reconstruction

We provide more qualitative comparisons on shape reconstruction in Fig. S.2. The results reinforce our observations regarding the limitations of baseline methods for challenging object shape and noisy pose initialization. HOLD fails to generate coherent object meshes when the initial object pose is noisy (rows 1, 3 of Fig. S.2), while the recovered hands lie far away from the "grasped" object. InterTrack struggles with generalization, particularly with the white box (row 3). Furthermore, InterTrack misaligns the object relative to the human, failing to preserve the correct spatial configuration required for a valid grasp (rows 1, 2). In contrast, our method (RHINO) faithfully reconstructs detailed object shapes and maintains better hand-object contact.

S.3.2. Novel–View Synthesis

We evaluate qualitatively on novel view synthesis using the WildRHINO dataset. The results in Fig. S.3 show that baselines have distinct failure modes, which RHINO successfully overcomes. HOLD [3] struggles with background reconstruction (see column 2). This is because HOLD operates in camera frame, so it lacks a global understanding of scene geometry. Thus, as the camera viewpoint changes, the background appears fragmented and noisy, failing to maintain temporal or spatial consistency. HSR [13] is capable of modeling the static background, but struggles significantly with the dynamic foreground elements; the rendered images suffer from "ghost" artifacts around the object (col. 3). Note in rows 1 and 3 that, although the object appears well-reconstructed, it is baked into the static background scene rather than being correctly modeled as a dynamic entity. In contrast, our RHINO framework yields high-fidelity renderings that differ negligibly from the input observations. By effectively decoupling the dynamic foreground from the static background, it preserves the structural integrity of the object even during fast motion or complex interactions.

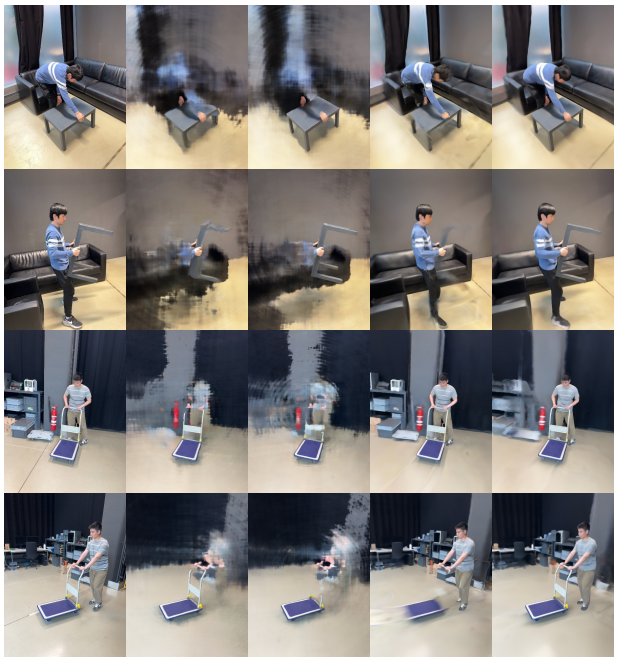

Figure S.3. **Evaluation on novel view synthesis** on WildRHINO. HOLD [3] fails to reconstruct the background scene due to operating in the camera frame. HSR [13] cannot capture dynamic manipulated objects as it assumes a static scene. Our RHINO framework decouples the dynamic foreground from the background, and yields high-fidelity renderings that closely resemble the input.

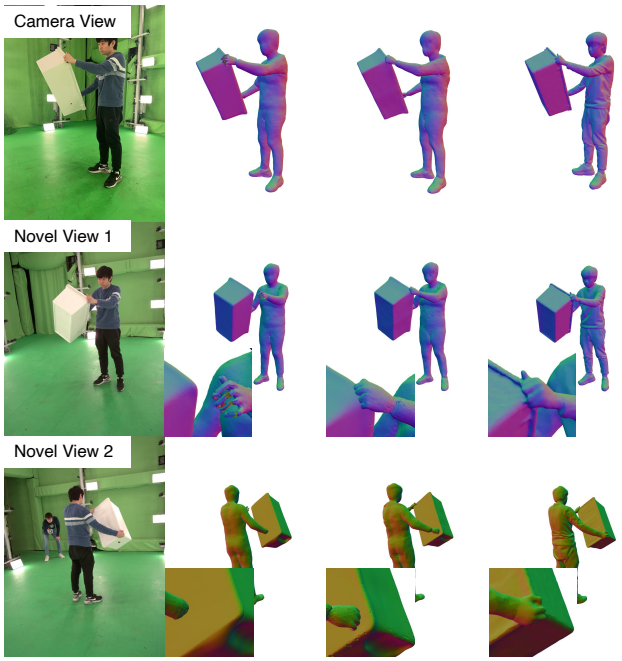

Figure S.4. **Ablation on the effect of contact** on BenchRHINO. Without contact optimization ("`w/o Contact Opt.`"), the reconstructed hands penetrate the object, or hover and fail to contact it (see zoomed-in insets). In contrast, our full method ("`RHINO (Ours)`") improves physical plausibility, recovering grasps that better align with the ground truth (``GT''').

S.3.3. Ablation on Contact

We evaluate our contact-based pose refinement in Fig. S.4. We compare our full framework ("`RHINO (Ours)`") against an ablated version where the physical contact losses (Eq. 12–13) are removed ("`w/o Contact Opt`").

Relying solely on visual cues often results in depth ambiguities and physical implausibility, particularly under heavy occlusion. While the ablated model ("`w/o Contact Opt`") may achieve good alignment in camera view, novel-view renderings have strong artifacts. Specifically, as shown in the zoomed-in insets of Fig. S.4, the hands often penetrate the object volume or hover above the surface. In contrast, our full model ("`RHINO (Ours)`") successfully resolves these ambiguities, producing tight, physically plausible grasps that closely match the ground truth ("`GT`").

S.3.4. Failure Modes

We identify the following typical failure modes of RHINO:

Insufficient Object Coverage. When the object is observed from a limited range of viewpoints (*e.g.*, always from the front), the SfM-based pose initialization may lead to incomplete or inaccurate object shape on unseen sides.

Rapid Motion. Rapid object or hand motion can cause severe motion blur and large inter-frame displacements, making feature matching unreliable and causing the motion decomposition (Eq. 4) yield noisy object trajectories.

Extreme Occlusion. When the object is almost entirely occluded by the body for many consecutive frames, the object's appearance and pose are under-constrained. The neural SDF may hallucinate geometry in unobserved regions.

Non-Rigid Objects. We assume rigid objects. For deformable objects (*e.g.*, cloth, soft toys), the rigid-body motion model is insufficient, causing reconstruction artifacts.

The above inform future work, toward handling articulated or deformable objects, leveraging temporal priors for rapid motions, and generative shape priors for occlusions.

S.4. Societal Impact

RHINO converts humans and objects into digital forms from a single RGB video, unlocking vast possibilities for augmented and virtual reality, assistive robotics, and learning from internet-scale videos. Our framework generates HOIs that can be animated to previously unseen poses.

Risks & Mitigation. The above capability carries an inherent risk of misuse, most notably for creating deep-fakes. It is crucial to address such concerns before this technology is integrated into products. We are committed to fostering applications that provide a clear benefit to society. While

we cannot eliminate the possibility of malicious use, we advocate for a policy of maximum transparency. To this end, openly discussing our methods and sharing both code and data is not only an ethical imperative but also a practical way to enable the development of countermeasures, thus mitigating the dangers of harmful applications.

References

[1] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-supervised interest point detection and description. In *Computer Vision and Pattern Recognition Workshops (CVPRw)*, 2018. 1

[2] Sai Kumar Dwivedi, Dimitrije Antić, Shashank Tripathi, Omid Taheri, Cordelia Schmid, Michael J. Black, and Dimitrios Tzionas. InteractVLM: 3D interaction reasoning from 2D foundational models. In *Computer Vision and Pattern Recognition (CVPR)*, 2025. 1

[3] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3D reconstruction of interacting hands and objects from video. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 4

[4] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *International Conference on Machine Learning (ICML)*, 2020. 2

[5] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3D with MASt3R. In *European Conference on Computer Vision (ECCV)*, 2024. 1

[6] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[7] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, and Ross Girshick et. el. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025. 1

[8] Tianhe Ren, Qing Jiang, Shilong Liu, Zhaoyang Zeng, Wenlong Liu, Han Gao, Hongjie Huang, Zhengyu Ma, Xiaoke Jiang, Yihao Chen, Yuda Xiong, Hao Zhang, Feng Li, Peijun Tang, Kent Yu, and Lei Zhanget. Grounding DINO 1.5: Advance the "edge" of open-set object detection. arXiv:2405.10300, 2024. 1

[9] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[10] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[11] Qingping Sun, Yanjun Wang, Ailing Zeng, Wanqi Yin, Chen Wei, Wenjia Wang, Haiyi Mei, Chi-Sing Leung, Ziwei Liu, Lei Yang, and Zhongang Cai. AiOS: All-in-one-stage expressive human pose and shape estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[12] Xianghui Xie, Jan Eric Lenssen, and Gerard Pons-Moll. InterTrack: Tracking human object interaction without object templates. In *International Conference on 3D Vision (3DV)*, 2025. 2, 3

[13] Lixin Xue, Chen Guo, Chengwei Zheng, Fangjinhua Wang, Tianjian Jiang, Hsuan-I Ho, Manuel Kaufmann, Jie Song, and Hilliges Otmar. HSR: Holistic 3D human-scene reconstruction from monocular videos. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3, 4