INTRINSIC ALIGNMENT BY GENERATION: MONOCULAR HAND–OBJECT RECONSTRUCTION VIA INTERACTION SCAFFOLDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Monocular hand–object reconstruction ultimately requires separate hand and object assets, but their relative pose and scale are inherently joint geometric quantities. This suggests the interaction is most naturally established in a shared geometry. Existing methods either incorporate hand geometry or interaction constraints into learned reconstruction while making limited use of object-centric image-to-3D priors, or exploit pretrained image-to-3D priors while deferring hand–object placement to test-time alignment optimization. JIGAR (Joint Interaction Generation and Asset Recovery) bridges this gap: its Joint Interaction Generation (JIG) stage adapts a pretrained image-to-3D model to jointly generate hand–object geometry as an interaction scaffold, which then guides separate asset recovery through in-place object regeneration and parametric hand recovery. This preserves the hand–object configuration while allowing the prior to refine object geometry. In this way, JIGAR combines pretrained generative shape priors with intrinsic hand–object alignment without requiring test-time optimization of the hand–object relative placement. Evaluations on SHOWMe, HOT3D, and HO-Cap show that JIGAR outperforms all evaluated baselines in object and interaction reconstruction, reducing hand-aligned object Chamfer distance by 42.3–63.2% relative to the strongest evaluated baseline across the three datasets. **The code is included in the supplementary material and will be open-sourced upon acceptance.**

1 INTRODUCTION

Reconstructing a hand together with the object it manipulates is important for understanding and reproducing human–object interactions, with applications in embodied perception, robot learning, and augmented and mixed reality. We study the challenging monocular setting: given a single RGB image, the goal is to recover separate 3D hand and object assets together with their relative pose and scale. This problem is inherently ambiguous due to missing depth information and severe mutual occlusion. More importantly, successful reconstruction requires not only plausible geometry for each individual asset, but also a coherent spatial configuration between them. This reveals a distinctive structure of the problem: the desired outputs are separate assets, whereas the interaction that relates them is inherently joint.

Recent monocular hand estimators and pretrained image-to-3D models provide pretrained priors for recovering the hand and object individually. Yet plausible individual assets do not by themselves determine their relative configuration. Existing methods address this missing relation through two broad approaches (Fig. 1(a,b)). Hand-aware Reconstruction incorporates hand geometry or interaction constraints to establish the relation during learned prediction (Ye et al., 2022; Chen et al., 2022; 2023; 2025c). Optimization-based Hand-Object Alignment, exemplified by EasyHOI (Liu et al., 2025) and FollowMyHold (Aytekin et al., 2026), uses pretrained image-to-3D models but resolves hand–object placement through explicit test-time optimization.

This contrast motivates a different question: can the hand–object relation itself be established by generation rather than recovered through explicit pose optimization? If the hand and object are generated as a single geometric sample, their relative translation, orientation, and scale are encoded directly in that sample, allowing pretrained generative priors to model geometry and the hand–

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

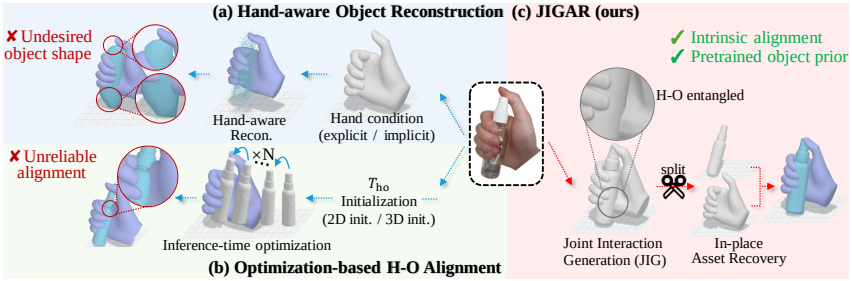


Figure 1: **JIGAR combines generative object priors with intrinsic hand–object alignment through joint interaction generation.** (a) Hand-aware Reconstruction incorporates hand geometry or interaction constraints into learned reconstruction, but makes limited use of object-centric image-to-3D priors. (b) Optimization-based Hand-Object Alignment recovers object placement through test-time pose and scale optimization, which can leave inaccurate hand–object alignment. (c) The JIG module jointly generates an interaction scaffold; then scaffold-conditioned in-place object regeneration and parametric hand recovery are used to obtain separate assets with coherent hand-relative placement. Hand and object surfaces in the joint scaffold can become locally entangled near contact (H-O tangle), requiring segmentation before asset recovery.

object relation together. This avoids treating relative pose as a separate variable inferred after object reconstruction and allows subsequent shape refinement to remain in the same frame. We refer to this property as intrinsic alignment by generation.

We realize this idea through JIGAR. Its Joint Interaction Generation (JIG) stage adapts a pretrained object-centric image-to-3D model to jointly generate hand–object geometry as an interaction scaffold. The model is conditioned on complementary hand renderings and trained with interaction-weighted supervision around the hand and contact regions. The scaffold establishes the hand-relative configuration in a shared frame. Because its hand and object surfaces may become locally entangled near contact, we partition it before asset recovery. We regenerate the object in place using the scaffold and an inpainted object observation, while retaining the parametric hand and using the generated hand component as a coordinate bridge. This separates interaction establishment from final asset recovery: joint generation establishes the relation, while in-place regeneration refines the object geometry in that frame (Fig. 1(c)).

Notably, although FollowMyHold also uses a coarse hand–object mesh, it uses this mesh to guide object generation and pose optimization at inference time. JIGAR instead learns to generate the interaction scaffold itself, establishing the hand–object relation for subsequent asset recovery without requiring test-time optimization of the hand–object relative placement.

We evaluate JIGAR on SHOWMe (Swamy et al., 2023), HOT3D (Banerjee et al., 2025), and HO-Cap (Wang et al., 2025a). JIGAR achieves the best object and hand-relative reconstruction scores among the evaluated methods without test-time optimization of the hand–object relative placement. Ablations examine the joint generation target, hand-render conditioning, and scaffold-guided recovery. On the common valid-output subsets, JIGAR reduces hand-aligned Chamfer distance by 42.3–63.2% relative to HORT and independently aligned object Chamfer distance by 3.4–28.4%.

Our contributions are threefold:

- **A joint interaction scaffold formulation for intrinsic hand–object alignment.** We represent hand–object configuration through jointly generated geometry in a shared frame, making relative position, orientation, and scale intrinsic to each generated sample and decoupling interaction establishment from final asset recovery.
- **A transferable adaptation strategy for pretrained image-to-3D generative models.** We condition joint generation on complementary renderings of the estimated hand and emphasize hand and contact geometry during training, enabling object-centric image-to-3D models to generate hand–object interaction scaffolds.
- **An empirical study of interaction-generation formulations.** Through controlled ablations and cross-backbone comparisons, we show the benefits of a joint hand–object gen-

eration target over hand-conditioned object-only generation and generation followed by explicit alignment.

2 Related Work

2.1 Hand-object Interaction Reconstruction

Hand–object interaction (HOI) reconstruction combines hand geometry, object geometry, and their spatial relation. Early methods reduce ambiguity with object templates, contact priors, or joint geometric constraints, including template-based HoMan, contact-field CPF, and joint mesh regularization (Hasson et al., 2021; Yang et al., 2021; Hasson et al., 2019). Video methods exploit temporal consistency and multi-frame visibility, with DiffHOI additionally using a hand- and category-conditioned diffusion prior (Huang et al., 2022; Fan et al., 2024; Wang et al., 2025b; On et al., 2025; Chen et al., 2026; Wang et al., 2026; Shi et al., 2026; Ye et al., 2023; 2024; Xu et al., 2026). We focus on the more ambiguous single-image setting.

Single-image methods can be broadly grouped into two approaches for establishing the hand–object relation. Hand-aware Reconstruction methods encode hand geometry or interaction constraints to establish the relation during learned prediction. They include joint mesh or interaction-field representations (HO and GF), hand-conditioned implicit or signed-distance fields (IHOI, AlignSDF, and gSDF), directed distance fields (DDF-HO), conditional diffusion (D-SCo), multi-view occlusion-aware supervision (MOHO), and point-cloud prediction (HORT) (Hasson et al., 2019; Karunratanakul et al., 2020; Ye et al., 2022; Chen et al., 2022; 2023; Zhang et al., 2023; Fu et al., 2024; Zhang et al., 2024; Chen et al., 2025c). Optimization-based Hand-Object Alignment methods, including EasyHOI (Liu et al., 2025) and FollowMyHold (Aytekin et al., 2026), use pretrained image-to-3D priors but resolve hand–object placement through explicit test-time optimization. Few existing methods combine intrinsic hand–object alignment with a pretrained object-generation prior. JIGAR addresses this gap through a jointly generated interaction scaffold.

2.2 Single-image 3D Object Reconstruction

Single-image reconstruction has evolved from supervised voxel, point-cloud, mesh, and implicit prediction (Choy et al., 2016; Groueix et al., 2018; Wang et al., 2018; Mescheder et al., 2019) to feed-forward and autoregressive generation (Hong et al., 2024; Tochilkin et al., 2024; Huang et al., 2025; Siddiqui et al., 2024; Chen et al., 2025a;b), diffusion-based reconstruction (Poole et al., 2023; Liu et al., 2023b; 2024; Long et al., 2024), and large-scale image-to-3D priors (Liu et al., 2023a; Xiang et al., 2025b; Tencent Hunyuan3D Team, 2025b; SAM 3D Team, 2025; Xiang et al., 2025a). JIGAR builds on Hunyuan3D-2.1 (Tencent Hunyuan3D Team, 2025a) but adapts its target to joint hand–object geometry, so that the interaction configuration is generated rather than recovered only after object reconstruction.

2.3 Monocular Hand Pose and Mesh Estimation

MANO (Romero et al., 2017) remains a standard parametric representation for hand shape and articulation. Single-image methods estimate MANO parameters or hand meshes (Boukhayma et al., 2019; Zimmermann et al., 2019), while HaMeR and WiLoR improve robustness in unconstrained images (Pavlakos et al., 2024; Potamias et al., 2025). JIGAR uses the estimated parametric hand as a conditioning and coordinate reference, while retaining it as the final hand asset.

3 Methodology

3.1 Problem Formulation and Overview

Given a single RGB image I, JIGAR recovers a hand mesh M_h^* and an object mesh M_o^* with plausible individual geometry and coherent relative configuration. Its two stages first establish the interaction with JIG and then recover the separate assets with scaffold-guided Asset Recovery.

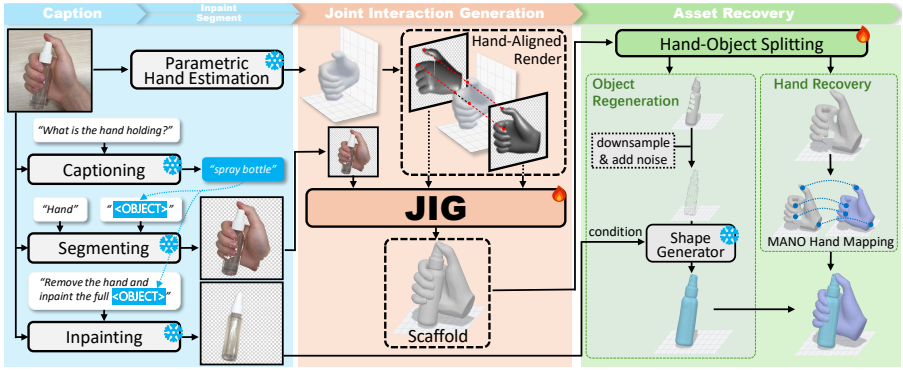


Figure 2: **JIGAR establishes the interaction jointly and recovers the assets separately.** Preprocessing supplies the interaction image, completed object image, and parametric hand. JIG generates the interaction scaffold, while scaffold-guided recovery separates the assets, regenerates the object, and transfers it to the predicted hand frame.

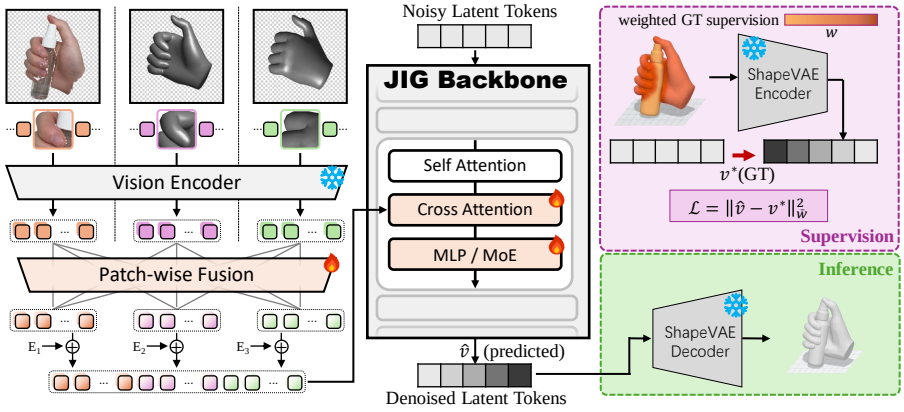


Figure 3: **JIG adapts a pretrained image-to-3D model to joint hand–object generation.** Complementary hand renderings condition the generator, which produces an interaction scaffold under hand- and contact-aware supervision.

In our implementation, Qwen3.8-Flash (Alibaba Cloud, 2026) identifies the manipulated object in I, SAM 3 (Carion et al., 2025) provides masks for the interaction image I_{HOI}, FLUX.2-dev (Black Forest Labs, 2025) removes the hand and completes the occluded object appearance to obtain I_o^{inp}, and WiLoR (Potamias et al., 2025) estimates the parametric hand $\hat{M}_h$ and camera parameters; preprocessing details are given in **Appendix A**.

As shown in Fig. 2, Joint Interaction Generation (JIG) produces the scaffold

$$S = G_{\theta_s}(I_{\text{HOI}}, f(\hat{M}_h)), \tag{1}$$

In Eq. 1, f constructs the hand condition from front and far renderings under the estimated camera (whose parameters are omitted for brevity). The scaffold is separated into hand and object components P_h^{scaf} and P_o^{scaf} in the same frame. The scaffold-guided Asset Recovery stage then regenerates the object in place from P_o^{scaf} and I_o^{inp}, retains $\hat{M}_h$ as the final hand, and uses P_h^{scaf} to transfer the recovered object to the WiLoR frame.

3.2 Joint Interaction Generation (JIG)

As shown in Fig. 3, we denote the JIG module by G_{θ_s} and build it on the flow-matching, ShapeVAE-based image-to-3D backbone of Hunyuan3D-2.1 (Tencent Hunyuan3D Team, 2025a). JIG changes the target from an isolated object to a joint hand–object union. For each sample, we apply the same hand- and camera-derived normalization to both ground-truth assets, preserving their relative translation, orientation, and scale, and encode the union with the frozen ShapeVAE. To separate

locally entangled hand and object surfaces, a geometry-based segmentor built on Point Transformer V3 (Wu et al., 2024) partitions the generated scaffold S into P_h^{scaf} and P_o^{scaf} without altering their shared frame or the configuration established by joint generation. Details of the segmentor are provided in **Appendix F**. The generated hand therefore serves as a geometric reference for the interaction rather than the final hand asset.

Hand-conditioned generation. The RGB observation alone is incomplete around occlusion and contact, so we augment it with front and far renderings of $\hat{M}_h$ under the estimated camera. The front rendering I_h^{front} records the nearest hand surface along each ray, while I_h^{far} exposes the complementary surface. All three images share the same crop and patch layout. The frozen DINOv2 encoder (Oquab et al., 2023) produces token streams C_{HOI}, C_h^{front}, and C_h^{far}, which are fused at corresponding image locations before cross-attention conditioning. For patch j, let c_j concatenate the three corresponding patch tokens. A lightweight fusion MLP F_ϕ predicts a residual for each stream $r \in \{\text{HOI}, \text{front}, \text{far}\}$ as $\delta_j^r = [F_\phi(c_j)]_r$. We update each token by $\tilde{c}_j^r = c_j^r + \delta_j^r + e_r$, where e_r is a learned stream-type embedding. This exchanges information across aligned patches while retaining each stream's identity. The resulting streams are concatenated as the cross-attention condition of the diffusion transformer (DiT) (Peebles & Xie, 2023).

During training, we emphasize hand and contact regions with a weighted flow-matching loss. We construct $z_t = (1-t)\epsilon + t z^{\text{gt}}$ with target velocity $v^* = z^{\text{gt}} - \epsilon$. Let ℓ_{bi} denote the velocity-prediction MSE of token i in sample b. We optimize

$$\mathcal{L}_{\text{FM}} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_i w_{bi}\ell_{bi}}{\sum_i w_{bi}}, \qquad w_{bi} = \begin{cases} \lambda_c, & i \in \mathcal{C}_b, \\ \lambda_h, & i \in \mathcal{H}_b \setminus \mathcal{C}_b, \\ 1, & \text{otherwise}, \end{cases} \tag{2}$$

where $\mathcal{H}_b$ denotes hand anchors, $\mathcal{C}_b \subseteq \mathcal{H}_b$ denotes contact-region hand anchors, and $\lambda_c > \lambda_h > 1$. The weighting emphasizes the relatively small hand and contact regions while per-sample normalization keeps the loss scale comparable across samples. JIG is trained on 90,518 filtered ObMan interactions (Hasson et al., 2019) and refined on 4,880 filtered DexYCB views (Chao et al., 2021); ground-truth MANO hands are used during training and WiLoR predictions at inference. Full schedules and adaptation details are provided in **Appendix F**.

3.3 Scaffold-guided Asset Recovery

The scaffold establishes the relative arrangement but is not required to provide the final object geometry. We therefore recover the object separately while retaining the estimated parametric hand and using the scaffold as their shared geometric reference.

Scaffold-conditioned in-place object regeneration. The object scaffold carries its coarse geometry and pose. We encode P_o^{scaf} directly in the scaffold frame with the frozen ShapeVAE, without independently re-centering, re-scaling, or canonicalizing it, and denote the resulting latent by z^{scaf}. Rather than restarting from pure noise, we partially perturb it as

$$z_s = s z^{\text{scaf}} + (1-s)\epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \qquad s = 0.35. \tag{3}$$

Starting from the partially perturbed latent in Eq. 3, the original Hunyuan3D-2.1 generator conditioned on I_o^{inp} produces $\tilde{M}_o$. Here s is the scaffold retention coefficient and $1-s$ is the effective noise fraction; it controls the trade-off between preserving coarse configuration and refining object geometry. We use $s = 0.35$ in the main experiments to balance these effects, with the retention sweep reported in **Appendix G**. Scaffold initialization anchors regeneration to the established interaction without subsequent optimization of the hand–object relative placement.

Hand recovery and coordinate transfer. We retain the WiLoR estimate $\hat{M}_h$ as the final hand because it provides a reliable articulated representation with fixed MANO topology (Romero et al., 2017); the generated hand scaffold serves only as a coordinate bridge. Although JIG preserves the input viewpoint in training, its scaffold can have a residual global pose error. We therefore perform residual rigid registration between the estimated hand and P_h^{scaf} in the scaffold frame. Let T_h denote the resulting coordinate mapping from the WiLoR frame to the scaffold frame.

This registration is a frame correction rather than object-pose estimation: the object and generated hand already share the scaffold frame. The final outputs are therefore $M_h^* = \hat{M}_h$ and $M_o^* = T_h^{-1}(\tilde{M}_o)$, expressed in the WiLoR camera frame while retaining the interaction established by JIG.

4 EXPERIMENTS

We evaluate whether jointly generating the interaction and then recovering separate assets improves both hand-relative configuration and object geometry. We first compare reconstruction accuracy on shared evaluation samples, then examine output coverage, contact, and physical behavior.

4.1 EXPERIMENTAL SETUP

Datasets and baselines. We evaluate on SHOWMe (Swamy et al., 2023), HOT3D (Banerjee et al., 2025), and HO-Cap (Wang et al., 2025a) using 500 independently processed RGB frames per dataset. Sampling and camera preprocessing are described in **Appendix A**; representative cases are visualized in **Appendix D**.

We compare Hand-aware Reconstruction methods (IHOI (Ye et al., 2022), AlignSDF (Chen et al., 2022), gSDF (Chen et al., 2023), and HORT (Chen et al., 2025c)) with Optimization-based Hand-Object Alignment methods (EasyHOI (Liu et al., 2025) and FollowMyHold (Aytekin et al., 2026)). For fairness, all methods receive WiLoR-derived hand inputs, with interface-specific preprocessing described in **Appendix A**. Methods that reconstruct the hand internally may still produce different final hand meshes. Hand-only results and runtime comparisons are reported in **Appendices C and E**.

Evaluation protocol. Before scoring, we resolve the monocular scale ambiguity for each prediction. We report CD-L1 for surface distance and F@5/F@10 for surface agreement at 5 and 10 mm thresholds under two alignment modes. Hand-aligned metrics evaluate the object in the recovered hand frame and therefore retain errors in the hand–object placement. Independently aligned metrics first estimate an independent similarity transform of the object, isolating its geometry from placement errors. For full-set robustness, OR denotes valid-output coverage, CR the fraction of valid outputs satisfying the contact criterion, SD the simulated object displacement, and IV the hand–object intersection volume.

Methods differ in which inputs yield valid reconstructions. To compare accuracy on identical samples, both reconstruction tables use only input samples for which all seven methods produce valid, evaluable reconstructions. This common intersection contains 457 frames on SHOWMe, 375 on HOT3D, and 317 on HO-Cap. Output coverage and physical behavior are evaluated separately in Sec. 4.3. Bold and underlined values indicate the best and second-best scores, respectively; higher F-scores, OR, and CR and lower CD, SD, and IV are preferred.

4.2 MAIN RESULTS

Object reconstruction in the hand-aligned frame. Table 1 evaluates the object with the hand as the spatial reference. JIGAR achieves the best CD and F-scores on all three datasets, reducing HORT's hand-aligned CD by 42.3–63.2% and substantially improving F@10 on SHOWMe and HOT3D. It also outperforms both optimization-based baselines, supporting scaffold-based recovery in an established interaction frame.

Object reconstruction after independent alignment. JIGAR remains best across all datasets after independent alignment (Table 2). The larger gain under hand-aligned evaluation supports improved hand–object alignment in addition to object geometry.

4.3 RECONSTRUCTION ROBUSTNESS

We evaluate output coverage, contact, simulation displacement, and hand–object intersection volume on the full 500-input set of each dataset (Table 3). Full simulation details are given in **Appendix A**.

Table 1: **Object reconstruction evaluated in the hand-aligned frame: JIGAR achieves the best scores across all three datasets.** The object follows the hand's rigid alignment, so scores measure both its geometry and its placement relative to the hand. CD is in mm; F@5 and F@10 measure surface agreement within 5 and 10 mm.

			SHOWMe ($N = 457$)			HOT3D ($N = 375$)			HO-Cap ($N = 317$)		
	Method	Venue	$CD_h \downarrow$	$F@5_h \uparrow$	$F@10_h \uparrow$	$CD_h \downarrow$	$F@5_h \uparrow$	$F@10_h \uparrow$	$CD_h \downarrow$	$F@5_h \uparrow$	$F@10_h \uparrow$
Hand-aware	IHOI	CVPR 2022	40.65	<u>0.206</u>	<u>0.384</u>	<u>58.96</u>	0.159	0.295	<u>46.27</u>	0.196	0.369
	AlignSDF	ECCV 2022	161.88	0.048	0.095	258.26	0.026	0.051	270.37	0.032	0.062
	gSDF	CVPR 2023	60.92	0.162	0.306	123.18	0.075	0.144	151.43	0.073	0.140
	HORT	ICCV 2025	<u>36.52</u>	0.204	0.380	67.66	<u>0.166</u>	<u>0.303</u>	81.09	<u>0.287</u>	<u>0.476</u>
optim.	EasyHOI	CVPR 2025	84.64	0.112	0.212	110.76	0.090	0.171	258.45	0.082	0.154
	FollowMyHold	3DV 2026	83.26	0.124	0.232	120.70	0.081	0.149	189.51	0.047	0.089
	JIGAR (ours)	–	**21.06**	**0.422**	**0.665**	**27.39**	**0.352**	**0.582**	**29.83**	**0.314**	**0.523**

Table 2: **Object reconstruction evaluated after independent alignment: JIGAR achieves the best scores across all three datasets.** This evaluation assesses the object's own reconstruction quality after independent similarity alignment. CD is in mm; F@5 and F@10 measure surface agreement within 5 and 10 mm.

			SHOWMe ($N = 457$)			HOT3D ($N = 375$)			HO-Cap ($N = 317$)		
	Method	Venue	CD ↓	F@5 ↑	F@10 ↑	CD ↓	F@5 ↑	F@10 ↑	CD ↓	F@5 ↑	F@10 ↑
Hand-aware	IHOI	CVPR 2022	23.87	0.325	0.568	26.47	0.268	0.487	26.57	0.281	0.506
	AlignSDF	ECCV 2022	26.46	0.285	0.528	28.38	0.256	0.473	35.75	0.201	0.386
	gSDF	CVPR 2023	20.10	0.381	0.643	27.28	0.266	0.484	36.18	0.212	0.396
	HORT	ICCV 2025	<u>14.59</u>	<u>0.505</u>	<u>0.760</u>	<u>17.84</u>	<u>0.421</u>	<u>0.685</u>	<u>19.57</u>	<u>0.373</u>	<u>0.663</u>
optim.	EasyHOI	CVPR 2025	21.37	0.436	0.645	24.56	0.327	0.551	33.65	0.224	0.411
	FollowMyHold	3DV 2026	24.96	0.405	0.586	24.95	0.338	0.561	36.74	0.215	0.398
	JIGAR (ours)	–	**11.47**	**0.606**	**0.833**	**12.77**	**0.583**	**0.816**	**18.90**	**0.446**	**0.664**

JIGAR produces valid outputs for 95.8–96.4% of inputs, with 99.6–100.0% of valid outputs satisfying the contact criterion; it also has the lowest SD on all three datasets and lower IV than HORT and both optimization-based baselines.

4.4 Ablation Studies

We evaluate formulation choices, JIG components, and asset recovery on the 375-image HOT3D common set used in Table 1; the results are summarized in Table 4. The backbone-replacement experiment uses each variant's valid outputs from the full 500-image list. All variants use the hand-aligned metrics of Sec. 4.1.

Formulation. On the same 375 HOT3D samples, joint hand–object generation outperforms object-only generation under both conditioning types, while 2D-render conditioning outperforms 3D-latent conditioning at the same target and budget (CD_h 31.49→27.39 mm; $F@10_h$ 0.527→0.582). The optimization-based row provides a system-level reference with explicit alignment.

Design and components. Removing either hand rendering causes a small drop, while removing both, patch fusion, or the interaction-weighted loss further degrades the results. For recovery, object regeneration, scaffold initialization, and residual hand registration each provide a measurable benefit, supporting both scaffold-guided regeneration and the coordinate bridge used for final asset transfer. We use $s = 0.35$ in the main experiments to balance scaffold preservation and object refinement; the complete retention sweep and transform-drift analysis are reported in **Appendix G**.

Backbone generality. To test backbone independence, we instantiate JIG on Hunyuan3D-2.1 and Step1X-3D (Li et al., 2025) with the same data, conditioning, targets, and evaluation protocol. We

Table 3: **Reconstruction robustness: JIGAR combines high output coverage with the highest contact rates and lowest simulation displacement.** Robustness is evaluated on each method's own valid outputs across the three datasets. OR measures output coverage over 500 inputs; CR measures contact among valid reconstructions. SD is short-horizon simulated object displacement (cm) over all valid outputs; IV is hand–object intersection volume (cm^3) among contacting outputs.

			SHOWMe				HOT3D				HO-Cap			
	Method	Venue	OR (%) ↑	CR (%) ↑	SD ↓	IV ↓	OR (%) ↑	CR (%) ↑	SD ↓	IV ↓	OR (%) ↑	CR (%) ↑	SD ↓	IV ↓
Hand-aware	IHOI	CVPR 2022	98.4	98.8	1.72	**5.14**	83.6	90.0	2.79	**3.54**	78.8	98.0	2.61	**3.98**
	AlignSDF	ECCV 2022	**98.8**	77.3	4.53	26.10	86.4	45.1	3.90	17.11	79.8	41.4	3.81	16.17
	gSDF	CVPR 2023	**98.8**	97.4	3.21	50.85	85.8	79.7	4.68	44.81	75.2	79.0	3.78	44.51
	HORT	ICCV 2025	**98.8**	99.6	2.88	30.49	86.4	91.9	3.58	15.26	80.4	87.3	2.60	15.48
optim.	EasyHOI	CVPR 2025	98.0	97.1	3.13	40.15	**96.6**	91.3	2.34	20.98	**99.0**	69.1	2.48	25.59
	FollowMyHold	3DV 2026	96.4	94.4	3.91	57.60	96.4	80.3	4.16	22.99	89.2	66.4	4.85	34.60
	JIGAR (ours)	–	95.8	**100.0**	**1.20**	9.04	95.8	**99.6**	**0.97**	3.89	96.4	**99.8**	**1.37**	5.46

Table 4: **Formulation and design/component ablations: joint generation and scaffold-guided recovery improve hand-aligned reconstruction.** All variants use the same HOT3D subset as the main results ($N = 375$). CD is in mm; F@5 and F@10 measure agreement within 5 and 10 mm.

(a) Formulation ablations

Variant	CD_h ↓	$F@5_h$ ↑	$F@10_h$ ↑
3D-latent conditioning			
Object-only	57.21	0.184	0.332
Joint	31.49	0.306	0.527
2D-render conditioning			
Object-only	50.26	0.210	0.351
JIGAR (joint)	**27.39**	**0.352**	**0.582**
Reference formulations			
Optimization-based Alignment	104.36	0.122	0.193

(b) Design and component ablations

	Variant	CD_h ↓	$F@5_h$ ↑	$F@10_h$ ↑
JIG	w/o front render	29.41	0.344	0.527
	w/o far render	29.47	0.340	0.516
	w/o front and far render	30.91	0.313	0.486
	w/o patch fusion	29.26	0.337	0.491
	w/o weighted loss	32.54	0.308	0.482
Recovery	w/o object regeneration	35.17	0.296	0.359
	w/o scaffold init. (pure noise)	147.25	0.114	0.136
	w/o residual hand registration	36.13	0.283	0.354
	JIGAR (full)	**27.39**	**0.352**	**0.582**

Table 5: **Backbone generality and adaptation.** Panel (a) compares complete JIGAR pipelines with JIG-adapted scaffold generators on the common 375-image HOT3D subset used by the main comparison. Panel (b) replaces the adapted scaffold generator with an original backbone and evaluates each complete pipeline on its own valid outputs from the fixed 500-image list; the valid-output count is shown in the evaluation-samples column. All rows report the complete JIGAR pipeline, and each row label specifies the scaffold generator. All metrics are hand-aligned whole-object metrics, and CD is in mm. Bold values indicate the best result within each panel.

Scaffold generator	Evaluation samples	CD-L1↓	F@5↑	F@10↑
(a) JIG-adapted backbones				
Hunyuan3D-2.1 (JIG-adapted)	Common ($N = 375$)	**27.39**	**0.352**	**0.582**
Step1X-3D (JIG-adapted)	Common ($N = 375$)	30.52	0.320	0.561
(b) Original-backbone replacement and adapted reference				
Hunyuan3D-2.1 (JIG-adapted)	**Own valid ($\mathbf{479/500}$)**	**28.36**	**0.341**	**0.567**
Hunyuan3D-2.1 (original)	Own valid ($257/500$)	68.76	0.103	0.194
Step1X-3D (original)	Own valid ($403/500$)	73.48	0.100	0.190

also replace the adapted scaffold generator with each original backbone while keeping the downstream JIGAR pipeline fixed.

Panel (a) shows that the interaction-scaffold formulation transfers to Step1X-3D with a moderate metric change. Replacing the adapted generator with either original backbone substantially lowers accuracy and valid-output coverage, mainly because scaffold hands can deviate from MANO and fail hand registration; panel (b) is therefore an end-to-end adaptation diagnostic rather than a matched-sample ranking.

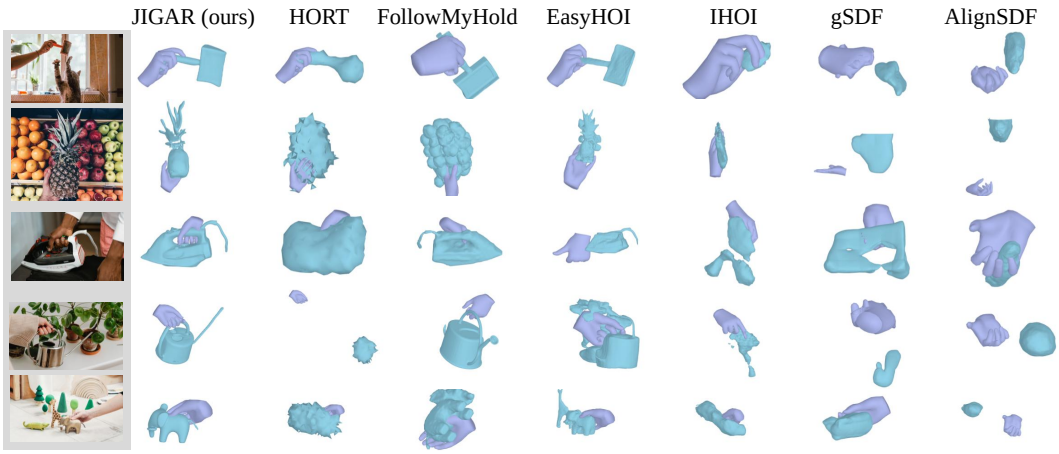


Figure 4: **JIGAR recovers distinctive object structure together with coherent hand-relative placement.** Each row shows the input and reconstructions from JIGAR and six baselines. Hands are purple and objects are cyan. Examples highlight thin handles, handle openings, and the watering can's spout.

4.5 Qualitative Results

Figure 4 illustrates the joint demands of object fidelity and interaction reconstruction. JIGAR recovers distinctive structures such as thin handles, openings, and elongated spouts, whereas HORT, IHOI, gSDF, and AlignSDF can exhibit incomplete or distorted geometry in these examples. EasyHOI and FollowMyHold recover recognizable objects in some cases but can exhibit placement or scale errors; JIGAR combines recognizable geometry with hand-relative placement through the shared scaffold.

5 Limitations

Limitations imposed by the task setting. A single image does not uniquely determine severely occluded surfaces or hidden internal structures. JIGAR recovers static surface geometry, while the current formulation does not infer object joints, part kinematics, or full physical dynamics.

Upstream dependence and physical validity. The pipeline depends on hand estimation, segmentation, and image completion, so errors in these inputs can propagate to the final reconstruction. Segmentation failures or severe inpainting hallucinations can cause the pipeline to fail and reduce valid-output coverage. JIGAR also imposes no explicit non-penetration or physical-contact constraint, so physically valid contact is not guaranteed in every output. Sensitivity to alternative upstream modules is quantified in **Appendix B**.

Dependence on the generative prior. The adapted 3D prior can hallucinate or smooth high-frequency object geometry. The scaffold establishes the coarse hand–object configuration, whereas final object details are supplied by the pretrained object generator. Representative examples are discussed in **Appendix H**.

6 Conclusion

We present JIGAR, a framework for monocular hand–object reconstruction that establishes the interaction through jointly generated geometry and then recovers separate assets. Controlled ablations support a joint generation target over hand-conditioned object-only generation and favor 2D-render conditioning over 3D latent conditioning under the evaluated training budget. On SHOWMe, HOT3D, and HO-Cap, JIGAR improves both hand-relative interaction reconstruction and object reconstruction over the evaluated baselines without requiring test-time optimization of the hand–object relative placement. High contact rates and low simulation displacement further support the coherence of the recovered interactions. These results support joint interaction generation as an effective way to combine pretrained shape priors with intrinsic hand–object alignment.

AI USE STATEMENT

We used generative AI tools to assist with manuscript writing, including language editing, improving clarity, and refining the organization and flow of the text.

REFERENCES

Alibaba Cloud. Qwen3.8-Flash released. https://www.alibabacloud.com/en/news/product/qwen38-flash-released-noi, 2026.

Ayce Idil Aytekin, Helge Rhodin, Rishabh Dabral, and Christian Theobalt. Follow my hold: Hand-object interaction reconstruction through geometric guidance. In *International Conference on 3D Vision (3DV)*, 2026.

Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Shangchen Han, Fan Zhang, Linguang Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, Jakob Julian Engel, and Tomas Hodan. HOT3D: Hand and object tracking in 3D from egocentric multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. URL https://facebookresearch.github.io/hot3d/.

Black Forest Labs. FLUX.2: Frontier visual intelligence. https://bfl.ai/blog/flux-2, 2025.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Adnane Boukhayma, Rodrigo de Bem, and Philip H. S. Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10843–10852, 2019.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9044–9053, 2021.

Yiwen Chen, Tong He, Di Huang, Weicai Ye, Sijin Chen, Jiaxiang Tang, Xin Chen, Zhongang Cai, Lei Yang, Gang Yu, Guosheng Lin, and Chi Zhang. MeshAnything: Artist-created mesh generation with autoregressive transformers. In *International Conference on Learning Representations (ICLR)*, 2025a.

Yongwei Chen, Yushi Lan, Shangchen Zhou, Tengfei Wang, and Xingang Pan. SAR3D: Autoregressive 3d object generation and understanding via multi-scale 3d VQVAE. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Yuantao Chen, Jiahao Chang, Chongjie Ye, Chaoran Zhang, Zhaojie Fang, Chenghong Li, and Xiaoguang Han. ForeHOI: Feed-forward 3d object reconstruction from daily hand-object interaction videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8868–8879, 2026.

Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. AlignSDF: Pose-aligned signed distance fields for hand-object reconstruction. In *European Conference on Computer Vision (ECCV)*, 2022.

Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gSDF: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12890–12900, 2023.

Zerui Chen, Rolandos Alexandros Potamias, Shizhe Chen, and Cordelia Schmid. HORT: Monocular hand-held objects reconstruction with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6046–6057, 2025c.

Christopher B. Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3D-R2N2: A unified approach for single and multi-view 3d object reconstruction. In *European Conference on Computer Vision (ECCV)*, 2016.

Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J. Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 494–504, 2024.

Bowen Fu, Gu Wang, Chenyangguang Zhang, Yan Di, Ziqin Huang, Zhiying Leng, Fabian Manhardt, Xiangyang Ji, and Federico Tombari. D-SCo: Dual-stream conditional diffusion for monocular hand-held object reconstruction. In *European Conference on Computer Vision (ECCV)*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/4261_ECCV_2024_paper.php.

Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. AtlasNet: A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 216–224, 2018.

Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11807–11816, 2019.

Yana Hasson, Gül Varol, Ivan Laptev, and Cordelia Schmid. Towards unconstrained joint hand-object reconstruction from RGB videos. In *International Conference on 3D Vision (3DV)*, pp. 659–668, 2021.

Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3d. In *International Conference on Learning Representations (ICLR)*, 2024.

Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia Conference Papers*, 2022. doi: 10.1145/3550469.3555401.

Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M. Rehg, and Varun Jampani. SPAR3D: Stable point-aware reconstruction of 3d objects from single images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, 2020.

Weiyu Li, Xuanyang Zhang, Zheng Sun, Di Qi, Hao Li, Wei Cheng, Weiwei Cai, Shihao Wu, Jiarui Liu, Zihao Wang, Xiao Chen, Feipeng Tian, Jianxiong Pan, Zeming Li, Gang Yu, Xiangyu Zhang, Daxin Jiang, and Ping Tan. Step1X-3D: Towards high-fidelity and controllable generation of textured 3d assets. *arXiv preprint arXiv:2505.07747*, 2025. URL https://arxiv.org/abs/2505.07747.

Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023a.

Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023b.

Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating multiview-consistent images from a single-view image. In *International Conference on Learning Representations (ICLR)*, 2024.

Yumeng Liu, Xiaoxiao Long, Zemin Yang, Yuan Liu, Marc Habermann, Christian Theobalt, Yuexin Ma, and Wenping Wang. EasyHOI: Unleashing the power of large models for reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7037–7047, 2025.

Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3D: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Luca Medeiros. Language segment-anything. `https://github.com/luca-medeiros/lang-segment-anything`, 2025.

Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4460–4470, 2019.

Jeongwan On, Kyeonghwan Gwak, Gunyoung Kang, Junuk Cha, Soohyun Hwang, Hyein Hwang, and Seungryul Baek. BIGS: Bimanual category-agnostic interaction reconstruction from monocular videos via 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17437–17447, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. URL `https://arxiv.org/abs/2304.07193`.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 2023.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

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

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

SAM 3D Team. SAM 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

Jin-Chuan Shi, Binhong Ye, Tao Liu, Junzhe He, Yangjinhui Xu, Xiaoyang Liu, Zeju Li, Hao Chen, and Chunhua Shen. AGILE: Hand-object interaction reconstruction from video via agentic generation. In *ACM SIGGRAPH Conference Papers*, 2026.

Yawar Siddiqui, Antonio Alliegro, Alexey Artemov, Tatiana Tommasi, Daniele Sirigatti, Vladislav Rosov, Angela Dai, and Matthias Nießner. MeshGPT: Generating triangle meshes with decoder-only transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19615–19625, 2024.

Anilkumar Swamy, Vincent Leroy, Philippe Weinzaepfel, Fabien Baradel, Salma Galaaoui, Romain Brégier, Matthieu Armando, Jean-Sebastien Franco, and Grégory Rogez. SHOWMe: Benchmarking object-agnostic hand-object 3d reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 1935–1944, 2023. URL https://openaccess.thecvf.com/content/ICCV2023W/ACVR/html/Swamy_SHOWMe_Benchmarking_Object-Agnostic_Hand-Object_3D_Reconstruction_ICCVW_2023_paper.html.

Tencent Hunyuan3D Team. Hunyuan3D 2.1: From images to high-fidelity 3d assets with production-ready PBR material. *arXiv preprint arXiv:2506.15442*, 2025a. URL https://arxiv.org/abs/2506.15442.

Tencent Hunyuan3D Team. Hunyuan3D 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025b.

Dmitry Tochilkin, David Pankratz, Zexiang Liu, Zixuan Huang, Adam Letts, Yangguang Li, Ding Liang, Christian Laforte, Varun Jampani, and Yan-Pei Cao. TripoSR: Fast 3d object reconstruction from a single image. *arXiv preprint arXiv:2403.02151*, 2024.

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

Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2Mesh: Generating 3d mesh models from single rgb images. In *European Conference on Computer Vision (ECCV)*, pp. 52–67, 2018.

Shibo Wang, Haonan He, Maria Parelli, Christoph Gebhardt, Zicong Fan, and Jie Song. MagicHOI: Leveraging 3d priors for accurate hand-object reconstruction from short monocular video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5957–5968, 2025b.

Zikai Wang, Zhilu Zhang, Yiqing Wang, Hui Li, and Wangmeng Zuo. ArtHOI: Taming foundation models for monocular 4d reconstruction of hand-articulated-object interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15998–16009, 2026.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer V3: Simpler faster stronger. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Wu_Point_Transformer_V3_Simpler_Faster_Stronger_CVPR_2024_paper.html.

Jianfeng Xiang, Xiaoxue Chen, Sicheng Xu, Ruicheng Wang, Zelong Lv, Yu Deng, Hongyuan Zhu, Yue Dong, Hao Zhao, Nicholas Jing Yuan, and Jiaolong Yang. Native and compact structured latents for 3d generation. *arXiv preprint arXiv:2512.14692*, 2025a.

Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025b.

Hao Xu, Yilin Liu, Yinqiao Wang, Chi-Wing Fu, and Niloy J. Mitra. CHOIR: Contact-aware 4d hand-object interaction reconstruction. *arXiv preprint arXiv:2605.20992*, 2026.

Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11097–11106, 2021.

Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3895–3905, 2022.

Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19660–19671, 2023.

Yufei Ye, Abhinav Gupta, Kris Kitani, and Shubham Tulsiani. G-HOP: Generative hand-object prior for interaction reconstruction and grasp synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1911–1920, 2024.

Chenyangguang Zhang, Yan Di, Ruida Zhang, Guangyao Zhai, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. DDF-HO: Hand-held object reconstruction via conditional directed distance field. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Chenyangguang Zhang, Guanlong Jiao, Yan Di, Gu Wang, Ziqin Huang, Ruida Zhang, Fabian Manhardt, Bowen Fu, Federico Tombari, and Xiangyang Ji. MOHO: Learning single-view hand-held object reconstruction with multi-view occlusion-aware supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9992–10002, 2024.

Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 813–822, 2019.

A Inference and Evaluation Protocol

This appendix describes the operations applied to an evaluation image and the protocols used to score the resulting reconstructions. Training data, optimization, and model-adaptation details are given separately in Appendix F.

A.1 Test-time preprocessing

Given a single RGB image, we first use Qwen3.8-Flash (Alibaba Cloud, 2026) to identify the manipulated object and provide semantic cues. SAM 3 (Carion et al., 2025) then segments the hand and object in the original image. Their union mask defines the isolated interaction image I_{HOI} used by the scaffold generator.

Using the same object cue, FLUX.2-dev (Black Forest Labs, 2025) removes the hand and completes the occluded object appearance. The completed image is used as the inpainted object observation I_o^{inp}. In parallel, WiLoR (Potamias et al., 2025) estimates the parametric MANO hand and camera parameters from the original image. The isolated interaction image and the estimated hand then enter the scaffold-generation stage.

A.2 Inference and asset-recovery settings

Scaffold generation uses JIG, our interaction-adapted version of Hunyuan3D-2.1 (Tencent Hunyuan3D Team, 2025a), together with its ShapeVAE. Object regeneration uses the pretrained object-centric Hunyuan3D-2.1 generator, conditioned on I_o^{inp} and initialized from the separated scaffold latent. After scaffold generation, the geometric segmentor separates hand and object components. The object component is encoded in its scaffold frame and regenerated with 50 flow-matching steps, guidance scale 5.5, and scaffold retention coefficient $s = 0.35$. The regenerated object is transferred to the WiLoR hand frame using the residual rigid registration described in Sec. 3.3. The transform-drift analysis in Appendix G varies only s.

A.3 Evaluation datasets and sampling

We evaluate on SHOWMe (Swamy et al., 2023), HOT3D (Banerjee et al., 2025), and HO-Cap (Wang et al., 2025a). We construct a fixed evaluation list for each dataset before running any reconstruction method. A valid case contains one RGB image, one target object, and one interacting hand. We retain cases with geometric hand–object contact, sufficient hand and object visibility, and acceptable image quality. We remove synchronized-camera duplicates and overly similar video frames while maintaining coverage across sequences, objects, and viewpoints. The resulting coverage is summarized in Table 6. Ground-truth geometry is used only for offline filtering, camera preprocessing, and final evaluation; it is never provided as an inference input.

Table 6: **Evaluation-set coverage.** Each dataset contributes 500 cases to the evaluation. The object count reports selected instances over the corresponding dataset inventory.

Dataset	Cases	Sequences / videos	Object instances (selected / total)
SHOWMe	500	91 videos	42 / 42 (100%)
HOT3D	500	82 sequences	29 / 33 (87.9%)
HO-Cap	500	43 sequences	63 / 64 (98.4%)

For HO-Cap, we limit the number of cases drawn from each single-hand sequence and enforce temporal and viewpoint separation. HOT3D is restricted to right-hand interactions without another hand entering the interaction region; we also limit per-sequence sampling and enforce viewpoint diversity. For SHOWMe, cases are selected across videos before filling the 500-case budget, with a minimum viewpoint difference between selected examples. The common valid-output subsets used for the two main reconstruction tables contain 457 SHOWMe frames, 375 HOT3D frames, and 317 HO-Cap frames.

The selected-object count denotes the number of instances that contribute at least one case to the frozen evaluation list, while the available count follows the corresponding dataset inventory. Object

instances are not selected according to reconstruction performance. An instance is included only when at least one of its cases satisfies the predefined single-hand, contact, visibility, image-quality, and de-duplication criteria; we do not add borderline cases solely to obtain complete object coverage.

Although HOT3D RGB images are captured with an F-Theta fisheye camera, its official calibration provides per-frame Fisheye624 intrinsics together with camera extrinsics and poses. We use this calibration and the official projection model to geometrically rectify each image into a virtual pinhole camera with a 1280×1280 image, a 60° horizontal field of view, and a centered principal point. The hand and object annotations are rotated into the same virtual-camera frame. This is an inverse geometric mapping, rather than a linear image resize or an approximation that treats fisheye parameters as pinhole intrinsics, so the official calibration is sufficient to obtain a geometrically consistent standard pinhole image and camera model for HOT3D. The rectification and virtual-view selection are fixed during offline dataset construction; ground-truth geometry is not provided to the reconstruction methods at inference.

A.4 Evaluation protocol

For each prediction, we resolve monocular scale ambiguity using the median ratio between ground-truth and predicted lengths over 20 hand bones and apply the resulting scale to both assets. We sample the reconstructed surfaces and compute symmetric CD-L1 and F-scores at 5 and 10 mm. CD is reported in millimeters, and each F-score is the harmonic mean of surface precision and recall at its distance threshold.

For hand-aligned interaction evaluation, we rigidly align the predicted hand to the ground-truth hand and apply the same transform to the object, without independently aligning the object. The resulting CD_h, $\mathrm{F@5}_h$, and $\mathrm{F@10}_h$ retain object-shape errors as well as errors in hand-relative position, orientation, and scale. For object evaluation, we independently align the predicted object to the reference with an ICP-based similarity transform that estimates rotation, translation, and scale.

The two reconstruction tables use only samples for which all seven methods produce valid, evaluable outputs. Robustness metrics are instead computed on each method's own valid outputs over the full 500-input set. Output rate (OR) is the fraction of inputs with valid reconstructions, and contact rate (CR) is the fraction of valid outputs whose minimum hand–object distance is below 1 cm after scale normalization. Simulation displacement (SD) measures the object's net displacement over 20 steps at $1/240$ s per step with the predicted hand fixed and gravity along the camera's negative Y axis. It is computed on all valid outputs without object alignment or penetration correction. Intersection volume (IV) is averaged over each method's contacting outputs.

B Sensitivity to Upstream Components

We replace one test-time upstream module at a time while keeping the trained JIGAR model and the remaining pipeline fixed. For the segmentation variant, LangSAM (Medeiros, 2025) receives the prompts "hand" and "manipulated object"; when multiple predictions are returned, we retain the hand–object pair with contact. The completion variant uses FLUX.1 Kontext [dev] (Black Forest Labs et al., 2025), the same checkpoint used by FollowMyHold (Aytekin et al., 2026). The hand-estimation variant replaces WiLoR with HaMeR (Pavlakos et al., 2024).

All variants are evaluated on the fixed 500-image HOT3D list. Metrics are computed over each variant's own valid outputs, so the valid-output count and the conditional reconstruction quality should be considered together (Table 7).

Replacing Qwen3.8-Flash and SAM 3 with LangSAM mainly reduces output coverage while leaving conditional reconstruction quality close to the main setting. Replacing FLUX.2-dev with FLUX.1 Kontext [dev] lowers both coverage and reconstruction quality, whereas replacing WiLoR with HaMeR has a small effect on either metric.

C Hand Reconstruction Results

We report hand-only reconstruction on the same 375 HOT3D samples used by the main comparison tables, namely the common set on which all seven methods produce valid outputs. For fairness,

Table 7: **Sensitivity to upstream test-time components on HOT3D.** Each variant changes one upstream module while keeping the remaining JIGAR pipeline fixed. Metrics are computed over each variant's own valid outputs from the fixed 500-image list; all metrics are hand-aligned whole-object metrics, with CD-L1 reported in mm.

Variant	Valid outputs	CD-L1↓	F@5↑	F@10↑
JIGAR (main setting)	479/500	28.36	0.341	0.567
LangSAM segmentation	403/500	28.76	0.337	0.564
FLUX.1 Kontext [dev] completion	442/500	32.61	0.314	0.529
HaMeR hand estimation	480/500	28.92	0.336	0.560

Table 8: **Hand reconstruction on the common HOT3D subset.** All methods are evaluated on the same 375 samples used by the main reconstruction tables. Errors are reported in mm.

Method	root MPJPE↓	root MPVPE↓	PA-MPJPE↓	PA-MPVPE↓
AlignSDF	158.10	152.87	18.86	19.45
gSDF	85.82	83.63	18.69	20.97
IHOI	42.92	41.24	11.62	12.12
FollowMyHold	56.89	55.26	8.91	9.45
EasyHOI	39.99	38.48	15.38	16.24
HORT	<u>14.13</u>	<u>13.91</u>	<u>6.40</u>	<u>6.80</u>
JIGAR (ours)	**14.10**	**13.88**	**6.38**	**6.77**

every method receives WiLoR-derived hand input in the main experiments. JIGAR, HORT, and FollowMyHold use WiLoR as their hand source; JIGAR and HORT directly retain the WiLoR hand as the final hand result. IHOI and EasyHOI are also supplied with WiLoR-based hand inputs, while gSDF and AlignSDF receive the WiLoR-based auxiliary inputs required by their interfaces. Methods that reconstruct the hand internally can therefore produce different final hand meshes even when their hand-related inputs originate from the same estimator.

We measure root-relative mean per-joint position error (MPJPE) and mean per-vertex position error (MPVPE), together with their Procrustes-aligned (PA) counterparts, in millimeters. Root-relative metrics remove the wrist translation before computing the error, whereas PA metrics additionally remove the best similarity transform. Lower values are better. The results are reported in Table 8.

JIGAR and HORT obtain nearly identical hand errors because both directly use the same WiLoR hand estimate as their final hand output. The differences in the remaining methods reflect their distinct internal processing and any hand reconstruction performed after WiLoR-based inputs are provided; they should therefore not be interpreted as differences in the quality of WiLoR itself.

D Evaluation Dataset Visualizations

Representative cases are shown in Fig. 5.

E Runtime Analysis

We measure the complete wall-clock time from a raw RGB image to the final reconstruction, including preprocessing, auxiliary prediction, generation, and post-processing; the aggregate results are reported in Table 9. Times are averaged over the 500 inputs in each dataset and then averaged across SHOWMe, HOT3D, and HO-Cap. All measurements use a single NVIDIA A800. Within each pipeline, a model is loaded once, all inputs are processed, and the next stage is then executed, so model loading and offloading overheads are amortized over the evaluation cases. These timings come from the same main experimental runs used for the reconstruction results.

The reported values are system-level runtimes rather than isolated core-model latencies. The original gSDF and AlignSDF interfaces assume auxiliary inputs that are unavailable at test time in our setting. We generate the required inputs through the same evaluation preprocessing pipeline, including WiLoR, so their runtimes include this input-completion cost. JIGAR is slower than the feed-forward hand-relative methods, whose complete pipelines take less than 4 seconds per input,



Figure 5: **Representative evaluation cases.** Each dataset column shows RGB observations together with the corresponding hand–object geometry, with the hand shown in purple and the object in cyan. From left to right, the columns correspond to HO-Cap, HOT3D, and SHOWMe. The examples cover diverse objects, grasps, and viewpoints.

but is faster than EasyHOI and FollowMyHold. This comparison also reflects that JIGAR does not repeatedly optimize the hand–object relative placement at test time.

Table 9: **Average end-to-end runtime.** Runtime is measured from the raw input image to the final reconstruction and reported in seconds per input. Lower is better.

	Method	Runtime (s) ↓
Hand-aware	IHOI	0.805
	AlignSDF	1.43
	gSDF	3.58
	HORT	0.121
optim.	EasyHOI	145
	FollowMyHold	556
	JIGAR (ours)	106

F Training and Model Adaptation

This appendix specifies how the generative prior and scaffold segmentor are adapted, including their data, conditioning, objectives, and optimization schedules. Test-time processing and evaluation are described in Appendix A.

F.1 Training Data and Input Construction

Synthetic ObMan data. We use synthetic hand–object interactions from ObMan (Hasson et al., 2019) with object meshes from ShapeNet. The initial object library contains 2,772 meshes from eight categories. We normalize each mesh, reconstruct a watertight surface, retain its largest connected component, and apply standard topology and geometry quality checks. We require a single connected, watertight, manifold mesh without boundary artifacts or other invalid geometry. The final library contains 1,768 retained meshes spanning six categories. Matching these instances back to ObMan yields 90,518 filtered synthetic hand–object interactions.

Real-world DexYCB data. For real-world refinement, we use the official $s0$/train split of DexYCB (Chao et al., 2021). We retain views that pass hand–object segmentation, grasp-validity, and occlusion-quality checks, remove synchronized-camera duplicates, subsample time within each sequence with a minimum spacing of approximately 30 frames, and select four qualified cameras per retained timestamp. This produces 1,230 timestamps from 735 sequences and 4,880 views covering all 20 object instances. Each view is cropped to 518×518 pixels around a saved hand-palm anchor. The hand and object masks define an HOI-masked RGB image: pixels outside their union are set to white, while pixels inside the union retain their original color. Camera intrinsics, crop transforms, aligned MANO parameters, and the object transform are retained for each view. The object geometry is stored as one canonical prototype per object together with its view-specific transform into the aligned frame.

Shared conditioning and targets. Synthetic and real-world samples use the same aligned hand–object coordinate convention and the same joint surface target. During training, the target is formed from the aligned MANO hand and object geometry, while DINOv2-Large encodes the image condition. In the render-conditioned stage, front and far hand renderings provide complementary conditioning; their corresponding features are fused by a shared projection and concatenated for cross-attention. Ground-truth MANO hands provide renderings during training; at inference, the corresponding renderings are produced from the WiLoR estimate described in Appendix A.

F.2 JIG Fine-tuning

Synthetic-data fine-tuning. JIG is adapted from Hunyuan3D-2.1 in two stages using the 90,518 filtered synthetic interactions and their joint hand–object targets. The first stage adapts the image-only generation pathway for one epoch with global batch size 2. The second stage starts from this checkpoint, adds the front and far rendering conditions, and fine-tunes the adapted model for seven epochs with global batch size 8. The backbone remains frozen, while parameter-efficient residual adapters are trained in both stages; the second stage additionally trains the render-conditioned attention and modality adapters.

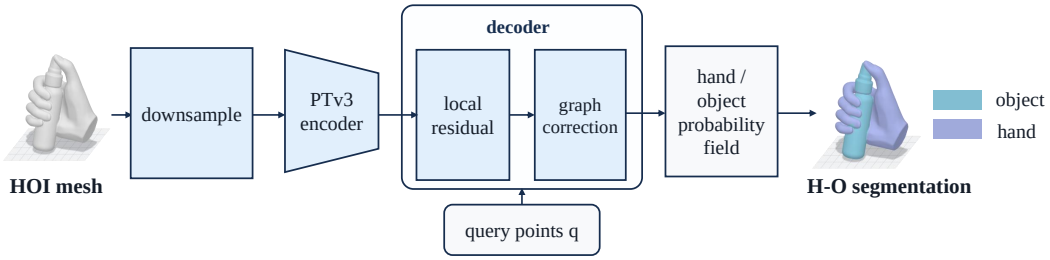


Figure 6: **Scaffold segmentation.** The input hand–object mesh is downsampled and encoded by a Point Transformer V3 encoder. Query points are processed by a continuous decoder with local residual and graph correction modules, producing a hand/object probability field and the final binary segmentation.

Both stages use AdamW with a warmup-cosine schedule and mixed-precision training. The synthetic stages use up to eight NVIDIA A800 GPUs and take about 14 hours in total.

Real-world data refinement fine-tuning. Every adapted variant, including the formulation and component ablations and the Step1X-3D model, starts this stage from its corresponding synthetic-training checkpoint. Each epoch mixes the 4,880 DexYCB views with 1,220 replayed ObMan samples. We train for three epochs using eight GPUs and a global batch size of eight. The rendering and parameter-efficient adaptation modules are updated while the pretrained backbone remains frozen. DexYCB is used only for training and is disjoint from all evaluation datasets.

F.3 Training Objective

Interaction-weighted supervision. The joint hand–object target is encoded by the ShapeVAE of Hunyuan3D-2.1 as a sequence of surface tokens. We reserve the first 779 tokens for one-to-one correspondences with the 779 vertices of our capped MANO hand mesh (Romero et al., 2017), and sample the remaining tokens from the joint surface. A hand anchor is marked as a contact anchor when its distance to the ground-truth object surface is below 3 mm. The final JIG run uses $\lambda_h = 2.35$ and $\lambda_c = 6.0$ in Eq. 2; other surface queries have weight one. The weighted loss therefore emphasizes hand and contact geometry directly in latent velocity space.

F.4 Scaffold Segmentation

Segmentation model. We design and train a binary hand/object geometric segmentation model for partitioning the generated scaffold. Figure 6 shows its inference path: the input HOI mesh is downsampled to 16,384 surface points with coordinates and normals, and a multiscale Point Transformer V3 (Wu et al., 2024) encoder extracts geometric features. Query points q are then processed by a continuous decoder that combines a local residual field with a graph-based correction module and outputs a hand/object probability field. For supervision, we select ObMan interactions and perturb their MANO hands within bounded ranges of pose, shape, and local geometry. These perturbations mimic the fact that the hand generated by JIG is only an approximation to a standard MANO hand. We then form the perturbed hand–object union; because the two source surfaces are known, each training query has a hand/object component label. We train on 47,769 such variants and hold out 6,222 for validation. A probability threshold of 0.5 yields the final hand/object partition.

F.5 Cross-backbone Adaptation

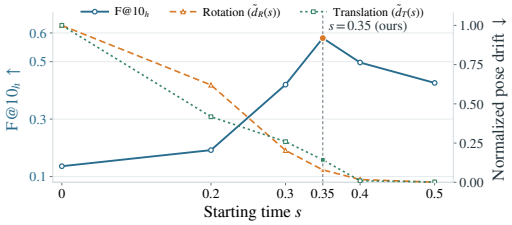
The Step1X-3D (Li et al., 2025) experiment uses its official geometry weights and VAE. We first train the image-only model for one epoch and then train for seven further epochs after adding the front/far rendering conditions. The data, sample split, camera inputs, ground-truth union meshes, and evaluation protocol match the Hunyuan3D-2.1 experiment. The adapted Step1X-3D checkpoint then receives the same DexYCB refinement as the Hunyuan3D-2.1 variants; the frozen image-only reference is not refined.

Table 10: **Scaffold retention: balancing scaffold preservation and geometric refinement.** Rotation and translation measure pose drift relative to the scaffold; the right panel visualizes normalized drift and F@10$_h$. The left panel reports the corresponding absolute values.

(a) Absolute metrics

Scaffold retention s	Rotation (°) ↓	Translation (cm) ↓	F@10$_h$ ↑
$s = 0$ (pure noise)	60.69	5.17	0.136
$s = 0.2$	37.64	2.16	0.192
$s = 0.3$	12.35	1.34	0.420
$s = 0.35$ (main setting)	4.79	0.74	**0.582**
$s = 0.4$	1.07	0.05	<u>0.497</u>
$s = 0.5$	0.03	0.01	0.426

(b) Trend visualization



For the original-backbone replacement experiment, we instead insert the unadapted Hunyuan3D-2.1 or Step1X-3D generator at the scaffold-generation stage and keep the remaining JIGAR pipeline fixed, including segmentation, object regeneration, hand registration, and output export. We evaluate the complete pipeline on the fixed 500-image HOT3D list. The original Hunyuan3D-2.1 and Step1X-3D variants produce 257 and 403 valid final reconstructions, respectively, whereas the fine-tuned JIGAR pipeline produces 479. The metrics in Table 5 are computed over each variant's valid outputs. For adapted backbones, hand registration failures are rare. In contrast, when the original backbones are used without adaptation, their scaffold hands can be severely distorted relative to the standard MANO shape, causing frequent downstream registration failures and accounting for much of the reduced valid-output coverage. Thus, these results measure the joint effect of backbone adaptation on reconstruction quality and end-to-end pipeline coverage.

G Additional Analysis

This appendix isolates the trade-off between preserving the scaffold's spatial arrangement and allowing the generative prior to refine object geometry. It complements the component and formulation ablations in the main paper.

G.1 Scaffold Retention and Transform Drift

We measure how regeneration changes the scaffold-to-object transform by varying the scaffold retention coefficient s. Smaller s introduces more noise into the scaffold latent and permits larger geometric changes; the effective noise fraction is $1 - s$. The pure-noise setting ($s = 0$) starts generation at $t = 0$ and integrates to $t = 1$. For each sample, we uniformly sample 3,000 points from the scaffold object using the same surface-sampling procedure as the main CD evaluation. We register these points to the regenerated object mesh with one-sided similarity ICP, using the nearest points on the target mesh as correspondences. This avoids penalizing newly completed surfaces in the reverse direction. The estimated transform is

$$x' = aR(x - c) + c + d,$$

where c is the scaffold point-cloud center, R is the rotation, a is the scale ratio, and d is the translation drift. Table 10 reports the rotation angle of R and the magnitude of d, averaged over the evaluated samples, alongside hand-aligned reconstruction accuracy.

At $s = 0.35$, JIGAR achieves the highest F@10$_h$ among the tested settings, with an estimated rotation drift of $4.79°$ and translation drift of 0.74 cm. Increasing s beyond this value reduces transform drift but lowers reconstruction accuracy, supporting a balance between preserving the coarse spatial arrangement and allowing geometric refinement.

H Failure Cases and Discussion

Figure 7 visualizes three representative failure modes. In the first row, the inpainting model supplies an incorrect appearance for the occluded object, which changes the visual evidence available to scaffold generation. In the second row, an incorrect hand chirality from the hand estimator produces

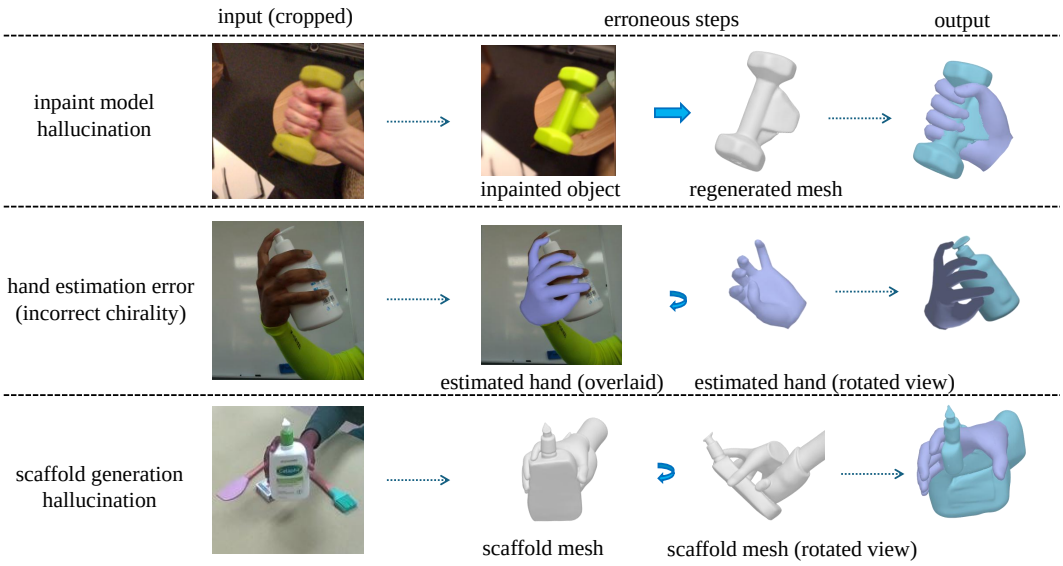


Figure 7: **Representative failure cases of JIGAR.** Each row shows a cropped input, an erroneous intermediate step, and the final output; hands are shown in purple and objects in cyan. The rows illustrate inpainting hallucination, incorrect hand chirality from hand estimation, and hallucinated scaffold geometry, respectively. Errors introduced before or during scaffold generation can therefore propagate to the recovered assets.

a geometrically inconsistent hand condition and propagates to the reconstructed interaction. Errors in upstream hand estimation, segmentation, or image completion can propagate to downstream reconstruction, as illustrated by the first two rows.

The third row shows a scaffold-generation failure in which the adapted generative prior hallucinates or smooths object structure. This is a consequence of the finite prior and the limited real-data refinement, rather than a failure of the asset-separation operation itself. JIGAR uses the scaffold primarily to establish coarse relative configuration and delegates final object detail to the pretrained object prior, so improved JIGAR can benefit from stronger 3D generation models without changing its formulation.

JIGAR does not impose an explicit physical simulator or non-penetration constraint. The interaction-weighted loss gives larger penalties to hand and contact anchors and thus provides an implicit contact-aware bias. This choice is compatible with the pretrained 3D generator but cannot guarantee physically valid contact in every case. Consistent with this tradeoff, JIGAR achieves the best SD and competitive IV in Table 3, although another method can be better on an individual robustness metric.