InstantHand: Feed-Forward 3D Gaussian Splatting for Animatable Hand Avatar Reconstruction from a Single Image

Chanho Kim **Muhammad Salman Ali** **Kyeonghwan Gwak** **Muhammad Shaheryar**
Incheol Park **Jeongwan On** **Seungryul Baek***
UNIST, South Korea
{coolho1129, salmanali, khgwak, shaheryar,
inchulnim, jeongwan.on, srbaek}@unist.ac.kr
Project page: https://coolho1129.github.io/InstantHand-page

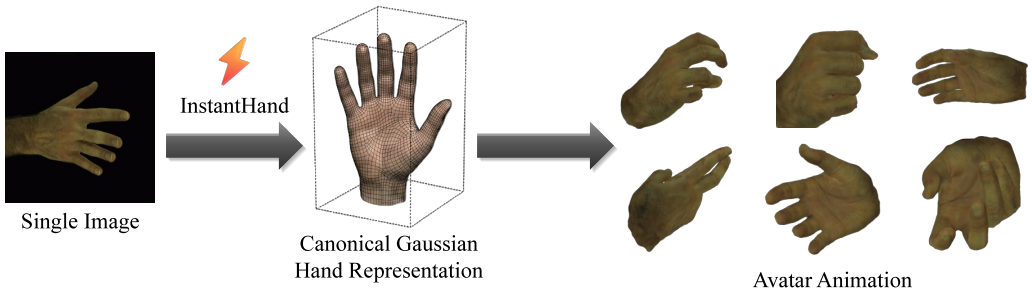


Figure 1: Single-image Feed-forward Gaussian Hand Avatar Reconstruction. InstantHand constructs an animatable 3D Gaussian hand avatar from a single image without per-subject optimization.

Abstract

Hand avatar reconstruction typically relies on multi-view capture or per-subject optimization, limiting its scalability. We introduce *InstantHand*, a feed-forward framework that constructs an animatable 3D Gaussian hand avatar from a single image using off-the-shelf MANO, camera, and mask estimates, without test-time optimization. Image-averaged losses provide insufficient supervision for alpha leakage, which thickens thin finger contours. To address this issue, we propose a *Soft Alpha Band* regularizer that concentrates supervision on silhouette boundaries. Furthermore, colors regressed from image features can drift toward a learned prior even in observed regions. Our *Reliability-gated Appearance Transfer* preserves observed appearance by transferring source colors according to each Gaussian's rendering contribution, while completing unobserved regions using the learned prior conditioned on global source identity. InstantHand constructs an avatar in 2.84 seconds and renders at 161 FPS, matching per-subject-optimized baselines on InterHand2.6M and remaining competitive in zero-shot evaluation on HanCo.

1 Introduction

The rapid evolution of digital human technologies is transforming immersive interaction across augmented and virtual reality, telepresence, gaming, and social communication. This progress has been driven by the combination of parametric human models, such as FLAME [1] for faces, SMPL [2]

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

and SMPL-X [3] for bodies, and MANO [4] for hands, with high-fidelity neural rendering techniques including NeRF [5] and 3D Gaussian Splatting (3DGS) [6]. Among these components, hands play a central role as the primary interface for fine-grained interaction and dexterous manipulation [7–11]. However, despite their importance, creating personalized, animatable, and photorealistic hand avatars from minimal input remains a key challenge for scalable digital human systems. While substantial progress has been made in single-image 3D avatar reconstruction for head avatars [12–14] and full-body avatars [15–18], high-fidelity hand avatar modeling from a single image remains comparatively underexplored and particularly challenging. Hands are highly articulated, thin, and frequently self-occluded [7, 8, 19], which makes reconstruction from sparse observations inherently ill-posed. Existing hand avatar pipelines address this ill-posedness through additional observations or per-subject optimization. Most methods rely on multi-view capture or monocular video with subject-specific fitting [20–22]; single-image approaches still require optimization-based inversion at inference [23]; and Gaussian-based methods remain dependent on multi-view or video-based reconstruction [24]. Consequently, scalable single-image hand avatar reconstruction without test-time optimization remains largely unresolved.

To eliminate the dependence on dense observations and costly test-time optimization, we aim to directly predict an animatable hand avatar from a single image, given off-the-shelf MANO, camera, and mask estimates, in a feed-forward manner. We adopt 3DGS [6, 25] as the output representation for its efficient rendering and controllable geometry and appearance. However, feed-forward Gaussian avatar methods developed for heads [14] and bodies [17] do not transfer directly to hands. First, hands are thin, low-volume, and highly articulated, so even minor errors in Gaussian position, scale, or orientation produce over-expanded or overlapping primitives, thickness artifacts, and degraded fidelity under novel poses. Second, appearance from a single image is ambiguous: existing feed-forward pipelines [14, 17] regress appearance directly from image features and tend to fall back on learned priors even in observed regions, weakening input-specific cues such as skin tone and fine texture, while unobserved regions must still be completed plausibly. Feed-forward hand avatar reconstruction therefore requires (i) geometry stabilization for thin, articulated structures and (ii) source-conditioned appearance modeling that preserves reliable input evidence while completing missing regions from priors.

To address these challenges, we propose InstantHand (Fig. 1), a feed-forward 3DGS framework for single-image animatable hand avatar reconstruction. For geometry, we constrain Gaussian placement, scale, and orientation around MANO anchors and supervise rendered alpha on the silhouette band to keep thin fingers compact. For appearance, we introduce Reliability-gated Appearance Transfer, where Gaussians supported by input observations are selectively enhanced with reliable local appearance cues to preserve input-specific details such as skin tone and fine textures. For unobserved regions, appearance is inferred using the learned prior together with a global identity representation, enabling consistent completion under occlusion. Our contributions can be summarized as follows:

- We present *InstantHand*, a feed-forward framework that reconstructs an animatable canonical 3D Gaussian hand avatar from a single image without per-subject optimization.
- We propose a Soft Alpha Band (SAB) regularizer that concentrates alpha supervision on silhouette boundaries, reducing alpha leakage and preserving thin finger structures.
- We introduce Reliability-gated Appearance Transfer, which preserves observed appearance based on each Gaussian's rendering contribution while completing unobserved regions using the learned prior and a global source identity.
- InstantHand achieves reconstruction quality comparable to per-subject-optimized baselines and competitive zero-shot performance, with avatar construction in 2.84 s and rendering at 161 FPS.

2 Related Work

Hand Pose and Shape Estimation. Monocular 3D hand pose and mesh estimation provides the geometric foundation for many hand-centric vision systems [26–37]. These methods either predict mesh vertices directly or regress the parameters of the MANO model [4] from RGB images [26, 27]. For example, METRO [29] and Mesh Graphormer [30] use transformers to regress the vertices of a MANO-topology hand mesh from a single image. More recent methods, including HaMeR [34] and WiLoR [37], regress MANO parameters together with camera parameters from unconstrained images.

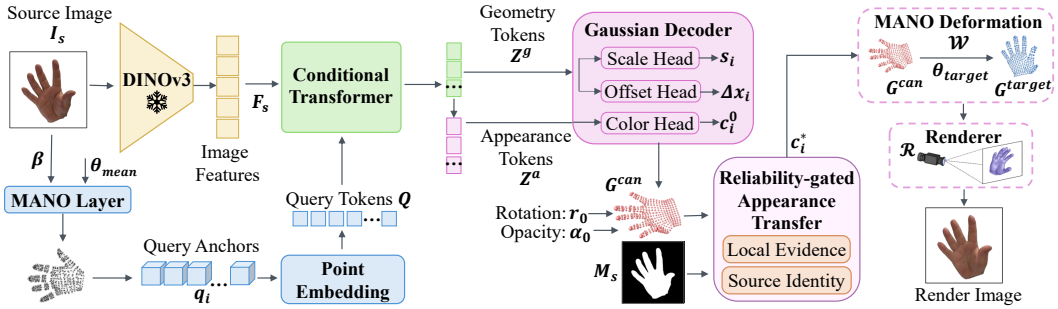


Figure 2: Overview of InstantHand. A frozen DINOv3 encoder provides image features $\mathbf{F}_s$ that serve as keys and values for a conditional transformer, whose query tokens are point-embedded MANO-surface anchors obtained from the MANO layer $(\boldsymbol{\beta}, \boldsymbol{\theta}_{\text{mean}})$. The resulting geometry and appearance tokens are decoded by scale, offset, and color heads into a canonical Gaussian hand $\mathbf{G}^{can}$ with fixed rotation $\mathbf{r}_0$ and opacity α_0. Reliability-gated Appearance Transfer then composes local source evidence and a global source identity under the source mask M_s to update only the Gaussian colors $(\mathbf{c}_i^0 \rightarrow \mathbf{c}_i^*)$, and the avatar is animated by MANO-driven deformation $\mathcal{W}$ and rendered by $\mathcal{R}$.

Our work takes off-the-shelf pose, camera, and mask estimates as input and focuses on reconstructing and rendering an animatable hand avatar.

Hand Avatar Reconstruction. Hand avatar reconstruction aims to recover subject-specific geometry and appearance for rendering under novel viewpoints and poses. Most existing approaches rely on multi-view images or monocular video sequences and employ neural implicit representations [5, 38] or explicit mesh-based pipelines [20–22, 39–41]. Representative methods include HandAvatar [20], HARP [22], and LiveHand [21], which adopt implicit, mesh-based, and multi-view neural rendering formulations, respectively. To reduce dependence on dense observations, single-image approaches such as OHTA [23] incorporate learned priors for one-shot reconstruction, but still require test-time fitting or inversion. Gaussian-based methods, including GaussianHand [24], use explicit 3D Gaussian primitives for efficient rendering. As summarized in Table S5 of the appendix, existing methods require either dense observations or per-subject optimization, regardless of representation. In contrast, our method constructs an animatable Gaussian hand avatar from a single image without test-time optimization.

Feed-Forward 3D Gaussian Avatar Reconstruction. Feed-forward avatar methods predict the parameters of animatable 3D Gaussians from one or a few images, avoiding per-subject optimization. This approach has been applied to full- and upper-body avatars by LHM [17] and GUAVA [42], and to head avatars by GAGAvatar [13] and LAM [14]. However, reconstructing an animatable hand avatar from a single image remains challenging, as thin finger boundaries are sensitive to alpha leakage and the source view provides incomplete appearance information (Sec. 1; Appendix D). We address these challenges with a Soft Alpha Band regularizer that targets boundary leakage and Reliability-gated Appearance Transfer that combines observed appearance with a learned prior and a global source identity.

3 Method

Given a single source image I_s, InstantHand reconstructs an animatable 3D hand in a single forward pass (Fig. 2). Source-image features are fused with MANO-surface query anchors to form per-Gaussian tokens (Sec. 3.1), which parameterize a canonical Gaussian representation with surface-anchored centers, learned scales, and base colors, while opacity and canonical rotation remain fixed (Sec. 3.2). Reliability-gated Appearance Transfer then refines only the Gaussian colors using contribution-aware local evidence and a global source appearance identity (Sec. 3.3). The resulting Gaussians are animated via MANO-driven deformation and rendered from arbitrary viewpoints. The two-stage training objective is described in Sec. 3.4.

3.1 Image–MANO Fusion for Geometry-aligned Tokens

Image features. We extract dense visual features from the source image using a frozen pretrained DINOv3 encoder [43]. Intermediate features are aggregated through a lightweight fusion head to

produce image tokens:

$$\mathbf{F}_s = E_{\text{img}}(I_s) \in \mathbb{R}^{N_f \times d_{\text{img}}}, \tag{1}$$

where N_f denotes the number of image tokens and d_{img} their dimension; each token captures localized visual evidence, and the resulting tokens are linearly projected to the transformer dimension d when used as keys and values.

MANO-surface query anchors. To anchor Gaussian primitives on an anatomically valid hand surface, we instantiate canonical query anchors from the MANO layer. Given the shape parameter $\boldsymbol{\beta}$ and the mean pose $\boldsymbol{\theta}_{\text{mean}}$, we define the query anchors as vertices of a subdivided canonical mesh and map them to query tokens via a point-embedding layer ϕ_q:

$$\{\mathbf{q}_i\}_{i=1}^{N} = \text{Vert}\left(\mathcal{S}_k\left(\text{MANO}(\boldsymbol{\beta}, \boldsymbol{\theta}_{\text{mean}})\right)\right), \quad \mathbf{Q} = \phi_q\left(\{\mathbf{q}_i\}_{i=1}^{N}\right) \in \mathbb{R}^{N \times d}. \tag{2}$$

Here, $\mathcal{S}_k$ denotes k-level mesh subdivision, with $k = 0$ corresponding to the original MANO mesh; we use $k = 3$ ($N = 49{,}281$; ablated in Appendix F.1). These MANO-surface query anchors provide a pose-normalized and anatomically structured prior for Gaussian primitives, constraining the reconstruction space before incorporating image evidence.

Geometry and appearance tokens. We fuse MANO-surface query tokens with image tokens using a conditional transformer (Fig. 2), implemented as two sequential streams Φ_g and Φ_a and inspired by recent multi-stream transformer designs [44]. Query tokens encode canonical surface anchors, while image tokens serve as keys and values that provide dense visual evidence from the source image. Given query tokens $\mathbf{Q}$ and image tokens $\mathbf{F}_s$, we compute geometry and appearance tokens as:

$$\mathbf{Z}^g = \Phi_g(\mathbf{Q}; \mathbf{F}_s), \qquad \mathbf{Z}^a = \Phi_a(\mathbf{Z}^g; \mathbf{F}_s), \tag{3}$$

where $\mathbf{Z}^g = \{\mathbf{z}_i^g\}_{i=1}^N$ and $\mathbf{Z}^a = \{\mathbf{z}_i^a\}_{i=1}^N$, with $\mathbf{Z}^g, \mathbf{Z}^a \in \mathbb{R}^{N \times d}$. The geometry tokens remain anchored to the canonical surface while aggregating local visual context, whereas the appearance tokens re-attend to image features under the same MANO-surface indexing, allowing fine-grained appearance cues to be captured independently of the geometry stream.

3.2 Canonical Gaussian Representation

Given the geometry and appearance tokens, we construct a canonical Gaussian hand representation:

$$\mathbf{G}^{can} = \left\{\left(\mathbf{x}_i, \mathbf{s}_i, \alpha_0, \mathbf{r}_0, \mathbf{c}_i^0\right)\right\}_{i=1}^{N}, \tag{4}$$

where $\mathbf{x}_i$ denotes the Gaussian center, $\mathbf{s}_i = (s_i, s_i, s_i)^\top$ is an isotropic scale, α_0 is a fixed opacity, $\mathbf{r}_0$ is a fixed canonical rotation, and $\mathbf{c}_i^0$ is the base RGB color. Rather than predicting unconstrained Gaussian parameters, we parameterize each Gaussian relative to a MANO-surface anchor $\mathbf{q}_i$ and decode the tokens with a lightweight Gaussian decoder comprising three prediction heads (Fig. 2): an offset head f_Δ and a scale head f_s applied to the geometry token $\mathbf{z}_i^g$, and a color head f_{rgb} applied to the appearance token $\mathbf{z}_i^a$:

$$\Delta\mathbf{x}_i = f_\Delta(\mathbf{z}_i^g), \quad s_i = \rho(f_s(\mathbf{z}_i^g)), \quad \mathbf{s}_i = (s_i, s_i, s_i)^\top, \quad \mathbf{x}_i = \mathbf{q}_i + \Delta\mathbf{x}_i, \quad \mathbf{c}_i^0 = f_{\text{rgb}}(\mathbf{z}_i^a). \tag{5}$$

Here, $\rho(\cdot)$ enforces a positive scalar scale, and $\Delta\mathbf{x}_i$ is constrained to remain within a local neighborhood of the anchor. This surface-anchored residual formulation keeps Gaussians close to the canonical MANO surface while allowing controlled deviations beyond the mesh. To reduce ambiguity from a single image, we fix opacity and canonical rotation and restrict each Gaussian to a learned isotropic scale, which removes unstable interactions among opacity, scale, and orientation and prevents the model from compensating for reconstruction errors through arbitrary Gaussian densities, rotations, or anisotropic stretching (ablation in Appendix F.2). The predicted base color $\mathbf{c}_i^0$ is geometry-aligned, but since it is inferred from compact tokens under reconstruction supervision, it may miss source-specific cues such as skin tone and fine texture, so we treat it as a stable anchor and refine it with explicit source-view evidence in Sec. 3.3.

3.3 Reliability-gated Appearance Transfer

As illustrated in Fig. 3, Reliability-gated Appearance Transfer adapts the canonical appearance to the source image while preserving the underlying 3D geometry. The key idea is to transfer source appearance according to how reliably each Gaussian is observed, using its rendering contribution as a soft reliability cue. Local updates therefore focus on Gaussians that contribute to the source-view rendering and are supported by the foreground mask, while a global source appearance identity token

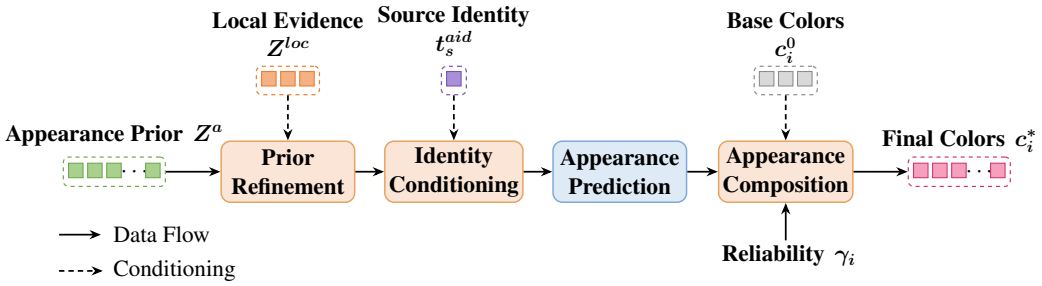


Figure 3: Reliability-gated Appearance Transfer. The Stage-1 appearance prior is refined with Gaussian footprint-aligned local evidence (Prior Refinement) and conditioned on the global source identity, which also generates a source-conditioned prior bank (Identity Conditioning). The composed representation is decoded into an observed color, an appearance gate, and a source-level shift (Appearance Prediction), and Appearance Composition blends these with the base colors under the reliability γ_i to produce the final Gaussian colors $\mathbf{c}_i^*$. Dashed arrows denote conditioning. Three further inputs of Eq. 11 are omitted for clarity. The source identity also enters Prior Refinement, and the local evidence and base colors also enter Identity Conditioning.

propagates image-level attributes, such as skin tone and illumination bias, across the full Gaussian set, including unobserved regions. The module comprises four steps: Prior Refinement, Identity Conditioning, Appearance Prediction, and Appearance Composition.

Gaussian contribution-aware source observation. We first animate the canonical Gaussian hand to the source pose and render it from the source camera:

$$\mathbf{G}^s = \mathcal{W}\left(\mathbf{G}^{can}; \boldsymbol{\theta}_s, \boldsymbol{\beta}\right), \qquad \left(\hat{I}_s, \{\Omega_i\}_{i=1}^N, \{C_i\}_{i=1}^N\right) = \mathcal{R}\left(\mathbf{G}^s; \mathbf{K}_s, \mathbf{T}_s\right). \tag{6}$$

Here, $\hat{I}_s$ denotes the rendered source view, Ω_i is the projected footprint of the i-th Gaussian, and C_i is its accumulated rendering contribution. This source-view rendering establishes a Gaussian-to-image correspondence and provides a soft measure of each Gaussian's participation in forming source-view pixels. Rather than relying on a single projected pixel, we extract local evidence using radius-adaptive Gaussian footprint aggregation. For an image or feature map X and an optional support mask M, we define:

$$\mathrm{Agg}_{\Omega_i}\left(X; M\right) = \frac{\sum_{\mathbf{p}\in\Omega_i} w_i(\mathbf{p}) M(\mathbf{p}) X(\mathbf{p})}{\sum_{\mathbf{p}\in\Omega_i} w_i(\mathbf{p}) M(\mathbf{p}) + \epsilon}. \tag{7}$$

Here, $w_i(\mathbf{p})$ denotes a Gaussian kernel centered at the projected center $\mathbf{u}_i$, with bandwidth determined by the projected radius. Setting $M = \mathbf{1}$ yields unmasked aggregation. This footprint-weighted sampling aligns the extracted RGB and feature evidence with the projected support of each Gaussian. We estimate the reliability of each projected Gaussian by combining its rendering contribution with source foreground support. Specifically, $m_i = \mathrm{Agg}_{\Omega_i}(M_s; \mathbf{1})$ measures foreground support within the projected footprint. We then define the contribution-aware observation reliability, which gates subsequent appearance transfer, as:

$$\gamma_i = \min\left(\frac{\sqrt{C_i}\, m_i}{\mathrm{P}_\zeta\left(\{\sqrt{C_j}\, m_j\}_{j\in\mathcal{O}}\right)}, 1\right) \cdot \mathbf{1}[\, i \in \mathcal{O}\,], \qquad \mathcal{O} = \{\, j \,:\, C_j > \kappa\, \mathrm{med}(C_+),\ m_j > \tau_m \,\}. \tag{8}$$

Here, $\mathcal{O}$ denotes the observed set, $\mathrm{med}(C_+)$ the median of nonzero rendering contributions, and $\mathrm{P}_\zeta(\cdot)$ the ζ-th percentile operator. We fix $(\kappa, \tau_m, \zeta) = (0.5, 0.5, 95)$ a priori without tuning on evaluation data. The reliability γ_i is computed deterministically and is not learned; performance is insensitive to these constants (Appendices A.4 and F.4). A Gaussian receives high reliability only when it both contributes to the source-view rendering and is supported by the foreground mask. Unlike the binary visible/invisible split that FlexiAvatar [45] uses to restrict per-subject optimization to observed body regions, γ_i provides graded reliability within $\mathcal{O}$, with a bounded discontinuity at its boundary due to the indicator function.

Contribution-aware local evidence and global source appearance identity. For each Gaussian, we aggregate source evidence over its projected footprint:

$$\bar{\mathbf{c}}_i^{src} = \mathrm{Agg}_{\Omega_i}\left(I_s; M_s\right), \qquad \bar{\mathbf{c}}_i^{ren} = \mathrm{Agg}_{\Omega_i}\left(\hat{I}_s; M_s\right), \qquad \mathbf{f}_i^x = \mathrm{Agg}_{\Omega_i}\left(\mathbf{F}_s^x; M_s\right). \tag{9}$$

Here, $\mathbf{F}_s^x$ denotes local image features used for footprint aggregation. The local evidence token is encoded as:

$$\mathbf{z}_i^{loc} = f_{loc}\left([\bar{\mathbf{c}}_i^{src}, \bar{\mathbf{c}}_i^{ren}, \bar{\mathbf{c}}_i^{src} - \bar{\mathbf{c}}_i^{ren}, \mathbf{f}_i^x, \gamma_i]\right). \tag{10}$$

We denote the set of local evidence tokens as $\mathbf{Z}^{loc} = \{\mathbf{z}_i^{loc}\}_{i=1}^N$. The residual term captures the discrepancy between the current Gaussian appearance and the source observation under consistent source-view alignment, while γ_i carries the reliability of the observation. Because a single source image observes only part of the hand, we aggregate reliable source information into a global source appearance identity token $\mathbf{t}_s^{aid}$ using masked global and local image features, source and canonical base color statistics, and reliability-weighted local residuals and evidence tokens (Appendix A.5). The token encodes source-level appearance properties, including skin tone, illumination bias, and global color distribution, without describing geometry.

Source-conditioned appearance and color composition. We combine three complementary sources of appearance information: local evidence, the Stage-1 appearance prior $\mathbf{Z}^a$, and the global source appearance identity. The Stage-1 appearance prior is updated using local evidence and the global identity (Prior Refinement), while the identity token is expanded into a source-conditioned prior bank for weakly observed regions (Identity Conditioning):

$$\tilde{\mathbf{Z}}^a = \mathcal{U}(\mathbf{Z}^a; \mathbf{Z}^{loc}, \mathbf{t}_s^{aid}), \qquad \mathcal{D}_s = \mathcal{B}(\mathbf{t}_s^{aid}), \quad \mathbf{h}_i = \mathcal{U}_c(\mathbf{z}_i^{loc}, \tilde{\mathbf{z}}_i^a, \mathbf{t}_s^{aid}, \mathbf{c}_i^0, \mathcal{D}_s). \tag{11}$$

This composition enables Gaussians that contribute to rendering to leverage spatially aligned local evidence, while weakly observed or unobserved Gaussians are guided by the base colors and the global source appearance identity. From the composed appearance token, the module predicts a source-observed color candidate, a learnable appearance gate, and an appearance displacement; the displacements are pooled into a source-level appearance shift (Appearance Prediction):

$$\left(\mathbf{c}_i^{obs}, g_i, \boldsymbol{\delta}_i\right) = f_{out}\left(\mathbf{h}_i\right), \qquad g_i \in [0,1], \qquad \boldsymbol{\Delta}_s^{aid} = \mathrm{Pool}\left(\{\boldsymbol{\delta}_i\}_{i=1}^N\right). \tag{12}$$

The predicted color is not directly assigned to the Gaussian. Instead, the observation reliability modulates the reliance on local source evidence (Appearance Composition):

$$\mathbf{c}_i^* = g_i\left[\gamma_i \mathbf{c}_i^{obs} + (1-\gamma_i)\,\mathbf{c}_i^0\right] + (1-g_i)\,\mathbf{c}_i^0 + \boldsymbol{\Delta}_s^{aid}. \tag{13}$$

Expanding Eq. 13 gives $\mathbf{c}_i^* = \mathbf{c}_i^0 + g_i\gamma_i(\mathbf{c}_i^{obs} - \mathbf{c}_i^0) + \boldsymbol{\Delta}_s^{aid}$, so γ_i and g_i act on the source evidence only through their product; they are nevertheless not interchangeable, since γ_i is closed-form and receives no gradient whereas g_i is learned from the evidence (Appendix A.4). This design separates local evidence transfer from global appearance propagation: local evidence enters only through the reliability-weighted term, whereas global appearance cues are consistently propagated across all Gaussians via the source-level shift. The composed colors $\mathbf{c}_i^*$ replace $\mathbf{c}_i^0$ in $\mathbf{G}^{can}$, and all non-color attributes are preserved. Novel-pose and novel-view rendering then deform $\mathbf{G}^{can}$, now carrying $\mathbf{c}_i^*$, with the same MANO-driven deformation and Gaussian renderer:

$$\mathbf{G}^{target} = \mathcal{W}\left(\mathbf{G}^{can}\big|_{\mathbf{c}_i^0 \leftarrow \mathbf{c}_i^*}; \boldsymbol{\theta}_{\mathrm{target}}, \boldsymbol{\beta}\right), \qquad \hat{I}_{\mathrm{target}} = \mathcal{R}\left(\mathbf{G}^{target}; \mathbf{K}_{\mathrm{target}}, \mathbf{T}_{\mathrm{target}}\right). \tag{14}$$

Since only the colors of $\mathbf{G}^{can}$ are replaced, the rendering geometry remains governed by the MANO-surface-anchored canonical representation and MANO-driven deformation, while the appearance is determined by the composed colors. Since the predicted colors are independent of viewing direction and driving pose, the model does not capture view- or pose-dependent effects such as self-shadowing. We discuss this limitation and potential extensions to view-dependent appearance modeling in Appendix G.2.

3.4 Training Objective

We train the model in two stages. The first stage learns a stable canonical Gaussian hand representation, while the second stage focuses on source-conditioned appearance modeling with the entire Stage-1 canonical predictor, including the base colors $\mathbf{c}_i^0$, kept fixed.

Canonical Gaussian training. In the first stage, we supervise rendered images using mask-aware photometric and perceptual losses, together with regularization on the canonical Gaussian geometry:

$$\begin{aligned} &\mathcal{L}_{\mathrm{can}} = \mathcal{L}_{\mathrm{rec}} + \lambda_{\mathrm{sil}}\mathcal{L}_{\mathrm{sil}} + \mathcal{L}_{\mathrm{geo}} + \mathcal{L}_{\mathrm{band}}, \\ &\mathcal{L}_{\mathrm{rec}} = \lambda_{\mathrm{rgb}}(\mathcal{L}_1 + \lambda_{\mathrm{lpips}}\mathcal{L}_{\mathrm{lpips}}), \quad \mathcal{L}_{\mathrm{geo}} = \lambda_{\mathrm{off}}\mathcal{L}_{\mathrm{off}} + \lambda_{\mathrm{lap}}\mathcal{L}_{\mathrm{lap}} + \lambda_{\mathrm{scale}}\mathcal{L}_{\mathrm{scale}}. \end{aligned} \tag{15}$$

Here, $\mathcal{L}_{\mathrm{rec}}$ enforces image reconstruction fidelity, $\mathcal{L}_{\mathrm{sil}}$ supervises silhouette consistency, and $\mathcal{L}_{\mathrm{geo}}$ regularizes residual Gaussian displacements, local smoothness, and Gaussian scales.

Table 1: Quantitative comparison on InterHand2.6M. Best, second-best, and third-best are marked **bold green**, underlined blue, and *italic orange*.

Method	Protocol	PSNR↑	SSIM↑	LPIPS↓
HARP [22]	Per-subj. opt.	19.82	0.761	0.225
LiveHand [21]	Per-subj. opt.	23.01	0.763	0.186
HandAvatar [20]	Per-subj. opt.	23.79	*0.820*	0.178
Handy [46]	Per-subj. opt.	*25.56*	0.794	*0.150*
OHTA [23]	Per-subj. opt.	26.11	0.864	**0.129**
BIGS* [47]	Per-subj. opt.	20.46	0.749	0.212
Ours	**Feed-forward**	**26.21**	**0.866**	0.133

Table 2: Single-image reconstruction and efficiency on HanCo. Time is the avatar-acquisition cost; FPS and memory are measured for 256×256 rendering (protocol in Appendix C.6). Best in **bold**.

Method	Protocol	Reconstruction			Efficiency		
		PSNR↑	SSIM↑	LPIPS↓	Time (s)↓	FPS↑	Mem. (GB)↓
HandAvatar	Per-subj. opt.	**20.32**	0.841	0.198	5579	0.32	14.80
OHTA	Per-subj. opt.	20.06	**0.843**	**0.186**	3360	0.26	47.80
BIGS*	Per-subj. opt.	17.71	0.731	0.246	1380	43.11	**1.50**
Ours	**Zero-shot feed-forward**	20.18	0.836	0.188	**2.84**	**160.71**	1.79

Soft Alpha Band regularization. We design a task-specific Soft Alpha Band regularizer $\mathcal{L}_{\text{band}}$ to stabilize thin articulated fingers. Given the ground-truth silhouette M and the rendered alpha map A, we define an outer band B_{out} via dilation and an inner region B_{in} via erosion. The regularizer encourages compact yet sufficient rendered-alpha support by penalizing alpha leakage near boundaries and insufficient alpha within the silhouette interior:

$$\mathcal{L}_{\text{band}} = \eta(n)(\lambda_{\text{out}}\mathcal{L}_{\text{out}} + \lambda_{\text{in}}\mathcal{L}_{\text{in}}), \quad \mathcal{L}_{\text{out}} = \text{Avg}_{B_{\text{out}}}[\psi([A - \tau_{\text{out}}]_+)], \quad \mathcal{L}_{\text{in}} = \text{Avg}_{B_{\text{in}}}[\psi([\tau_{\text{in}} - A]_+)]. \tag{16}$$

Here, $\psi(\cdot)$ denotes a robust Huber penalty and $\eta(n)$ is a weighting coefficient scheduled over the optimization step n. The outer-band term suppresses alpha spill around finger boundaries, reducing thickened geometry artifacts, while the inner-region term maintains sufficient coverage inside the hand silhouette.

Source-conditioned appearance training. In the second stage, we freeze the entire Stage-1 canonical predictor, including the base-color head f_{rgb}, and train only the source-conditioned appearance module. The objective combines source-view reconstruction, feature-statistic alignment, and novel-view supervision:

$$\mathcal{L}_{\text{app}} = \lambda_{s1}\mathcal{L}_1^{\text{src}} + \lambda_{sp}\mathcal{L}_{\text{lpips}}^{\text{src}} + \lambda_{\text{stat}}\mathcal{L}_{\text{stat}} + \lambda_{n1}\mathcal{L}_1^{\text{novel}} + \lambda_{np}\mathcal{L}_{\text{lpips}}^{\text{novel}}. \tag{17}$$

The source-view losses enforce consistency between the rendered output and the input image; $\mathcal{L}_{\text{stat}}$ aligns foreground feature statistics between the source image and the rendered source view, and the novel-view losses maintain appearance consistency under MANO-driven deformation. All loss terms are defined in full and all loss weights of both stages are reported in Appendices A.1 and A.2.

4 Experiments

Training Data and Preprocessing. We train on InterHand2.6M [48], HO3Dv3 [49, 50], Goliath [51], Hand Appearance [22], and PALM [52]; each sample pairs a source image with multiple target frames of the same subject. For the appearance stage we add synthetic source–target pairs generated following Zhao et al. [53] to widen source appearance variation. Images are cropped to hand-centered 256×256 regions with intrinsics updated accordingly; where annotations are unavailable, masks and MANO/camera parameters are estimated with SAM 3 [54] and WiLoR [37] (details in Appendix B).

4.1 Comparisons with Prior Methods

Evaluation Protocol and Metrics. We evaluate on InterHand2.6M [48] under the single-image reconstruction protocol of prior hand avatar benchmarks [20, 23]: each method reconstructs a subject-

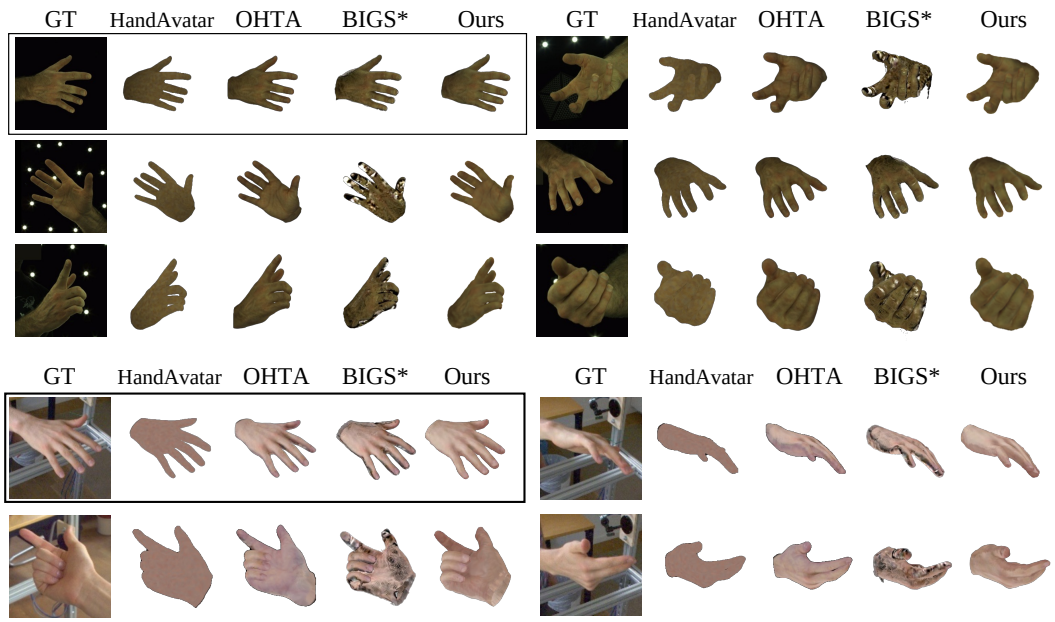


Figure 4: Qualitative comparison with prior methods on InterHand2.6M [48] (top three rows) and HanCo [55] (bottom two rows). Columns: GT, HandAvatar, OHTA, BIGS*, Ours; the black box marks the input image.

specific hand from one source image and renders novel poses and views. Zero-shot generalization is evaluated on HanCo [55] without HanCo training or test-time optimization. All methods share the same source frame and identical target poses and views; we report PSNR, SSIM, and LPIPS over hand-centered regions defined by bounding-box ray masks. The complete protocol, per-baseline input protocol, and provenance of all baseline numbers are given in Appendix C.

Quantitative Comparisons. Table 1 reports results on InterHand2.6M: without any test-time fitting, InstantHand matches the strongest per-subject-optimized baseline (26.21 vs. 26.11 dB PSNR, 0.866 vs. 0.864 SSIM) and attains the second-best LPIPS. Table 2 presents zero-shot cross-dataset evaluation and efficiency on HanCo, where the baselines are fitted to each test subject and our model is applied without HanCo training subjects or optimization; InstantHand stays within 0.14 dB PSNR, 0.007 SSIM, and 0.002 LPIPS of the best fitted baselines at over 400$\times$ lower avatar-acquisition cost. BIGS* denotes our single-hand adaptation of BIGS [47] (Appendix E.1).

Qualitative Comparisons. Figure 4 presents qualitative comparisons on InterHand2.6M and HanCo. Under the single-image setting, the fitted baselines exhibit over-smoothed geometry, distorted fingers, or texture artifacts. InstantHand retains coherent hand geometry, well-defined finger boundaries, and fine details such as wrinkles and skin texture. On HanCo, these results are obtained zero-shot, whereas the baselines are fitted to each subject. Figure 5 applies InstantHand to fully estimated inputs and to edited versions of the same source images. Because each edited image shares the mask, MANO, and camera of its original, comparing their renderings shows how the input appearance is carried to the avatar. Both the local patterns (top) and the global skin tone (bottom) follow the hand through novel poses and views. Figure S9 in the appendix shows novel-pose animations of eight unseen subjects, including the in-the-wild input of Figure 5. The supplementary video (Appendix K) further includes 360° rotations and continuous novel-pose animations.

4.2 Ablation Study

Component ablation. Table 3 reports ablations on InterHand2.6M. Removing the Soft Alpha Band regularizer reduces reconstruction quality for thin structures. Reliability gating (γ) and the learned appearance gate (g_i) play complementary roles: γ weights a Gaussian according to its source-view contribution and foreground support, while g_i predicts how strongly to use the sampled appearance evidence. Removing either produces a substantial drop. Although they appear only as a product in Eq. 13, retraining each variant does not allow the remaining gate to compensate for the removed one. Disabling appearance transfer entirely (the Stage-1 model) yields 26.01 dB, outperforming

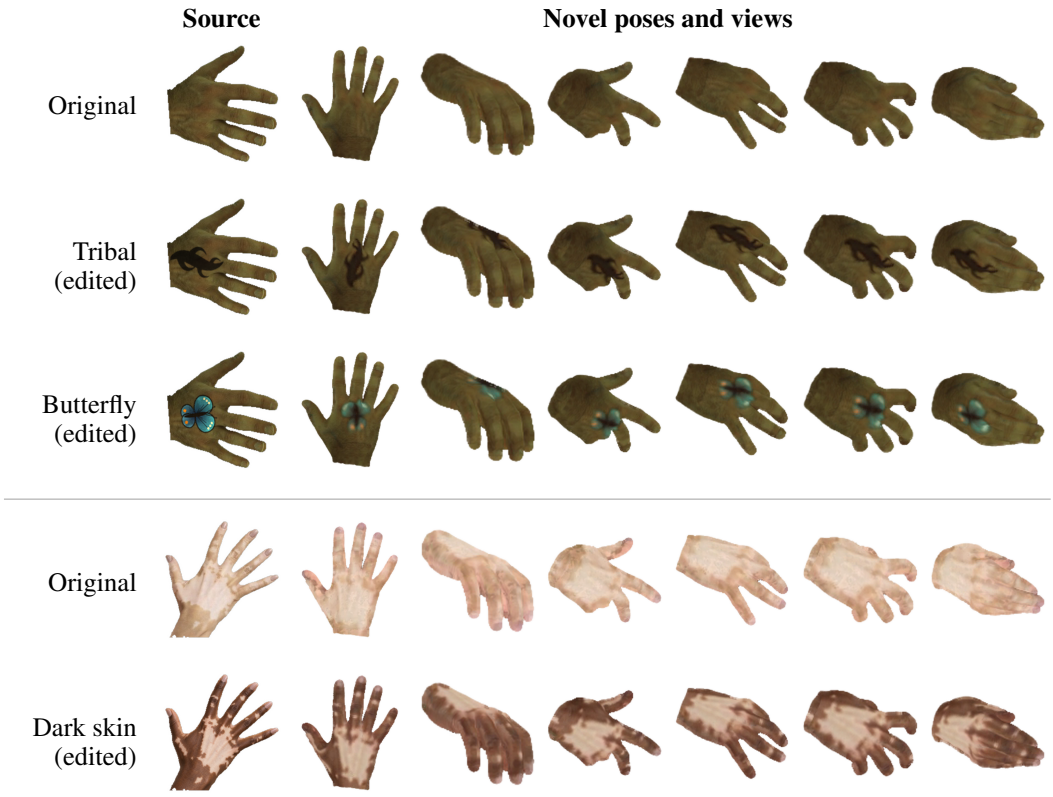


Figure 5: Avatars reconstructed from fully estimated inputs (SAM 3 mask, WiLoR MANO and camera) for original and edited source images. Top: an unseen InterHand2.6M test subject, with hand-drawn tattoos composited onto the source image. Bottom: an in-the-wild image released with OHTA [23], with its skin tone darkened by an image generation model. Edited rows reuse the mask, MANO, and camera of the original row, and each column shows the same pose and view.

Table 3: Ablation and source-cue robustness on InterHand2.6M. Ablation rows remove one component and retrain both stages. Estimated-cue rows keep the trained full model and replace its source inputs with estimates.

Setting	PSNR↑	SSIM↑	LPIPS↓
w/o SAB ($\mathcal{L}_{\text{band}}$)	23.89	0.825	0.166
w/o reliability gating (γ)	23.18	0.829	0.171
w/o appearance gate (g_i)	24.52	0.833	0.152
w/o appearance transfer	26.01	0.865	0.134
w/o synthetic data	26.13	0.864	0.134
Ours (GT cues)	**26.21**	**0.866**	**0.133**
Estimated source cues			
SAM 3 mask	26.12	0.865	0.133
WiLoR MANO	26.07	0.863	0.134
Fully estimated	25.99	0.863	0.134

both single-gate variants; thus, appearance transfer is beneficial only when both gates are present. Synthetic data provides a modest additional gain. Further analyses are provided in Appendices F and H.

Robustness to source-cue errors. Table 3 (bottom) evaluates sensitivity to estimated source inputs. We replace the ground-truth source mask with a SAM 3 estimate [54], and the source MANO parameters and camera with WiLoR estimates [37], keeping the trained model and evaluation protocol fixed. Target poses and cameras remain ground truth in every setting. When all source inputs

Table 4: Controlled comparison with feed-forward avatar recipes on InterHand2.6M. Each recipe swaps its template (FLAME, SMPL-X) for MANO and keeps its published parameterization under otherwise matched training (Appendix E.2); boundary metrics, distinct from the training bands, follow Appendix C.4.

Configuration	PSNR↑	SSIM↑	LPIPS↓	Bound. F↑	Leakage↓	Int. cov.↑
Ours (constrained + SAB)	**26.21**	**0.866**	**0.133**	**0.698**	**0.234**	0.999
Constrained, w/o SAB	23.89	0.825	0.166	0.366	0.555	1.000
Head-recipe port [14]	23.60	0.809	0.182	0.299	0.608	1.000
Body-recipe port [17]	23.36	0.805	0.182	0.267	0.627	1.000

are estimated from a single RGB image, PSNR decreases by only 0.22 dB, while SSIM and LPIPS remain essentially unchanged. The decrease arises from the mask and MANO estimates; replacing the source camera with its estimate causes no further PSNR drop. Controlled input perturbations reduce PSNR by at most 0.16 dB. Appendix F.3 provides the per-cue breakdown (Table S8) and perturbation protocol, while Appendix H.5 reports results on unseen FreiHAND images with fully estimated inputs.

Source-fidelity diagnostic. Aggregate novel-pose metrics combine source-observed and unobserved regions, obscuring the fidelity of source appearance reconstruction. To isolate the effect of the appearance module, we render each avatar in its source view while keeping all non-color Gaussian attributes fixed at their canonical values, so metric differences reflect only the color update. With each evaluation frame serving in turn as the source, the module improves source-view PSNR by 1.96 dB on InterHand2.6M and 3.41 dB on zero-shot HanCo (Table S2; protocol in Appendix C.5).

Comparison with feed-forward avatar recipes. Table 4 examines whether avatar designs developed for heads and bodies transfer to hands under matched conditions. Compared with InstantHand, the head-recipe port loses 2.61 dB, reduces the boundary F-score from 0.698 to 0.299, and increases alpha leakage from 0.234 to 0.608, while interior coverage remains at 1.000. This pattern localizes the failure to thin, articulated contours rather than the hand interior, consistent with Sec. 1. The head-recipe port also has a lower boundary F-score and greater leakage than the constrained model without SAB, indicating that both the constrained parameterization and boundary-band supervision contribute to boundary quality. Even with boundary supervision retained, freeing per-axis scale, rotation and opacity jointly, or all three reduces PSNR by 0.47–0.71 dB (Appendix F.2). The body-recipe port performs similarly to the head-recipe port, within 0.24 dB in PSNR and with a boundary F-score of 0.267 versus 0.299, despite its canonical-space regularizers.

Pose and temporal diagnostics. Grouping test frames by pose change relative to the source reveals that quality degrades primarily with finger articulation: PSNR drops by 2.69 dB on InterHand2.6M and 1.93 dB on HanCo between the mildest and strongest articulation thirds. Changes in viewing direction have a smaller effect, with drops of at most 0.97 dB. Under continuous pose driving on zero-shot HanCo, frame-to-frame changes in our renderings exceed those in the ground truth by 4.7% on average and at most 8.0% per camera. All views and frames are rendered from a single fixed set of canonical Gaussians, providing a shared representation across views; we assess cross-view and temporal consistency empirically through these diagnostics and the supplementary video. Appendix J provides the full analyses and comparisons with baselines evaluated on the same frames.

5 Conclusion

We presented *InstantHand*, a feed-forward framework for reconstructing an animatable 3D Gaussian hand avatar from a single image without per-subject optimization. A MANO-anchored canonical representation, together with the Soft Alpha Band regularizer, preserves compact, thin articulated structures, while Reliability-gated Appearance Transfer integrates observed source appearance according to each Gaussian's rendering contribution and completes unobserved regions using the learned prior and a global source identity. Once constructed, the avatar renders at 161 FPS. Its Gaussian colors and scales remain fixed across driving poses, preserving appearance consistency. InstantHand constructs avatars over 400$\times$ faster than optimization-based baselines while maintaining competitive reconstruction quality in zero-shot evaluation on HanCo and under fully estimated inputs.

Acknowledgments and Disclosure of Funding

This work is supported by NRF grants (No. RS-2025-00521013 30%, No. RS-2025-02216916 10%) and IITP grants (No. RS-2020-II201336 Artificial intelligence graduate school program (UNIST) 10%; No. RS-2025-25442824 AI Star Fellowship Program (UNIST) 10%; No. RS-2022-II220264 Comprehensive video understanding and generation with knowledge-based deep logic neural network 10%; No. RS-2025-25442149 LG AI STAR Talent Development Program for Leading Large-Scale Generative AI Models in the Physical AI Domain 10%), funded by the Korean government (MSIT). This work is also supported by the InnoCORE program of the Ministry of Science and ICT (26-InnoCORE-01 10%, UNIST 1.260006.01 10%).

References

[1] Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 2017.

[2] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), November 2015. doi: 10.1145/2816795.2818013.

[3] 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 *CVPR*, 2019.

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

[5] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023.

[7] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *Int. J. Comput. Vision*, 118(2):172–193, June 2016. doi: 10.1007/s11263-016-0895-4.

[8] 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 *CVPR*, 2019.

[9] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020.

[10] 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 *CVPR*, 2021.

[11] MD Khalequzzaman Chowdhury Sayem, Mubarrat Tajoar Chowdhury, Yihalem Yimolal Tiruneh, Muneeb A. Khan, Muhammad Salman Ali, Binod Bhattarai, and Seungryul Baek. HandVQA: Diagnosing and improving fine-grained spatial reasoning about hands in vision-language models. In *CVPR*, 2026.

[12] Xuangeng Chu, Yu Li, Ailing Zeng, Tianyu Yang, Lijian Lin, Yunfei Liu, and Tatsuya Harada. GPAvatar: Generalizable and precise head avatar from image(s). In *ICLR*, 2024.

[13] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable Gaussian head avatar. In *NeurIPS*, 2024.

[14] Yisheng He, Xiaodong Gu, Xiaodan Ye, Chao Xu, Zhengyi Zhao, Yuan Dong, Weihao Yuan, Zilong Dong, and Liefeng Bo. LAM: Large avatar model for one-shot animatable Gaussian head. In *SIGGRAPH Conference Papers*, 2025. doi: 10.1145/3721238.3730706.

[15] Yiyu Zhuang, Jiaxi Lv, Hao Wen, Qing Shuai, Ailing Zeng, Hao Zhu, Shifeng Chen, Yujiu Yang, Xun Cao, and Wei Liu. IDOL: Instant photorealistic 3D human creation from a single image. In *CVPR*, 2025.

[16] Lingteng Qiu, Shenhao Zhu, Qi Zuo, Xiaodong Gu, Yuan Dong, Junfei Zhang, Chao Xu, Zhe Li, Weihao Yuan, Liefeng Bo, Guanying Chen, and Zilong Dong. AniGS: Animatable Gaussian avatar from a single image with inconsistent Gaussian reconstruction. In *CVPR*, 2025.

[17] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, and Liefeng Bo. LHM: Large animatable human reconstruction model for single image to 3D in seconds. In *ICCV*, 2025.

[18] Geonhee Sim and Gyeongsik Moon. PERSONA: Personalized whole-body 3D avatar with pose-driven deformations from a single image. In *ICCV*, 2025.

[19] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, 2017.

[20] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand Avatar: Free-pose hand animation and rendering from monocular video. In *CVPR*, 2023.

[21] Akshay Mundra, Mallikarjun B R, Jiayi Wang, Marc Habermann, Christian Theobalt, and Mohamed Elgharib. LiveHand: Real-time and photorealistic neural hand rendering. In *ICCV*, 2023.

[22] Korrawe Karunratanakul, Sergey Prokudin, Otmar Hilliges, and Siyu Tang. HARP: Personalized hand reconstruction from a monocular RGB video. In *CVPR*, 2023.

[23] Xiaozheng Zheng, Chao Wen, Zhuo Su, Zeran Xu, Zhaohu Li, Yang Zhao, and Zhou Xue. OHTA: One-shot hand avatar via data-driven implicit priors. In *CVPR*, 2024.

[24] Lizhi Zhao, Xuequan Lu, Runze Fan, Sio Kei Im, and Lili Wang. GaussianHand: Real-time 3D Gaussian rendering for hand avatar animation. *IEEE Transactions on Visualization and Computer Graphics*, 31(9):6484–6496, 2025.

[25] Muhammad Salman Ali, Chaoning Zhang, Marco Cagnazzo, Giuseppe Valenzise, Enzo Tartaglione, and Sung-Ho Bae. Compression in 3D Gaussian splatting: A survey of methods, trends, and future directions. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

[26] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In *CVPR*, 2019.

[27] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via GAN and mesh model for estimating 3D hand poses interacting objects. In *CVPR*, 2020.

[28] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, 2020.

[29] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021.

[30] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh Graphormer. In *ICCV*, 2021.

[31] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. MobRecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, 2022.

[32] Hoseong Cho, Chanwoo Kim, Jihyeon Kim, Seongyeong Lee, Elkhan Ismayilzada, and Seungryul Baek. Transformer-based unified recognition of two hands manipulating objects. In *CVPR*, 2023.

[33] Zhishan Zhou, Shihao Zhou, Zhi Lv, Minqiang Zou, Yao Tang, and Jiajun Liang. A simple baseline for efficient hand mesh reconstruction. In *CVPR*, 2024.

[34] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024.

[35] Haoye Dong, Aviral Chharia, Wenbo Gou, Francisco Vicente Carrasco, and Fernando De la Torre. Hamba: Single-view 3D hand reconstruction with graph-guided bi-scanning mamba. In *NeurIPS*, 2024.

[36] Elkhan Ismayilzada, MD Khalequzzaman Chowdhury Sayem, Yihalem Yimolal Tiruneh, Mubarrat Tajoar Chowdhury, Muhammadjon Boboev, and Seungryul Baek. QORT-Former: Query-optimized real-time transformer for understanding two hands manipulating objects. In *AAAI*, 2025.

[37] Rolandos Alexandros Potamias, Jinglei Zhang, Jiankang Deng, and Stefanos Zafeiriou. WiLoR: End-to-end 3D hand localization and reconstruction in-the-wild. In *CVPR*, 2025.

[38] Ali Haider, Muhammad Salman Ali, Maryam Qamar, Tahir Khalil, Soo Ye Kim, Jihyong Oh, Enzo Tartaglione, and Sung-Ho Bae. I-INR: Iterative implicit neural representations. In *AAAI*, 2026.

[39] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. LISA: Learning implicit shape and appearance of hands. In *CVPR*, 2022.

[40] Zhiyang Guo, Wengang Zhou, Min Wang, Li Li, and Houqiang Li. HandNeRF: Neural radiance fields for animatable interacting hands. In *CVPR*, 2023.

[41] Pratik Kalshetti and Parag Chaudhuri. Intrinsic hand avatar: Illumination-aware hand appearance and shape reconstruction from monocular RGB video. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

[42] Dongbin Zhang, Yunfei Liu, Lijian Lin, Ye Zhu, Yang Li, Minghan Qin, Yu Li, and Haoqian Wang. GUAVA: Generalizable upper body 3D Gaussian avatar. In *ICCV*, 2025.

[43] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michael Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Herve Jegou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=2NlGyqNjns`.

[44] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

[45] Yihalem Yimolal Tiruneh, Muhammad Salman Ali, Uyoung Jeong, Muneeb A. Khan, MD Khalequzzaman Chowdhury Sayem, Allanur Bayramgeldiyev, Binod Bhattarai, and Seungryul Baek. FlexiAvatar: Unified 3D Gaussian human avatars under arbitrary body visibility. In *ECCV*, 2026.

[46] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3D hand shape and appearance model. In *CVPR*, 2023.

[47] 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 *CVPR*, 2025.

[48] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3D interacting hand pose estimation from a single RGB image. In *ECCV*, 2020.

[49] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020.

[50] Shreyas Hampali, Sayan Deb Sarkar, and Vincent Lepetit. HO-3D_v3: Improving the accuracy of hand-object annotations of the HO-3D dataset. *arXiv preprint arXiv:2107.00887*, 2021.

[51] Julieta Martinez, Emily Kim, Javier Romero, Timur Bagautdinov, Shunsuke Saito, Shoou-I Yu, Stuart Anderson, Michael Zollhöfer, Te-Li Wang, Shaojie Bai, Chenghui Li, Shih-En Wei, Rohan Joshi, Wyatt Borsos, Tomas Simon, Jason Saragih, Paul Theodosis, Alexander Greene, Anjani Josyula, Silvio Mano Maeta, Andrew I. Jewett, Simon Venshtain, Christopher Heilman, Yueh-Tung Chen, Sidi Fu, Mohamed Ezzeldin A. Elshaer, Tingfang Du, Longhua Wu, Shen-Chi Chen, Kai Kang, Michael Wu, Youssef Emad, Steven Longay, Ashley Brewer, Hitesh Shah, James Booth, Taylor Koska, Kayla Haidle, Matt Andromalos, Joanna Hsu, Thomas Dauer, Peter Selednik, Tim Godisart, Scott Ardisson, Matthew Cipperly, Ben Humberston, Lon Farr, Bob Hansen, Peihong Guo, Dave Braun, Steven Krenn, He Wen, Lucas Evans, Natalia Fadeeva, Matthew Stewart, Gabriel Schwartz, Divam Gupta, Gyeongsik Moon, Kaiwen Guo, Yuan Dong, Yichen Xu, Takaaki Shiratori, Fabian Prada, Bernardo R. Pires, Bo Peng, Julia Buffalini, Autumn Trimble, Kevyn McPhail, Melissa Schoeller, and Yaser Sheikh. Codec Avatar Studio: Paired human captures for complete, driveable, and generalizable avatars. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[52] Zicong Fan, Edoardo Remelli, David Dimond, Fadime Sener, Liuhao Ge, Bugra Tekin, Cem Keskin, and Shreyas Hampali. PALM: A dataset and baseline for learning multi-subject hand prior. In *3DV*, 2026.

[53] Zhuoran Zhao, Linlin Yang, Pengzhan Sun, Pan Hui, and Angela Yao. Analyzing the synthetic-to-real domain gap in 3D hand pose estimation. In *CVPR*, 2025.

[54] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, 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 Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *ICLR*, 2026.

[55] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *German Conference on Pattern Recognition*, 2021.

[56] Bowen Cheng, Ross Girshick, Piotr Dollár, Alexander C. Berg, and Alexander Kirillov. Boundary IoU: Improving object-centric image segmentation evaluation. In *CVPR*, 2021.

[57] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. GaussianAvatar: Towards realistic human avatar modeling from a single video via animatable 3D Gaussians. In *CVPR*, 2024.

[58] Gyeongsik Moon, Takaaki Shiratori, and Shunsuke Saito. Expressive whole-body 3D Gaussian avatar. In *ECCV*, 2024.

[59] Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *International Conference on Medical Imaging with Deep Learning (MIDL)*, 2019.

[60] Jeongwan On, Muhammad Salman Ali, Muneeb A. Khan, Sunwoo Park, Inwoong Moon, Hyung Jin Chang, Jaekwang Kim, Seong Jong Ha, and Seungryul Baek. Multi-THuMBS: Multi-person tracking of 3D human meshes beyond video shots. In *ECCV*, 2026.

[61] 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 *ICCV*, 2019.

[62] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *CVPR*, 2023.

Appendix

This appendix gives the full implementation, data, and evaluation details of InstantHand. Secs. A and B describe the model and the training data, and Sec. C specifies every evaluation protocol used in the main paper and in this appendix. Sec. D extends the related work, Sec. E describes the baseline adaptations, and Sec. F reports additional ablations. Secs. G–I contain the discussion (including the ethical considerations in Sec. G.3), additional qualitative results, and the failure cases and limitations. Sec. J analyzes the results by pose difficulty and over time and examines the reliability and opacity of the canonical avatar. Sec. K describes the supplementary video, and Sec. L summarizes reproducibility. For a quick overview, we recommend the supplementary video (Sec. K), which shows 360° rotations and continuous novel-pose driving, together with the novel-pose animation of unseen subjects from fully estimated inputs (Fig. S9) and the failure cases (Fig. S10).

A Implementation Details

A.1 Training Protocol

We train InstantHand in two stages, canonical representation learning followed by source-conditioned Gaussian appearance modeling. Both stages use AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$) with a learning-rate schedule of 2,000 warmup steps, 10,000 hold steps, and 8,000 cosine-decay steps, gradient clipping at norm 1.0, and bfloat16 mixed precision. Training is distributed over four NVIDIA RTX A6000 GPUs.

Stage 1: canonical representation learning. We train the MANO-anchored canonical Gaussian representation for 10 epochs of 2,000 optimization steps each (20,000 steps in total), with a per-GPU batch size of 6, gradient accumulation over 2 steps, a learning rate of 3×10^{-4} that decays to $0.1\times$ its peak, and a weight decay of 0.05. This stage takes about 25 hours. The objective of Eq. 15 of the main paper combines mask-aware RGB reconstruction, LPIPS, silhouette supervision, Laplacian, offset, and scale regularization, and the Soft Alpha Band (SAB) regularizer of Eq. 16, with weights $\lambda_{\text{rgb}} = 1.0$, $\lambda_{\text{lpips}} = 0.2$, $\lambda_{\text{sil}} = 0.6$, $\lambda_{\text{lap}} = 10{,}000$, $\lambda_{\text{off}} = 0.1$, and $\lambda_{\text{scale}} = 0.75$, and with $\lambda_{\text{out}} = 0.25$ and $\lambda_{\text{in}} = 0.05$ for SAB.

Stage 2: source-conditioned appearance modeling. We initialize from the Stage-1 checkpoint and *freeze the entire Stage-1 canonical predictor*, including the base-color head f_{rgb}. The canonical geometry and base colors are therefore fixed, and only the source-conditioned appearance module of Sec. 3.3 of the main paper is optimized. We train for 10 epochs of 2,000 steps each with a per-GPU batch size of 4, a learning rate of 1×10^{-4} that decays to $0.2\times$ its peak, and a weight decay of 0.01. This stage takes about 13 hours. The objective of Eq. 17 of the main paper combines source-view reconstruction, novel-view reconstruction, and feature-statistic consistency, with $\lambda_{s1} = 1.0$, $\lambda_{sp} = 0.5$, $\lambda_{n1} = 1.0$, $\lambda_{np} = 0.25$, and $\lambda_{\text{stat}} = 0.05$. All loss terms are defined in Sec. A.2.

Data preparation. Each training sample consists of one source image and six target frames of the same subject, which provides supervision under varying poses and viewpoints. All images are cropped around the hand using the bounding box of the 2D keypoints, enlarged by a factor of 1.3, and resized to 256×256, and the camera intrinsics are adjusted accordingly. When available, we use the provided masks, MANO parameters, and camera annotations. Otherwise, we estimate hand masks with SAM 3 [54] and MANO and camera parameters with WiLoR [37]. Left-hand samples are flipped horizontally to keep a consistent canonical hand orientation. The same flip is applied during training and inference, so a single network serves both hands. Dataset construction and sampling are detailed in Sec. B.

Canonical Gaussian parameterization values. The fixed and bounded attributes of $\mathbf{G}^{can}$ in Eqs. 4 and 5 of the main paper take the following values. The opacity is fixed at $\alpha_0 = 1$, and the canonical rotation is fixed to the identity quaternion $\mathbf{r}_0 = (1, 0, 0, 0)$. The scale activation ρ is a clamped exponential, $s_i = \rho\big(f_s(\mathbf{z}_i^g)\big) = \text{clamp}\big(\exp(f_s(\mathbf{z}_i^g))\,,\, 5 \times 10^{-4},\, s_{\max}\big)$ in canonical MANO units, with $s_{\max} = 0.008$. The offset head f_Δ ends in the per-axis bounded activation $0.008\,(\text{sigmoid}(\cdot) - 0.5)$ on the output of its final linear layer. This bound keeps each center within ± 4 mm of its anchor, well below the finger width, so the offsets can refine the local surface but cannot

move a center across a finger boundary. Both bounds lie below the finger scale, and the boundary leakage that remains even under these constraints is what the Soft Alpha Band targets (Sec. A.3). Sec. F.2 quantifies the cost of relaxing each of these constraints.

A.2 Loss Definitions

For completeness, we define every term of Eqs. 15 and 17 of the main paper. Throughout, $\hat{I}$ denotes a render and I the corresponding ground-truth image. All losses are computed at 256×256 resolution, and the reconstruction terms are restricted to the foreground region given by the provided mask.

Stage-1 terms (Eq. 15). $\mathcal{L}_1$ is a mask-aware L_1 reconstruction term on the target views, $\mathcal{L}_{\text{lpips}}$ the perceptual loss on the same renders, and $\mathcal{L}_{\text{sil}}$ an L_1 silhouette term on the rendered alpha against the ground-truth mask. The three geometry regularizers of $\mathcal{L}_{\text{geo}}$ are

$$\mathcal{L}_{\text{lap}} = \frac{1}{N}\sum_{i=1}^{N}\Big\|\Delta\mathbf{x}_i - \tfrac{1}{\deg(i)}\sum_{j\sim i}\Delta\mathbf{x}_j\Big\|_2^2, \qquad \mathcal{L}_{\text{off}} = \frac{1}{3N}\sum_{i=1}^{N}\|\Delta\mathbf{x}_i\|_2^2, \tag{S1}$$

$$\mathcal{L}_{\text{scale}} = \frac{1}{N}\sum_{i=1}^{N}\big[\, s_i - s_{\text{th}} \,\big]_+^2, \tag{S2}$$

where $j \sim i$ ranges over the $\deg(i)$ one-ring neighbors of anchor i on the subdivided canonical mesh and $[\cdot]_+ = \max(\cdot, 0)$. $\mathcal{L}_{\text{lap}}$ penalizes the uniform graph Laplacian of the offset field, $\mathcal{L}_{\text{off}}$ keeps the offsets small, and $\mathcal{L}_{\text{scale}}$ penalizes only scales above the threshold $s_{\text{th}} = 0.004$ in canonical MANO units, which is half the hard ceiling $s_{\max} = 0.008$ of Sec. A.1. $\mathcal{L}_{\text{band}}$ is defined in Sec. A.3.

Stage-2 reconstruction terms (Eq. 17). The source-view terms $\lambda_{s1}\mathcal{L}_1^{\text{src}} + \lambda_{sp}\mathcal{L}_{\text{lpips}}^{\text{src}}$ are computed between the rendered source view and the source image, and the novel-view terms $\lambda_{n1}\mathcal{L}_1^{\text{novel}} + \lambda_{np}\mathcal{L}_{\text{lpips}}^{\text{novel}}$ apply the same masked L_1 and LPIPS losses to novel-view renderings against their ground-truth target frames.

Feature-statistic consistency $\mathcal{L}_{\text{stat}}$. This term matches foreground-masked statistics of features from the frozen DINOv3 encoder, the same encoder that conditions our model. Let $F_\ell(\cdot)$ denote channel ℓ of the DINOv3 feature map of an image, N_{ch} the number of feature channels, and M the ground-truth hand mask downsampled to the feature resolution. The per-channel masked mean and standard deviation are

$$\mu_\ell(X; M) \;=\; \frac{\sum_{\mathbf{p}} M(\mathbf{p})\, F_\ell(X)(\mathbf{p})}{\sum_{\mathbf{p}} M(\mathbf{p})}, \tag{S3}$$

$$\sigma_\ell(X; M) \;=\; \sqrt{\frac{\sum_{\mathbf{p}} M(\mathbf{p})\,\big(F_\ell(X)(\mathbf{p}) - \mu_\ell(X; M)\big)^2}{\sum_{\mathbf{p}} M(\mathbf{p})}}. \tag{S4}$$

The loss is the mean absolute difference of the two statistics over channels,

$$\mathcal{L}_{\text{stat}}(X, Y) = \frac{1}{N_{\text{ch}}}\sum_{\ell}\big|\mu_\ell(X; M_X) - \mu_\ell(Y; M_Y)\big| + \frac{1}{N_{\text{ch}}}\sum_{\ell}\big|\sigma_\ell(X; M_X) - \sigma_\ell(Y; M_Y)\big|. \tag{S5}$$

In Stage 2, it is applied between the rendered source view and the source image with $\lambda_{\text{stat}} = 0.05$. The source-image features are detached, so gradients flow only through the render. The term regularizes the global foreground appearance, such as tone and contrast, without requiring pixel alignment.

A.3 Soft Alpha Band Details

The bands B_{out} and B_{in} of Eq. 16 of the main paper are constructed from the ground-truth silhouette M by morphological dilation and erosion. B_{out} is the 2-pixel dilation band outside M, and B_{in} is the interior eroded by 2 pixels, with all widths defined at 256×256 resolution. We use the tolerances $\tau_{\text{out}} = 0.03$ and $\tau_{\text{in}} = 0.60$ and smooth-L_1 (Huber) penalties $\psi(\cdot)$ whose quadratic-to-linear transition lies at 0.03 for the outer term and at 0.05 for the inner term. The tolerances are deliberately asymmetric. Alpha in the outer band is penalized above a near-zero level, whereas the

interior is only required to stay above 0.60, because leakage is the failure mode being targeted while interior coverage only needs a floor.

The schedule $\eta(n)$ over the optimization step n is zero for the first 5,000 steps and then increases linearly to 1 between steps 5,000 and 12,000 (25% and 60% of the 20,000 Stage-1 steps). The band penalty is thus activated only after a coarse geometry has formed. Applying it from initialization destabilizes early optimization, when the rendered silhouette is still far from the target.

Both bands are constructed per frame from the ground-truth silhouette. Under crossed fingers or near-contact poses, every region inside M, including contact regions between fingers, is therefore untouched by the outer penalty, which acts only on true background. Since hands are opaque, no valid translucency is suppressed, and the interior floor τ_{in} prevents erosion near contact boundaries. Figure S4 (Sec. H.1) shows the qualitative effect of SAB.

A.4 Observation Reliability Details

The observation reliability of Eq. 8 of the main paper fuses each Gaussian's accumulated rendering contribution C_i and foreground support m_i into a confidence $\gamma_i \in [0, 1]$ through a square-root transform, median-relative thresholding, and percentile normalization.

Design requirements. The reliability is designed to satisfy four requirements. (i) γ_i should increase monotonically with the evidence a Gaussian actually contributes to source pixels. (ii) It should be insensitive to changes in projected hand size, camera distance, and image scale. (iii) It should be robust to the heavy-tailed distribution of rendering contributions. (iv) It should be bounded in $[0, 1]$, so that the reliability-weighted term in Eq. 13 of the main paper interpolates convexly between observed and base colors. Each component of Eq. 8 addresses one of these requirements with a standard robust-statistics tool.

Square-root transform. The distribution of $\{C_i\}$ is heavy-tailed, as a small number of Gaussians near the silhouette and on prominent surface regions accumulate disproportionately large contributions. Using C_i directly would let these few Gaussians dominate the normalization and push the confidence of moderately observed Gaussians toward zero. The square root, the classical variance-stabilizing transform for nonnegative count-like mass, compresses this tail (requirement iii). As a result, γ_i reflects *whether* a Gaussian is reliably observed rather than *how strongly* it dominates rasterization.

Median-relative thresholding. Thresholding contributions relative to the median of the nonzero contributions, rather than with an absolute cutoff, adapts to each source image and is therefore insensitive to the global scale of the rendering contributions (requirement ii). Absolute thresholds are brittle because the total contribution mass varies with hand pose, projected hand size, and camera distance. A fixed sigmoid or softmax over C_i would instead require a temperature, which is not scale-invariant across images.

Percentile normalization. Normalizing $\sqrt{C_i}\, m_i$ by its ζ-th percentile over $\mathcal{O}$ instead of its maximum amounts to quantile normalization with winsorization at the ζ-th percentile (requirements iii and iv). It is robust to outliers that survive the variance-stabilizing transform, and clipping to $[0, 1]$ gives a bounded confidence that is comparable across poses and image scales.

Hyperparameters. With $(\kappa, \tau_m, \zeta) = (0.5, 0.5, 95)$, a Gaussian enters $\mathcal{O}$ if its rendering contribution exceeds half the median nonzero contribution and more than half of its projected footprint lies on the hand, and the scores are normalized by the 95th percentile. Both thresholds are set to the midpoint 0.5, and the 95th percentile is a conventional choice for robust normalization. Sec. F.4 reports the sensitivity around these values.

Graded weights and the observed-set boundary. Unlike a binary visibility mask, γ assigns *graded* weights within $\mathcal{O}$ and is zero outside it. Gaussians at the edge of $\mathcal{O}$ contribute only weakly to the source image, so the transition from source-driven to prior-driven color takes place where the source evidence is weakest. Alpha blending across many overlapping Gaussians with different γ values usually smooths this transition in the rendered image. When the observed tone differs from that of the prior, for example because shading is baked into the source (Sec. I), the transition can remain visible as a seam, most noticeably where the boundary crosses a thin structure. Sec. J.3 quantifies the color step at this boundary. The source-level shift $\mathbf{\Delta}_s^{aid}$ is a single per-channel shift

shared by all Gaussians (Sec. A.5) and therefore introduces no spatial variation of its own. The 360° rotations in the supplementary video (Sec. K) show the transition between the source-observed and prior-completed sides directly.

Role separation. As noted after Eq. 13 of the main paper, the reliability γ_i and the appearance gate g_i act on the source evidence only through their product $g_i\gamma_i$, but the two factors differ in kind. γ_i is a closed-form estimate of how reliably a Gaussian is observed. It is computed from the rasterizer outputs and the source mask, receives no gradient, and does not predict appearance. Appearance itself is predicted by the learned modules, namely the local evidence tokens, the global source identity token, and the source-conditioned prior bank $\mathcal{D}_s$, and the appearance gate g_i decides how much of the observed color to adopt where evidence exists. We do not learn γ_i, since a learned confidence could be satisfied from dataset priors rather than from the observation process. A single learned scalar in place of $g_i\gamma_i$ would lose the one factor that training cannot adjust. The ablations confirm that neither factor absorbs the other. With every variant retrained through the full two-stage pipeline, removing γ costs 3.03 dB and removing g_i costs 1.69 dB (Table 3 of the main paper).

A.5 Architecture Details

This section specifies every learned module in Eqs. 1–13 and Figs. 2 and 3 of the main paper. In Fig. 2, the conditional transformer consists of Φ_g and Φ_a, the point embedding is ϕ_q, and the scale, offset, and color heads are f_s, f_Δ, and f_{rgb}. In Fig. 3, Prior Refinement is $\mathcal{U}$, Identity Conditioning comprises $\mathcal{B}$ and $\mathcal{U}_c$, Appearance Prediction is f_{out}, and Appearance Composition is Eq. 13.

Image encoder and fusion head. E_{img} applies a frozen DINOv3 ViT-L/16 backbone (patch size 16) to the 256×256 crop, giving a 16×16 grid of patch tokens. The patch tokens of blocks 4, 11, 17, and 23 are each passed through a 1×1 convolution, concatenated along channels, and fused by a second 1×1 convolution into a $d_{\mathrm{img}} = 1{,}024$-channel map at patch resolution. This map is flattened into the $N_f = 256$ image tokens $\mathbf{F}_s$ of Eq. 1. Only this fusion head is trained (in Stage 1), and the backbone stays frozen throughout.

Query embedding and transformer streams (Eqs. 2–5). The query anchors $\mathbf{q}_i$ of Eq. 2 are the $N = 49{,}281$ vertices of the $k = 3$ subdivided $\mathrm{MANO}(\boldsymbol{\beta}, \boldsymbol{\theta}_{\mathrm{mean}})$ mesh. The point-embedding layer ϕ_q encodes each anchor with sinusoidal features at eight octaves per axis, concatenates the raw coordinates, and applies a linear layer with LayerNorm to obtain a $d = 512$-dimensional token. Φ_g consists of two SD3-style joint-attention blocks (8 heads, feed-forward ratio 4) in which the N query tokens and the N_f image tokens, linearly projected from d_{img} to d, attend to one another. The image stream is updated in the first block and serves only as keys and values in the second, and a final LayerNorm yields $\mathbf{Z}^g$. Φ_a is a single cross-attention layer without a feed-forward block (queries $\mathbf{Z}^g$, keys and values $\mathbf{F}_s$, 8 heads) that produces $\mathbf{Z}^a$. The Gaussian decoder of Eq. 5 applies one shared trunk MLP (two hidden layers of width 512, SiLU) separately to $\mathbf{z}_i^g$ and $\mathbf{z}_i^a$, followed by one linear output layer per attribute, namely f_Δ (3 values), f_s (1 value, replicated to the isotropic scale), and f_{rgb} (3 values, sigmoid), with the output bounds of Sec. A.1.

Footprint aggregation and local evidence (Eqs. 7, 9, and 10). Agg_{Ω_i} samples a 7×7-pixel window centered at the projected Gaussian center $\mathbf{u}_i$ with Gaussian weights $w_i(\mathbf{p}) = \exp\!\big(-\|\mathbf{p} - \mathbf{u}_i\|^2 / 2\sigma_i^2\big)$, where σ_i is half the projected radius reported by the rasterizer, with a minimum of 0.5 pixels. The same window is applied to I_s, $\hat{I}_s$, M_s, and $\mathbf{F}_s^x$. $\mathbf{F}_s^x$ is produced by a ResNet-18 U-Net (ImageNet-pretrained encoder) at $1/4$ of the input resolution (64×64) with 64 channels. It is trained in Stage 2 and is separate from the DINOv3 encoder. f_{loc} is a two-layer MLP ($3 + 3 + 3 + 64 + 1 \to d_a \to d_a$, GELU, LayerNorm) followed by a residual MLP block ($d_a \to 4d_a \to d_a$, GELU), with $d_a = 384$.

Global source appearance identity. The global source appearance identity token $\mathbf{t}_s^{aid}$ is computed once per source image from the concatenation of (i) the M_s-masked spatial mean of $\mathbf{F}_s$ (d_{img} values), (ii) the M_s-masked spatial mean of $\mathbf{F}_s^x$ (64), (iii) the M_s-masked mean and standard deviation of I_s (6), (iv) the mean and standard deviation of the base colors $\mathbf{c}_i^0$ over all N Gaussians (6), (v) the γ-weighted mean and standard deviation over Gaussians of the residuals $\bar{\mathbf{c}}_i^{src} - \bar{\mathbf{c}}_i^{ren}$ (6), and (vi) the γ-weighted mean of the local evidence tokens $\mathbf{z}_i^{loc}$ (d_a). A two-layer MLP (GELU, LayerNorm) maps this vector to $\mathbf{t}_s^{aid} \in \mathbb{R}^{d_a}$.

Prior Refinement and Identity Conditioning (Eq. 11). Prior Refinement $\mathcal{U}$ projects $\mathbf{Z}^a$ to width d_a with a two-layer MLP and applies two cross-attention blocks (pre-LayerNorm on queries and memory, 8 heads, residual feed-forward of ratio 4) whose memory is the $N+1$ tokens $\mathbf{Z}^{loc} \cup \{\mathbf{t}_s^{aid}\}$. In Identity Conditioning, the bank $\mathcal{D}_s = \mathcal{B}(\mathbf{t}_s^{aid})$ consists of 64 learned prototypes in $\mathbb{R}^{d_a}$, each shifted by the same two-layer MLP image of $\mathbf{t}_s^{aid}$ (additive conditioning). $\mathcal{U}_c$ then forms the query $\mathrm{MLP}\big([\mathbf{z}_i^{loc}, \tilde{\mathbf{z}}_i^a, \mathbf{t}_s^{aid}, \mathrm{MLP}(\mathbf{c}_i^0)]\big)$ $(4d_a \to d_a \to d_a)$ and applies four cross-attention blocks of the same form whose memory is the 65 tokens $\{\mathbf{t}_s^{aid}\} \cup \mathcal{D}_s$, giving $\mathbf{h}_i$.

Appearance Prediction and Composition (Eqs. 12 and 13). f_{out} consists of three two-layer MLP heads ($d_a \to d_a \to$ output) that produce $\mathbf{c}_i^{obs} = \mathrm{sigmoid}(\cdot)$, $g_i = \mathrm{sigmoid}(\cdot)$, and $\boldsymbol{\delta}_i = 0.1\tanh(\cdot)$. The output bias of the g_i head is initialized to -1.5 so that transfer starts close to the base colors. Pool in Eq. 12 is the unweighted mean over all N Gaussians, and the composed color of Eq. 13 is clipped to $[0, 1]$. The cross-attention blocks and the residual MLP of the appearance module use dropout 0.1. Stage 1 trains the fusion head, ϕ_q, Φ_g, Φ_a, and the Gaussian decoder, and Stage 2 trains the appearance module, including the ResNet-18 U-Net.

B Datasets and Sampling

B.1 Training Datasets

We train InstantHand using a mixture of real and synthetic hand datasets. The real datasets include PALM [52], InterHand2.6M [48], HO3Dv3 [49, 50], Goliath [51], and Hand Appearance (HARP) [22]. For Goliath and InterHand2.6M, left and right hands are treated as separate subjects after preprocessing. In total, the real training datasets contain 304 subjects and 339,532 frames. For InterHand2.6M, we train only on the official training split, so the evaluation subject (`test/Capture0`) and all HanCo subjects are unseen during training. Figure S1 shows representative training examples from the real datasets (top row) and from the synthetic appearance set of Sec. B.2 (bottom row). Hand regions are shown segmented on a white background for visualization.

PALM and InterHand2.6M provide large-scale multi-view observations with diverse poses and identities, while HO3Dv3 introduces hand–object interactions and partial occlusions. Goliath and HARP provide additional appearance diversity. Synthetic hand data is used only for training the source-conditioned Gaussian appearance module, as described in Sec. B.2.



Figure S1: Examples of real (top) and synthetic (bottom) training data. Synthetic samples are used only in the source-conditioned appearance stage (Sec. B.2).

B.2 Synthetic Hand Data Generation

For the source-conditioned appearance stage, we augment training with synthetic source–target pairs generated following Zhao et al. [53]. In addition to realistic hand appearances, we sample diverse hand colors to expose the appearance module to a broader range of source appearances (bottom row

of Fig. S1). This augmentation is used only for training the source-conditioned Gaussian appearance module, and the canonical Gaussian representation stage is trained on real hand datasets only.

Each synthetic sample is converted into the same source–target format as real training data. A single source image provides the appearance condition, while target frames provide supervision under different poses and viewpoints. All synthetic images are cropped, resized, and processed using the same preprocessing pipeline as real images.

B.3 Dataset Sampling Strategy

We use a subject-centric sampling strategy to construct source–target training pairs. Each dataset is organized by subject, and each subject contains one or more groups. A group denotes a subject-internal observation bucket, such as a camera, sequence, session, or multi-camera unit depending on the dataset. For each training iteration, we first sample a dataset according to a stage-specific dataset weight, then uniformly sample a subject from the selected dataset. A source frame is sampled from the subject-level source pool, and six target frames are sampled from the same subject according to dataset-specific self, same-group, and cross-group target counts.

Table S1: Dataset sampling policies used for training. Targets per source image are reported as self / same-group / cross-group.

Stage	Dataset	Weight	Source stride	Targets
Canonical	PALM	0.45	1	1 / 1 / 4
Canonical	InterHand2.6M	0.30	1	1 / 2 / 3
Canonical	HO3Dv3	0.10	10	1 / 3 / 2
Canonical	Goliath	0.10	5	1 / 5 / 0
Canonical	HARP	0.05	5	1 / 5 / 0
Appearance	PALM	0.35	1	1 / 1 / 4
Appearance	InterHand2.6M	0.30	1	1 / 2 / 3
Appearance	Synthetic	0.15	1	1 / 1 / 4
Appearance	HO3Dv3	0.08	10	1 / 3 / 2
Appearance	Goliath	0.08	5	1 / 5 / 0
Appearance	HARP	0.04	5	1 / 5 / 0

The self target corresponds to the source frame itself and provides source-view supervision. Same-group targets are sampled from the same subject group as the source, while cross-group targets are sampled from different groups of the same subject. For multi-view datasets, cross-group targets usually correspond to different cameras or observation groups. For monocular datasets, they correspond to different subject-internal sequences when available. This strategy keeps identity and hand shape consistent while exposing the model to varying poses and viewpoints. Table S1 summarizes the sampling weights and target-frame composition used in each stage.

C Evaluation Protocol

We follow the single-image animatable hand reconstruction protocol established by prior hand avatar benchmarks [20, 23], including their sequence, designated source-frame, cropping, and masking choices, to ensure direct comparability with reported baselines. The InterHand2.6M source frame is the one provided with the official OHTA evaluation configuration. For HanCo, whose protocol leaves the source frame unspecified, we fix one common source frame for all methods (Sec. C.2). Given a single source image, the model reconstructs a canonical Gaussian hand representation, which is then rendered under novel target poses and viewpoints for evaluation. All methods receive the identical source image, ground-truth annotations, and evaluation targets.

C.1 InterHand2.6M Evaluation

We follow the evaluation protocol used in HandAvatar [20] and OHTA [23]. We use the following sequence and, as the source image, the example provided in the official OHTA repository:

Sequence	`test/Capture0/ROM03_RT_No_Occlusion`
Source image	`test_Capture0_ROM03_RT_No_Occlusion_cam400272_image15012.jpg`

Following HandAvatar, the hand region is cropped using the provided evaluation bounding box. The bounding box is converted into a square region with a $1.3\times$ enlargement factor, and the cropped image is resized to 256×256. Unless otherwise specified, the same cropping protocol is used for all experiments. We use the ground-truth camera parameters and MANO parameters provided by the dataset, and the foreground evaluation masks are ray masks generated following the HandAvatar evaluation protocol. The source image itself is excluded from quantitative evaluation. Table 1 of the main paper reports the results under this protocol. The corresponding evaluation with *estimated* source-side cues is reported in Table 3 of the main paper and, with boundary-level metrics, in Sec. F.3.

C.2 HanCo Evaluation

For HanCo we follow the zero-shot cross-dataset evaluation setting of OHTA [23]. We use sequence 0191 and evaluate frames from cam3, cam5, cam6, and cam7, where the cameras are positioned above the hands. Following OHTA, we do not use cameras below the hands due to inconsistent lighting conditions and color calibration. We use the 76th and last frame of cam6 (frame index 75) as the common source image for all methods, together with the ground-truth MANO parameters and hand masks provided by HanCo. Unlike the InterHand2.6M protocol, the HanCo evaluation set retains all 76 frames of each of the four cameras (304 frames), so the source frame itself is one of the evaluation frames. It is retained for every compared method, so relative comparisons are unaffected, and it contributes one of the 304 frames to the aggregate metrics.

Preprocessing and evaluation-mask generation are identical to the InterHand2.6M protocol of Sec. C.1: the hand region is cropped to a square with a $1.3\times$ enlargement factor and resized to 256×256, and metrics are computed over ray masks generated following the HandAvatar evaluation protocol, applied consistently to all methods. The only difference is the bounding-box source: since HanCo does not provide the HandAvatar evaluation bounding boxes, the square crop is instead derived from the minimum and maximum coordinates of the ground-truth foreground mask. Because the OHTA HanCo protocol designates the held-out sequence (0191) but not the source frame, absolute numbers are not directly comparable to previously reported HanCo values. All methods are therefore re-evaluated by us from the identical source frame under identical inputs (Sec. C.7). Table 2 of the main paper reports the results under this protocol.

C.3 Metric Computation

Following the HandAvatar protocol, pixels outside the ray mask are set to the background color in both the rendered and the ground-truth image, and PSNR, SSIM, and LPIPS are computed on the resulting images, which reduces background bias. Because our rasterizer ignores the principal point, we render with a centered principal point and map the render to the original crop with the exact homography, so all methods are scored on the same pixels. All evaluations are performed at 256×256 resolution.

C.4 Boundary-aware Silhouette Metric Definitions

Boundary F-score and boundary IoU are computed between the binarized rendered alpha ($A > 0.5$) and the ground-truth mask, with boundaries extracted as 1-pixel contours. The boundary F-score is the standard contour-matching F-measure: predicted and ground-truth boundary pixels are matched within a distance tolerance of 2 pixels via distance transforms. Boundary IoU is a variant of the boundary-restricted IoU of Cheng et al. [56]: each 1-pixel contour is dilated by 3 pixels and the IoU is computed between the two resulting bands. The band is thus symmetric about the silhouette rather than restricted to the mask interior, so it also registers outward alpha leakage, which an interior-only band ignores. Alpha leakage is the mean excess of the rendered alpha above a tolerance of 0.1 over the outer band, constructed as a 5-pixel dilation of the ground-truth silhouette excluding the silhouette itself, and interior coverage is the mean rendered alpha over the interior eroded by 5 pixels. The evaluation bands are deliberately wider than, and distinct from, the 2-pixel training bands of Sec. A.3, so the boundary evaluation does not simply re-measure the region that the SAB loss optimizes. All metrics are computed at 256×256 and, unless stated otherwise, averaged over the InterHand2.6M novel-pose/novel-view evaluation set. They are reported in Table 4 of the main paper and in Tables S8 and S11, the latter also on the HanCo evaluation set.

Table S2: Self-reconstruction with geometry fixed (Sec. C.5). w/o transfer renders the Stage-1 base colors, and Ours renders the composed colors.

Dataset	Setting	PSNR↑	SSIM↑	LPIPS↓
InterHand2.6M	w/o transfer	26.58	0.869	0.133
	Ours	**28.54**	**0.903**	**0.108**
HanCo (zero-shot)	w/o transfer	20.47	0.839	0.187
	Ours	**23.88**	**0.898**	**0.153**

C.5 Self-reconstruction Protocol

Self-reconstruction is a controlled source-view appearance diagnostic: it measures how faithfully the input image is reflected in the avatar when no pose or viewpoint gap is involved. Each frame of the evaluation data is used in turn as the input: the avatar is reconstructed from that frame and rendered back into the same pose and camera, so the target is the input itself. All non-color attributes are fixed to the Stage-1 canonical values, so any metric difference between the w/o transfer and Ours rows of Table S2 is attributable *solely* to the source-conditioned color update. The w/o transfer row renders the Stage-1 base colors $\mathbf{c}_i^0$ (the w/o appearance transfer configuration of Table 3 in the main paper), and the Ours row renders the composed colors $\mathbf{c}_i^*$ (Eq. 13 of the main paper). Table S2 reports the average over the evaluation frames of each benchmark (the InterHand2.6M sequence of Sec. C.1 and the HanCo sequence of Sec. C.2), with the same cropping protocol and ray masks as Sec. C.3.

This diagnostic complements the novel-pose/novel-view ablation in Table 3 of the main paper by isolating whether the color update follows the observed source appearance while geometry is fixed. Figure S6 (Sec. H.3) shows representative frames.

C.6 Timing and Efficiency Measurement

The 2.84 s reported in the main paper is the wall-clock *avatar-network* inference time on a single NVIDIA A100 (80 GB), covering image encoding, MANO-anchored token computation, Gaussian decoding, source-view rendering, and source-conditioned appearance composition. It excludes disk I/O, data loading, input preprocessing, one-time model loading, and CUDA initialization, and it is averaged over repeated runs after warm-up (standard deviation 0.007 s).

Cost breakdown. Building the canonical Gaussians alone takes 1.38 s under the same protocol. Source-conditioned appearance composition, including the source-view rendering it requires, therefore accounts for the remaining 1.46 s, about half of the avatar-network time.

End-to-end raw image to avatar. Because the reported 2.84 s assumes that MANO, camera, and mask inputs are already available, we additionally report the complete raw-image-to-avatar cost on the same GPU (Table S3). Upstream estimation adds approximately 0.46 s (0.38 s for the SAM 3 foreground mask and 0.08 s for WiLoR MANO and camera estimation), giving an end-to-end construction time of approximately 3.30 s. Upstream estimation is thus about 14% of the end-to-end cost and is not the bottleneck. This entire cost is incurred *once per avatar*, for example when a user's hand is enrolled at the start of an AR/VR or telepresence session, and subsequent animation re-runs neither network and renders at 161 FPS. The construction time is therefore not a per-frame latency, and we do not claim real-time acquisition.

All baseline measurements in Table 2 of the main paper (per-subject fitting time, rendering speed, and peak memory) are taken on the same NVIDIA A100 (80 GB). Relative to our 2.84 s avatar-network inference, the fitting times of HandAvatar, OHTA, and BIGS* correspond to speedups of $1964\times$, $1183\times$, and $486\times$ ($5579/2.84$, $3360/2.84$, and $1380/2.84$). The "over $400\times$" figure quoted in the main paper is the most conservative of these.

Rendering speed and memory. FPS measures novel-pose rendering (MANO-driven deformation and rasterization) at 256×256 with batch size 1 over the HanCo evaluation frames, excluding data loading: 161 FPS (≈ 6.2 ms per frame). Peak GPU memory, including the CUDA context, is 1.79 GB during rendering and 4.78 GB during avatar construction. Since all compared methods receive the same ground-truth inputs (Sec. C.7), upstream cost is identical across the compared systems and orthogonal to the reported comparison.

Table S3: Raw-image-to-avatar runtime on a single NVIDIA A100 (80 GB) under the protocol of Sec. C.6.

Stage	Component	Time
Upstream estimation	SAM 3 foreground mask [54]	0.38 s
	WiLoR MANO + camera [37]	0.08 s
	subtotal	*0.46 s*
Avatar construction	Canonical Gaussian construction	1.38 s
	Source-conditioned appearance composition	1.46 s
	subtotal (Table 2 of the main paper)	*2.84 s*
End-to-end	raw image → animatable avatar	**≈3.30 s**
Animation	novel-pose rendering (per avatar, repeated)	160.71 FPS

Table S4: Input protocol for every compared method.

Method	Original input	Input here
HARP	Monocular video	Single frame
LiveHand	Multi-view video	Single frame
HandAvatar	Monocular video	Single frame
Handy	Single image	Single frame
OHTA	Single image	Single frame
BIGS*	Monocular video	Single frame
Ours	Single image	Single frame

C.7 Baseline Input Protocols

Table S4 states, for every compared method, its originally intended input modality and the input actually provided in our evaluation. All methods receive the identical single source frame and identical ground-truth MANO, camera, and mask annotations, following the single-image protocol established by prior hand avatar benchmarks [20, 23], and no method receives additional frames. For Handy, a parametric shape and appearance model built from large-scale 3D hand scans, the single-frame evaluation follows the setting established by OHTA [23]: the provided MANO annotations are converted into Handy's format, and its texture latent code is fitted to the single source frame via differentiable rendering.

Provenance of baseline numbers. The provenance of the reported baseline numbers differs between the two benchmarks. On InterHand2.6M, we adopt the single-frame results of HARP, LiveHand, HandAvatar, Handy, and OHTA established under this protocol by Zheng et al. [23], and evaluate our method under the identical protocol. BIGS* postdates that work, so its InterHand2.6M numbers are measured by us under the same protocol using the single-hand adaptation of Sec. E.1. On HanCo, the exact evaluation configuration behind previously reported numbers (evaluation frames, source frame, and mask generation) is not fully specified publicly, so directly citing them would not yield a controlled comparison. We therefore re-evaluate every compared method ourselves under the fully specified protocol of Sec. C.2, using the same source frame, ground-truth annotations, crops, and masks for all methods. All qualitative comparisons (Fig. 4 of the main paper) are rendered by us from our own runs of each baseline under the protocols of Secs. C.1 and C.2, so the qualitative and quantitative comparisons share the same source frames, crops, and masks.

C.8 Protocols of the Additional Analyses

The pose-stratified and temporal analyses of Sec. J reuse the renders, ground truth, and ray masks of the main evaluation (Secs. C.1 and C.2), and only the reporting changes. The canonical diagnostics use the avatars reconstructed from the two protocol source frames.

Pose strata. For every evaluation frame we measure two changes of the driving pose relative to the source pose. The *palm-orientation change* is the geodesic distance between the source and target global hand rotations, each expressed in the frame of its own camera, so it captures both the rotation of the hand and the change of camera. The *finger-articulation change* is the mean geodesic distance over the 15 finger joints of MANO, excluding the global rotation. The frames are split into

equal-count terciles (Easy, Medium, Extreme) of each measure, and every method is scored on the same frames. For the palm terciles, we also report the median change of the viewing direction in the hand frame, which measures how far the camera has moved around the hand irrespective of in-plane roll. Tables S12 and S13 (Sec. J.1) report the results.

Temporal stability. On the continuous HanCo sequence we form, per camera, the 75 pairs of consecutive evaluation frames (300 pairs in total). For a pair $(t, t{+}1)$ with ray masks M_t and M_{t+1}, the frame-to-frame change of the prediction is

$$d_{\mathrm{pred}} = \frac{1}{|M_t \cup M_{t+1}|} \sum_{\mathbf{p} \in M_t \cup M_{t+1}} \frac{1}{3} \left\| \hat{I}_{t+1}(\mathbf{p}) - \hat{I}_t(\mathbf{p}) \right\|_1, \tag{S6}$$

where the L_1 norm runs over the three color channels of the same composited 8-bit images used for the metrics. We define d_{gt} identically on the ground truth, and the excess variation is $E_t = d_{\mathrm{pred}} - d_{\mathrm{gt}}$. For each camera, which records one continuous video, we average E_t over its consecutive pairs and report the result relative to the mean ground-truth change of that camera. The summary over cameras averages the magnitudes of these per-camera values, so that opposite biases of different cameras do not cancel. As a calibration, we add to our own renders a per-frame brightness offset inside the ray mask, drawn uniformly within $\pm 1\%$, $\pm 2\%$, or $\pm 4\%$, and recompute E_t. Table S14 (Sec. J.2) reports the results.

Canonical boundary diagnostic. The Gaussians are partitioned by their reliability into a high-reliability set ($\gamma_i \geq 0.5$), a transition set, and a low-reliability set ($\gamma_i < 0.1$). Neighbors are the six nearest Gaussians in canonical space, and color differences are CIE76 ΔE values in CIELAB between the final colors $\mathbf{c}_i^*$. We compare three kinds of neighbor pairs. Pairs within the high-reliability set measure the natural texture variation and serve as the reference, pairs that straddle the threshold $\gamma_i = 0.5$ measure the transition, and *strict* pairs connect a high-reliability and a low-reliability Gaussian directly. Sec. J.3 reports the results.

Reliability and opacity renderings. The reliability maps (Fig. S12) render γ_i as a scalar field with the same rasterizer, and the alpha-weighted mean of γ along each ray is color-mapped after compositing, so blending never produces colors outside the color map. The opacity maps (Fig. S13) show the accumulated alpha A of the render, the quantity behind the boundary metrics of Sec. C.4. The avatar is posed with the source pose and rendered from the source view and from an orbit of the source camera about its vertical image axis through the centroid of the posed Gaussians.

D Extended Related Work

Hand avatar reconstruction. A large body of work relies on multi-view inputs or monocular video sequences, combined with neural implicit representations or explicit mesh-based pipelines [20–22, 39–41]. HandAvatar [20] models pose-dependent appearance using implicit representations with self-occlusion-aware texture modeling. HARP [22] employs a mesh-based formulation augmented with learned reflectance maps to capture high-frequency details, while LiveHand [21] leverages multi-view video and neural rendering to synthesize photorealistic hand appearance. Despite achieving high visual fidelity, these approaches depend on dense observations and per-subject optimization, leading to substantial computational overhead and limited scalability. Single-image methods such as OHTA [23] introduce data-driven priors for one-shot hand avatar reconstruction, but still require test-time fitting or inversion. Gaussian-based approaches such as GaussianHand [24] adopt explicit 3D Gaussian primitives for efficient rendering but continue to depend on multi-view supervision or temporal sequences and often require subject-specific optimization. As summarized in Table S5, existing methods remain constrained by dense inputs or per-subject optimization, whether they use implicit fields, meshes, or Gaussian primitives.

Feed-forward 3D Gaussian avatar reconstruction. Feed-forward avatar methods regress the parameters of animatable 3D Gaussians from one or a few images, replacing per-subject optimization with amortized inference. Among them, LHM [17] extends Gaussian representations to animatable full-body avatars, while GUAVA [42] focuses on upper-body reconstruction. For head avatars, approaches such as GAGAvatar [13] and LAM [14] demonstrate high-fidelity rendering via feed-forward Gaussian prediction. Extending these designs to hands exposes the two failure modes

Table S5: Comparison of representative 3D hand avatar reconstruction methods under their original input settings. BIGS*: our single-hand adaptation of BIGS (Sec. E.1). §GaussianHand reports results from as few as 5–20 camera views.

Method	Input	Feed-forward	Repr.
LiveHand [21]	Multi-view video	✗	NeRF-like
GaussianHand§ [24]	Multi-view video	✗	3DGS
HandAvatar [20]	Monocular video	✗	Implicit
HARP [22]	Monocular video	✗	Mesh
BIGS* [47]	Monocular video	✗	3DGS
OHTA [23]	Single image	✗	Implicit
Ours	**Single image**	✓	**3DGS**

discussed in Sec. 1 of the main paper: thin, articulated structures are sensitive to small geometric errors, and appearance regressed from image features over-relies on learned priors under single-view ambiguity. The controlled recipe ports of Sec. E.2 quantify the former.

E Baseline Adaptations

E.1 BIGS-based Hand Reconstruction Baseline

BIGS [47] originally targets bimanual hand–object interaction reconstruction from monocular videos, and therefore does not directly match our single-image, single-hand evaluation setting. To construct a reference optimization-based 3DGS baseline, we adapt BIGS to a single-hand setting by retaining only the Gaussian hand avatar component and removing the object and second-hand branches.

The adapted baseline represents the hand in a canonical Gaussian space. Following the Gaussian hand avatar modeling design, Gaussian attributes are decoded from a triplane feature representation using MLP heads: the geometry head predicts Gaussian offsets, rotations, and scales, the appearance head predicts colors and opacities, and the deformation head predicts MANO-related skinning weights. The canonical Gaussians are then transformed to posed space using MANO-based linear blend skinning and rendered with differentiable Gaussian rasterization.

The original formulation is designed for monocular videos without known camera poses, where the camera coordinate system and hand trajectory must be estimated or aligned. In our benchmark setting, ground-truth annotations are available, so for controlled fitting and evaluation we use the ground-truth MANO parameters, camera parameters, and hand masks whenever available. This removes errors from external hand-pose, camera, and mask estimation, and focuses the comparison on the reconstruction and rendering formulation itself. Accordingly, the fixed virtual-camera construction and the 2D-keypoint-based global trajectory alignment of the original pipeline are bypassed, and the temporal smoothness regularization defined across consecutive video frames is inactive under single-frame fitting.

The adapted baseline is fitted to the single source frame by minimizing a masked reconstruction objective restricted to the foreground hand region, combining an L_1 term and an SSIM term, together with a silhouette loss on the rendered opacity and a regularizer that encourages the predicted skinning weights to remain close to the corresponding MANO skinning priors. Fitting uses Adam for 15,000 iterations, with learning rates of 1.6×10^{-4} for the Gaussian position parameters, 1.0×10^{-4} for the remaining Gaussian attribute parameters, and 1.0×10^{-3} for the triplane and MLP decoder weights.

During evaluation, the fitted BIGS-based baseline is rendered using the same target MANO poses, target cameras, hand-centered crops, and foreground masks as InstantHand. Since this protocol adapts BIGS to a single-hand benchmark and differs from the original bimanual hand–object interaction setting, we report it as a reference optimization-based 3DGS baseline (BIGS*) rather than as BIGS in its original setting. Its results appear in Tables 1 and 2 of the main paper, and its input protocol in Table S4.

E.2 Controlled Recipe Ports of One-shot Head and Body Avatar Methods

A natural question is whether existing feed-forward one-shot *head* and *full-body* avatar frameworks can simply be retargeted to hands. To answer this empirically rather than by argument, we construct

controlled hand adaptations of both families and evaluate them under matched conditions. All quantitative results are reported in Table 4 of the main paper. This section specifies the two ports and states what the resulting gap does and does not show.

Matched-condition protocol. In both ports, the training data, MANO anchors, image encoder, optimization schedule, and compute budget are identical to ours, and only the representation and supervision recipe differ. Free per-Gaussian parameters are initialized from our constrained solution so that the comparison does not confound optimization difficulty with initialization. Both ports are evaluated under the InterHand2.6M protocol of Sec. C.1 with the boundary metrics of Sec. C.4.

Head recipe [14]. We replace FLAME with MANO while retaining the method's free per-Gaussian opacity, rotation, and anisotropic scale, together with its direct image-feature-based RGB prediction. Our boundary-aware supervision and our reliability-gated color pathway are both removed, so this port evaluates the published parameterization rather than a hybrid.

Body recipe [17]. Directly reproducing LHM [17] is infeasible at our scale. It uses a 0.5–1B-parameter model trained on approximately 300K videos with 32–64 A100 GPUs, and its released model requires full-body images and SMPL-X parameters rather than cropped hand images and MANO parameters. We therefore construct a controlled recipe-level port. SMPL-X anchors are replaced by MANO anchors, and the freely regressed Gaussian attributes and canonical-space regularization of Qiu et al. [17] are retained as released. Writing $s_i^{(1)} \geq s_i^{(2)} \geq s_i^{(3)}$ for the sorted per-axis scales of a freed Gaussian, the as-spherical-as-possible term takes its released ratio-hinge form, $\max\big(s_i^{(1)}/s_i^{(3)} - \tau_{\mathrm{ASAP}},\, 0\big)$ with $\tau_{\mathrm{ASAP}} = 5$, at the released hand weighting $\lambda_{\mathrm{ASAP}} = 10^5$. Because this term is a ratio of scales, it transfers to MANO unchanged. The position-anchoring term $\frac{1}{N}\sum_i \max(\|\Delta\mathbf{x}_i\|_2 - \tau_{\mathrm{ACAP}},\, 0)$ and the offset range are stated in Qiu et al. [17] relative to a $1.8\,\mathrm{m}$ body ($\tau_{\mathrm{ACAP}} = 1.8/32\,\mathrm{m}$, weight 10^3, range $1.0\,\mathrm{m}$). We apply the same numbers as fractions of the template extent ($\approx 0.19\,\mathrm{m}$ for MANO): $\tau_{\mathrm{ACAP}} = \mathrm{extent}/32 \approx 5.9\,\mathrm{mm}$, $\lambda_{\mathrm{ACAP}} = 10^3 \cdot 1.8/\mathrm{extent} \approx 9.5 \times 10^3$, and an offset range of $\mathrm{extent} \cdot (1.0/1.8) \approx 0.106\,\mathrm{m}$. This choice preserves the range-to-threshold ratio of the original exactly. This soft anchoring replaces the hard offset bound and L2 offset prior of the head recipe, since the two are one design choice. The mask weight follows Eq. 7 of Qiu et al. [17] (0.5), and the backbone, data, anchors, schedule, and compute are matched to ours. At the weights of Eq. 10 of Qiu et al. [17] (50/10), the anisotropy hinge never activates (mean scale ratio $\approx 3 < \tau_{\mathrm{ASAP}}$), and the anchoring term contributes $\approx 0.25\%$ of the loss, so the free attributes are effectively unregularized. We therefore report the hand weighting released in the official implementation of Qiu et al. [17], under which the anchoring term is active. We report this as a recipe port, not as a reproduction of the original system at its original scale. Matched renderings of both ports are shown in Fig. S5.

Reading the reported gap. Both ports are worse than our constrained model without SAB in boundary F-score and leakage (boundary F 0.267–0.299 vs. 0.366, leakage 0.608–0.627 vs. 0.555). The two ports are independent configurations, each retaining its own published parameterization, yet they score close to each other (PSNR within $0.24\,\mathrm{dB}$, boundary F 0.267 vs. 0.299). Their anchoring terms differ in kind. The head recipe uses hard anchoring, an L2 prior within a $\pm 4\,\mathrm{mm}$ bound, and the body recipe uses soft anchoring, a hinge that is inactive within $\tau_{\mathrm{ACAP}} \approx 5.9\,\mathrm{mm}$. At our framing, this leaves a free shell of $\approx 6\,\mathrm{px}$, three times the 2-px training band. Neither term observes the rendered silhouette, however, and neither recovers boundary precision. Supervision on the silhouette itself does. Under the constrained parameterization, SAB raises the boundary F-score from 0.366 to 0.698 (Table 4 of the main paper and Fig. S4). The contribution of the attribute constraints with boundary supervision held fixed is isolated in Sec. F.2.

F Additional Ablation Studies

Unless otherwise specified, all variants are evaluated on InterHand2.6M using the same protocol as the main paper.

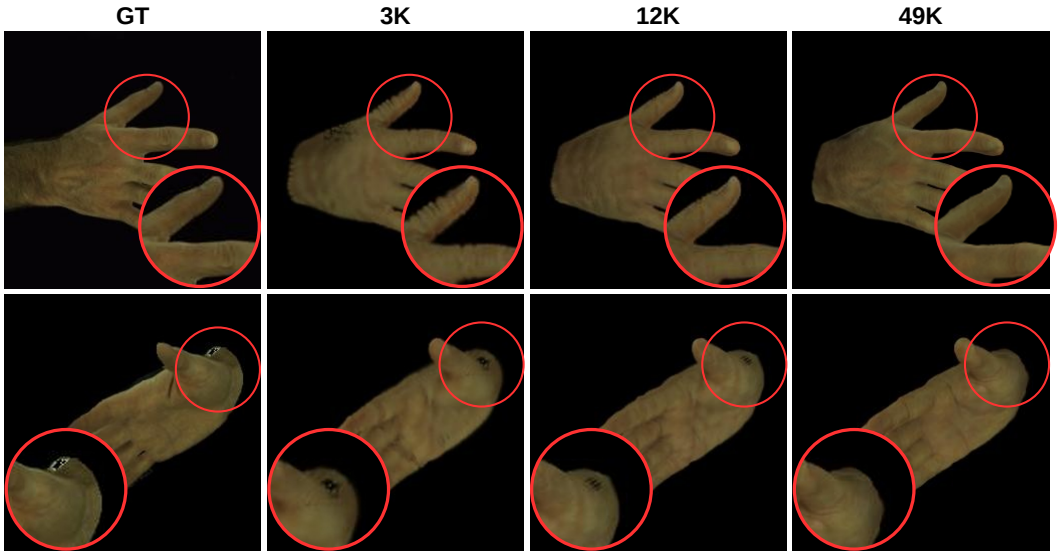


Figure S2: Qualitative effect of the number of Gaussians. Columns: ground truth (GT) and the 3K, 12K, and 49K settings of Table S6. Red circles mark regions magnified in the corner insets.

Table S6: Ablation on the number of Gaussians (MANO subdivision level). All variants are full models differing only in the subdivision level.

k	#Gaussians	PSNR↑	SSIM↑	LPIPS↓
1	3,093	26.11	0.859	0.168
2	12,337	25.92	0.860	0.147
3	49,281	**26.21**	**0.866**	**0.133**

F.1 Effect of the Number of Gaussians

InstantHand anchors Gaussian primitives on a subdivided MANO surface ($\mathcal{S}_k$ in Eq. 2 of the main paper), so the subdivision level directly controls the density of query anchors and therefore the number of canonical Gaussians. Subdivision levels $k = 1$, $k = 2$, and $k = 3$ produce 3,093, 12,337, and 49,281 Gaussian anchors, respectively, and the main model uses $k = 3$. A lower subdivision level provides a compact anchor set but may underrepresent thin fingers, hand boundaries, and local surface details, while a higher level provides denser surface coverage. Each variant is trained with the full two-stage pipeline and evaluated as a complete model under the novel-pose/novel-view protocol, and only the subdivision level differs.

Table S6 and Fig. S2 show the effect. The figure compares full-model reconstructions, rendered on the same black background, with ground truth. SSIM and LPIPS improve consistently with denser anchors (LPIPS $0.168 \rightarrow 0.147 \rightarrow 0.133$), while PSNR stays within 0.29 dB. In the sparse 3K setting, the gaps between neighboring splats appear as serrated finger contours (red circles), and surface detail is coarser. Such local, high-frequency defects barely move pixel-averaged PSNR but are captured by LPIPS. With denser anchors the contours become smooth and continuous, and at 49K the finger silhouettes are sharp. Training at the next level, $k = 4$ (196,993 anchors), runs out of memory on our training GPUs (RTX A6000, 48 GB) and was therefore not feasible with our resources. Rendering the trained $k = 3$ prediction at $k = 4$ density without retraining is examined in Sec. I (Table S11).

F.2 Gaussian Attribute Constraints

Our canonical parameterization deliberately fixes opacity and canonical rotation and constrains scales to be isotropic (Sec. A.1), as in prior monocular Gaussian-avatar designs [57, 58]. This subsection isolates the cost of relaxing each constraint in the hand setting.

Matched protocol. Boundary supervision is retained in *every* configuration of Table S7, with all other settings unchanged and per-axis scales initialized isotropically, so the differences isolate the

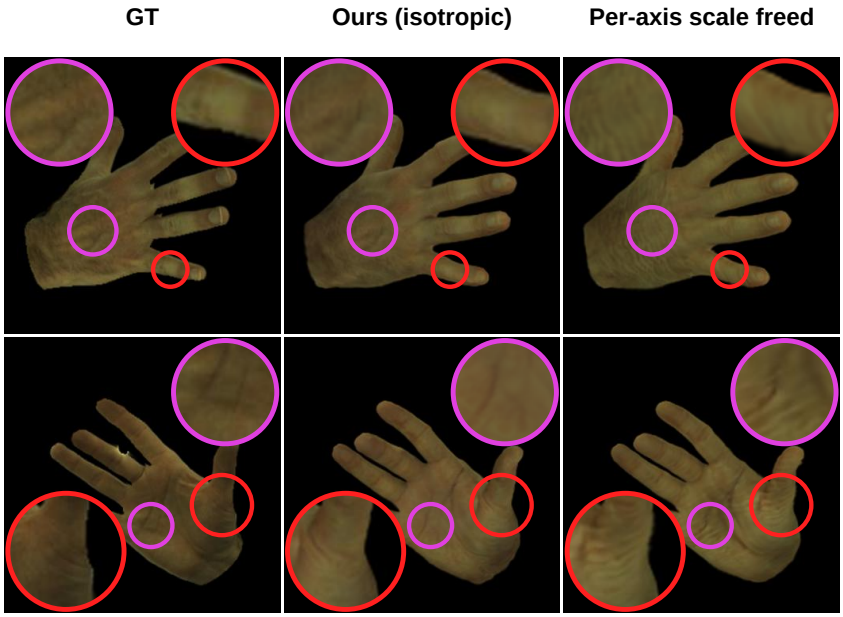


Figure S3: Artifacts under per-axis Gaussian scales. Ground truth, the isotropic model (ours), and the per-axis-scale-freed variant of Table S7 on identical frames. Magenta circles mark interior regions and red circles silhouette regions, each magnified in the matching corner inset.

Table S7: Effect of relaxing the Gaussian attribute constraints (InterHand2.6M). Boundary supervision is retained in every row.

Configuration	PSNR↑	SSIM↑	LPIPS↓
Ours (isotropic; rotation, opacity fixed)	**26.21**	**0.866**	**0.133**
Per-axis scale freed	25.58	0.856	0.142
Rotation and opacity freed	25.74	0.862	**0.133**
All three freed	25.50	0.856	0.148

attribute parameterization. The effect of boundary supervision under the constrained parameterization is reported separately in Table 4 of the main paper, where removing SAB lowers PSNR by 2.32 dB, to 23.89 dB.

Under matched supervision, freeing per-axis scale alone reduces PSNR by 0.63 dB. Freeing rotation and opacity reduces PSNR by 0.47 dB. Because rotation does not alter the covariance of an isotropic Gaussian, this row primarily measures the effect of learning opacity. Freeing all three produces the largest degradation, 0.71 dB, and worsens SSIM and LPIPS as well. Curved finger geometry is instead captured by increasing the density of small isotropic Gaussians (LPIPS $0.168 \rightarrow 0.133$ from 3K to 49K, Table S6). Rendering the trained 49K prediction at 197K density with scales $0.6\, s_i$, without retraining, lowers it further to 0.129 (Table S11).

Artifacts under per-axis scales. Because the metric gap alone does not identify *how* the anisotropic variant fails, we compare the variant with only per-axis scales freed and the isotropic model on identical frames, so that observed differences are attributable specifically to the scale parameterization. Figure S3 shows two artifact types, with boundary supervision retained in both models. The magenta circles fall on the dorsum (top) and the palm (bottom), and the red circles on a fingertip and the thumb. The isotropic model preserves the vein and crease structure of the ground truth, whereas the per-axis variant replaces it with parallel bands and widens the silhouette fringe. *(i) Striated interiors.* Broad, smooth regions such as the back of the hand exhibit quasi-parallel, regularly spaced bands, producing a corduroy-like ripple consistent with elongated splats becoming similarly aligned. These patterns are distinguishable from anatomical texture, since veins branch and follow the hand geometry whereas the artifacts remain approximately parallel and uniformly spaced. The distortions occur *within* the rendered surface rather than as holes or transparent regions, which is why opacity-based

Table S8: Sensitivity to source-side MANO and mask errors on InterHand2.6M. (a) Controlled perturbations around the ground-truth source cues. (b) Ground-truth source cues replaced by SAM 3 and WiLoR outputs. The last row also replaces the source camera.

(a) Controlled perturbations around ground truth

Source perturbation	PSNR↑	SSIM↑	LPIPS↓	Bound. F↑	B-IoU↑	Leakage↓
Clean GT inputs	26.21	0.866	0.133	0.698	0.562	0.234
Pose noise ($\sigma_{\text{pose}} = 5^\circ$)	26.12	0.864	0.134	0.698	0.562	0.234
Pose noise ($\sigma_{\text{pose}} = 10^\circ$)	26.05	0.863	0.134	0.698	0.562	0.234
Mask erosion (2 px)	26.28	0.867	0.132	0.704	0.565	0.229
Mask dilation (2 px)	26.24	0.864	0.134	0.697	0.561	0.237

(b) Estimated source-cue substitution

Source mask	Source MANO	PSNR↑	SSIM↑	LPIPS↓	Bound. F↑	Leakage↓
GT	GT	26.21	0.866	0.133	0.698	0.234
SAM 3	GT	26.12	0.865	0.133	0.695	0.239
GT	WiLoR	26.07	0.863	0.134	0.689	0.235
SAM 3	WiLoR	25.99	0.862	0.135	0.686	0.239
SAM 3 + WiLoR + WiLoR camera		25.99	0.863	0.134	0.685	0.240

measures such as interior coverage barely register them, while LPIPS degrades from 0.133 to 0.142. *(ii) Feathered silhouettes.* The per-axis model produces a wider semi-transparent fringe around the hand boundary, consistent with elongated Gaussians extending across the true contour, so finger boundaries appear smeared while the matched isotropic renders remain substantially sharper.

F.3 Sensitivity to Source-side MANO and Mask Errors

We evaluate how errors in the source-side MANO parameters and hand mask affect novel-pose/novel-view rendering on InterHand2.6M. All experiments use the protocol-designated source frame of Sec. C.1, fixed network weights, and the same evaluation protocol and metrics as the main paper. Only the specified source cue is changed: the source RGB image and crop, target MANO parameters and cameras, and ground-truth target evaluation masks remain fixed. Results are averaged over the same target frames. The estimated-cue substitution rows appear in Table 3 (bottom) of the main paper, and Table S8 reports the complete study with boundary-level metrics.

Controlled perturbations. For pose perturbations we add independent zero-mean Gaussian noise directly to the 45 articulated MANO axis–angle components of the source frame,

$$\boldsymbol{\theta}' = \boldsymbol{\theta} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(0, \sigma_{\text{pose}}^2), \tag{S7}$$

where $\sigma_{\text{pose}} \in \{5^\circ, 10^\circ\}$ is the component-wise standard deviation. The global wrist orientation, translation, shape, and camera remain unchanged, and each pose-noise condition is averaged over three random draws. For mask perturbations we apply two iterations of binary erosion or dilation to the source mask before the otherwise identical source preprocessing, and the source image and crop remain unchanged.

Estimated source-cue substitution. We separately replace the ground-truth source cues with outputs from practical upstream estimators. SAM 3 [54] replaces the source hand mask, and WiLoR [37] replaces the source MANO pose and shape fields. The mask–MANO combinations form a 2×2 substitution study that separates the individual and combined effects of the estimated cues. We additionally report the *fully estimated* configuration, in which the jointly regressed WiLoR camera also replaces the ground-truth source camera, so that every source-side input originates from a single RGB image. This camera differs from the calibrated ground truth in both intrinsics and extrinsics. Unlike the controlled perturbations, these estimator outputs contain structured and spatially correlated errors.

Table S8(a) shows that controlled pose perturbations produce gradual but small appearance changes: relative to clean ground-truth inputs, $\sigma_{\text{pose}} = 5^\circ$ and 10° reduce PSNR by 0.09 and 0.16 dB and SSIM by 0.002 and 0.003, while boundary F-score, boundary IoU, and leakage are unchanged at the reported precision, as expected since the source pose enters only the appearance pathway. Two-pixel

Table S9: Inference-time sensitivity to observation-reliability hyperparameters. One parameter is varied at a time with all weights fixed, and bold marks the default.

κ	τ_m	ζ	PSNR↑	LPIPS↓
(a) InterHand2.6M (novel-pose/novel-view)				
0.25	0.5	95	26.21	0.133
0.5	**0.5**	**95**	**26.21**	**0.133**
1.0	0.5	95	26.20	0.133
0.5	0.3	95	26.22	0.133
0.5	0.7	95	26.19	0.133
0.5	0.5	90	26.22	0.134
0.5	0.5	99	26.19	0.132
(b) HanCo (zero-shot, sequence 0191)				
0.25	0.5	95	20.16	0.187
0.5	**0.5**	**95**	**20.18**	**0.188**
1.0	0.5	95	20.21	0.188
0.5	0.3	95	20.18	0.188
0.5	0.7	95	20.18	0.188
0.5	0.5	90	20.17	0.188
0.5	0.5	99	20.20	0.187

mask erosion and dilation change PSNR by at most 0.07 dB and the boundary F-score by at most 0.006, so the model is insensitive to uniform mask morphology at this scale.

Structured estimator errors produce a somewhat larger but still limited gap. With ground-truth MANO fixed, replacing the source mask with SAM 3 decreases PSNR by 0.09 dB and boundary F by 0.003. With the ground-truth mask fixed, replacing the source MANO with WiLoR decreases PSNR by 0.14 dB and boundary F by 0.009. Unlike pose noise, which leaves the boundary metrics unchanged, WiLoR also replaces the shape $\boldsymbol{\beta}$ that places the canonical anchors (Eq. 2 of the main paper), which is consistent with its change in boundary F. Using both estimated cues decreases PSNR by 0.22 dB and boundary F by 0.012. At the reported precision, the combined degradation is close to the sum of the individual changes, which suggests limited interaction between the two source-cue errors. Replacing the source camera as well, which changes the source intrinsics and extrinsics, leaves PSNR unchanged at the reported precision and moves each boundary metric by only 0.001. The fully estimated configuration thus sits 0.22 dB below clean ground-truth inputs.

Qualitative results under fully estimated inputs on unseen images are shown in Figs. S8 and S9.

F.4 Inference-time Sensitivity of Observation Reliability

We keep the trained checkpoint fixed and vary the reliability hyperparameters of Eq. 8 of the main paper one at a time around the default $(\kappa, \tau_m, \zeta) = (0.5, 0.5, 95)$. For each setting, γ is recomputed once from the Stage-1 source-view render and the complete appearance-composition forward pass is rerun, while no network weights or canonical Gaussian attributes are updated. This is therefore a local inference-time sensitivity analysis rather than a comparison of separately retrained models. We run the sweep under two protocols, the InterHand2.6M protocol used for the other ablations and the held-out HanCo protocol of Sec. C.2, so that the analysis covers both subject-level and dataset-level generalization.

Table S9 shows a plateau rather than a peak. Across the seven settings, PSNR spans 0.03 dB (26.19–26.22) and LPIPS 0.002, even though κ varies by a factor of four and τ_m by more than a factor of two. For scale, removing reliability gating entirely costs 3.03 dB (Table 3 of the main paper), roughly two orders of magnitude more than any hyperparameter change here. The behavior of reliability gating is therefore governed by its structure, namely square-root compression, median-relative thresholding, and percentile normalization (Sec. A.4), rather than by the exact constants. The same insensitivity holds under dataset-level transfer: on the held-out HanCo sequence the same seven settings span 0.05 dB PSNR (20.16–20.21) and 0.001 LPIPS. In practice, reliability gating requires no per-dataset tuning.

F.5 Rationale for Two-stage Training

The two-stage schedule enforces a structural separation between canonical representation learning and source-conditioned appearance adaptation. Stage 1 learns the canonical predictor, and Stage 2 freezes it in full, including the base-color head f_{rgb}, while optimizing only the source-conditioned appearance module of Sec. 3.3 of the main paper (training details in Sec. A.1). By construction, source-view appearance supervision in Stage 2 cannot alter the canonical geometry or base colors, so novel-pose rendering starts from exactly the Stage-1 representation.

The frozen canonical branch also gives the appearance module a well-defined prior to fall back on: wherever source evidence is missing or unreliable, the composed color reverts to the Stage-1 base color plus the source-level shift $\mathbf{\Delta}_s^{aid}$ shared by all Gaussians. The full model improves source-view fidelity (Table S2), and observation reliability controls what reaches novel views. A color taken from the source is only as trustworthy as the evidence behind it. When transfer is not weighted by γ, novel-pose PSNR falls from 26.21 to 23.18 dB (Table 3 of the main paper), which is consistent with colors of weakly observed Gaussians being carried to novel views. The two stages thus keep the two roles separate: Stage 1 provides a stable prior for every Gaussian, and Stage 2 learns only where, and how strongly, source evidence should replace it, together with a source-level shift that propagates global appearance cues to every Gaussian.

G Discussion

G.1 Relation to Boundary Losses

Boundary-focused objectives and metrics are established in segmentation, where regional losses are complemented by contour-based losses [59] and by boundary-aware evaluation [56]. Rendered alpha, however, differs from the pixel-labeling setting those losses assume, and each element of SAB (Eq. 16 of the main paper and Sec. A.3) addresses one of the resulting differences. Per-pixel alpha is not an independent prediction but a composite of many projected 3D primitives, so band supervision acts through the rasterizer on primitive centers and scales rather than relabeling pixels. A one-sided outer penalty admits a degenerate optimum with no segmentation analogue, namely shrinking the hand until nothing crosses the contour. The paired interior coverage term closes this loophole, so leakage cannot be "fixed" by shrinking the hand. Soft alpha at boundaries is correct anti-aliased rendering rather than an error, and the tolerances and the robust penalty leave it unpenalized. Finally, the schedule $\eta(n)$ reflects the 3D setting by activating the penalty only once the bands are geometrically meaningful for the evolving primitives. All SAB ablations keep the constrained parameterization of Sec. A.1 fixed, so the boundary-aware silhouette evaluation (Sec. C.4 and Table 4 of the main paper) measures the effect of SAB under identical constraints.

G.2 View-dependent Appearance, Illumination, and Multi-frame Extension

Our appearance model predicts view-independent Gaussian colors and does not employ view-dependent representations such as spherical harmonics (SH). The two address different problems. The ambiguity we target is observational. A single source image observes essentially one side of the hand, so the core difficulty is deciding where source evidence is reliable and how to complete unobserved regions, which is what reliability-gated transfer and the global source identity address. A similar observational problem arises for human avatars built from monocular video, where FlexiAvatar [45] optimizes only the visible body regions and completes entirely unseen regions with diffusion-generated texture. View-dependent appearance modeling instead captures viewpoint- and illumination-dependent reflectance after the underlying appearance has been recovered, and it typically requires multi-view observations to be constrained. Under single-image supervision, per-Gaussian SH coefficients would be severely underdetermined. The two mechanisms are therefore complementary rather than alternatives: view-dependent terms could be attached to our composed colors without changing the geometry stabilization or the reliability-gated transfer.

A related limitation is that shading baked into the source image is transferred as albedo and thus carried to novel poses. This is as much a property of the setting as of the model: a single image does not determine illumination and albedo separately. Explicit albedo–shading decompositions such as those of HandAvatar [20] and OHTA [23] model self-occlusion shading learned from video or multi-identity data, but the scene illumination of a single source image, such as its exposure or color

Table S10: Multi-frame evidence integration at inference time (InterHand2.6M). The same trained model is used without retraining or additional parameters, and frames beyond the first are selected to maximize the union coverage of observed Gaussians ($\gamma_i > 0$).

Source frames	Coverage	PSNR↑	SSIM↑	LPIPS↓
1	0.45	26.21	0.866	0.133
2	0.68	26.53	0.867	0.131
4	0.71	26.52	0.867	0.132
8	0.96	**26.57**	**0.868**	**0.130**

cast, is not identifiable from that image alone. The per-subject baselines, which are fitted to the same source frame, are not exempt. In Part 4 of the supplementary video (Sec. K), they also depart from the ground truth in tone and shading, with dark discoloration for OHTA and a uniform tone without shading for HandAvatar. Disentangling illumination from intrinsic appearance is therefore likewise future work, and the corresponding failures are documented in Sec. I. Finally, the formulation already extends to multiple source frames at inference time, without retraining or additional parameters. Since γ is computed per source frame from the same rasterizer by-products, each additional frame is mapped onto the shared canonical Gaussians, the most reliable observation is selected for each Gaussian, and the colors are composed with Eq. 13 of the main paper. Related work similarly registers observations from disjoint viewpoints into a shared 3D space for multi-person mesh tracking across video shots [60]. Frames beyond the first are selected to maximize the union coverage, the fraction of canonical Gaussians with nonzero reliability ($\gamma_i > 0$) in at least one source frame, and they span cameras $12°$–$178°$ apart as well as different poses and time steps. Table S10 reports the results under the InterHand2.6M protocol of Sec. C.1.

A single complementary view that covers the unseen palm or dorsum improves all three metrics ($+0.32$ dB PSNR). The gain follows newly observed surface rather than the number of frames. From two to four frames, coverage grows by only 0.03 and PSNR is effectively unchanged (-0.01 dB), whereas from four to eight frames coverage grows by 0.25 and all three metrics improve again. The gain per unit of coverage nevertheless diminishes near full coverage, from $+0.32$ dB for the first 0.23 of added coverage to $+0.05$ dB for the last 0.25, for two reasons. First, overlapping observations of an already covered region are handled by reliability-gated selection rather than learned fusion. The reliability separates observed from unobserved Gaussians within a single view but is not designed to rank several valid observations of the same region, so near-ties are resolved in favor of the base view. Second, the last uncovered regions are visible only at grazing angles, where they receive low reliability and occupy little area in the evaluation views. Learned multi-view fusion, trained with explicit multi-view or temporal conditioning, could combine overlapping observations instead of selecting among them and is a natural direction for future work.

G.3 Ethical Considerations

Hand imagery can act as a soft biometric: skin texture, vein-like patterns, and nail features may be identifying, and a reconstructed avatar could in principle support impersonation in hand-centric interfaces. All training and evaluation data come from published research datasets used under their licenses, and our model neither identifies individuals nor is trained for recognition. Feed-forward hand avatar reconstruction may benefit virtual reality, telepresence, digital human animation, and human–computer interaction by enabling efficient avatar creation from minimal input. At the same time, as with other avatar and generative modeling technologies, reconstructed avatars could be misused for identity impersonation or deceptive synthetic media. Deployments that capture user hands should obtain informed consent and safeguard reconstructed avatars as biometric-adjacent data.

H Additional Qualitative Results

H.1 Qualitative Effect of the Soft Alpha Band

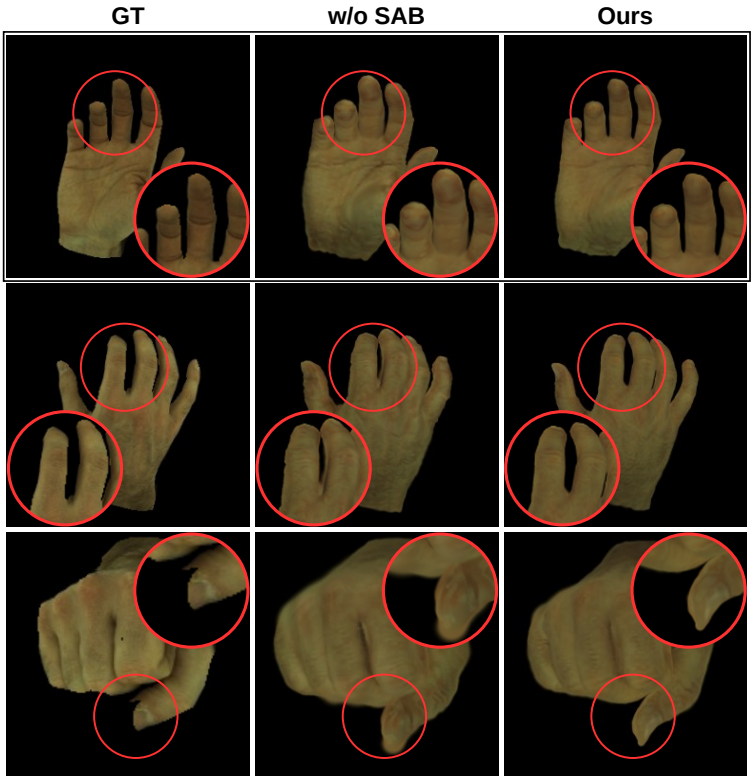


Figure S4: Qualitative effect of the Soft Alpha Band (SAB). Columns: ground truth (GT), without SAB (w/o SAB), and with SAB (Ours). The black box marks the input frame, and the remaining rows are novel poses. Red circles mark boundary regions magnified in the corner insets.

Figure S4 visualizes the boundary-level failure mode diagnosed in the main paper. Since the "w/o SAB" model retains the same constrained parameterization, the comparison isolates the effect of the regularizer. Without SAB, residual leakage concentrates in the thin band around finger boundaries and appears as blurred, apparently thickened contours, most visibly where adjacent fingers meet (red circles), whereas the full model keeps these silhouettes sharp. This qualitative gap is consistent with Table 4 of the main paper, where SAB raises the boundary F-score from 0.366 to 0.698 and more than halves alpha leakage ($0.555 \rightarrow 0.234$), while interior coverage remains essentially unchanged (1.000 vs. 0.999). SAB therefore suppresses leakage without eroding valid hand regions, rather than improving the metrics by shrinking the silhouette.

H.2 Qualitative Comparison with the Recipe Ports



Figure S5: Qualitative comparison with the recipe ports of Table 4 of the main paper. InterHand2.6M novel-pose/novel-view frames, with the source camera in the top row. Red circles mark inter-finger gaps and contours magnified in the corner insets.

Figure S5 shows matched renderings of the two recipe ports (Sec. E.2) alongside ours, on four frames chosen to expose inter-finger gaps and finger contours from the dorsal, palmar, and lateral sides. Each red circle marks the same location in every column. Both ports render the hand interior well, consistent with their interior coverage of 1.000, but blur the silhouette into a translucent fringe: the gaps between adjacent fingers partially fill with alpha, and finger and thumb contours widen, most visibly between the ring and little fingers (second row) and at the thumb–index junction (top row). The two ports look alike, in line with their close boundary F-scores of 0.299 and 0.267. Ours keeps the gaps open and the contours sharp on the same frames. The constrained model without SAB, which lies between the ports and ours in boundary F-score and leakage, is compared separately in Fig. S4.

H.3 Qualitative Self-reconstruction Diagnostics

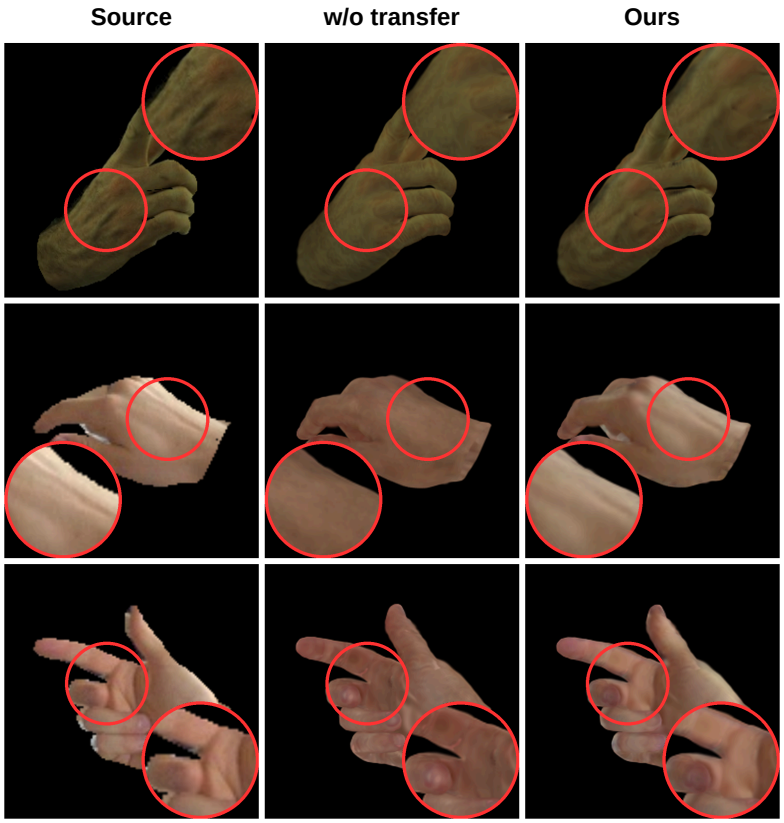


Figure S6: Qualitative source-view appearance diagnostic with geometry fixed (protocol in Sec. C.5). Columns: source observation (Source), Stage-1 base colors (w/o transfer), and composed colors (Ours). Red circles mark regions magnified in the insets.

Figure S6 shows the diagnostic of Table S2 on representative frames. Because all non-color attributes are identical between w/o transfer and Ours, every visible difference is attributable to the source-conditioned appearance module: relative to the Stage-1 base colors, the full model recovers the observed skin tone and source-observed local texture (circled regions), in line with the source-view gains of Table S2.

H.4 Branch-level Diagnostics of the Appearance Module

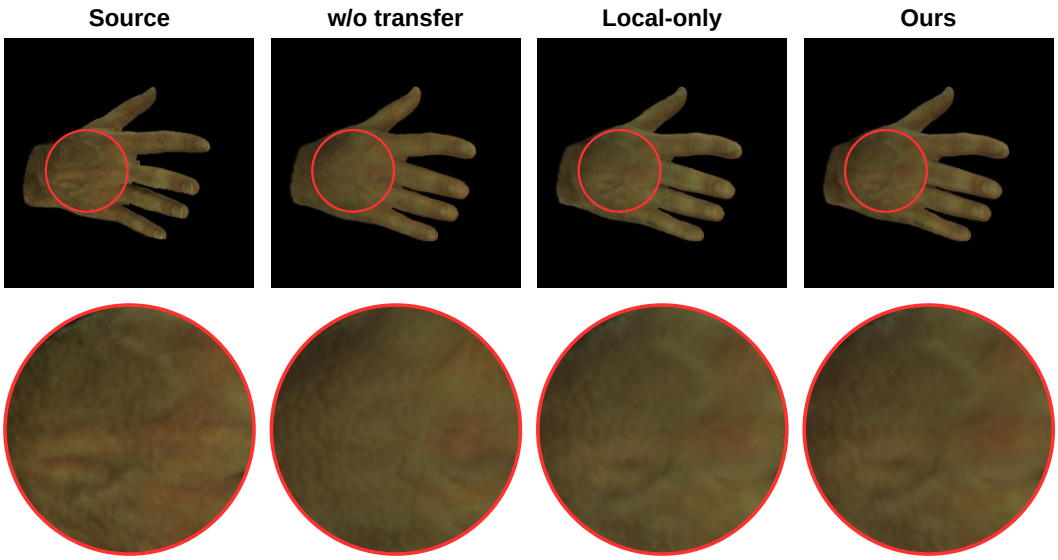


Figure S7: Branch-level analysis of source-conditioned local detail transfer. Columns: source observation (Source), Stage-1 base colors (w/o transfer), the local-only variant (Local-only), and the full model (Ours) under the same source pose and view. The red circle marks the region magnified below.

Complementing the module-level evaluation above, we dissect the source-conditioned appearance module into branch-isolated variants. These are diagnostic probes of the appearance composition, not separate model configurations for the main comparison. The w/o transfer variant uses the Stage-1 base colors without source-conditioned refinement. The local-only variant uses Gaussian footprint-aligned local evidence and is designed to transfer source-observed local details such as skin wrinkles and vein-like texture. The full model combines footprint-aligned local observations, rendering-contribution-based reliability, the global source identity, and the Stage-1 base colors.

Figure S7 compares the variants under the same source pose and view, so that source-observed details can be traced from the input to each branch. The circled region on the dorsum contains subject-specific texture, such as vein-like patterns and skin creases, that the w/o transfer variant attenuates. The local-only branch and the full model both recover it from the source observation and are nearly indistinguishable in this view, consistent with the local-evidence pathway carrying source-observed detail in the full model. The contribution of the global source identity cannot be seen from the source view, since it matters most for Gaussians that the source does not observe. It appears instead on the prior-completed sides in the palm-facing renderings of Fig. S9 and the 360° rotations of the supplementary video (Sec. K). This branch analysis is qualitative, and its quantitative counterparts are Table 3 of the main paper and Table S2.

H.5 Generalization to Unseen Images under Fully Estimated Inputs

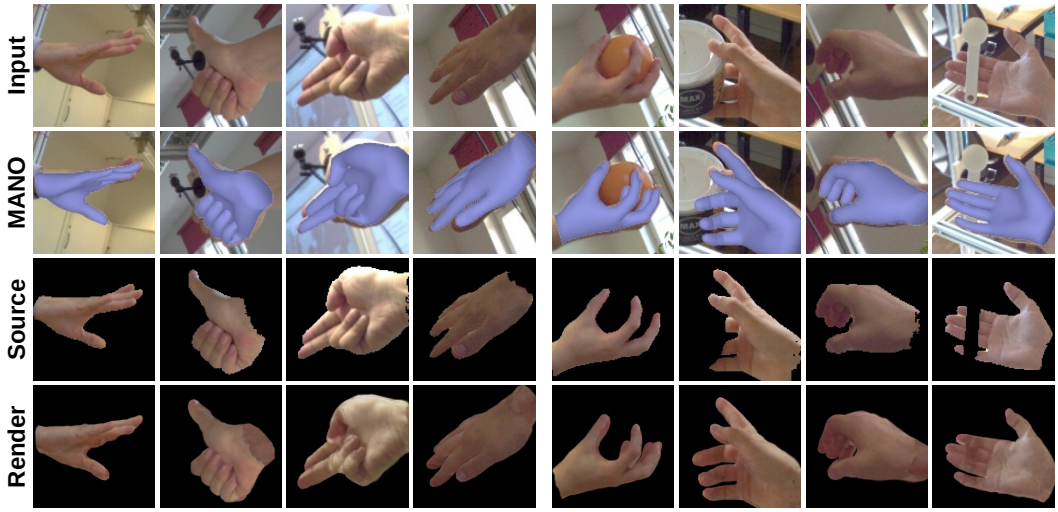


Figure S8: Qualitative generalization to unseen FreiHAND [61] test images under fully estimated inputs. Rows: input frame (Input), WiLoR [37] pose overlay (MANO), SAM 3 [54] source crop (Source), and our render (Render). No ground-truth annotations or test-time optimization are used.

We show source-view reconstructions of unseen FreiHAND [61] test images to evaluate generalization beyond the benchmark datasets of the main paper. Unlike in the diagnostic of Sec. C.5, every input here is estimated. Given only a single input image, our method reconstructs plausible canonical Gaussian hand avatars in a fully feed-forward manner without subject-specific optimization or test-time fitting.

For these results, MANO parameters and camera estimates are obtained using WiLoR [37], and hand masks using SAM 3 [54]. The reconstructed avatars are rendered with the estimated camera parameters. Figure S8 shows eight representative examples. Each column documents every estimated input alongside the output, and FreiHAND's ground-truth MANO annotations are not used. Failure cases are shown in the same format in Fig. S10 (Sec. I). The left four columns show unoccluded hands, and the right four involve hand–object interaction. The examples cover side views, thin finger structures, articulated and near-closed poses, and diverse skin tones. The overlays also reveal that the estimated poses are imperfect, as fingertips visibly deviate from the mesh in several columns. Nevertheless, the reconstructed avatars remain faithful to the observed appearance. In the rightmost column, the held object crosses the fingers near their bases and splits the estimated mask into disjoint pieces. The avatar still completes the masked-out finger segments with coherent appearance, because appearance is anchored on the canonical MANO surface and reliability gating falls back to the source-conditioned appearance prior where source evidence is missing. Unlike the benchmark evaluations, these results use estimated rather than ground-truth pose, camera, and mask inputs, so they illustrate the behavior of the full pipeline under real, uncontrolled upstream estimation errors. The corresponding quantitative comparison is Table 3 of the main paper and Table S8(b). Novel-pose animation of avatars reconstructed under the same fully estimated inputs is shown in Fig. S9 (Sec. H.6).

H.6 Novel-pose Animation from a Single Image



Figure S9: Novel-pose animation from a single image. In each block, the top row shows the driving MANO poses and each subject row the source image followed by our renderings in those poses. All inputs are fully estimated. Best viewed zoomed in.

Figure S8 shows how faithfully an avatar reproduces its own source view, whereas Fig. S9 shows whether avatars reconstructed in the same way remain consistent once they are driven away from it. Eight subjects are reconstructed with the fully estimated pipeline of Sec. H.5, using WiLoR [37] pose and camera and a SAM 3 [54] mask, with no ground truth and no test-time optimization. From top to bottom, their sources are an in-the-wild image released with OHTA [23], FreiHAND [61], FreiHAND, a HARP [22] subject excluded from the training split, FreiHAND, ARCTIC [62] (not part of the training data), FreiHAND, and ARCTIC. The subjects are split into two blocks of four,

and each block is driven by its own six poses, rendered from a camera fixed in the MANO frame. The two blocks share no poses. Within a block, every column therefore shows all four subjects in the same pose and from the same viewpoint. The top row of each block renders the driving MANO mesh from the same camera, and the second block includes palm-facing views.

Subject-specific texture is carried across poses without drifting toward a generic hand. The vitiligo pattern (row 1), the tattoo (row 4), and, more faintly, the motion-capture markers (row 8) stay attached to their anatomical locations under finger articulation and global rotation, including on the wrist and on fingers that are foreshortened or partially hidden in the source view. Fine detail is preserved only coarsely, as the line work of the tattoo is blurred and the markers appear as small light dots that are best seen in the supplementary video (Sec. K). Surfaces that are unobserved in the source are completed by the source-conditioned prior and remain coherent across views, as the palm-facing columns of the second block show for dorsal-side inputs. The second row deliberately uses the palm-facing input of the rightmost column of Fig. S8, whose SAM 3 mask is corrupted with missing finger segments. We use the mask as is, so the dorsal appearance of this avatar is entirely completed from the prior. The avatar nevertheless remains animatable and coherent, which complements the mask-failure analysis of Sec. I. No per-frame refinement is involved, since all Gaussian attributes are fixed after the single feed-forward pass. The identity consistency across columns is therefore a property of the canonical representation rather than of per-pose fitting. The same eight avatars are shown under continuous driving in the supplementary video, and four of them are also shown in full 360° rotation (Sec. K).

I Failure Cases and Limitations

Representative failure cases are shown in Fig. S10, in the input-documentation format of Fig. S8: the input frame (Input), the WiLoR [37] pose overlay (MANO), the SAM 3 [54] source crop (Source), and our reconstruction (Render).

First, our method inherits the inherent ambiguity of single-image hand reconstruction. Under severe self-occlusion the occluded surface is never observed, so its appearance is completed from the source-conditioned prior rather than transferred from evidence, and thin local structures that are only weakly observed are reconstructed with reduced fidelity. The upstream pose estimate on which the avatar is anchored also becomes less reliable in such views. Where the boundary of the observed set crosses a thin or strongly foreshortened part, such as the thumb in a dorsal view, source-driven and prior-completed colors meet on the same structure. If the observed tone differs from the prior, as when shading is baked into the source (the third point below), the transition appears as a visible seam rather than a gradual blend. Sec. A.4 discusses this boundary, and Sec. J.3 measures where such seams can occur (Fig. S12).

Second, all Gaussian attributes are fixed after the single forward pass. This keeps the identity consistent across poses, but the avatar does not model pose-dependent appearance. Wrinkles that form and shading that changes under strong articulation are not synthesized, whereas wrinkles visible in the source are transferred as texture. The representation also ends where the MANO template ends, and the avatar has no forearm. When the hand is viewed along the forearm, as in the 360° rotations of the supplementary video, the open wrist boundary is visible, including the inner side of the template surface on which the Gaussians lie. The MANO-based baselines likewise end at the wrist. The anchor set is fixed as well, and the isotropic scales are predicted once in the canonical pose. Under strong flexion, skinning stretches the surface on the convex side of a joint and spreads neighboring anchors apart, but the splats are not enlarged to follow this stretch. Where the spacing exceeds the splat footprint, neighboring rows of splats separate, and the lattice becomes visible as fine striations at the knuckles. Enlarging the splats relative to the anchor spacing would close these gaps, but the same splats also form the silhouette. Rendering the prediction on the next subdivision level, which halves the anchor spacing, without shrinking the scales (row "$k = 4$, s_i" of Table S11) doubles the footprint-to-spacing ratio. This raises alpha leakage from 0.234 to 0.293 and lowers the boundary F-score from 0.698 to 0.661, which is exactly the leakage that the scale ceiling and SAB are designed to suppress. The striations are a local effect, and since curvature is handled by density rather than by anisotropy (Sec. F.2), a finer subdivision level reduces them. Training at $k = 4$, however, runs out of memory on our training GPUs (RTX A6000, 48 GB) and was not feasible with our resources.



Figure S10: Representative appearance failure cases. Columns: Input, MANO, Source, and Render (the row labels of Fig. S8). Red circles mark the affected regions in the third and bottom rows.

These striations can nevertheless be largely removed at inference without retraining. We keep the trained $k = 3$ network and its 49,281 predicted Gaussians and place Gaussians on all 196,993 anchors of the next subdivision level ($k = 4$, continuing Table S6). The 49,281 anchors shared with $k = 3$ keep their predicted attributes, and each of the 147,712 new anchors lies at the midpoint of a $k = 3$ edge and takes the average of all predicted attributes of the two endpoints, including the composed colors $\mathbf{c}_i^*$. Since the anchor spacing halves, every scale s_i is then multiplied by a single fixed factor, and the $k = 4$ Gaussians are skinned with the $k = 4$ weights and rendered as usual. This densification is a fixed post-processing step on the predicted Gaussians. It involves no per-subject or per-image optimization and uses neither the source image nor ground truth. Table S11 reports the sweep over the factor. Each of $0.5\,s_i$, $0.6\,s_i$, and $0.7\,s_i$ improves LPIPS, boundary F, and alpha leakage on both benchmarks, while PSNR changes by -0.04 to -0.06 dB on InterHand2.6M and by $+0.12$ to $+0.30$ dB on HanCo, so the conclusion does not depend on the exact factor within this range. We use $0.6\,s_i$, chosen from the InterHand2.6M results without using HanCo, for Fig. S11 and the supplementary video (Part 4, Sec. K). Figure S11 shows three HanCo evaluation frames spanning the fist at the start of the sequence, selected by pose rather than by any metric. The jagged ground-truth boundary comes from the HanCo mask. In the $k = 3$ rendering, the anchor lattice appears as striations at the knuckles and near the wrist, where the avatar ends at the MANO template boundary, and the densified rendering removes most of them. Feeding the $k = 4$ anchors to the network directly instead worsens all five metrics of Table S11 on both benchmarks, so the gain comes from the rendering density rather than from the network. The overhead is small. Measured in the same run as the $k = 3$ avatar, the rendering speed stays within 1%, the Gaussian parameters grow from 3.4 MB to 13.4 MB, and the GPU memory added by rendering a frame grows from 13 MB to 38 MB. For comparison, the rendering peak of Sec. C.6, which also includes the CUDA context and the network, is 1.79 GB. Faint striations can remain at the most strongly flexed knuckles (bottom row of Fig. S11), and the factor is a single global constant rather than adapted to the local anchor spacing. All other results in this paper use the $k = 3$ avatar.

Table S11: Inference-time densification (Sec. I). The same trained $k = 3$ network is used throughout. "$k = 4$ input" feeds the $k = 4$ anchors to the network. The other $k = 4$ rows interpolate the $k = 3$ prediction onto the $k = 4$ anchors and multiply every scale s_i by the stated factor, and "$k = 4$, s_i" keeps the scales unchanged. Boundary metrics follow Sec. C.4. Bold marks the setting used for Fig. S11 and the supplementary video.

	InterHand2.6M					HanCo (zero-shot)				
Setting	PSNR↑	SSIM↑	LPIPS↓	Bnd. F↑	Leak.↓	PSNR↑	SSIM↑	LPIPS↓	Bnd. F↑	Leak.↓
$k = 3$ (Tables 1, 2)	26.21	0.866	0.133	0.698	0.234	20.18	0.836	0.188	0.617	0.314
$k = 4$ input	25.98	0.862	0.137	0.671	0.247	19.75	0.829	0.190	0.559	0.373
$k = 4$, s_i	25.86	0.862	0.136	0.661	0.293	19.68	0.831	0.189	0.542	0.381
$k = 4$, $0.85\, s_i$	26.06	0.864	0.132	0.697	0.244	20.04	0.837	0.184	0.613	0.321
$k = 4$, $0.7\, s_i$	26.15	0.865	0.130	0.711	0.205	20.30	0.840	0.181	0.668	0.269
$\boldsymbol{k = 4, 0.6\, s_i}$	**26.17**	**0.865**	**0.129**	**0.708**	**0.185**	**20.41**	**0.842**	**0.179**	**0.693**	**0.239**
$k = 4$, $0.5\, s_i$	26.15	0.863	0.129	0.701	0.168	20.48	0.842	0.178	0.708	0.215

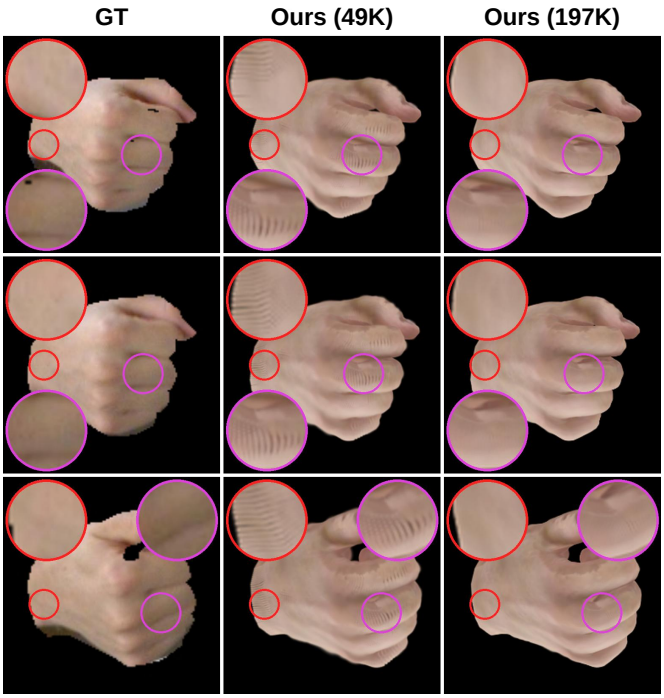


Figure S11: Inference-time densification. Columns: ground truth, the $k = 3$ avatar used throughout the paper (49K Gaussians), and the same prediction densified onto the $k = 4$ anchors (197K, scales $0.6\, s_i$, Table S11). Rows: three HanCo evaluation frames. Magenta circles mark knuckles and red circles the wrist, each magnified in the matching corner inset.

Third, Fig. S10 collects appearance failures that arise while the hand itself is fully observed. These failures fall into two regimes, which the input columns separate explicitly. In the first regime (top three rows), the Source crop retains the evidence that the failure concerns, so the failure lies in the appearance composition itself. The top two rows share one cause. Illumination and intrinsic appearance are entangled in the composed colors, because shading baked into the source is carried as albedo rather than separated from it (Sec. G.2). This entanglement has two signatures. A strongly overexposed, washed-out source is rendered with a more saturated skin tone than observed, and an underexposed tone is absorbed into the albedo and, since all attributes are fixed after the forward pass, carried to every novel pose. These two rows are left unmarked because the illumination shift affects the whole hand. The third row concerns a rare cosmetic attribute that the transfer does not carry. Dark nail pigment inside the source crop is composed as diffuse skin discoloration around the fingertips and nails rather than reproduced as painted nails. The nails in the source crop and the fingertip blotches they become are circled at the same image location in the Source and Render columns.

Table S12: Pose-stratified evaluation of our method. Terciles follow Sec. C.8, and ranges are in degrees.

	InterHand2.6M				HanCo (zero-shot)			
Tercile	Range	PSNR↑	SSIM↑	LPIPS↓	Range	PSNR↑	SSIM↑	LPIPS↓
Palm-orientation change								
Easy	0–79	26.37	0.863	0.133	0–82	20.67	0.842	0.169
Medium	80–115	25.67	0.858	0.140	82–128	19.70	0.826	0.212
Extreme	116–179	26.61	0.876	0.126	129–180	20.15	0.840	0.183
Finger-articulation change								
Easy	0–11	27.37	0.888	0.106	0–22	20.89	0.851	0.177
Medium	11–17	26.56	0.875	0.119	22–43	20.67	0.845	0.174
Extreme	17–56	24.68	0.834	0.175	43–56	18.96	0.812	0.213
All		26.21	0.866	0.133		20.18	0.836	0.188

Table S13: Articulation-stratified comparison on the finger-articulation terciles of Table S12.

	PSNR↑				LPIPS↓			
Tercile	Ours	OHTA	HandAvatar	BIGS*	Ours	OHTA	HandAvatar	BIGS*
InterHand2.6M								
Easy	27.37	27.23	24.90	21.30	0.106	0.103	0.150	0.168
Medium	26.56	26.43	24.21	20.38	0.119	0.115	0.163	0.191
Extreme	24.68	24.66	22.24	19.67	0.175	0.169	0.223	0.277
HanCo (ours zero-shot, baselines fitted to the source frame)								
Easy	20.89	20.12	19.88	18.60	0.177	0.177	0.194	0.224
Medium	20.67	20.53	20.77	17.99	0.174	0.172	0.185	0.228
Extreme	18.96	19.53	20.32	16.51	0.213	0.209	0.216	0.287

In the second regime (bottom row), the failure is inherited from a corrupted mask input, consistent with our assumption of externally provided segmentation. The source crop is torn at the wrist, and evidence along the tear boundary leaks into the composed appearance (the circles mark the wrist tear and the resulting brown discoloration). The contrast with the rightmost column of Fig. S8 delimits where reliability gating operates. When fragmentation removes a region outright, reliability gating routes it to the source-conditioned appearance prior and the region is completed coherently, as the second row of Fig. S9 shows for one such fragmented input. Evidence that survives at a defect boundary, however, is locally indistinguishable from valid observation and is accepted as such. In all cases shown, the MANO-anchored geometry remains plausible, the torn wrist silhouette of the bottom row is itself completed by the anchors, and the failures are confined to the appearance branch.

Separating illumination from intrinsic appearance is left for future work (Sec. G.2). The cosmetic-attribute failure reflects the limited coverage of such attributes in existing hand avatar datasets, and the mask-inherited failure diminishes as upstream segmentation improves. Extending feed-forward Gaussian hand avatars to more diverse real-world conditions and hand–object interaction scenarios remains an important direction for future work.

J Additional Analyses

This section analyzes the evaluation results along three axes that aggregate metrics do not separate, namely pose difficulty, temporal stability under continuous driving, and the reliability, opacity, and boundary continuity of the canonical avatar. The protocols are given in Sec. C.8, and the pose-stratified and temporal numbers use the same renders and ground truth as Tables 1 and 2 of the main paper.

J.1 Pose-stratified Evaluation

Table S12 splits both benchmarks into equal-count terciles of the palm-orientation and finger-articulation change from the source pose. On both benchmarks, the evaluation subject is unseen during training, and HanCo is also unseen as a dataset (Sec. B.1). The terciles therefore show

Table S14: Excess temporal variation on HanCo (Eq. S6, 75 consecutive-frame pairs per camera). Ours is zero-shot, and the baselines are fitted to the source frame. GT change: mean change between consecutive ground-truth frames. Method columns give E_t relative to it, where 0 means as much change as the footage and positive values mean additional variation. cam6 is the source view, and the last row is the mean of the absolute per-camera values.

Camera	GT change	Ours	OHTA	HandAvatar	BIGS*
cam3	0.0362	$+1.1\%$	-5.1%	-17.6%	$+37.8\%$
cam5	0.0319	$+7.3\%$	$+5.0\%$	-8.7%	$+50.5\%$
cam6 (source view)	0.0360	$+2.6\%$	-2.2%	-23.3%	$+30.7\%$
cam7	0.0412	$+8.0\%$	$+2.2\%$	-15.6%	$+34.9\%$
Mean of $\lvert\cdot\rvert$	0.0363	4.7%	3.6%	16.3%	38.5%

how the quality of an avatar reconstructed from a single image of an unseen hand, without test-time optimization, changes as the pose and view move away from that image. The effect of palm orientation is non-monotonic. On both datasets, the middle tercile is the lowest, 0.70 and 0.97 dB below the easiest on InterHand2.6M and HanCo, respectively. In this tercile, the median viewing direction has turned by 84° relative to the hand on both datasets, so the hand is seen nearly edge-on. The extreme tercile is viewed mostly from the side that the dorsal source images do not observe, with a median viewing change of 106° on InterHand2.6M and 121° on HanCo. This tercile is 0.24 dB above the easiest on InterHand2.6M and 0.52 dB below it on HanCo. Finger articulation has a larger effect, concentrated in the extreme tercile. Relative to the easiest tercile, quality decreases by 0.81 and 0.22 dB in the middle tercile on InterHand2.6M and HanCo, respectively, and by 2.69 and 1.93 dB in the extreme tercile.

Table S13 scores all methods on the same articulation terciles of both datasets. On InterHand2.6M, the per-frame renders of the baselines average to their numbers in Table 1 of the main paper within 0.01 dB. On HanCo, our avatar is reconstructed zero-shot, whereas each baseline is optimized on the source frame of the subject (Table 2 of the main paper). In LPIPS, the extreme articulation tercile is the hardest for every method on both datasets, not only for ours. On InterHand2.6M, all four methods also lose PSNR from the easiest to the extreme tercile, by 1.63–2.69 dB, and ours remains the highest in PSNR in every tercile. On HanCo, BIGS* and ours lose 2.09 and 1.93 dB, whereas OHTA and HandAvatar change by -0.59 and $+0.44$ dB. The PSNR gain of HandAvatar is not matched by LPIPS, which worsens from 0.194 to 0.216 over the same terciles. Its renderings have a uniform tone without shading (Sec. G.2) and change less than the footage over time (Sec. J.2). Such smooth renderings keep the squared pixel error low even when the appearance of the hand changes, so PSNR can rise while LPIPS worsens. In the extreme tercile, ours is below OHTA and HandAvatar in PSNR, at 18.96 dB against 19.53 and 20.32 dB. Our larger drop on HanCo is consistent with the limitation that our colors are fixed after the forward pass, so wrinkles and shading that appear under strong articulation are not synthesized (Sec. I). These data, however, do not isolate the cause, since on InterHand2.6M OHTA and HandAvatar, which model self-occlusion shading (Sec. G.2), lose about as much PSNR as ours (2.57 and 2.66 dB).

J.2 Temporal Stability

Table S14 reports temporal stability on the continuous HanCo sequence. Our avatar is evaluated zero-shot, as neither the subject nor the dataset is seen during training and no test-time optimization is performed. Every time step except that of the source frame shows a novel pose, and three of the four cameras show a novel view. Under these conditions, our renderings remain within 8.0% of the ground-truth change on every camera, with a mean absolute excess of 4.7%. To calibrate this value, we add a per-frame brightness jitter of $\pm 1\%$ to our own renderings. The excess pooled over all pairs then increases from $+4.8\%$ to $+12.4\%$, and further to $+22.0\%$ and $+43.1\%$ for $\pm 2\%$ and $\pm 4\%$. A 1% flicker thus adds 7.6 points, which exceeds our mean absolute excess of 4.7%. By contrast, the per-subject 3DGS fit BIGS* exceeds the ground-truth change by 30.7–50.5% across the four cameras, which is consistent with view-specific appearance noise from single-view optimization. OHTA, which optimizes its avatar on this subject for 3360 s (Table 2 of the main paper), is closest to zero, with a mean absolute excess of 3.6%, although its sign differs across cameras. Its stability is therefore comparable to that of our feed-forward avatar. HandAvatar changes less than the footage on every camera, by 8.7–23.3%, consistent with the uniform tone without shading noted in Sec. G.2.

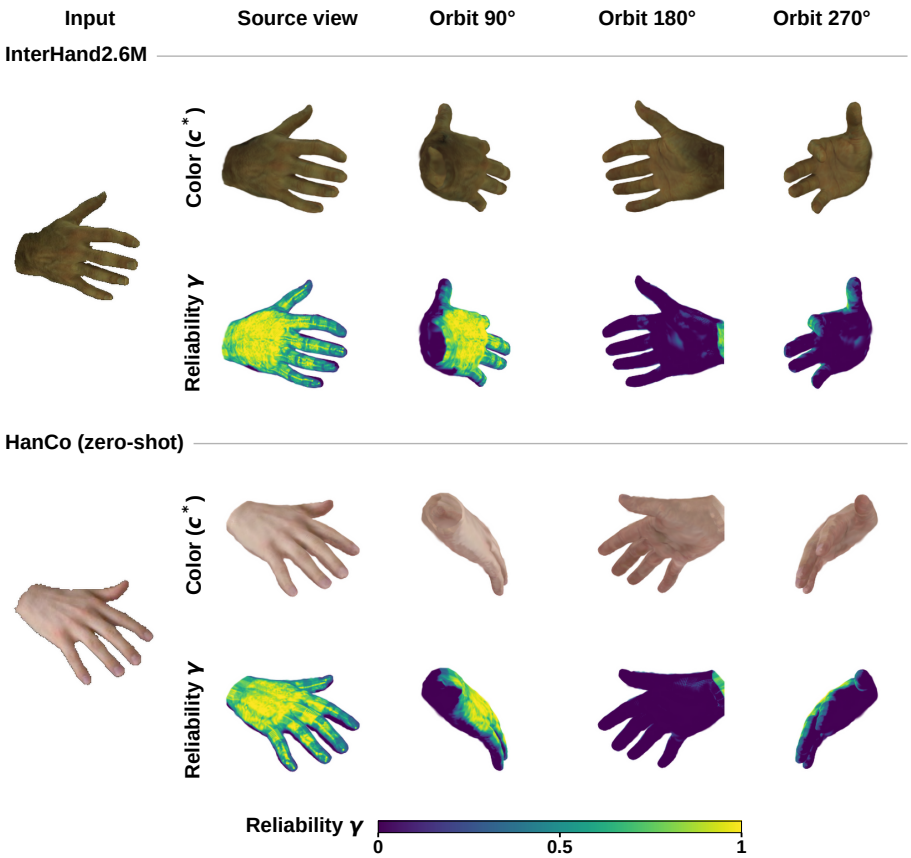


Figure S12: Canonical color and observation reliability. Final colors $\mathbf{c}^*$ (top) and reliability γ (bottom) of each canonical avatar, rendered from the source view and from an orbit of the source camera about the hand, with the same viewpoints and crops in both rows.

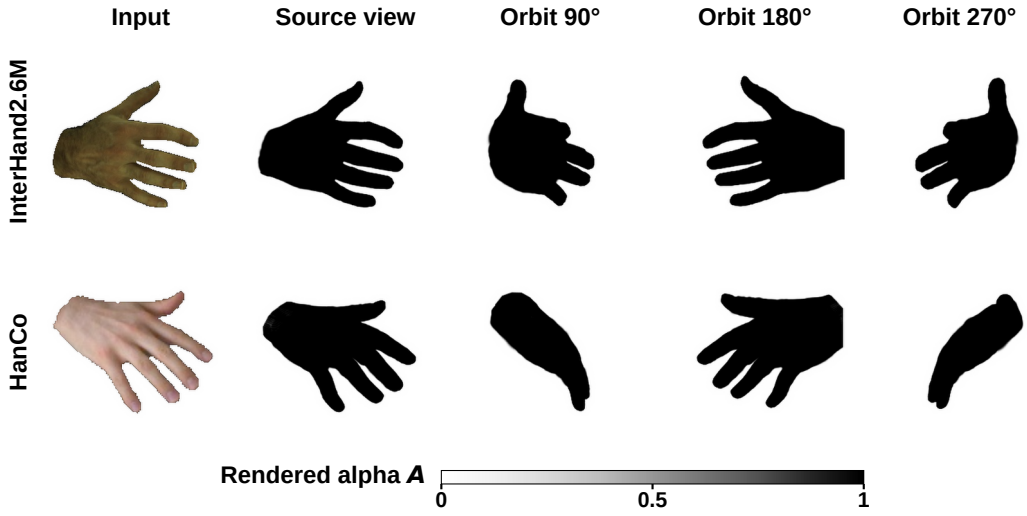


Figure S13: Rendered opacity (accumulated alpha A) of the canonical avatars from the views of Fig. S12. The straight cut in the $180°$ view is the wrist opening of the MANO template.

Because the canonical Gaussians are fixed after the single forward pass, the representation itself cannot drift, and the remaining variation arises from the pose-driven deformation and the rendering.

J.3 Reliability, Opacity, and Boundary Continuity on the Canonical Avatar

Figure S12 shows that γ is high on the dorsal surface visible in the input, decreases at grazing angles, and is exactly zero on the side that the input does not observe. Specifically, 55.1% of the Gaussians on InterHand2.6M and 61.6% on HanCo have $\gamma_i = 0$, and these Gaussians are colored by the Stage-1 prior and the source-level shift alone. Because $\mathbf{\Delta}_s^{aid}$ is shared by all Gaussians, the color update of these Gaussians is spatially uniform, with a standard deviation below 10^{-7} across Gaussians. The module therefore cannot create a color boundary inside the unobserved region, and a seam between transferred and prior-completed colors can arise only where observed and unobserved Gaussians meet.

At the reliability threshold $\gamma_i = 0.5$, the color step is small. The median ΔE between neighbors on either side of the threshold is 1.75 on InterHand2.6M and 1.91 on HanCo, which is $1.05\times$ and $1.32\times$ the variation between neighbors inside the high-reliability set (1.67 and 1.45) and below the commonly cited just-noticeable difference of about 2.3. The step is larger where a high-reliability Gaussian directly neighbors a low-reliability one without a transition band. This occurs for 79 pairs on InterHand2.6M and 209 pairs on HanCo, or 0.9% and 3.1% of the straddling pairs, with a median ΔE of 8.8 and 10.0. These pairs mark where the seams discussed in Sec. I can appear. Both steps are larger on HanCo, at $1.32\times$ against $1.05\times$ the reference and 10.0 against 8.8 for the strict pairs. This is consistent with a larger local correction on HanCo, where the mean $\|\mathbf{c}_i^* - \mathbf{c}_i^0 - \mathbf{\Delta}_s^{aid}\|$ over the Gaussians with $g_i\gamma_i > 0$ is 0.163, against 0.056 on InterHand2.6M. The unobserved side, in contrast, receives only $\mathbf{\Delta}_s^{aid}$, which on HanCo reaches the bound of 0.1 per channel set by $\boldsymbol{\delta}_i = 0.1\tanh(\cdot)$ (Sec. A.5) and amounts to a uniform brightening.

Figure S13 shows that the rendered opacity is equally high in every view. The mean of A over the opaque region ($A \geq 0.5$) is 0.99 in every view of both datasets. Semi-transparent pixels ($0.05 \leq A < 0.5$) amount to 3.5–5.2$\%$ of the opaque area, and this fraction is similar on the observed and unobserved sides, for example 4.5% in the source view and 3.9% at 180° on InterHand2.6M. The silhouettes show neither the translucent fringe of the recipe ports (Fig. S5) nor holes on the unobserved side.

K Supplementary Video

We provide a single supplementary video (`InstantHand_Supplementary_Video.mp4`, 3 min 9 s) in four parts. In every part, all of our renderings are projections of one fixed set of canonical Gaussians per subject, with no view-dependent parameters, no image-space refinement, and no per-frame fitting or update. The avatar representation therefore cannot drift over time: frame-to-frame changes in the renderings come only from the driving pose and the rendering itself, and Sec. J.2 quantifies them on HanCo. Temporal stability and multi-view consistency are difficult to convey with still frames, so we encourage readers to view the video directly.

Parts 1–2: benchmark subjects. Part 1 follows the InterHand2.6M protocol of Sec. C.1 and Part 2 the zero-shot HanCo protocol of Sec. C.2, where the source subject is unseen and no test-time optimization is performed. In both parts the avatar is reconstructed from the protocol-designated source frame with its ground-truth MANO, camera, and mask. Each part shows (i) the single source input image together with the reconstructed canonical avatar, (ii) full 360° yaw and pitch rotations of the canonical 3D avatar, and (iii) continuous novel-pose driving. In the rotations, the prior-completed side comes into view as the hand turns away from the source view, so the transition between the source-observed and prior-completed sides can be inspected directly. Each driving pose is rendered synchronously from six fixed cameras: the source camera, the reference camera in which the driving sequence is defined, and four fixed novel cameras at azimuth offsets of $+60^\circ$, $+120^\circ$, $+150^\circ$, and -60° from that reference. Both parts are driven by the same 35.2 s MANO pose sequence, assembled from a diverse set of hand gestures with short holds and interpolation between them. The sequence is disjoint from the evaluation frames of either dataset, covers a wider range of poses than either evaluation sequence, and shows the InterHand2.6M and HanCo avatars in identical poses. No ground-truth images exist for these poses, so the driving segments are qualitative. All quantitative novel-pose results (Tables 1, 2, and 3 of the main paper) are computed on the evaluation sequences.

Part 3: unseen subjects under fully estimated inputs. Part 3 shows the eight unseen subjects of Fig. S9 (Sec. H.6), reconstructed from fully estimated inputs (SAM 3 mask and WiLoR MANO and camera, with no ground-truth annotations). It first shows 360° yaw rotations, together with a small pitch sweep, of the canonical avatar for four subjects: a dorsal ARCTIC input, the palm-facing FreiHAND input whose fragmented SAM 3 mask is used as is (row 2 of Fig. S9), the dorsal in-the-wild OHTA input with a vitiligo pattern, and a FreiHAND input with a darker skin tone. In each, the side not visible in the source is completed from the source-conditioned prior, so the rotations expose the transition between the observed and completed sides. Part 3 then drives all eight subjects, two at a time, with a shared MANO pose sequence of diverse hand gestures, rendered from three cameras fixed in the MANO frame at azimuths of 0°, $+45^\circ$, and $+90^\circ$.

Part 4: comparison with ground truth and per-subject baselines. Part 4 replays the HanCo evaluation sequence of Table 2 of the main paper from the source camera, placing the ground-truth frames beside our zero-shot renderings and those of HandAvatar [20], OHTA [23], and BIGS*, which are fitted to the same source frame, in the column order of Fig. 4 of the main paper. Each column header lists the avatar-acquisition time of the method from Table 2 of the main paper. Every frame of the sequence is shown in order, and all methods are driven by the same frames, so the four renderings differ only in the method that produced them. All panels are shown at the native 256×256 evaluation resolution without resampling. HanCo is recorded at 5 Hz [55], so consecutive evaluation frames are 0.2 s apart and the replay, which shows each frame for 0.1 s, advances in visible steps in every column, including the ground truth. Part 4 closes with the same frames rendered by the same trained avatar with its Gaussians densified at inference (Sec. I, 197K instead of 49K Gaussians, scales $0.6\, s_i$), beside the ground truth and the $k = 3$ rendering used throughout the paper. Part 4 is the only part that can be checked against real footage, because the 360° rotations and gesture sequences of Parts 1–3 were never recorded for these subjects.

L Reproducibility Statement

The full model is specified in the main paper (Eqs. 1–17). All loss terms are defined in Sec. A.2, the Soft Alpha Band construction and schedule in Sec. A.3, the design rationale and thresholds of reliability gating (Eq. 8) in Sec. A.4, and the network architecture in Sec. A.5. Notation is consistent between the main paper and the appendix. All real training datasets and both evaluation datasets are publicly available (Sec. B), with preprocessing, annotation sources, subject and frame counts, and the complete sampling policy documented in Secs. A.1–B.3. The synthetic training data are generated following the published protocol of Zhao et al. [53]. All loss weights, learning rates, schedules, and batch sizes are reported with their final values in Sec. A.1. Evaluation sequences, designated source frames, crops, masks, metric computation, boundary-aware silhouette definitions, the scope of the self-reconstruction diagnostic, timing exclusions and the end-to-end runtime breakdown, per-baseline input protocols, and the provenance of all baseline numbers are specified in Secs. C–E. The protocols of the pose-stratified, temporal, and canonical analyses are given in Sec. C.8. The source-side MANO and mask sensitivity protocol and the inference-time reliability sensitivity settings are specified in Secs. F.3 and F.4, and the inference-time densification in Sec. I. All such variants keep the trained checkpoint fixed. Training uses four NVIDIA RTX A6000 GPUs (approximately 25 and 13 hours for the two stages), and all timing and efficiency measurements use a single NVIDIA A100 (80 GB) under the protocol of Sec. C.6. The reported 2.84 s is the avatar-network inference time once common MANO, camera, and mask inputs are available, and the corresponding end-to-end raw-image-to-avatar time, including SAM 3 and WiLoR estimation, is approximately 3.30 s (Table S3). Code, pretrained models, and the evaluation scripts are available at `https://github.com/coolho1129/InstantHand`, and videos and further results are on the project page, `https://coolho1129.github.io/InstantHand-page/`.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction claim a feed-forward single-image hand avatar with two components (Soft Alpha Band and Reliability-gated Appearance Transfer), explicitly scoped to off-the-shelf MANO, camera, and mask estimates. These claims are supported by the method in Section 3, the comparisons, ablations, source-cue robustness study, and controlled recipe-port comparison in Section 4, and the additional ablations and qualitative analyses in Appendices F and H.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations and failure cases are discussed in Appendix I. They include the inherent ambiguity under severe self-occlusion, the absence of pose-dependent wrinkles and shading because all Gaussian attributes are fixed after the forward pass, the open wrist boundary of the MANO template, the entanglement of illumination and albedo, rare cosmetic attributes that the appearance transfer does not carry, and failures inherited from corrupted upstream masks. Appendix G.2 discusses related limitations, namely the lack of view-dependent appearance, the transfer of shading baked into the source as albedo, and learned multi-view fusion as future work. The pose-stratified analysis (Appendix J.1) shows that the quality cost concentrates in the strongest-articulation tercile, and the canonical diagnostic (Appendix J.3) quantifies the color step where high- and low-reliability Gaussians meet directly.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not include theoretical results requiring formal assumptions or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The model is specified in Section 3 and the experimental setup in Section 4. Appendices A–C give the full training protocol, loss definitions and weights, Soft Alpha Band construction, the closed-form reliability estimator, dataset construction and sampling policy, and every evaluation protocol including designated source frames, crops, and masks. Appendix E specifies the baseline adaptations and recipe ports, and Appendix L summarizes reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All real training datasets and both evaluation datasets are publicly available (Appendix B). The synthetic appearance data are not released, but they are generated following the published protocol of Zhao et al. [53]. The code, pretrained models, and evaluation scripts are available at `https://github.com/coolho1129/InstantHand` with reproduction instructions. Implementation, preprocessing, and evaluation details are given in Appendices A–C and summarized in Appendix L.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Training schedules, optimizer settings, per-GPU batch sizes, all loss weights with their final values, the fixed reliability hyperparameters and their sensitivity, dataset sampling policies, evaluation sequences and designated source frames, cropping and masking,

and metric definitions are given in Section 4 and Appendices A–C. The configurations of every ablation and port are given in Appendices E–F.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: All metrics are averaged over fixed evaluation sets (194 InterHand2.6M and 304 HanCo target frames) from single training runs, following the protocol of prior hand avatar benchmarks. We do not report error bars over repeated training runs, because retraining each configuration through the full two-stage pipeline costs approximately 38 hours on 4 GPUs. The variation across evaluation frames is instead analyzed by pose tercile (Appendix J.1) and by camera (Appendix J.2).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Training uses four NVIDIA RTX A6000 GPUs (48 GB each) for approximately 25 and 13 hours in the two stages, with per-GPU batch sizes, gradient accumulation, and mixed-precision settings given in Appendix A.1. Ablations and recipe ports are retrained under the same schedule and budget (Appendices E.2 and F). Inference time, FPS, and memory are measured on a single NVIDIA A100 (80 GB) under the protocol of Appendix C.6 and reported in Table 2 of the main paper, and the end-to-end raw-image-to-avatar cost is given in Table S3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research follows the NeurIPS Code of Ethics. It uses publicly available research datasets under their licenses together with synthetic data generated following a published protocol, does not involve human subjects, and discusses ethical considerations in Appendix G.3.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Appendix G.3 discusses positive applications (virtual reality, telepresence, digital human animation, and human–computer interaction) and negative impacts (hand imagery as a soft biometric, identity impersonation, and deceptive synthetic media), together with the recommendation that deployments obtain informed consent and treat reconstructed avatars as biometric-adjacent data.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a scraped dataset or a general-purpose generative model. The released model reconstructs a hand avatar from a user-provided image and is trained on published research datasets and synthetic data generated following a published protocol. The associated misuse risks and mitigation recommendations are discussed in Appendix G.3.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [No]

    Justification: All datasets, pretrained models, and baseline methods are cited in Sections 2 and 4 and Appendices B and E, and are used under their respective research licenses, but the license name and terms of use of each asset are not listed individually.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper does not introduce a new dataset or benchmark. The code, pre-trained models, and evaluation scripts are released at `https://github.com/coolho1129/InstantHand` together with a README that documents installation, data preparation, training, and evaluation (Appendix L).

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human-subject studies requiring IRB approval.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as part of the core methodology or experimental pipeline.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.