

# HumanNOVA: Photorealistic, Universal and Rapid 3D Human Avatar Modeling from a Single Image

## Supplementary Material

### 1. Appendix

In this Supplementary Material, we provide additional details that were not included in the main manuscript due to space constraints. Specifically, we include more discussions (Section 2), more details on the experiment setup (Section 3), and more visual results (Section 4).

### 2. More Discussions

**Quality of our generated data.** Our generated dataset contains two components, *i.e.*, the synthetic subset and the real-world subset. For the synthetic subset, we render synthetic multi-view images in Blender from rigged SynBody [8] assets. Because SynBody meshes and textures are of high quality and Blender’s renderer preserves their visual fidelity, the resulting images maintain the same level of detail and realism as the source assets. For the real-world subset, this subset is generated by fitting multi-view capture data to 3DGS. We then re-render the fitted 3DGS back to the capture viewpoints and compare them with the corresponding ground-truth images, achieving an average of 36.23 / 0.9881 / 16.57 (PSNR / SSIM / LPIPS).

**Design of real-world data generation.** We argue that good initialization is critical in the real-world data generation stage. COLMAP initialization is the common practice in 3DGS, but it performs poorly on multi-view human capture data. To illustrate this, we performed a test on 10 randomly selected samples. Our improved initialization based on SMPL-X vertices enables significantly fewer training iterations, reducing optimization time from 40 minutes to 4 minutes. The performance on the initialization method is shown in Table 1.

Table 1. **Impact of initialization in real-world data generation.**  $\uparrow$  and  $\downarrow$  represent the higher the better, and the lower the better, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
COLMAP	16.49	0.9025	68.79
Ours	<b>36.38</b>	<b>0.9886</b>	<b>16.36</b>

**Comparison with animation-based methods.** Methods like SHERF [2] and LHM [7] fall under a related but distinct task setting: they aim to build generalizable, animatable human avatars from a single image. Since these methods focus primarily on pose-driven animation, they typically rely on ground-truth, high-quality SMPL/SMPL-X poses during both training and testing to ensure precise alignment.

Table 2. **Comparison with animation-based methods.**  $\uparrow$  and  $\downarrow$  represent the higher the better, and the lower the better, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
SHERF [2]	16.83	0.9037	87.99
LHM [7]	17.75	0.9083	76.85
Ours	<b>22.29</b>	<b>0.9360</b>	<b>42.42</b>



Figure 1. Visual results on the effectiveness of the SMPL prior.



Figure 2. Visual results on the robustness of HumanNOVA under inaccurate SMPL estimates.

In contrast, our task operates under a more challenging input condition: only a single image is available. That is to say, SMPL(-X) parameters must be estimated using off-the-shelf tools. For SHERF and LHM, the SMPL(-X) pose is derived from 4D Humans (same as our method) and its provided demo, respectively. We follow SHERF’s instructions from their Github repo and train it on the corresponding dataset. For LHM, since it does not release training and testing code, we directly utilize its pre-trained weights and its demo code to perform inference. As shown in Table 2, our method achieves superior performance with notable gains across multiple datasets and input viewpoints. For LHM and SHERF, they need to animate the avatar from the canonical space to the target pose space, during which any pose misalignment will directly exist in the rendering results, leading to performance degradation.

**Generalization.** For evaluation, we use a single model and report its performance across all benchmarks. To further assess the generalization capability of HumanNOVA, we conduct a leave-one-out evaluation on the CustomHuman dataset. As shown in Table 3, the performance under the

Table 3. **Generalization capability of HumanNOVA.** We conduct a leave-one-out evaluation with the CustomHuman dataset.  $\uparrow$  and  $\downarrow$  represent the higher the better, and the lower the better, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Leave-CustomHuman-Out	21.99	0.9344	44.22
Ours	<b>22.29</b>	<b>0.9360</b>	<b>42.42</b>

Table 4. **More ablations on HumanNOVA framework on the CustomHuman dataset.**  $\uparrow$  and  $\downarrow$  represent the higher the better, and the lower the better, respectively.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Ours	<b>22.29</b>	<b>0.9360</b>	<b>42.42</b>
Visual encoder (DINOv2 $\rightarrow$ Sapiens)	21.98	0.9327	46.52
Fusion layer (4 $\rightarrow$ 2)	21.42	0.9301	50.65
Supervision views (4 $\rightarrow$ 2)	22.07	0.9344	45.18

leave-one-out setting remains comparable to that of the full model, indicating that HumanNOVA generalizes well to unseen settings.

**More analysis on SMPL mesh prior.** We provide more visualization about the effect of the mesh prior in Figure 1. It can be observed that the base LRM architecture (w/o mesh prior) fails to predict feasible human structures, while HumanNOVA performs robustly when the input is an in-the-wild image. In Figure 2, we demonstrate two representative in-the-wild examples (backgrounds are removed) where the estimated SMPL is visibly inaccurate. It can be observed that HumanNOVA remains resilient to this noise and adaptively prioritizes the appearance cues. However, our approach can still fail in the extreme case where the SMPL estimation completely breaks.

**More ablations on HumanNOVA.** We conduct additional ablation studies on the CustomHuman dataset to analyze the impact of different design choices in HumanNOVA. As shown in Table 4, utilizing Sapiens [4] as the visual encoder still achieves comparable performance with DINOv2. Reducing the number of fusion layers from 4 to 2 causes the most significant performance drop, especially in LPIPS, suggesting that sufficient cross-modal fusion depth is critical for human avatar modeling. In addition, decreasing the number of supervision views from 4 to 2 also results in inferior performance, which demonstrates the importance of richer multi-view supervision for improving reconstruction quality.

**Broader Impact.** We are dedicated to fast and photorealistic 3D human avatar modeling from a single image. However, these technological advancements also raise concerns regarding potential misuse, including deceptive practices, harassment, and privacy violations. The lowered barrier to generating realistic human reconstructions could facilitate the creation of harmful or unauthorized content, thereby intensifying privacy risks. This highlights the need for careful

consideration of both technical and regulatory measures.

### 3. More Experiment Setup Details

**Our data generation.** For synthetic data generation, we utilize all 1,000 unique characters officially released by SynBody [8] for public download. These assets exhibit high diversity in terms of ethnicity, body shape, and clothing. Specifically, the characters span a wide range of skin tones, and wear diverse outfits based on approximately 68 clothing templates, including dresses, T-shirts, coats, pants, and more. For real-world data generation, the Gaussian optimization only lasts 4,000 iterations with the densification performed between iterations 400 and 1,500. For other parameters, we follow the original settings [3]. Note that both data generation strategies are scalable. The data size of the generated real-world data is 22k assets, while the generated synthetic data contains 78k assets. We generate an average of 26 views per asset, with camera positions randomly distributed on a sphere. The azimuth angles range from  $0^\circ$  to  $360^\circ$ , and the elevation angles range from  $-45^\circ$  to  $60^\circ$ . These views are randomly sampled without enforcing any preference for frontal views during the training of HumanNOVA.

**HumanNOVA training.** The input image resolution of HumanNOVA is  $512 \times 512$  and we crop a  $180 \times 180$  patch as supervision. The NeRF [6] MLP contains 10 layers with the width set as 60 and SiLU [1] is utilized as the activation function. The number of samples per ray is set as 128. For datasets with 3D scans such as THuman, CustomHuman, and 2K2K, we follow a unified preprocessing pipeline. We place each 3D mesh under a canonical camera setup and render 36 multi-view images at 10-degree intervals along a horizontal  $360^\circ$  circle. These rendered RGB images are then used as the supervision signals during training. Importantly, we do not use the original 3D meshes themselves for supervision, but only the rendered images.

**Patch selection.** We utilize patches from the image to provide the supervision signals. Its selection process is formulated as follows. Given an image  $I \in \mathbb{R}^{H \times W \times C}$  and a corresponding binary mask  $M \in \{0, 1\}^{H \times W}$ , we aim to sample a square patch of size  $r \times r$  (e.g.,  $r = 180$ ) such that the region contains a sufficient proportion of foreground pixels. Let  $R_{(i,j)} \subseteq M$  denote a square region of size  $r \times r$  with its top-left corner at position  $(i, j)$ . The foreground ratio of region  $R_{(i,j)}$  is computed as:

$$\alpha_{(i,j)} = \frac{1}{r^2} \sum_{u=0}^{r-1} \sum_{v=0}^{r-1} M_{i+u, j+v}.$$

We define the candidate set of regions as:  $\mathcal{S} = \{(i, j) \mid \alpha_{(i,j)} \geq \tau, i, j \in [0, H - r] \text{ stride } s\}$ . Here,  $\tau$  is the mask threshold (e.g.,  $\tau = 0.05$ ), and  $s$  is the stride used to slide the sampling window (e.g.,  $s = 10$ ). For each valid region  $(i, j) \in \mathcal{S}$ , we assign a sampling weight proportional to the



Figure 3. **Visualization of our training data.** (Best viewed in color.) The first three rows correspond to real-world generated data, while the remaining rows are generated synthetic data.

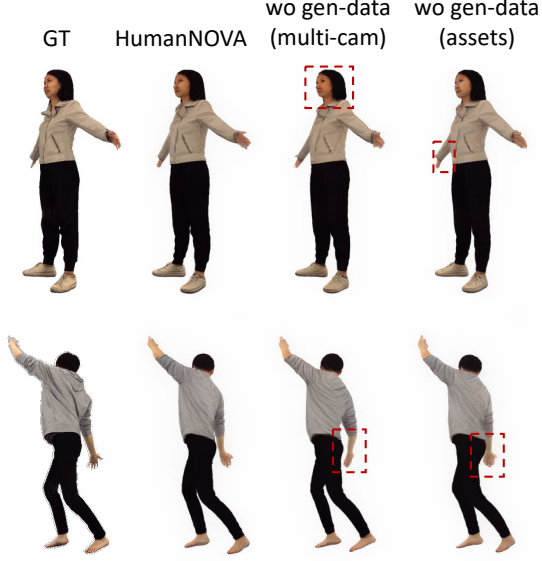


Figure 4. **Visualization of ablation studies on generated data type.** (Best viewed in color.) Removing the generated data (multi-cam or assets) negatively affects the model performance.

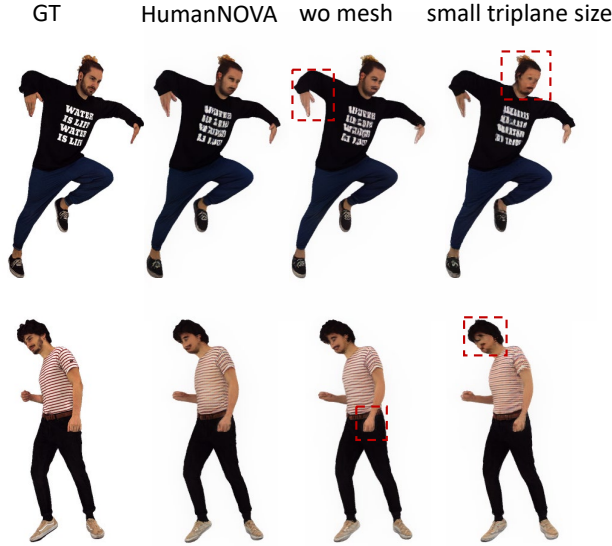


Figure 5. **Visualization of ablation studies on model design.** (Best viewed in color.) Excluding the mesh prior or reducing the triplane resolution reduces the quality of the output.

number of foreground pixels:

$$w_{(i,j)} = \sum_{u=0}^{r-1} \sum_{v=0}^{r-1} M_{i+u,j+v}.$$

The weights are normalized into a valid probability distribu-

tion:

$$p_{(i,j)} = \frac{w_{(i,j)}}{\sum_{(i',j') \in \mathcal{S}} w_{(i',j')}}.$$

Finally, a region  $(i^*, j^*) \sim p$  is sampled, and the corresponding image patch is extracted as  $P = I[i^* : i^* + r, j^* : j^* + r, :]$ . This strategy encourages sampling of patches rich in foreground content while preserving diversity.

**Evaluation.** We utilize 50 assets from each dataset (CustomHuman, Thuman2, and 2K2K) for evaluation. Both CustomHuman and THuman2 feature diverse clothing styles, including loose garments and layered outfits, as well as a wide range of body poses, making them particularly challenging for accurate 3D reconstruction. In contrast, the 2K2K dataset primarily consists of humans in upright, standing poses. For additional 3D geometry evaluation, we extract the isosurface based on Marching Cubes [5] to convert HumanNOVA’s implicit representations into meshes. During evaluation, CD is measured in centimeters (cm), providing a precise indication of surface accuracy. F-Score is computed with a threshold of 0.01 meters.

## 4. More Visual Results

In Figure 3, we visualize a sample from our training data. The first three rows correspond to real-world generated data, while the remaining rows are generated synthetic data. Together, they provide the foundational training data that empowers HumanNOVA to learn robust and generalizable 3D human representations.

We provide visualization of our ablation studies in Figure 4 and Figure 5. Figure 4 illustrates the impact of different training data configurations. Removing gen-data (multi-cam) leads to less photorealistic results or less accurate structure. In contrast, removing gen-data (assets) weakens the model’s capability to perceive and reconstruct human poses (see the right shoulder). Figure 5 shows the effects of varying model settings. Excluding the mesh prior degrades the structural quality of the output, while reducing the triplane resolution compromises the model’s capacity to represent fine-grained details (see the red box).

To better showcase the results, we include additional in-the-wild reconstruction examples and comparisons with previous methods in the supplementary video. In the video, the reconstruction results are presented with 360-degree rotation. From the video, it could be observed that our method is generally not affected by the Janus problem. This is because we model the 3D human directly as a whole in 3D space, instead of decomposing the task into separate front/back generation followed by heuristic merging, which is a common cause of inconsistent geometry or appearance across views (e.g., SiTH).



## References

- [1] Stefan Elfving, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018. [2](#)
- [2] Shoukang Hu, Fangzhou Hong, Liang Pan, Haiyi Mei, Lei Yang, and Ziwei Liu. Sherf: Generalizable human nerf from a single image. In *ICCV*, pages 9352–9364, 2023. [1](#)
- [3] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4):1–14, 2023. [2](#)
- [4] Rawal Khrodar, Timur Bagautdinov, Julieta Martinez, Su Zhaoen, Austin James, Peter Selednik, Stuart Anderson, and Shunsuke Saito. Sapiens: Foundation for human vision models. In *ECCV*, pages 206–228, 2024. [2](#)
- [5] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353, 1998. [4](#)
- [6] 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. *ECCV*, pages 1–25, 2020. [2](#)
- [7] Lingteng Qiu, Xiaodong Gu, Peihao Li, Qi Zuo, Weichao Shen, Junfei Zhang, Kejie Qiu, Weihao Yuan, Guanying Chen, Zilong Dong, et al. Lhm: Large animatable human reconstruction model from a single image in seconds. In *ICCV*, pages 1–15, 2025. [1](#)
- [8] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, et al. Synbody: Synthetic dataset with layered human models for 3d human perception and modeling. In *ICCV*, pages 20282–20292, 2023. [1](#), [2](#)