Supplementary
HandDreamer: Zero-Shot Text to 3D Hand Model Generation using Corrective Hand Shape Guidance

Green Rosh Prateek Kukreja* Vishakha SR * Pawan Prasad B H
Samsung R&D Institute India Bangalore

1. Supplementary

We provide proofs and derivations for the results presented in the main paper in section A.1 We provide additional implementation details in section A.2. Additional studies are provided in section A.3, additional quantitative results in section A.4 and more results in section A.5. We also provide multi-view videos in the multimedia supplementary attachment.

2. A.1 Proofs and Derivations

In this section, we provide proofs and derivations for the theorems and definitions defined in the main paper.

2.1. Proof for Theorem 1

Theorem 1. *Let x^v_{latent} and x^v_{init} denote the set of views rendered from an ideal latent 3D model (m^{latent}_{3D}) and an initial 3D model (m^{init}_{3D}) respectively. Then the expected absolute score of m^{init}_{3D} w.r.t m^{latent}_{3D} is:*

$$|S_\phi| = \left| \mathbb{E}_v \left[\frac{-\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t} \left[(\mathcal{E}(x^v_{init}) - \mathcal{E}(x^v_{latent})) + \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}} \epsilon \right] \right] \right| \tag{1}$$

where $\mathcal{E}(.)$ is the encoder of Stable Diffusion and $\bar{\alpha}_t$ denotes forward noise parameters of the diffusion model. Let $z^{latent}_t = \sqrt{\bar{\alpha}_t}\mathcal{E}(x^v_{latent})$ denote the mode towards which a view (v) should converge into. Since $p_\phi(z_t|y,t)$ is Gaussian, z^{latent}_t is also the mode of a locally Gaussian distribution within p_ϕ. Next we present the following lemma.

Lemma 1. *The score function of a multivariate Gaussian distribution $\mathcal{N}(\mathbf{x};\mu,\mathbf{\Sigma})$ is given as:*

$$s(\mathbf{x}) = -\Sigma^{-1}(\mathbf{x}-\mu) \tag{2}$$

where, μ and $\mathbf{\Sigma}$ denote the mean and covariance metrics respectively.

Proof:

Score of a probability distribution is defined as the derivative of the log likelihood as follows:

$$s(x) = \nabla_x \log p(x) \tag{3}$$

Since $p(x)$ is a multivariate Gaussian with dimensionality d,

$$\begin{aligned} s(\mathbf{x}) &= \nabla_\mathbf{x} \log(\mathcal{N}(\mathbf{x};\mu,\mathbf{\Sigma})) \\ &= \nabla_\mathbf{x} \log \left[\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) \right) \right] \\ &= -\nabla_\mathbf{x} \left(\frac{d}{2}\log(2\pi) + \frac{1}{2}\log|\Sigma| \right) \\ &\quad - \nabla_\mathbf{x} \left(-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu) \right) \end{aligned} \tag{4}$$

The first term of the above equation vanishes to 0 since it is independent of $\mathbf{x}$. The argument of the gradient of the second term is of quadratic form and hence it reduces to the following:

$$\begin{aligned} s(\mathbf{x}) &= -\frac{1}{2}.2.\Sigma^{-1}(\mathbf{x}-\mu) \\ &= -\Sigma^{-1}(\mathbf{x}-\mu) \end{aligned} \tag{5}$$

Next, we proceed to derive theorem 1 defined in Eq. 9.

Proof:

Since z^{latent}_t is the mode of locally isotropic Gaussian with $\Sigma = (\sqrt{1-\bar{\alpha}_t})\mathbf{I}$ at noise timestep t, the expected absolute score at any point z_t w.r.t to the local probability distribution is given by lemma 1.

$$|\mathbb{E}_v[s(z_t)]| = \left| \mathbb{E}_v \left[-\Sigma^{-1}(z_t - \mu) \right] \right| \tag{6}$$

Using $\Sigma = (\sqrt{1-\bar{\alpha}})\mathbf{I}$ and $\mu = z^{latent}_t$, since the mean and mode are same for Gaussian,

$$|\mathbb{E}_v[s(z_t)]| = \left| \mathbb{E}_v \left[-\frac{1}{\sqrt{1-\bar{\alpha}_t}}(z_t - z^{latent}_t) \right] \right| \tag{7}$$

Using $z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$ and $z^{latent}_t = \sqrt{\bar{\alpha}_t}z^{latent}_0$,

*Equal Contribution

$$
\begin{aligned}
|\mathbb{E}_v[s(z_t)]| &= \left|\mathbb{E}_v\left[-\frac{1}{\sqrt{1-\bar{\alpha}_t}}(z_t - z_t^{latent})\right]\right| \\
&= \left|\mathbb{E}_v\left[-\frac{1}{\sqrt{1-\bar{\alpha}_t}}(\sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon - \sqrt{\bar{\alpha}_t}z_0^{latent})\right]\right| \\
&= \left|\mathbb{E}_v\left[-\frac{1}{\sqrt{1-\bar{\alpha}_t}}(\sqrt{\bar{\alpha}_t}(z_0 - z_0^{latent}) + \sqrt{1-\bar{\alpha}_t}\epsilon)\right]\right| \\
&= \left|\mathbb{E}_v\left[\frac{-\sqrt{\bar{\alpha}_t}}{\sqrt{1-\bar{\alpha}_t}}(z_0 - z_0^{latent}) + \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\epsilon)\right]\right|
\end{aligned} \tag{8}
$$

Using a slight abuse of notations to denote $\mathbb{E}_v[s(z_t)]$ as S_ϕ, the expected score of initial 3D model (m_{3D}^{init}), and reparameterizing $z_0 = \mathcal{E}(x_{init}^v)$ and $z_0^{latent} = \mathcal{E}(x_{latent}^v)$ for view v, we arrive at theorem 1 as follows:

$$
|S_\phi| = \left|\mathbb{E}_v\left[\frac{-\sqrt{\bar{\alpha}_t}}{1-\bar{\alpha}_t}\left[(\mathcal{E}(x_{init}^v) - \mathcal{E}(x_{latent}^v)) + \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\bar{\alpha}_t}}\epsilon\right]\right]\right| \tag{9}
$$

As argued in section 4 of the main paper, this results shows that an ideal initial conditions minimizes the semantic difference $|\mathcal{E}(x_{init}^v) - \mathcal{E}(x_{latent}^v)|$, which motivates us to use MANO based hand model for a low-score initialization.

2.2. Derivation of CHS loss weighing term

In this section, we provide derivation for the CHS loss weighing term defined in Eq. 6 of the main paper.

$$
\lambda_t^{chs} = \lambda_{max}^{chs}\left[\frac{t - t_{min}}{t_{max} - t_{min}}\right] + \lambda_{min}^{chs}\left[\frac{t_{max} - t}{t_{max} - t_{min}}\right] \tag{10}
$$

We first define λ_t^{chs} annealing as a function of optimization iterations as follows:

$$
\lambda_t^{chs} = \lambda_{max}^{chs} - (\lambda_{max}^{chs} - \lambda_{min}^{chs})\sqrt{\frac{i}{i_{max}}} \tag{11}
$$

where, i and i_{max} denote the current and maximum optimization iterations respectively. This is similar in form to that of square-root time annealing proposed by [16] as follows:

$$
\begin{aligned}
t &= t_{max} - (t_{max} - t_{min})\sqrt{\frac{i}{i_{max}}} \\
\Rightarrow \sqrt{\frac{i}{i_{max}}} &= \frac{t_{max} - t}{t_{max} - t_{min}}
\end{aligned} \tag{12}
$$

Substituting Eq. 12 into Eq. 11 we get,

$$
\begin{aligned}
\lambda_t^{chs} &= \lambda_{max}^{chs} - (\lambda_{max}^{chs} - \lambda_{min}^{chs})\frac{t_{max} - t}{t_{max} - t_{min}} \\
&= \lambda_{max}^{chs}\left[1 - \frac{t_{max} - t}{t_{max} - t_{min}}\right] + \lambda_{min}^{chs}\left[\frac{t_{max} - t}{t_{max} - t_{min}}\right] \\
&= \lambda_{max}^{chs}\left[\frac{t - t_{min}}{t_{max} - t_{min}}\right] + \lambda_{min}^{chs}\left[\frac{t_{max} - t}{t_{max} - t_{min}}\right]
\end{aligned} \tag{13}
$$

As explained in the main paper, we use this equation to adjust the weight of CHS loss function so that more weights is given at higher noise-timesteps so that the geometry does not degrade too much. We empirically choose λ_{max}^{chs}, λ_{min}^{chs}, t_{max} and t_{min} as 15000, 1000, 600 and 300 respectively.

3. A.2 Additional Implementation Details

3.1. Hand Shape Initialization

Obtaining hand silhouette groundtruth: As explained in the main paper, we use hand silhouette mask obtained from MANO mesh to initialize the NeRFs in stage 1. To this end, we first obtain hand mesh in diverse articulations using the code provided by MANO [9]. The hand mesh is then placed in a virtual 3D environment using open3D library [14]. We place virtual cameras, corresponding to the sampled viewpoints in the same environment. Next, we obtain the depth map of the hand mesh, as observed by the virtual cameras using open3D APIs. Finally, we convert the depth map into a binary map to use as the hand silhouette ground truth from the required viewpoint.

3.2. Skeleton based SDS

We use ControlNet v1.1 model [12], conditioned on OpenPose skeleton [2] for SDS based optimization. We obtain hand skeleton from a given view point using the codes provided by MANO [9]. However, this skeleton excludes the fingertips. Hence, we add fingertip keypoints using vertex information from MANO hand mesh, resulting in a 3D hand skeleton of dimensionality (21×3). Next, we transform this skeleton into OpenPose format and project it onto the 2D view space of the virtual camera, to generate the control input.

4. A.3 Additional Studies

In section 5.2 of the main paper, we justified CHS annealing by the observation that SDS tends to perform more geometric updates at higher noise t (lower iteration) and more texture updates at lower t (higher iteration). In Fig. 1 we provide two examples on this observation. It can be seen that the geometry of the 3D model is optimized more in the earlier iterations and the texture in the later iterations. This empirically justifies the proposed CHS loss annealing, wherein we provide higher weightage to MANO prior in the

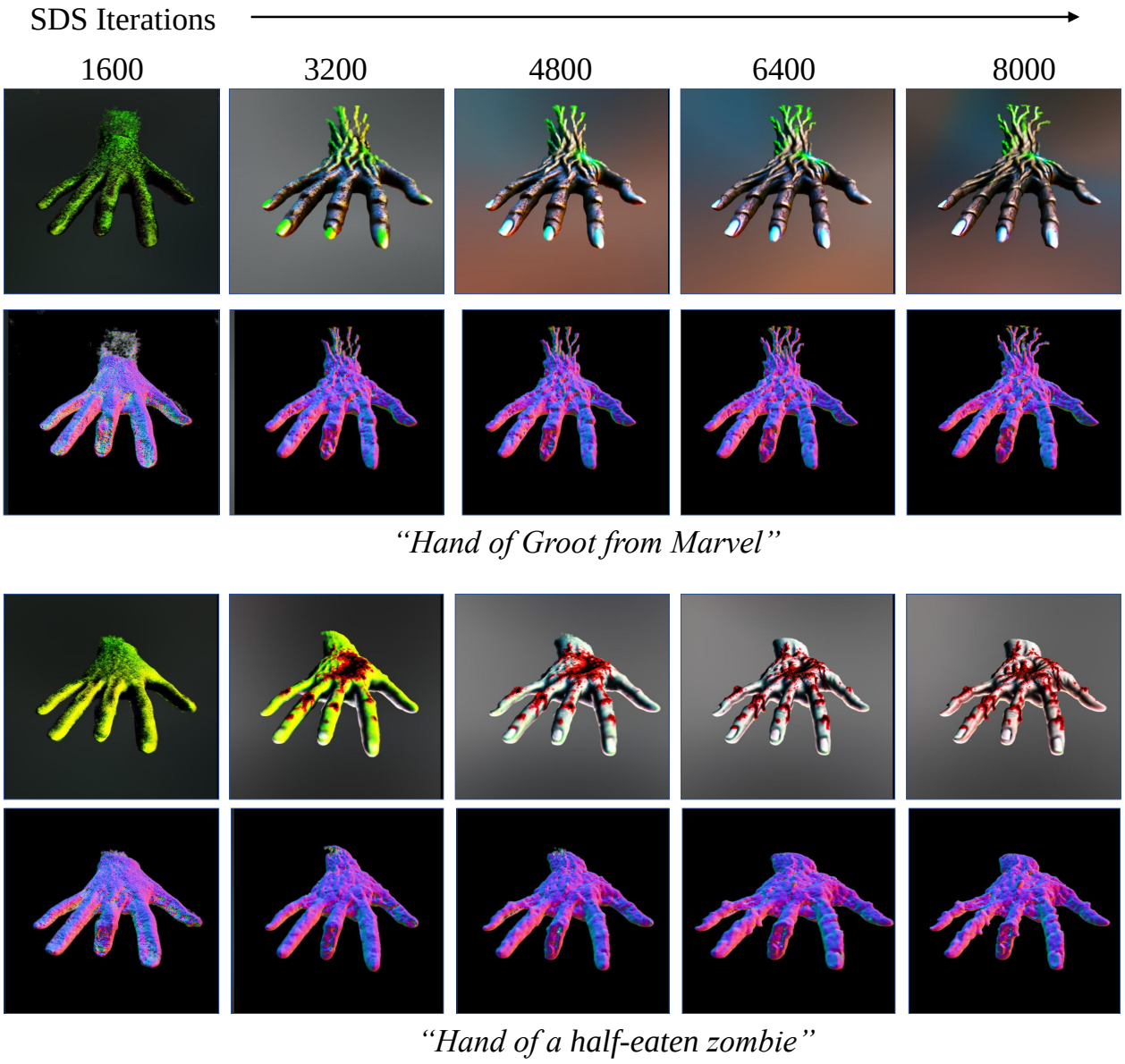


Figure 1. During the earlier iterations, the SDS optimizes geometry more and in the later iterations, texture is optimized more. It can be seen that the surface maps do not change much after 3200 iterations of SDS. This observation encourages us to anneal λ_t^{chs} in Corrective Hand Shape guidance loss.

earlier iterations to ensure that the geometry optimization is stabilized.

5. A.4 Additional Quantitative Analysis

In this section, we provide analysis of both mean and standard deviation of the proposed method compared to state-of-the-art in Table. 1. While methods such as OHTA [13] and Fantasia3D [3] generates results with lower standard deviation for CLIP L14, their mean scores are much lower compared to our method. Further, state of the art method CFD [11] achieves the second best mean scores on all the metrics, but they have a high standard deviation in their results. It can be seen our method (HandDreamer) achieves the best mean value for all the 3 metrics with a low standard deviation. This shows that our method generates the best results consistently over several prompts.

6. A.5 Additional Results

We provide additional results from our method for multiple viewpoints for several prompts in Figures 2 to 7. It can be

seen that our method generates high fidelity 3D models for a variety of text prompts. We have also provided videos in the multimedia supplementary.

We also provide additional comparative studies against state-of-the art methods ESD'24 [10], CFD'25 [11] and dreamDPO'25 [15] in Figs. 8 to 12. It can be seen that while ESD generates Janus artifacts with protruding fingers, CFD and dreamDPO generates results with erroneous number of fingers and lower details. On the other hand, our method is able to generate high-fidelity and geometrically accurate 3D hand model outputs.

References

[1] Yukang Cao, Yan-Pei Cao, Kai Han, Ying Shan, and Kwan-Yee K Wong. Dreamavatar: Text-and-shape guided 3d human avatar generation via diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 958–968, 2024. 5

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):172–186, 2019. 2

[3] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22246–22256, 2023. 3, 5

[4] Xin Huang, Ruizhi Shao, Qi Zhang, Hongwen Zhang, Ying Feng, Yebin Liu, and Qing Wang. Humannorm: Learning normal diffusion model for high-quality and realistic 3d human generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4568–4577, 2024. 5

[5] Yukun Huang, Jianan Wang, Ailing Zeng, He Cao, Xianbiao Qi, Yukai Shi, Zheng-Jun Zha, and Lei Zhang. Dreamwaltz: Make a scene with complex 3d animatable avatars. *Advances in Neural Information Processing Systems*, 36:4566–4584, 2023. 5

[6] Artem Lukoianov, Haitz Sáez de Ocáriz Borde, Kristjan Greenewald, Vitor Guizilini, Timur Bagautdinov, Vincent Sitzmann, and Justin M Solomon. Score distillation via reparametrized ddim. *Advances in Neural Information Processing Systems*, 37:26011–26044, 2024. 5

[7] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12663–12673, 2023. 5

[8] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 5

[9] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*, 2022. 2

[10] Peihao Wang, Dejia Xu, Zhiwen Fan, Dilin Wang, Sreyas Mohan, Forrest Iandola, Rakesh Ranjan, Yilei Li, Qiang Liu, Zhangyang Wang, et al. Taming mode collapse in score distillation for text-to-3d generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9037–9047, 2024. 4, 12, 13, 14, 15, 16

[11] Runjie Yan, Yinbo Chen, and Xiaolong Wang. Consistent flow distillation for text-to-3d generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 3, 4, 5, 12, 13, 14, 15, 16

[12] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2

[13] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 799–810, 2024. 3, 5

[14] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3d: A modern library for 3d data processing. *arXiv preprint arXiv:1801.09847*, 2018. 2

[15] Zhenglin Zhou, Xiaobo Xia, Fan Ma, Hehe Fan, Yi Yang, and Tat-Seng Chua. Dreamdpo: Aligning text-to-3d generation with human preferences via direct preference optimization. In *Forty-second International Conference on Machine Learning*, 2025. 4, 12, 13, 14, 15, 16

[16] Junzhe Zhu, Peiye Zhuang, and Sanmi Koyejo. Hifa: High-fidelity text-to-3d generation with advanced diffusion guidance. *arXiv preprint arXiv:2305.18766*, 2023. 2

Method	CLIP L14 ↑	FID ↓	HPSv2 ↑
DreamFusion'22 [8]	25.12 ±2.41	344.19 ±45.54	0.187 ±0.035
LatentNerf'23 [7]	24.34 ±3.61	316.42 ±34.02	0.189 ±0.026
Fantasia3D'23 [3]	20.93 ±1.21	329.31 ±62.28	0.198 ±0.013
DreamWaltz'23 [5]	23.96 ±3.08	265.11 ±37.32	0.222 ±0.021
DreamAvatar'24 [1]	20.02 ±2.12	329.85 ±50.02	0.215 ±0.025
HumanNorm'24 [4]	23.01 ±2.56	327.42 ±34.32	0.177 ±0.021
SDI'24 [6]	26.32 ±3.12	297.12 ±35.32	0.192 ±0.013
OHTA'24 [13]	22.59 ±0.93	467.51 ±21.01	0.181 ±0.017
CFD'25 [11]	<u>26.62 ±5.12</u>	<u>262.83 ±44.32</u>	<u>0.223 ±0.021</u>
HandDreamer (Ours)	**28.63 ±1.49**	**254.62 ±34.12**	**0.241 ±0.012**

Table 1. Quantitative comparisons. Our method outperforms the other methods on all the metrics while generating results in low standard deviation.

“A hand with leather gloves”

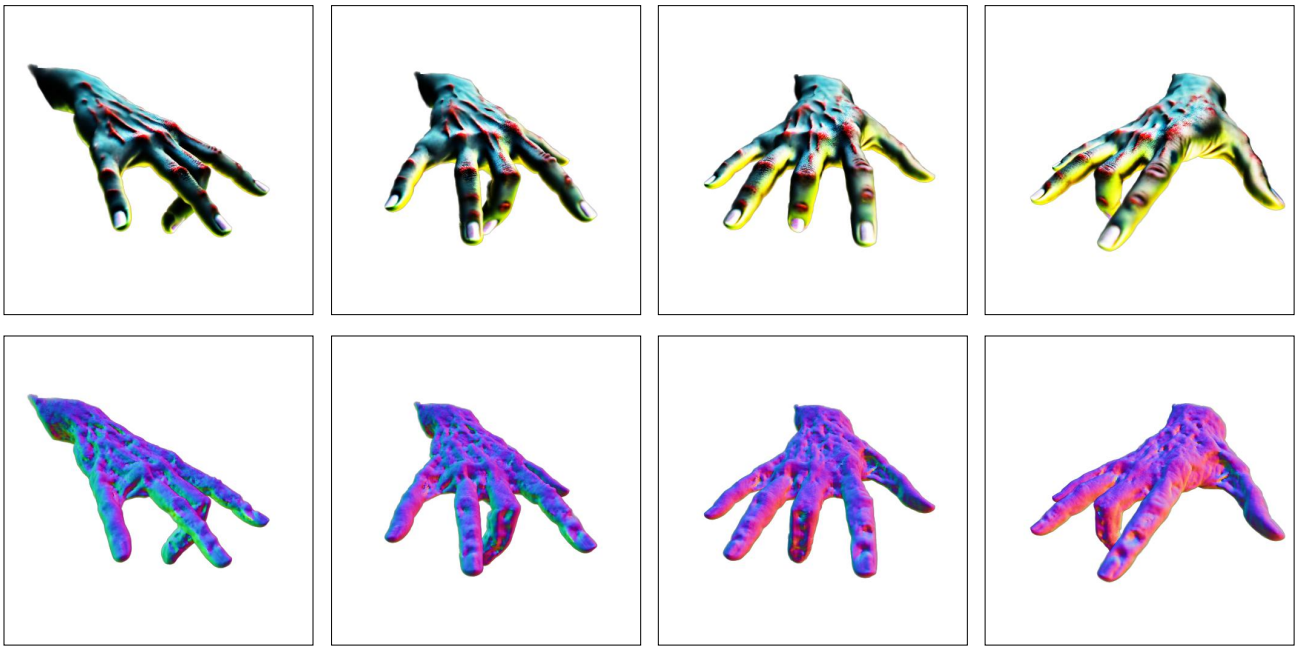
“A dark-skinned hand”

Figure 2. Results from the proposed HandDreamer method. Top row: Rendered images. Bottom Row: Surface maps

“Hand of a wooden toy”

“Hand of a medieval knight”

Figure 3. Results from the proposed HandDreamer method. Top row: Rendered images. Bottom Row: Surface maps



“An alien hand with scars”

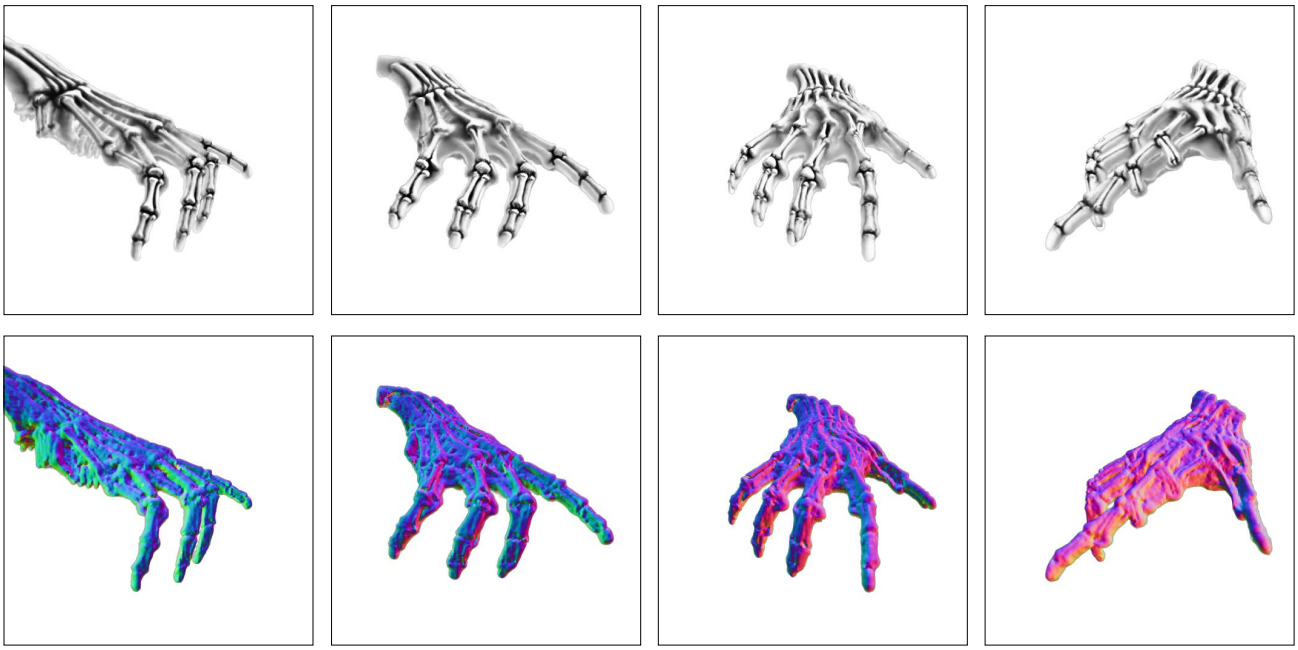
“Hand of a stormtrooper from Star Wars”

Figure 4. Results from the proposed HandDreamer method. Top row: Rendered images. Bottom Row: Surface maps

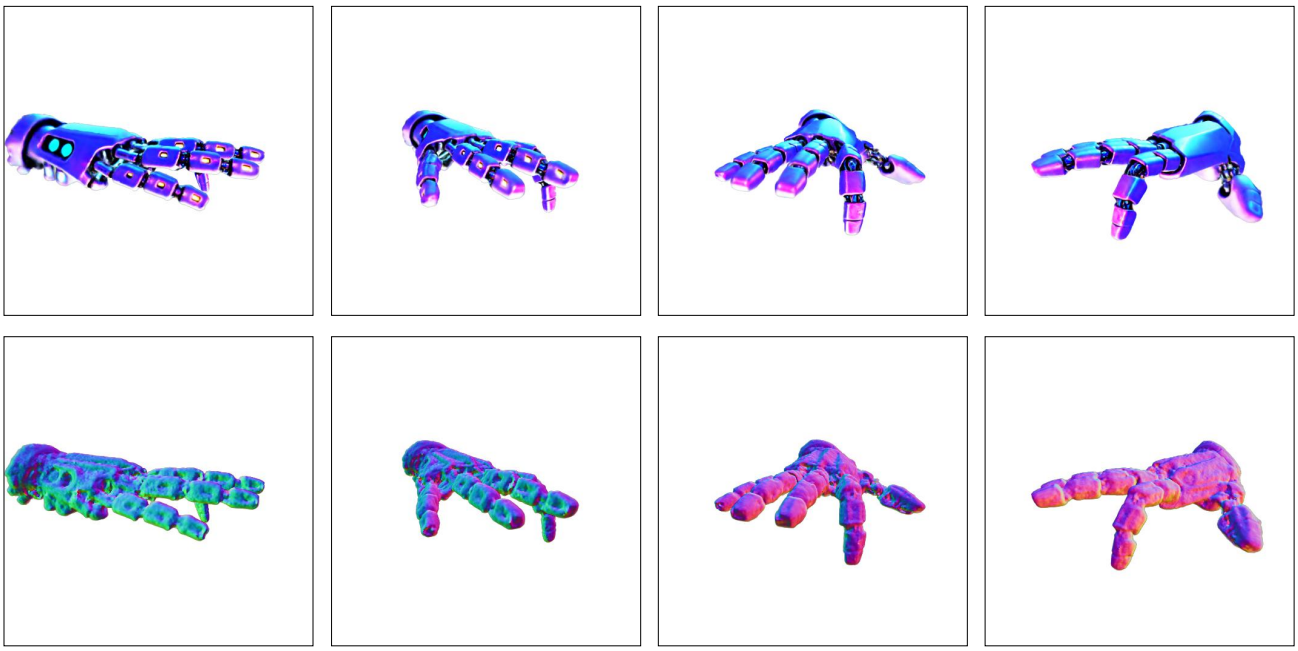
“Hand of Kratos from God of War”

“Hand with tattoos”

Figure 5. Results from the proposed HandDreamer method. Top row: Rendered images. Bottom Row: Surface maps

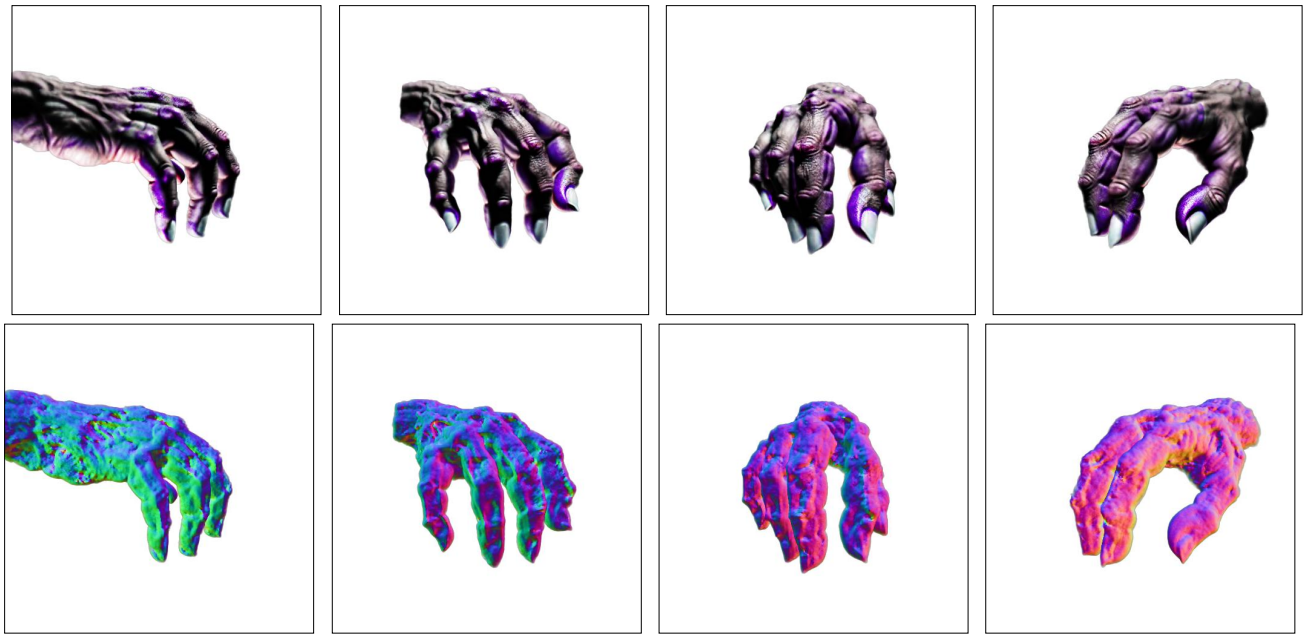


“A skeletal hand”



“Blue Robot hand”

Figure 6. Results from the proposed HandDreamer method. Top row: Rendered images. Bottom Row: Surface maps



"Hand of Thanos from Marvel"

"Hand of C3PO from Star Wars"

Figure 7. Results from the proposed HandDreamer method. Top row: Rendered images. Bottom Row: Surface maps

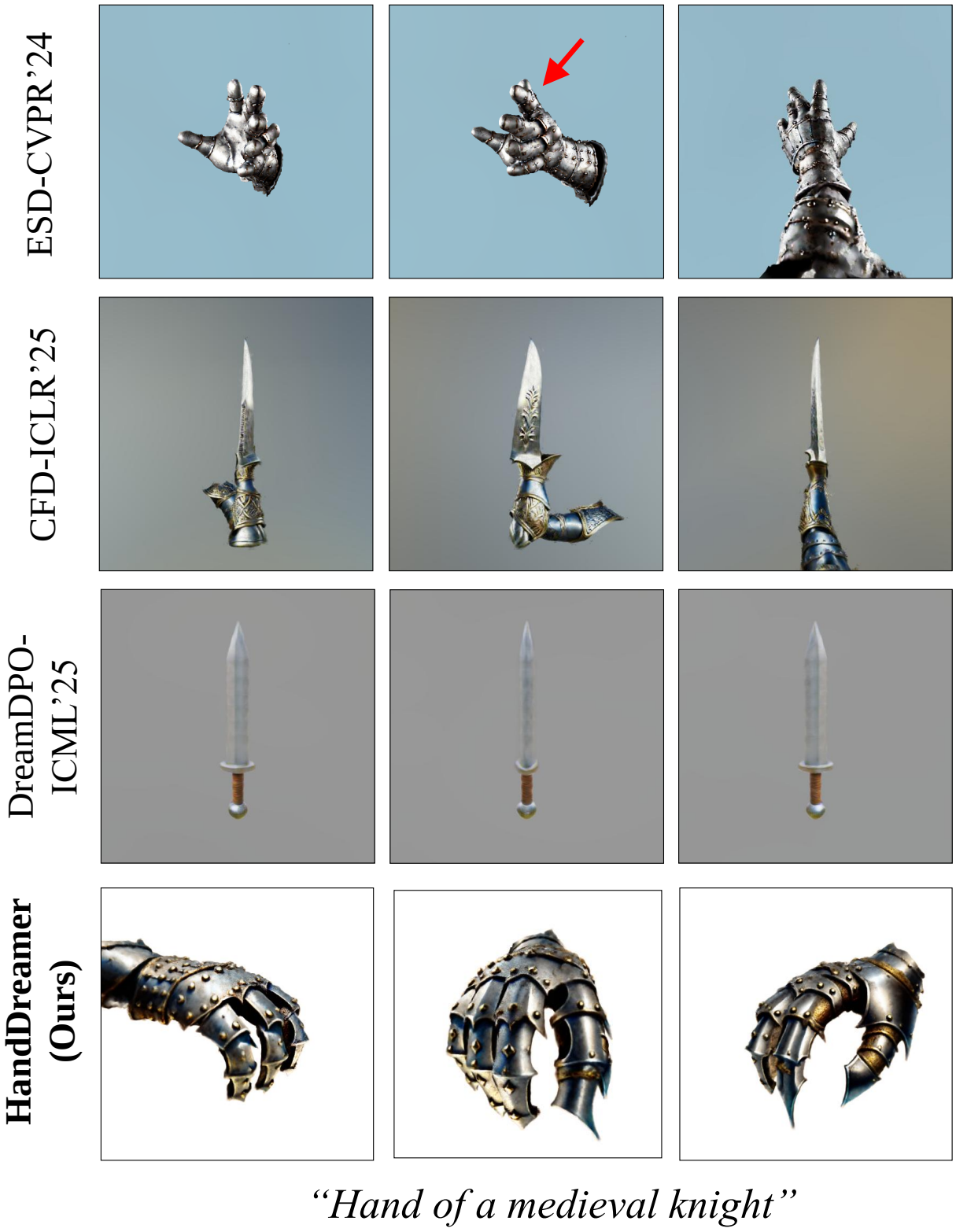


Figure 8. Comparisons against state-of-the-art methods: ESD'24 [10], CFD'25 [11], DreamDPO [15]. ESD generates Janus effects (red arrows) while CFD and dreamDPO fail to generate a hand

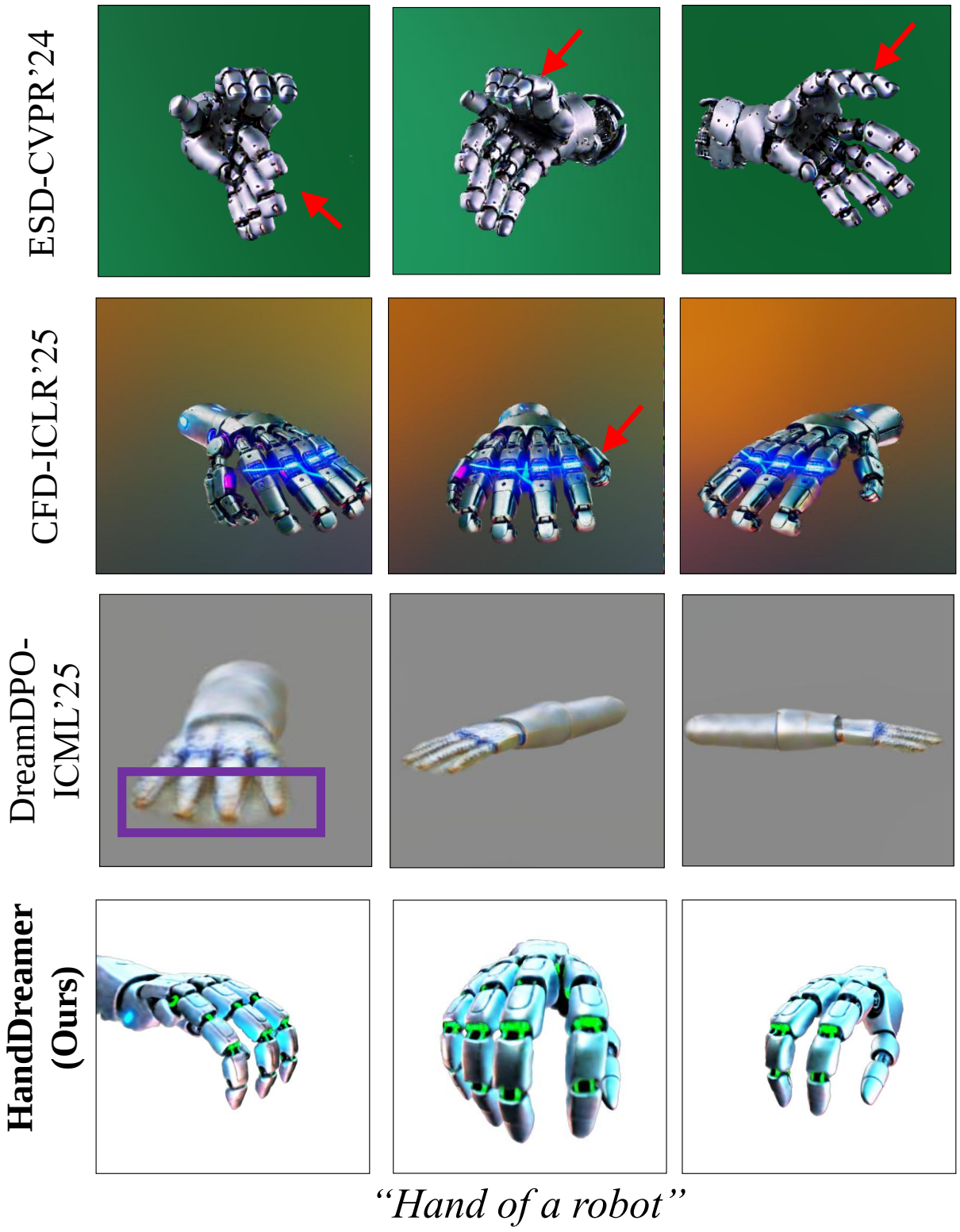


Figure 9. Comparisons against state-of-the-art methods: ESD'24 [10], CFD'25 [11], DreamDPO [15]. Janus effects (ESD) and extra finger artifacts (CFD) shown in red arrows. Missing fingers (dreamDPO) denoted in violet box.

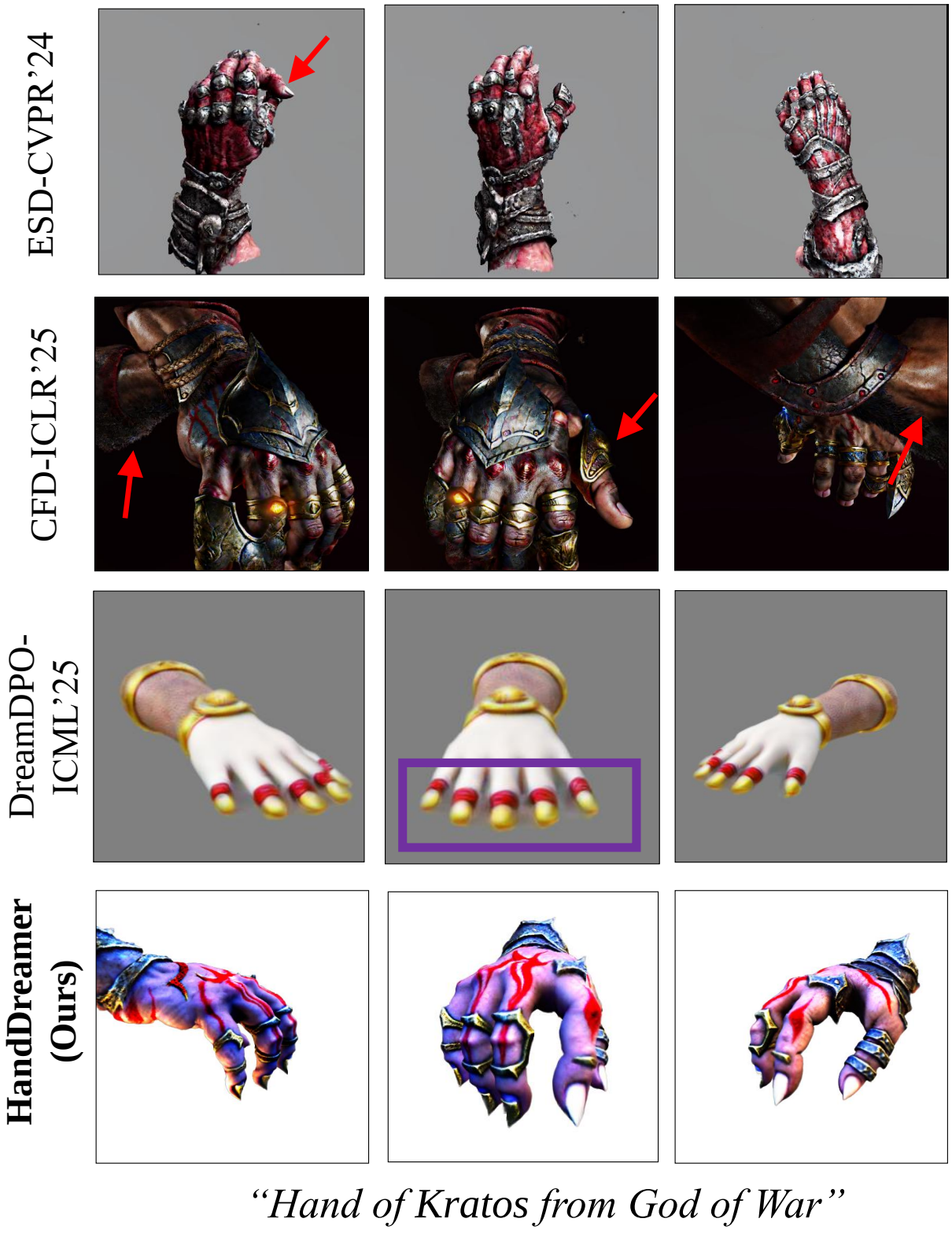


Figure 10. Comparisons against state-of-the-art methods: ESD'24 [10], CFD'25 [11], DreamDPO [15]. ESD generates Janus artifacts (red arrows) and CFD generates extra fingers and arms (red arrows). DreamDPO generates low-fidelity hands with same length for all fingers (violet box)

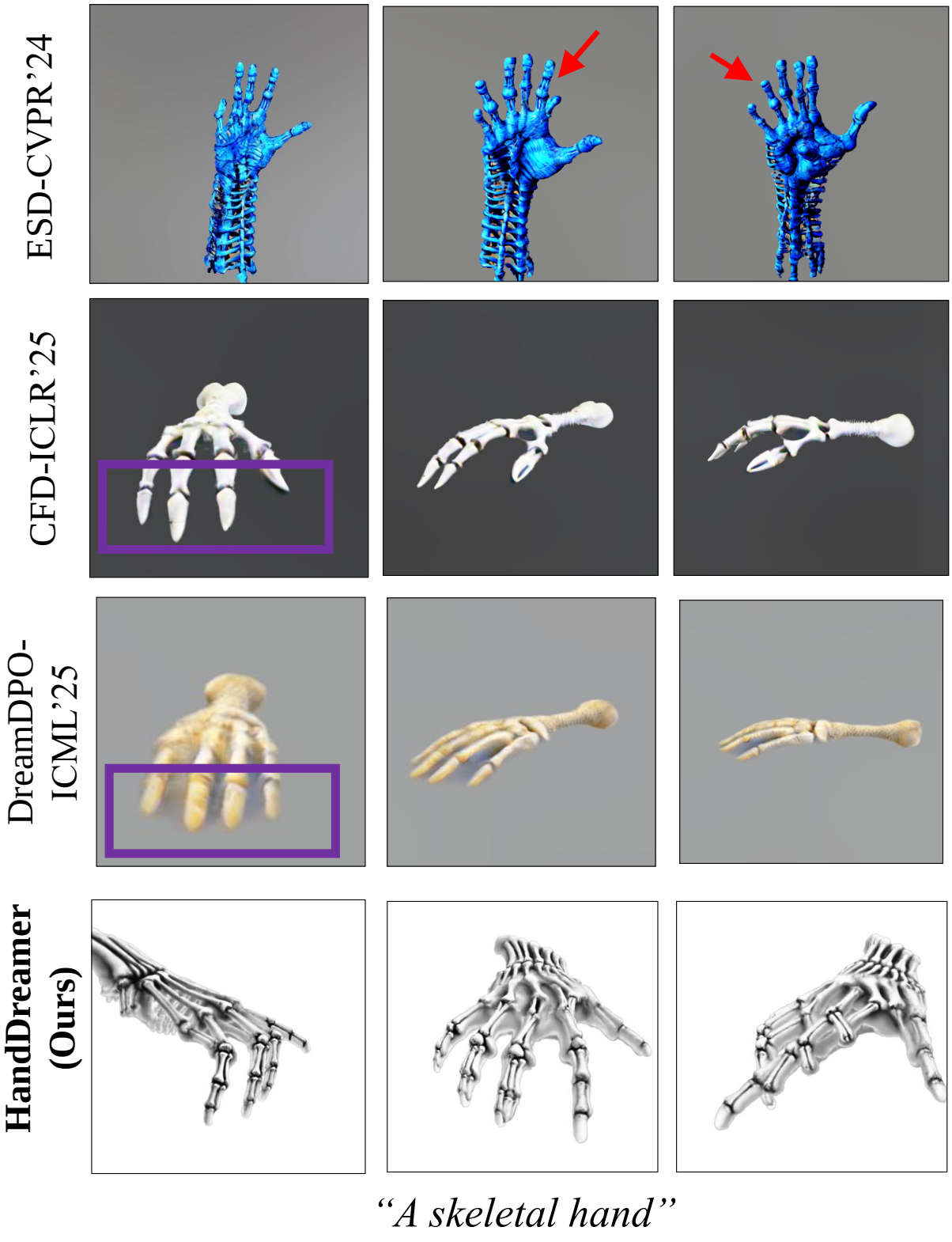


Figure 11. Comparisons against state-of-the-art methods: ESD'24 [10], CFD'25 [11], DreamDPO [15]. Janus effects shown in red arrows (ESD). Missing fingers denoted in violet box (CFD, dreamDPO).

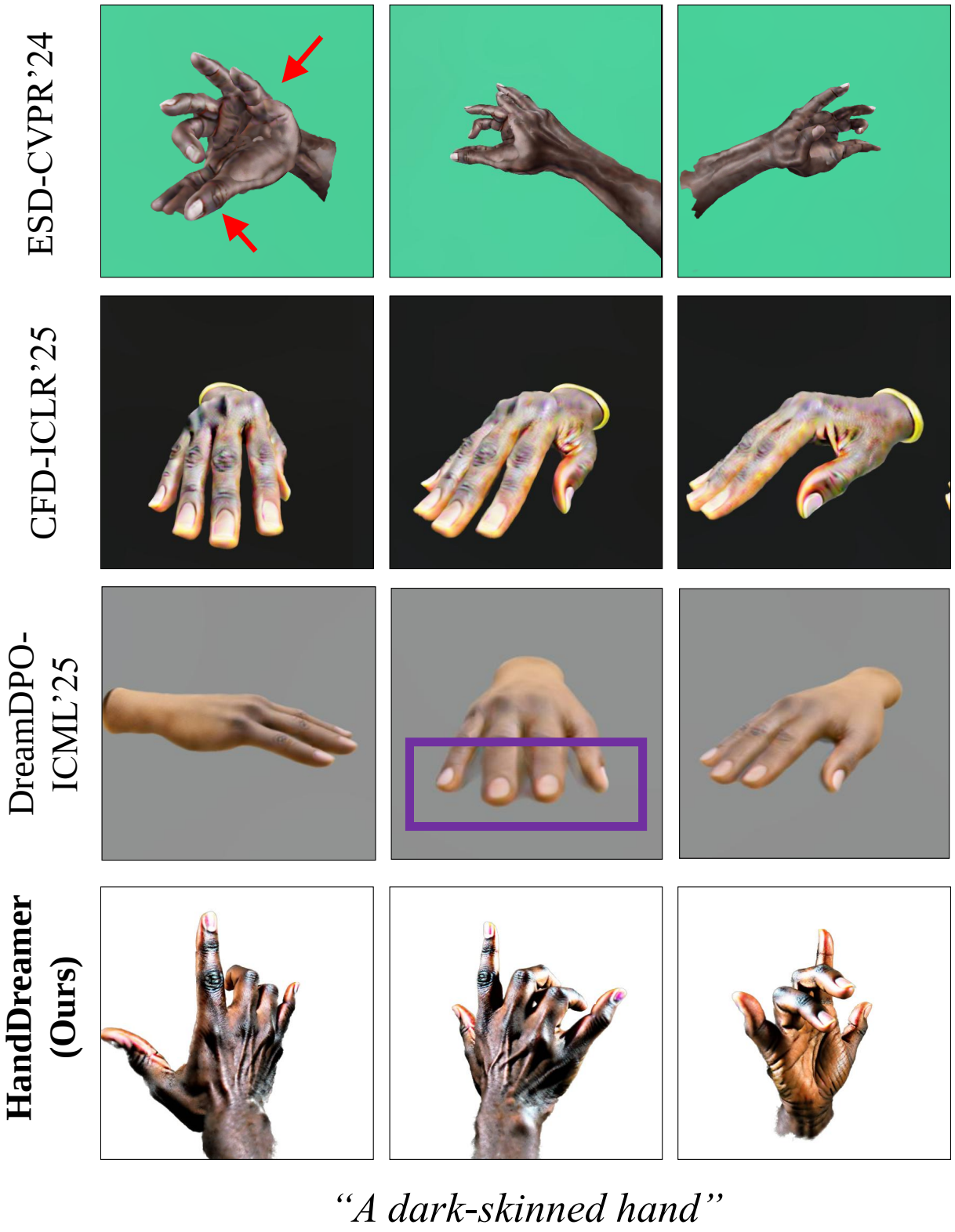


Figure 12. Comparisons against state-of-the-art methods: ESD'24 [10], CFD'25 [11], DreamDPO [15]. Janus effects shown in red arrows (ESD). Missing fingers denoted in violet box (dreamDPO).