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

Authors:

Green Rosh K S (Speaker)

Prateek Kukreja*

Vishakha SR*

Pawan Prasad B H

Samsung Research India - Bangalore

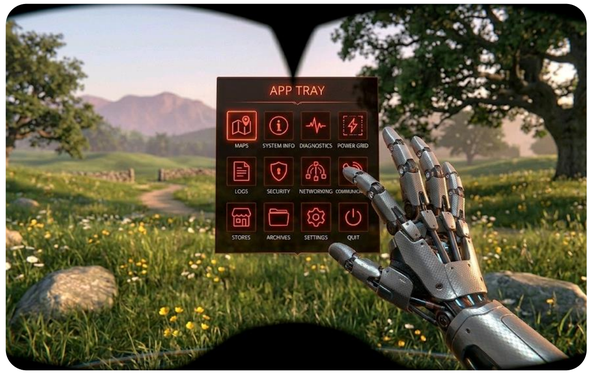
*Equal Contribution

Problem Statement

3D Hand Models

Customizable 3D hands are essential in immersive applications such as FPS games and VR interactions

Skeletal hands in FPS game

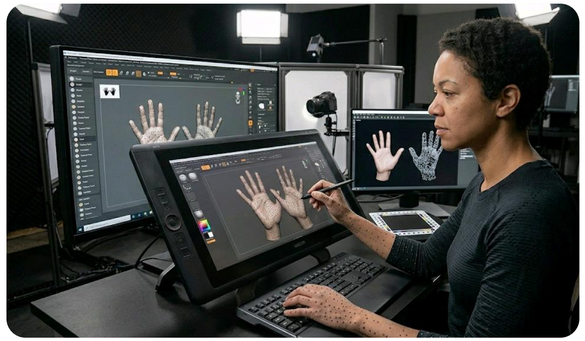


Robotic hands in VR

Challenges

- Requires expensive photogrammetry rigs
- Requires graphics expertise

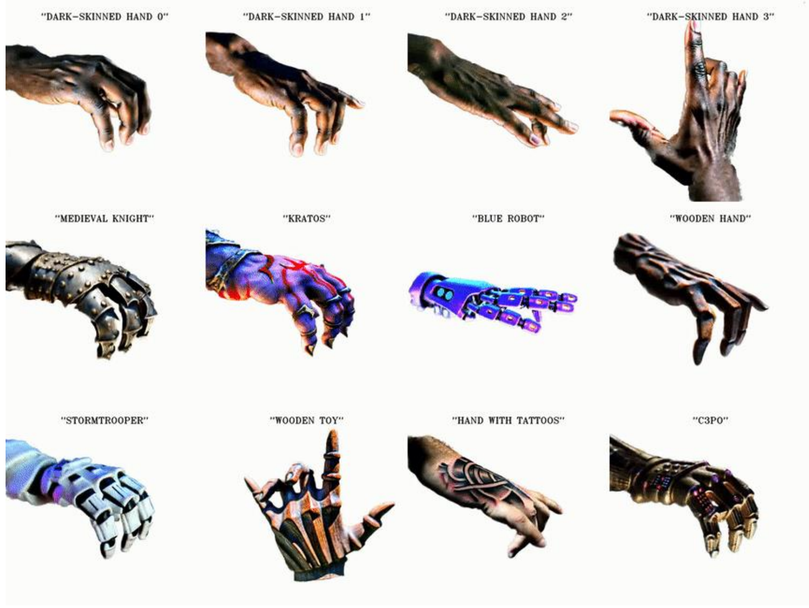
Photogrammetry rig capture



Graphics Expertise

3D hands from text

- Fast generation of 3D hands
- Easy user customization
- Democratization of 3D hand avatar creation



HandDreamer generates complex hands with detailed texture from text prompts

Representative images generated using Google Gemini

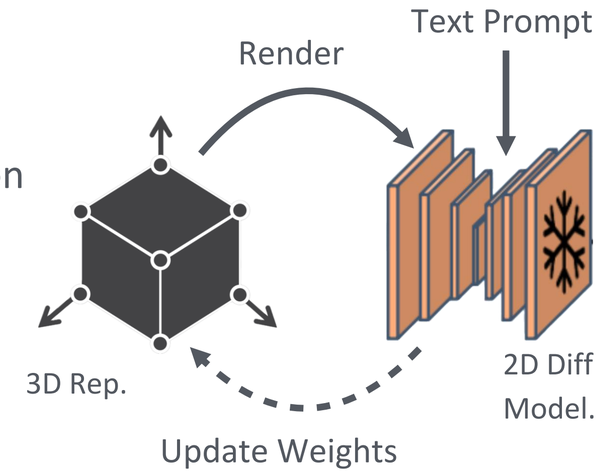
Representative images generated using Google Gemini

3D Generation from Text

Existing Methods

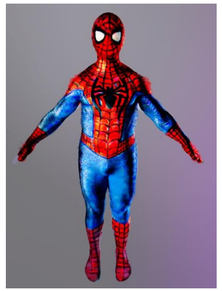
Text to 3D methods

- Score Distillation Sampling (SDS)
- Update 3D Model using frozen diffusion model
- No large 3D datasets required
- Convergence in under 3 hours



Text to human avatar methods

- Uses SMPL priors with SDS
- Accurate general human anatomy
- Articulated 3D human models



DreamAvatar

DreamWaltz

Challenges for hand generation

Text to 3D methods

- Janus Artifacts and view inconsistencies
- Anatomically wrong hand generation

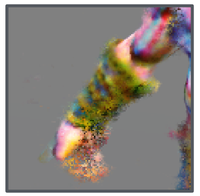
ProlificDreamer

ESD

CFD

Text to human avatar methods

- Generates hand regions with less details

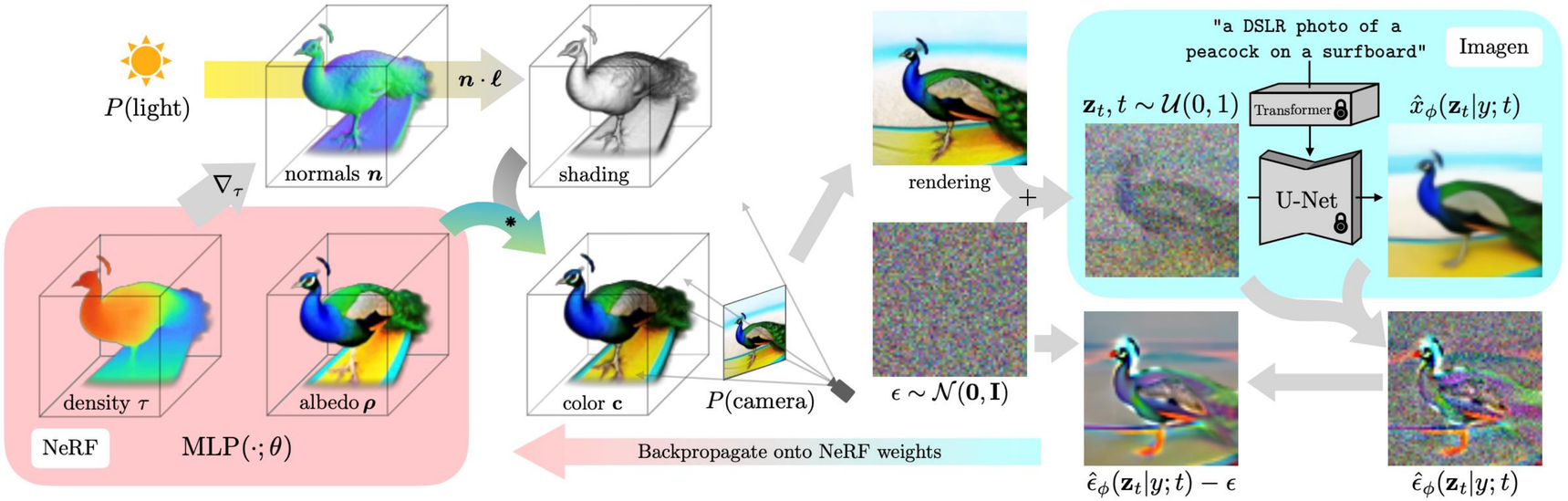


DreamAvatar

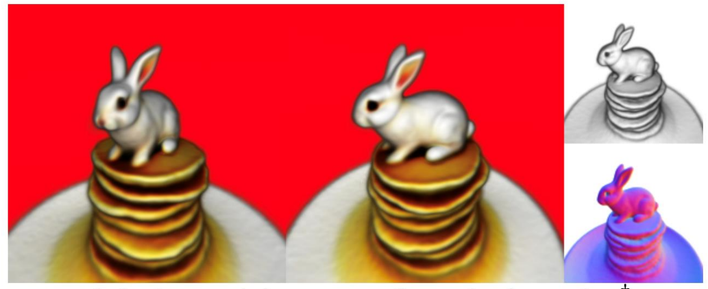
DreamWaltz

Score Distillation Sampling

Image sourced from DreamFusion

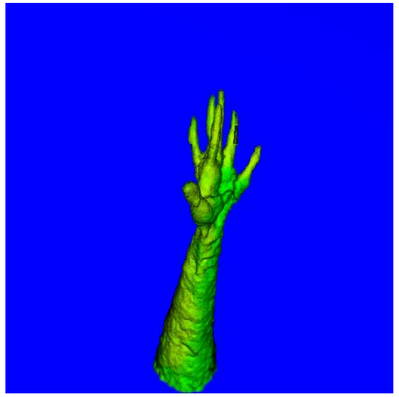


$$\nabla_\theta \mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}\left[w(t)(\epsilon_\phi(\mathbf{z}_t; y, t) - \epsilon)\frac{\partial \mathbf{z}}{\partial \theta}\right]$$



a baby bunny sitting on top of a stack of pancakes†

DreamFusion



DreamFusion

LatentNerf

Fantasia3D

HiFA

HiFA

CFD

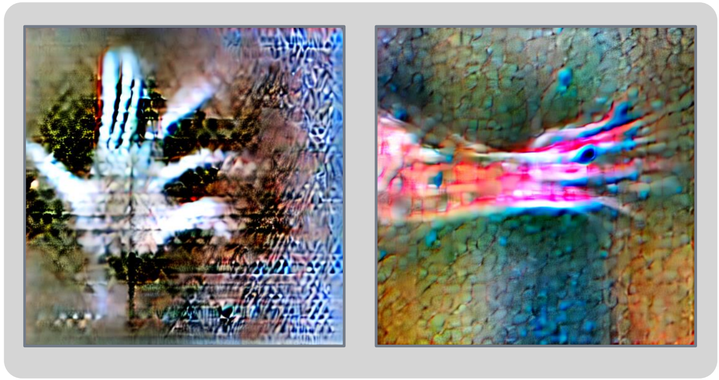
Challenges

Origin of View Inconsistencies

Large number of modes

"Hand of a robot, front view" (Obtained from Stable Diffusion)

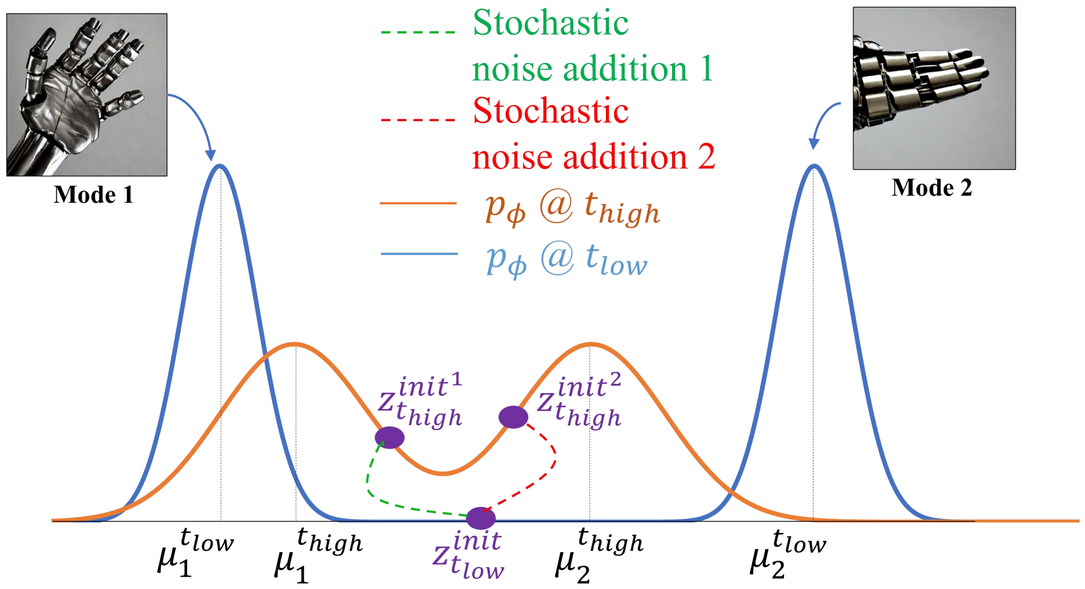
- Probability landscape defined by the text prompt contains large number of modes
- Exacerbated for highly articulated objects such as hands
- SDS gradients are inconsistent for same view-point, leading to geometric distortions



Gradients pointing in different directions for same prompt and view

Origin of View Inconsistencies

Initialization with high score



- Current methods typically use random or spherical initialization
- This is OOD from required 3D hand model and has high score ($\Delta \log p$) at lower t
- Requires evaluation at higher t for reliable gradients
- Noise stochasticity pushes gradients towards different modes for same prompt
- Results in view-inconsistencies and geometric artifacts

View inconsistencies can be minimized using an initialization with low score for w.r.t all views for low values of t

Solving High Score Initialization

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}, denoted as $|S_\phi|$ is:*

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

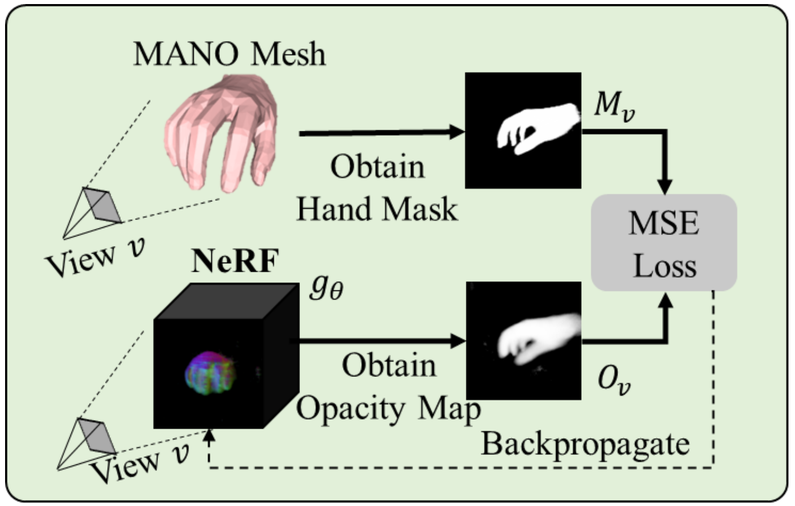
- Minimize $\| z^v_{init} - z^v_{latent} \|$ for low score initialisation
- The initialisation should be semantically closer to required output
- We use hand shape initialization using MANO model for low score

Hand Shape Initialization

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}, denoted as $|S_\phi|$ is:*

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

- Minimize $\| z^v_{init} - z^v_{latent} \|$ for low score initialisation
- The initialisation should be semantically closer to required output
- We use hand shape initialization using MANO model for low score



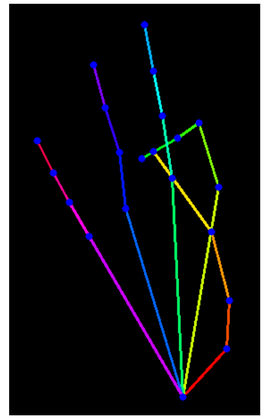
Minimize L2 distance between silhouette masks from MANO and opacity maps from 3D Representation (NeRF)

Opacity Map

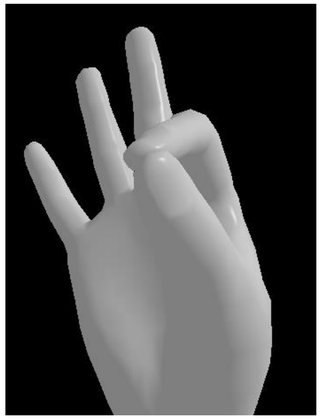
$$O_{v,\mathbf{p}} = \sum_{i=1}^{T} \left(\prod_{j=1}^{i-1} \exp\left(-\sigma_j \delta_j\right) \right) \left(1 - \exp\left(-\sigma_i \delta_i\right)\right)$$

Minimizing Modes

- Use hand skeleton to guide SDS
- Hand skeleton reduces mode ambiguity by encoding pose and view point together
- Novel corrective hand shape guidance loss for constant mode correction
- Time step annealing to avoid ambiguities at lower time steps

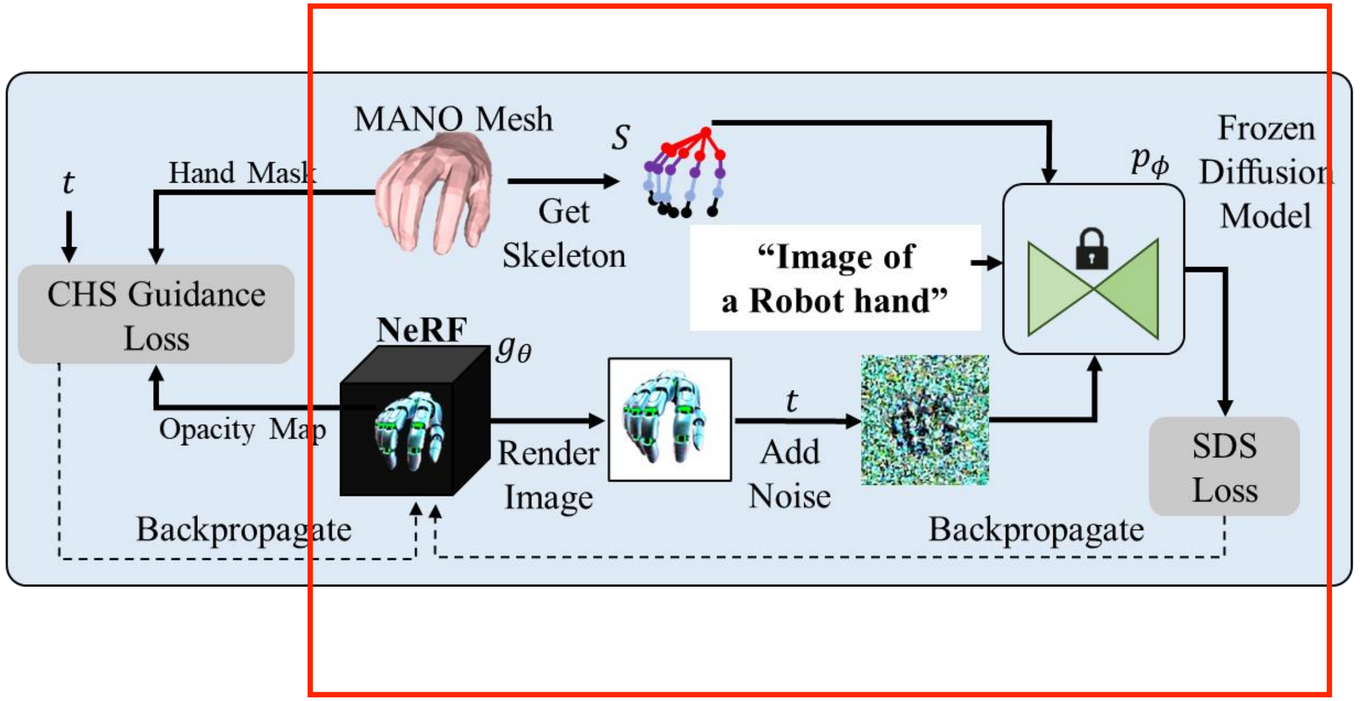


Skeleton guided SDS



Corrective hand shape guided mode correction

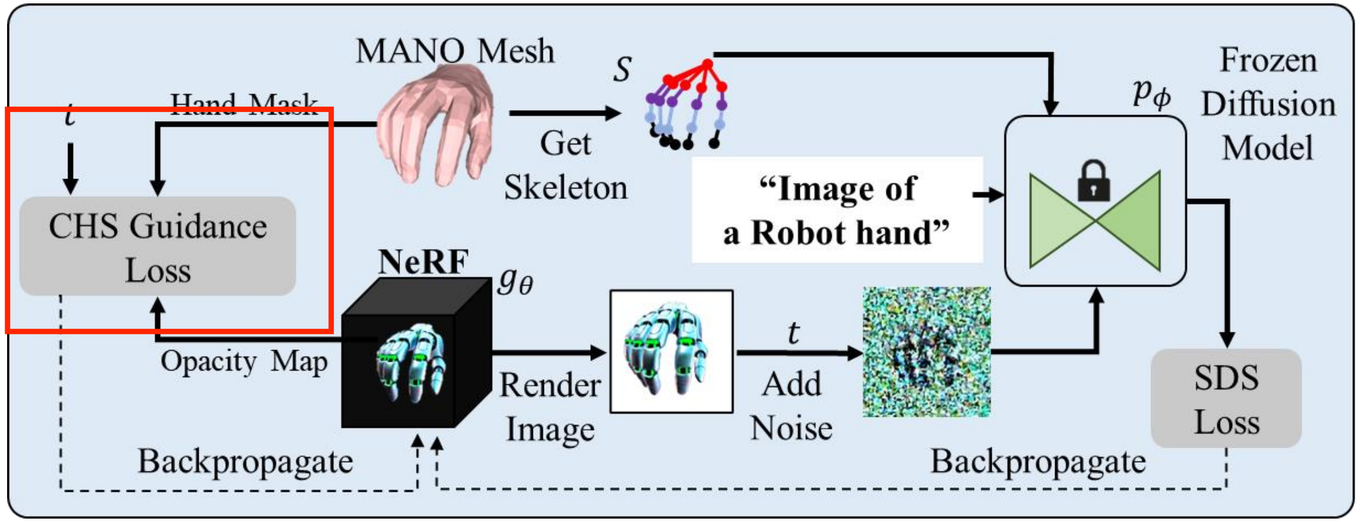
Hand Guided SDS



- Render from NeRF for view v and add noise
- Get skeleton for view v from MANO mesh
- Give noisy image, skeleton & text prompt a ControlNet trained on hand skeletons for pose and view guidance
- Compute SDS gradients and update NeRF weights

$$\nabla_\theta \mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}\left[w(t)(\epsilon_\phi(\mathbf{I}_t; y, t, S) - \epsilon)\frac{\partial \mathbf{I}}{\partial \theta}\right]$$

Corrective Hand Shape Guidance



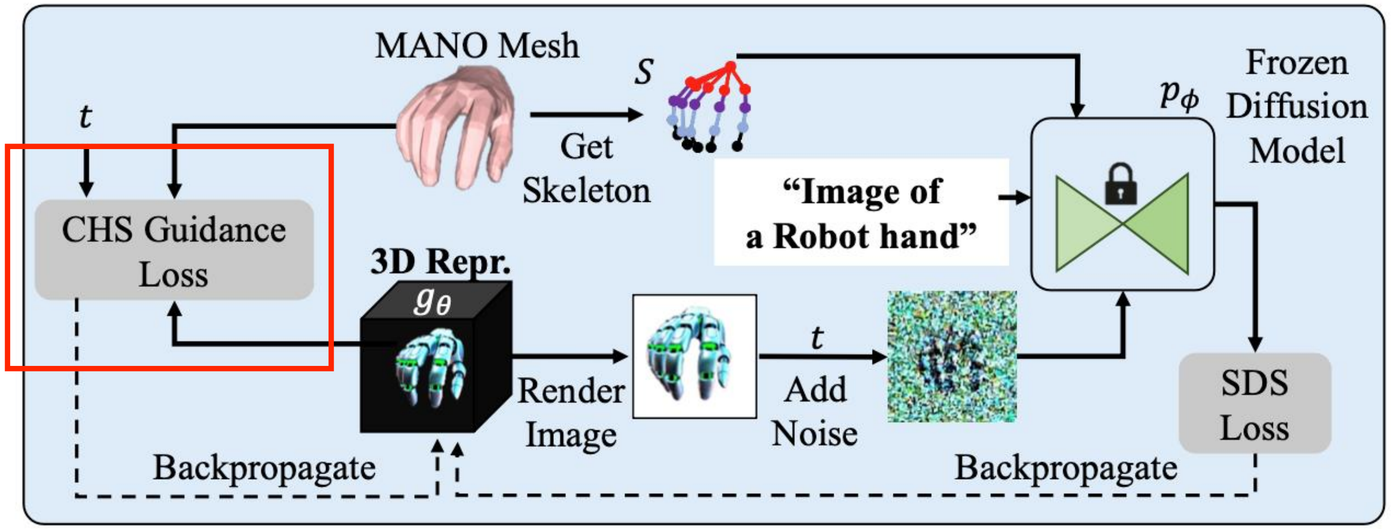
$$\mathcal{L}_{chs}(t) = \lambda_t^{chs} . \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \|O_v - M_v\|_2$$

where λ_t^{chs} is annealed as follows:

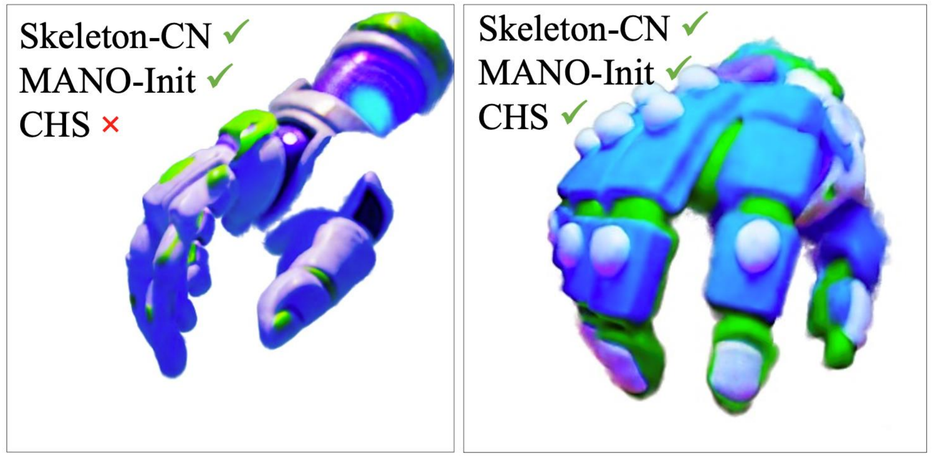
$$\lambda_t = \lambda_{max}^{chs} \left[\frac{t - t_{min}}{t_{max} - t_{min}} \right] + \lambda_{min}^{chs} \left[\frac{t_{max} - t}{t_{max} - t_{min}} \right]$$

- Only SDS lead to geometric distortions in side view due self-occlusion of fingers
- Additional hand shape guidance provided using MANO shape prior (CHS Loss)
- L2 Distance between current opacity mask and MANO hand shape minimised
- Ensures the optimisation do not converge too much away from a rough hand shape
- Geometric updates more at higher t and texture updates at lower t
- Anneal weight of CHS loss for higher weightage at earlier iterations with higher t

Corrective Hand Shape Guidance



"Hand of buzz light year"



- Only SDS lead to geometric distortions in side view due self-occlusion of fingers
- Additional hand shape guidance provided using MANO shape prior (CHS Loss)
- L2 Distance between current opacity mask and MANO hand shape minimised
- Ensures the optimisation do not converge too much away from a rough hand shape
- Geometric updates more at higher t and texture updates at lower t
- Anneal weight of CHS loss for higher weightage at earlier iterations with higher t

Results

“Hand of a wooden toy”

“Hand of a Stormtrooper from Star Wars”

“A dark-skinned hand”

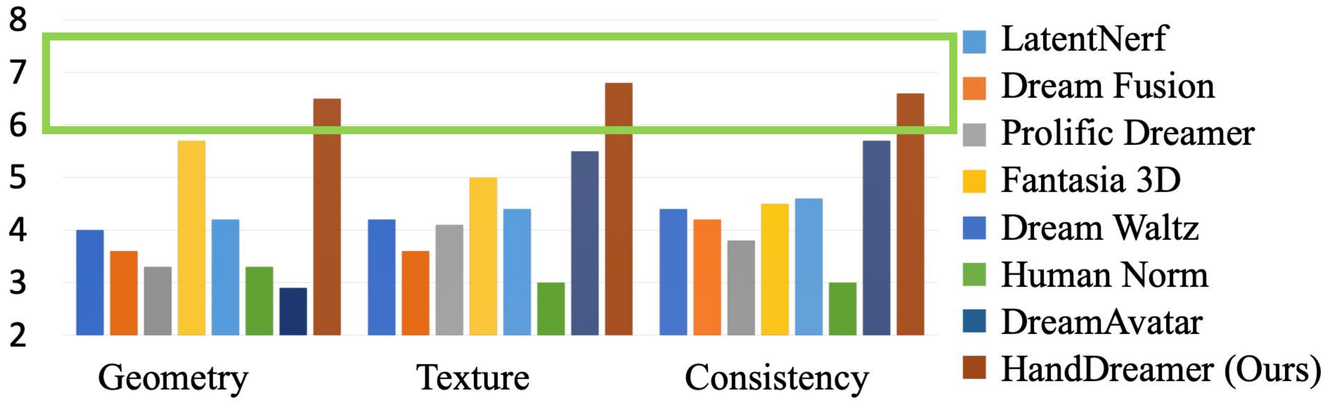
HandDreamer generates geometrically accurate 3D hand models from text prompts with diverse textures and articulations

Comparisons

Comparisons

Method	CLIP L14 ↑	FID ↓	HPSv2 ↑
DreamFusion'22 [35]	25.12	344.19	0.187
LatentNerf'23 [29]	24.34	316.42	0.189
Fantasia3D'23 [5]	20.93	329.31	0.198
DreamWaltz'23 [17]	23.96	265.11	0.222
DreamAvatar'24 [2]	20.02	329.85	0.215
HumanNorm'24 [15]	23.01	327.42	0.177
SDI'24 [27]	26.32	297.12	0.192
OHTA'24 [63]	22.59	467.51	0.181
CED'25 [57]	26.62	262.83	0.223
HandDreamer (Ours)	**28.63**	**254.62**	**0.241**

Quantitative Comparisons



User preference studies

Conclusions

Conclusions

- HandDreamer: The first method for zero shot 3D hand generation from text prompts
- Analysis: View inconsistencies due to large number of modes in probability landscape of the text prompt
- Solution: MANO based low score initialization, Hand shape and pose based guidance
- Results: Significantly outperforms text-to-3D methods visually and quantitatively

Limitations

- Automated hand articulation currently not supported
- Convergence takes ~1 hour to complete

Future Work

- Support automated articulation using articulated NeRF models
- Faster convergence using better initialisation approaches