3DOT: Texture Transfer for 3DGS Objects from a Single Reference Image

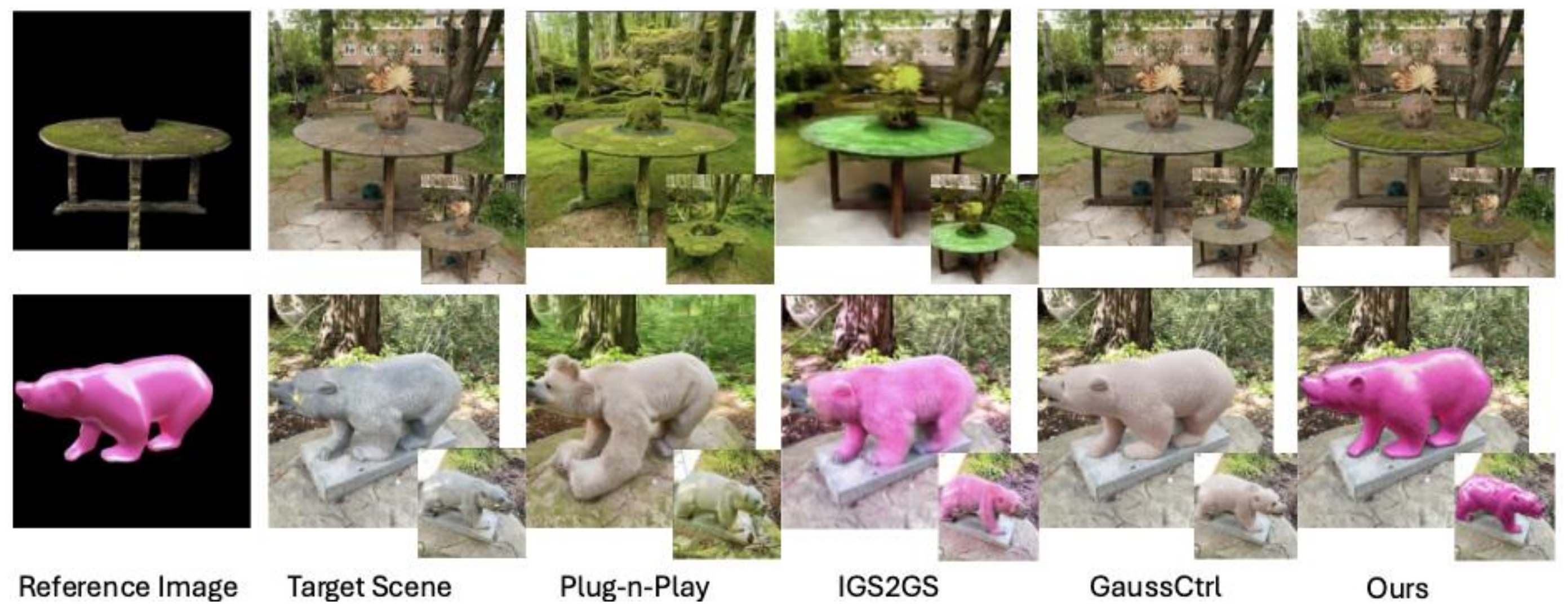
Xiao Cao, Beibei Lin, Bo Wang, Zhiyong Huang, Robby T. Tan

Project Page: https://massyzs.github.io/3DOT_web/
Email: xiaocao@u.nus.edu

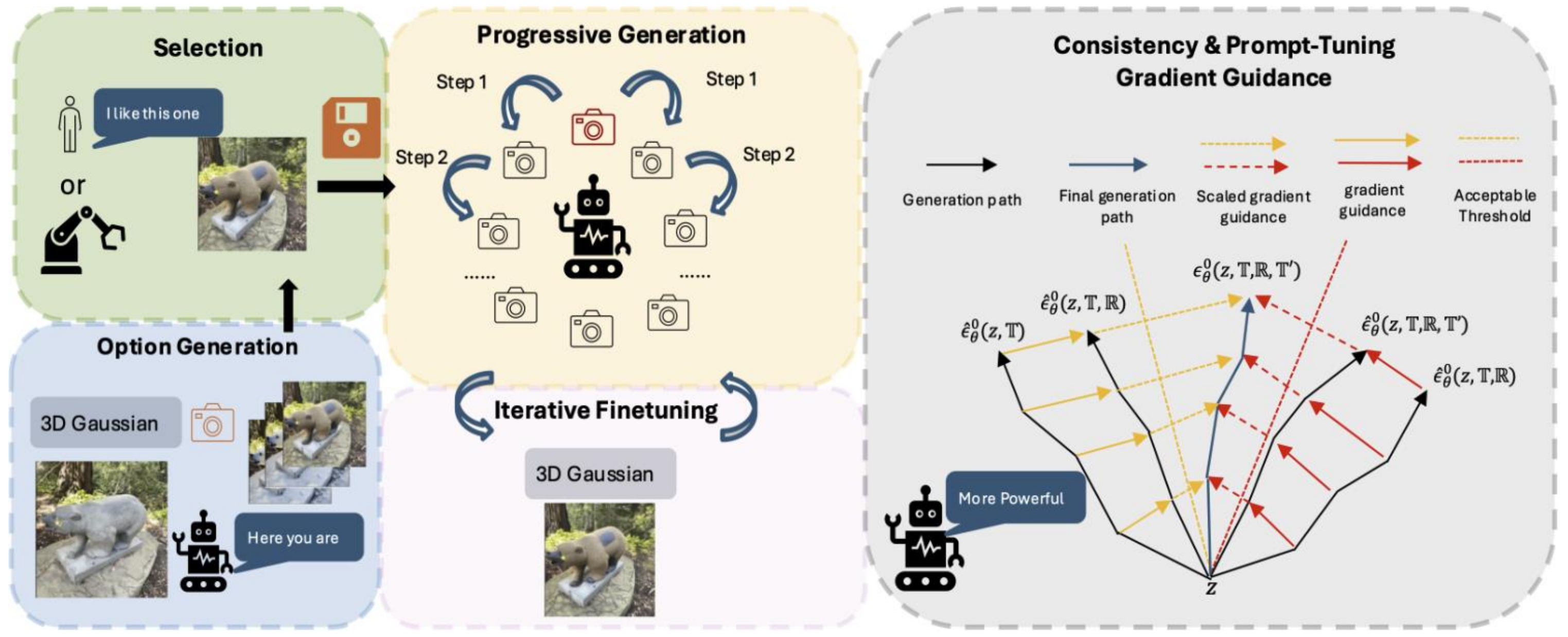
Problem Statement

Transferring texture from a 2D image to a 3D object is a valuable yet underexplored capability in 3D editing. It enables efficient texture manipulation and benefits applications such as virtual reality, CG films, and 3D games. Existing methods may be adjusted to achieve this functionality, while suffer from several problems.

- 2D methods perform texture transfer by finetuning a diffusion model. The resulting 3D object often suffers from view inconsistency and identity loss.
- 3D editing methods , especially text-driven ones, guide editing using prompts derived from reference images via VLM. However, these prompts are coarse and miss fine-grained features, resulting in identity mismatch and inconsistent appearance across views.



Multi-view Consistent 3D Editing Pipeline


Selection
I like this one
or
Option Generation
3D Gaussian
Here you are
Progressive Generation
Step 1
Step 1
Step 2
Step 2
......
......
Iterative Finetuning
3D Gaussian
Consistency & Prompt-Tuning Gradient Guidance
Generation path
Final generation path
Scaled gradient guidance
gradient guidance
Acceptable Threshold
$\epsilon_\theta^0(z,\mathbb{T},\mathbb{R},\mathbb{T}')$
$\hat{\epsilon}_\theta^0(z,\mathbb{T},\mathbb{R})$
$\hat{\epsilon}_\theta^0(z,\mathbb{T})$
$\hat{\epsilon}_\theta^0(z,\mathbb{T},\mathbb{R},\mathbb{T}')$
$\hat{\epsilon}_\theta^0(z,\mathbb{T},\mathbb{R})$
z
More Powerful

Workflow:

(1) User provides images with desired texture / User utilizes provided depth-based diffusion to generate texture and manually select the satisfactory one.

(2) The reference images are fed into progressive generation stage. The generative model with the help of consistency and prompt-tuning gradient guidance, edit image progressively.

(3) The multi-view consistent edited images are used to finetune the 3DGS model. This process is performed iteratively — i.e., the finetuned 3DGS is rendered and edited again, followed by further fine-tuning.

Consistency Gradient Guidance

To enhance cross-attention effectiveness under this constraint, we propose a consistency-aware gradient guidance mechanism inspired by classifier-free guidance, modifying the noise estimate to amplify cross-view signals without additional training.

To be specific, given a target view I_i and reference set $R_i = \{I_T, I_{i-1}, F(I_T)\}$, we define the denoising prediction as:

$$\begin{aligned}\epsilon_\theta^t(z_\lambda,\mathbb{T},\mathbb{R}) = {} & \epsilon_{\hat{\theta}}^t(z_\lambda) \\ & + w_{\mathbb{T}}\big(\epsilon_\theta^t(z_\lambda,\mathbb{T},\mathbb{R}) - \epsilon_\theta^t(z_\lambda,\mathbb{R})\big) \\ & + w_{\mathbb{R}}\big(\epsilon_\theta^t(z_\lambda,\mathbb{T},\mathbb{R}) - \epsilon_{\hat{\theta}}^t(z_\lambda,\mathbb{T})\big),\end{aligned}$$

The intuition is to compare the editing difference between images generated with and without cross-attention mechanism and amplify this feature.

Prompt-tuning Gradient Guidance

Text prompts provide only coarse control during diffusion, often leading to identity loss and inconsistent texture fidelity. Learning new token precisely describe a texture is hard. We learn a new token that captures the texture discrepancy between the unedited 3D object and the reference image, and to use this token to guide denoising toward the desired style.

We first extract the texture difference in Clip image space and optimize a new token by cosine similarity.

$$\Delta_{\hat{\mathbb{I}}_\tau \rightarrow \mathbb{I}_\tau} = \mathbf{CLIP}(\hat{\mathbb{I}}_\tau) - \mathbf{CLIP}(\mathbb{I}_\tau)$$

$$L_{\text{clip}} = \text{cosine}(\Delta_{\hat{\mathbb{I}} \quad \rightarrow \mathbb{I}} \ , \hat{\mathbb{T}})$$

To further reduce the domain gap between image clip space with diffusion feature space, we further optimize the token by diffusion loss.

$$L_{\text{diff}} = \epsilon_\theta(z_\lambda,\mathbb{T}',\mathbb{R}) - \epsilon_\theta'(z_\lambda,\mathbb{T}',\mathbb{R})$$

Overall Denoising Formulation

The above two components can be easily cooperated into denoising process together with classifier-free negative guidance. The overall formulation can be concluded as:

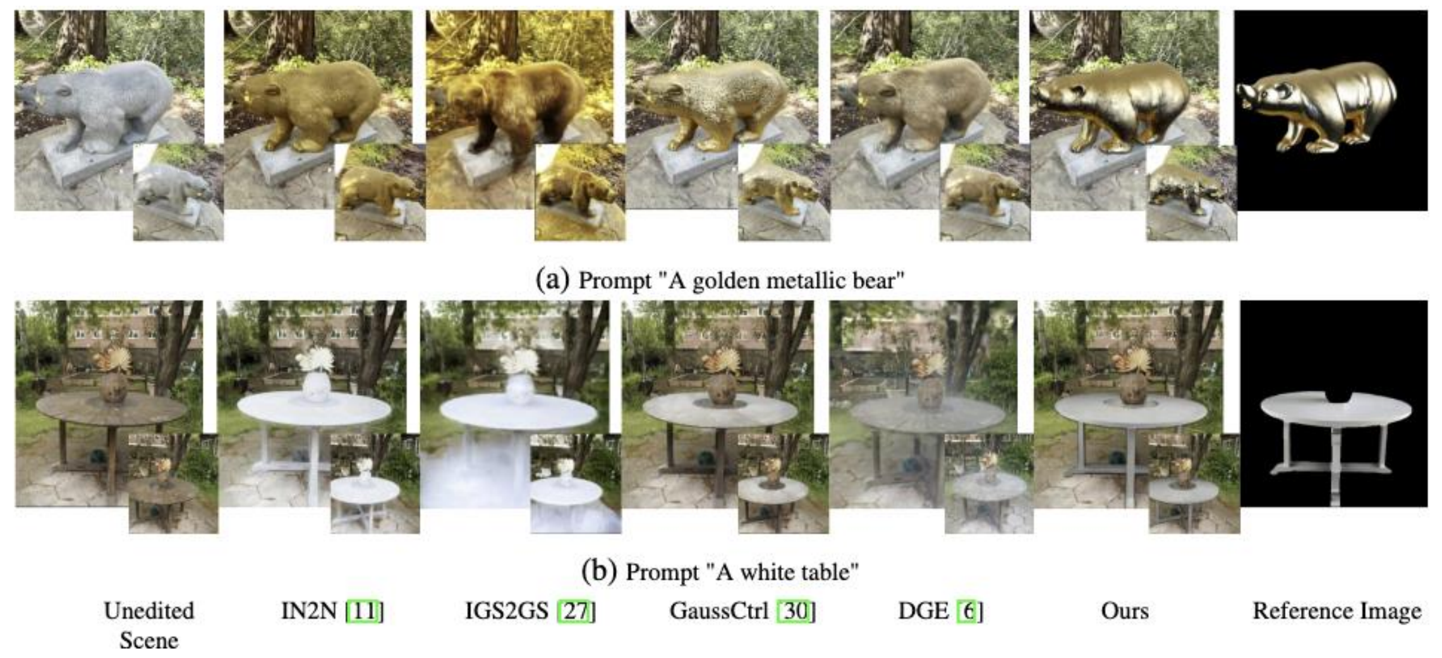
$$\begin{aligned}\epsilon_\theta^t(z_\lambda,\mathbb{T},\mathbb{R},\mathbb{T}') = {} & \epsilon_{\hat{\theta}}^t(z_\lambda) \\ & + w_{\mathbb{T}}\big(\epsilon_\theta^t(z_\lambda,\mathbb{T},\mathbb{R}) - \epsilon_\theta^t(z_\lambda,\mathbb{R})\big) \\ & + w_{\mathbb{R}}\big(\epsilon_\theta^t(z_\lambda,\mathbb{T},\mathbb{R}) - \epsilon_{\hat{\theta}}^t(z_\lambda,\mathbb{T})\big) \\ & + w_{\mathbb{T}'}\big(\epsilon_\theta^t(z_\lambda,\mathbb{T}',\mathbb{R}) - \epsilon_{\hat{\theta}}^t(z_\lambda,\mathbb{T},\mathbb{R})\big)\end{aligned}$$

Experiments

Qualitative Results

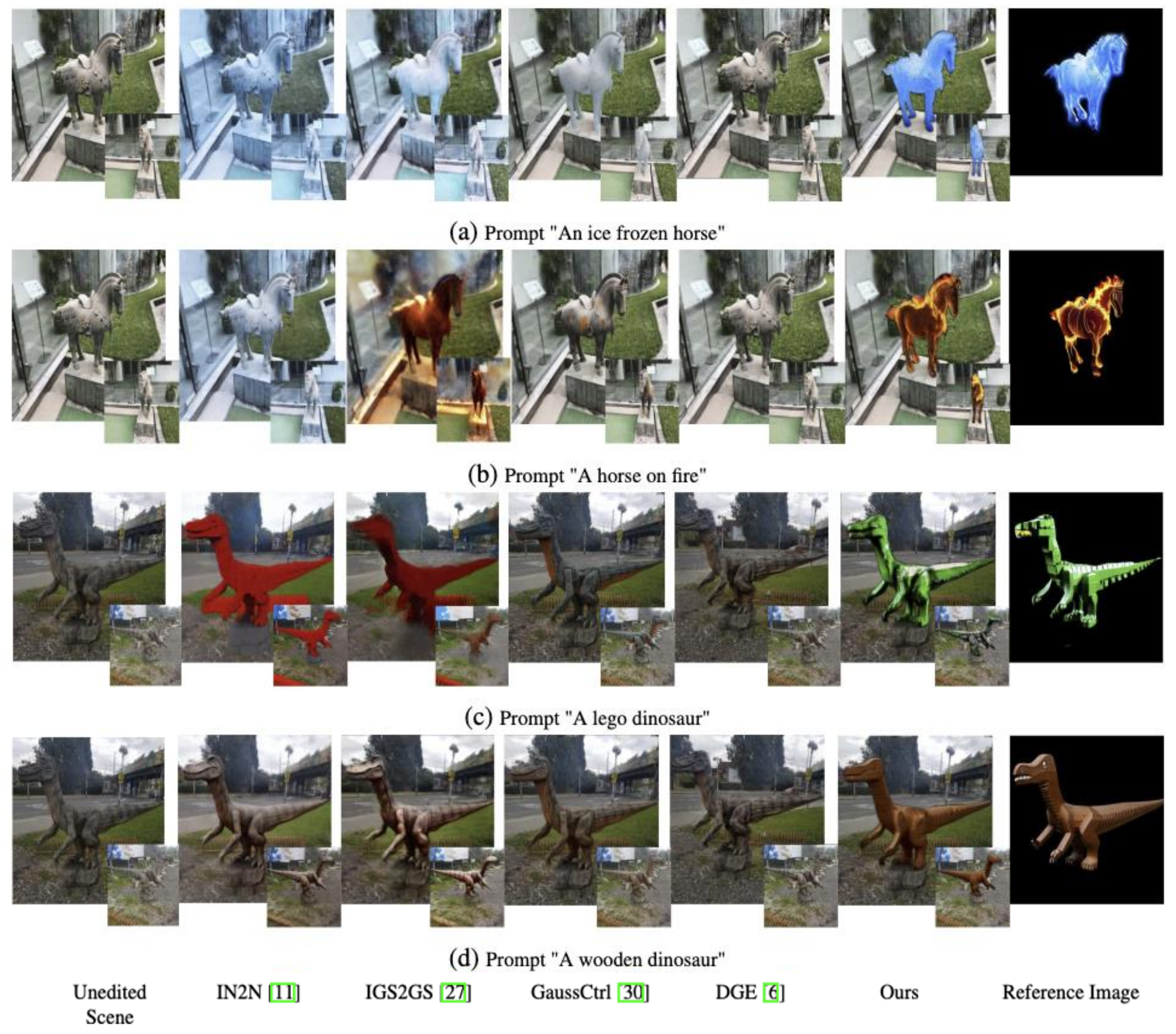
We conduct editing on easy cases (simple texture), hard cases (complicated texture) and facial cases.

Easy Cases


(a) Prompt "A golden metallic bear"

(b) Prompt "A white table"

Hard Cases

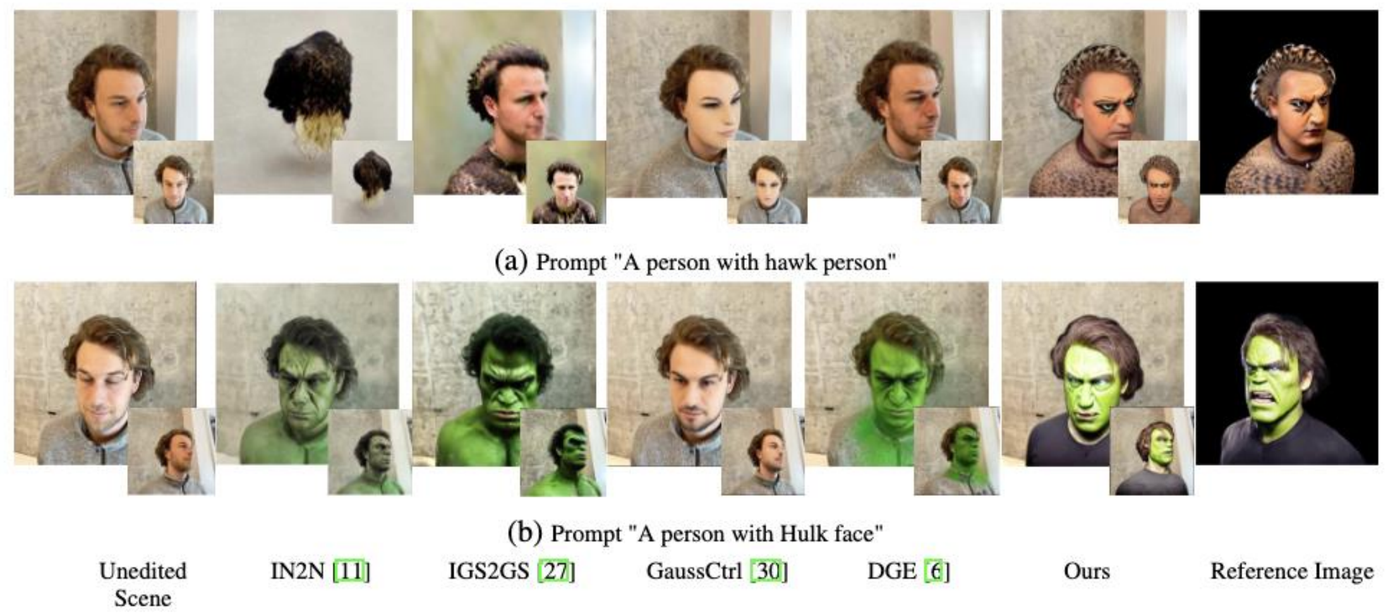

(a) Prompt "An ice frozen horse"

(b) Prompt "A horse on fire"

(c) Prompt "A lego dinosaur"

(d) Prompt "A wooden dinosaur"

Facial Cases


(a) Prompt "A person with hawk person"

(b) Prompt "A person with Hulk face"

Quantitative Results

3DOT outperforms baselines for all metrics.

Metrics	IN2N	IGS2GS	GaussCtrl	DGE	Ours
CLIP Score↑	0.8917	0.8908	0.8638	0.8572	**0.9333**
Lpips(Alex)↓	0.1708	0.1683	0.1692	0.1713	**0.1166**
Lpips(VGG)↓	0.1676	0.1594	0.1591	0.1603	**0.1247**
Vision-GPT ↑	45.5	52	48	54	**76**
User study↑	2.0375	2.4375	2.3750	2.0000	**4.5750**