GenesisTex: Adapting Image Denoising Diffusion to Texture Space

Supplementary Material

1. Implementation Details

Detailed Parameters. We set 3 sets of camera viewpoints $\mathcal{C}^{(sampling)}$, $\mathcal{C}^{(inpainting)}$ and $\mathcal{C}^{(img2img)}$ for texture space samling, Inpainting epoch and Img2Img epoch, respectively. We use the same viewpoints configuration for all the inputs, as shown in Tab. 1. We disable 'guess mode', *i.e.*, we did not apply depth control to the unconditional guidance side of the classifier-free guidance because guess mode tends to produce unnatural colors. Following the original DDIM [10], in the denoising process we set $\sigma_i = \sqrt{(1-\alpha_{i-1})/(1-\alpha_i)}\sqrt{1-\alpha_i/\alpha_{i-1}}$ and use a linear time schedule $\{t_i\}_{i=T}^{0}$.

Rendering Settings. We modify nvdiffrec [6], which is based on the differentiable rendering pipeline implemented using nvdiffrast [3], to implement the rendering function $\mathcal{R}$. We set the BSDF (bidirectional scattering distribution function) type to 'k_d' to ignore the influence of lighting, as in this work, we focus on generating the content of textures rather than decoupling materials from lighting. The visual results presented in the main paper also use 'k_d' as the BSDF type. In Fig. 1, we show the rendering results with '*diffuse*' as the BSDF type, using a museumplein environment light. It can be observed that some inconsistent light-dark relationships appear in '*diffuse*' rendering results. We will explore generating texture maps that comply with the Physically Based Rendering (PBR) workflow in future work.

elevation	azimuth	elevation	azimuth
$\mathcal{C}^{(sampling)}$			
0°	0°	0°	180°
0°	90°	0°	270°
$\mathcal{C}^{(inpainting)}$			
90°	0°	60°	315°
0°	45°	60°	90°
0°	315°	60°	270°
0°	135°	60°	135°
0°	225°	60°	225°
60°	0°	60°	180°
60°	45°		
$\mathcal{C}^{(img2img)}$			
0°	180°	0°	270°
0°	135°	0°	45°
0°	225°	0°	315°
0°	90°	0°	0°

Table 1. Viewpoints settings.

Figure 1. Renderings with '*diffuse*' BSDF.

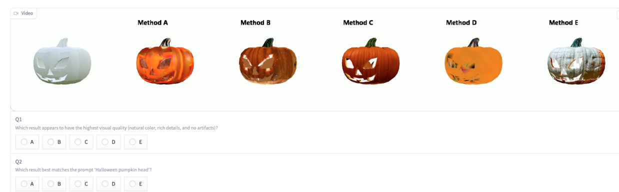

Figure 2. Screenshot of user study.

Loss \ Win	Ours	Text2Tex	TEXTure
Ours	-	20	24
Text2Tex	60	-	43
TEXTure	56	37	-

Table 2. Results of pairwise user study.

User Study. To compare with the baseline methods, we conduct a user study as part of the evaluation. We implement a survey using Gradio [1], which is a webpage-based tool. The survey randomly present 10 groups of generated results to each participant. A screenshot of the survey for a group of generated results is displayed in Fig. 2, which includes six videos and two questions:

1. *Which result appears to have the highest visual quality (natural color, rich details, and no artifacts)?*
2. *Which result best matches the prompt '[prompt]'?*

For each group of results displayed in the videos, we ensure that their order is randomly shuffled to prevent bias. Responses where all answers have the same selection and responses with completely identical answers are considered invalid. After filtering, we obtain a total of 35 valid surveys.

We also conduct a pairwise comparison test with two competitive methods, as shown in Tab. 2. We employ the Bradley-Terry model to analyze the results of the pairwise user study. The estimated Bradley-Terry model parameters $p_{\texttt{Ours}}, p_{\texttt{Text2Tex}}, p_{\texttt{TEXTure}}$ are $1.91, 0.66, 0.79$ respectively, which indicates that ours is the strongest.

2. More Results

We highly recommend readers to visit our project homepage[1] to view the result videos.

2.1. Comparison Results

We provide more visual comparisons between our method and state-of-the-art baselines [2, 4, 5, 8] in the video titled '*Comparisons*'. In Fig. 5-10, we show some multi-view renderings from the video. It is clear from these comparisons that our method outperforms the baseline approaches in terms of both visual quality and alignment with the input prompt.

2.2. Ablation Results

In texture space sampling, we leverage dynamic alignment and style consistency to ensure consistency across multiple viewpoints. To verify the effectiveness of these two operations on the results, we present a visual comparison of the generated results under different consistency settings in the video titled '*Consistency Ablations*'. In Fig. 11-12, we show some multi-view renderings from the video. It can be observed that style consistency greatly affects the global style harmony, while dynamic alignment can resolve multi-view conflicts.

2.3. Stable Diffusion XL Generation Results

In the main paper, we utilized `Stable Diffusion v1.5` [9] as the image diffusion model. To further enhance the quality of generated textures, we conducted an experiment to explore the effectiveness of GenesisTex using `Stable Diffusion XL` [7] for texture synthesis. The results are showcased in the video titled '*Texturing with Stable Diffusion XL*'. Figure 3 and Fig. 4 displays some multi-view rendering images extracted from the video. It can be observed that our method, leveraging Stable Diffusion XL, produces textures with remarkably high detail quality and minimal artifacts. This experiment highlights the potential of our approach when applied with more powerful image diffusion models.

References

[1] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*, 2019. 1

[2] Dave Zhenyu Chen, Yawar Siddiqui, Hsin-Ying Lee, Sergey Tulyakov, and Matthias Nießner. Text2tex: Text-driven texture synthesis via diffusion models. *arXiv preprint arXiv:2303.11396*, 2023. 2

[3] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 1

[4] 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. 2

[5] Oscar Michel, Roi Bar-On, Richard Liu, Sagie Benaim, and Rana Hanocka. Text2mesh: Text-driven neural stylization for meshes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13492–13502, 2022. 2

[6] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 1

[7] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2

[8] Elad Richardson, Gal Metzer, Yuval Alaluf, Raja Giryes, and Daniel Cohen-Or. Texture: Text-guided texturing of 3d shapes. *arXiv preprint arXiv:2302.01721*, 2023. 2

[9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[10] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1

[1] https://cjeen.github.io/GenesisTexPaper/

Figure 3. Generation results with Stable Diffusion XL I.

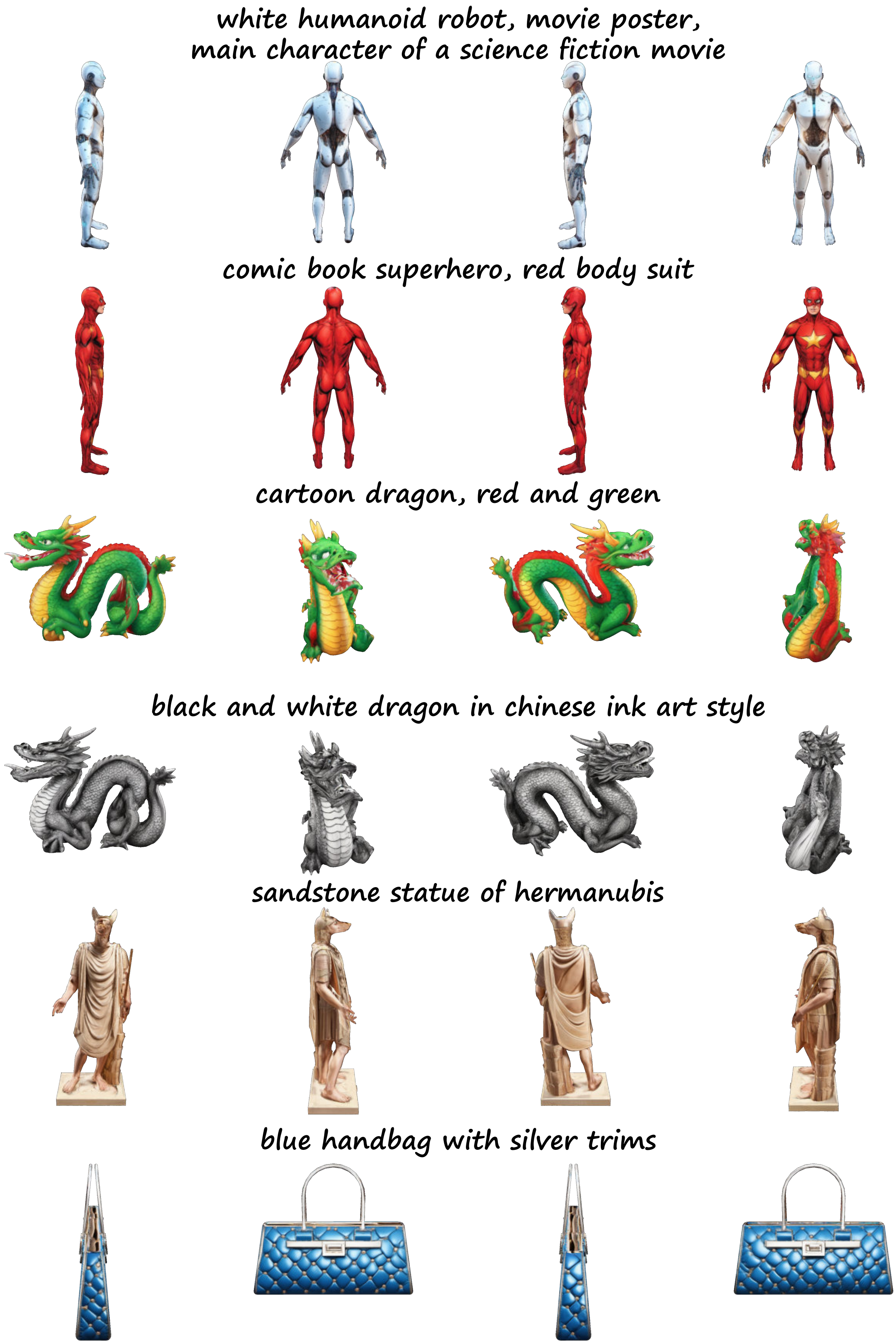

Figure 4. Generation results with Stable Diffusion XL II.

Figure 5. More qualitative comparisons I.

Figure 6. More qualitative comparisons II.

Figure 7. More qualitative comparisons III.

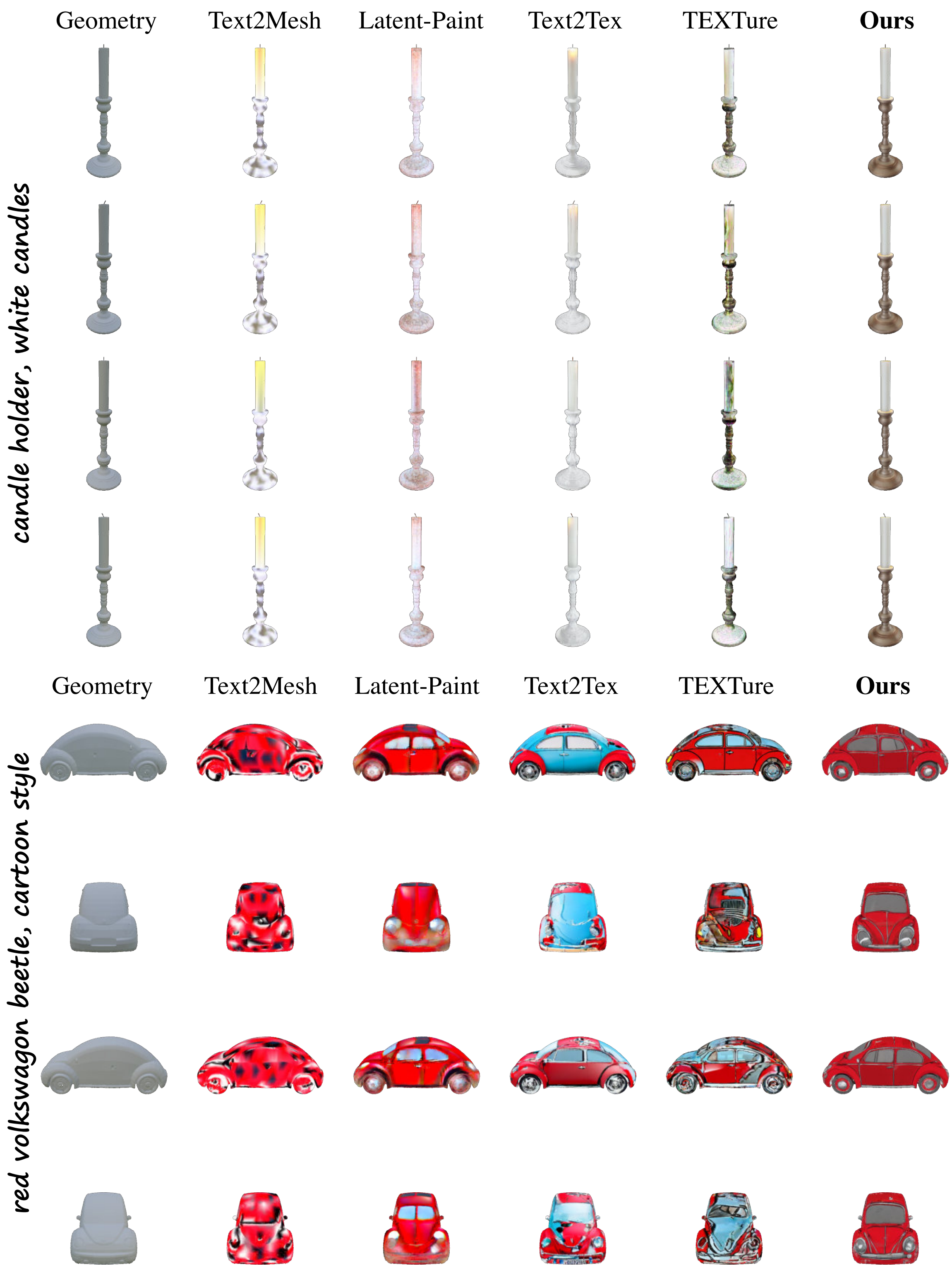

Figure 8. More qualitative comparisons IV.

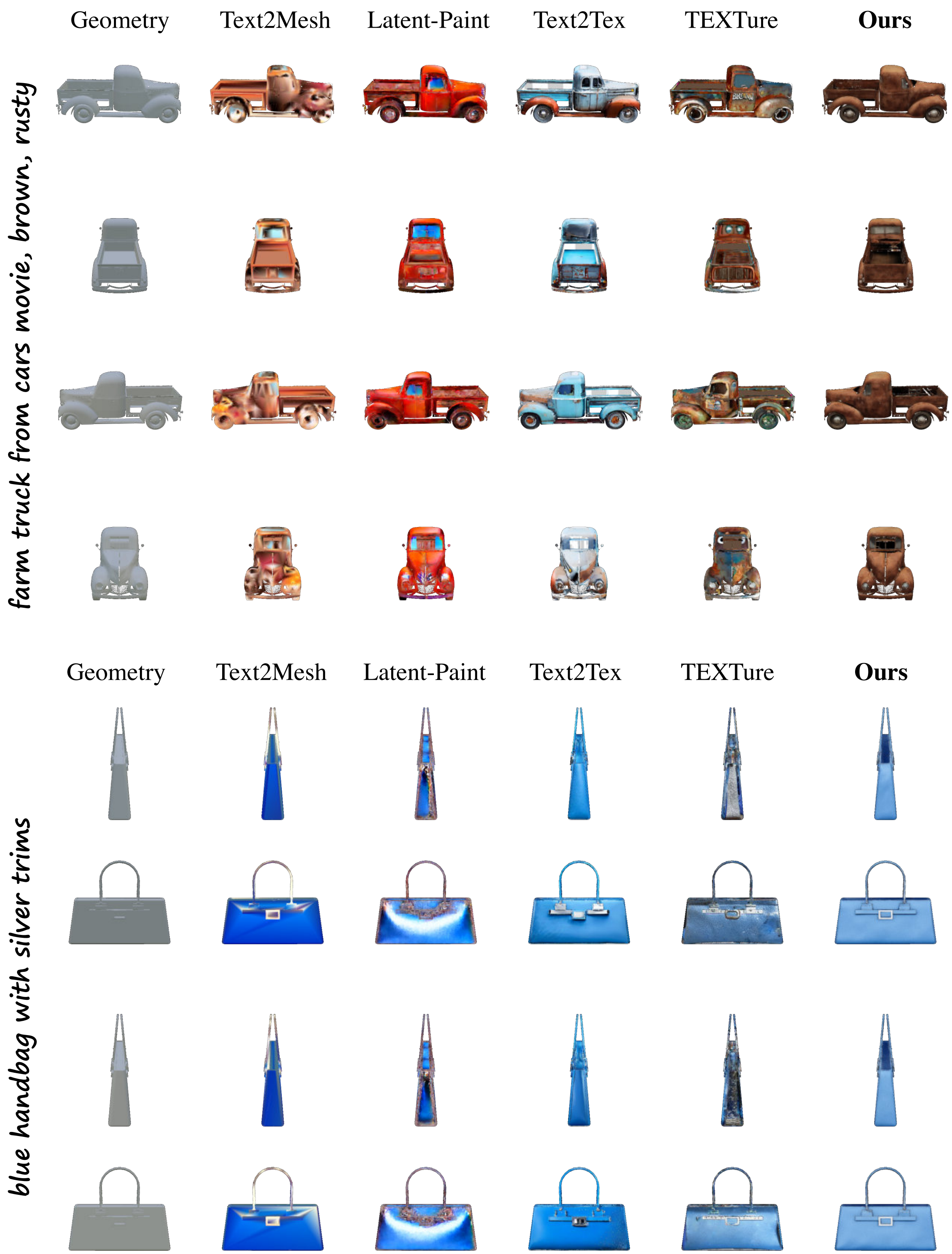

Figure 9. More qualitative comparisons V.

Figure 10. More qualitative comparisons VI.

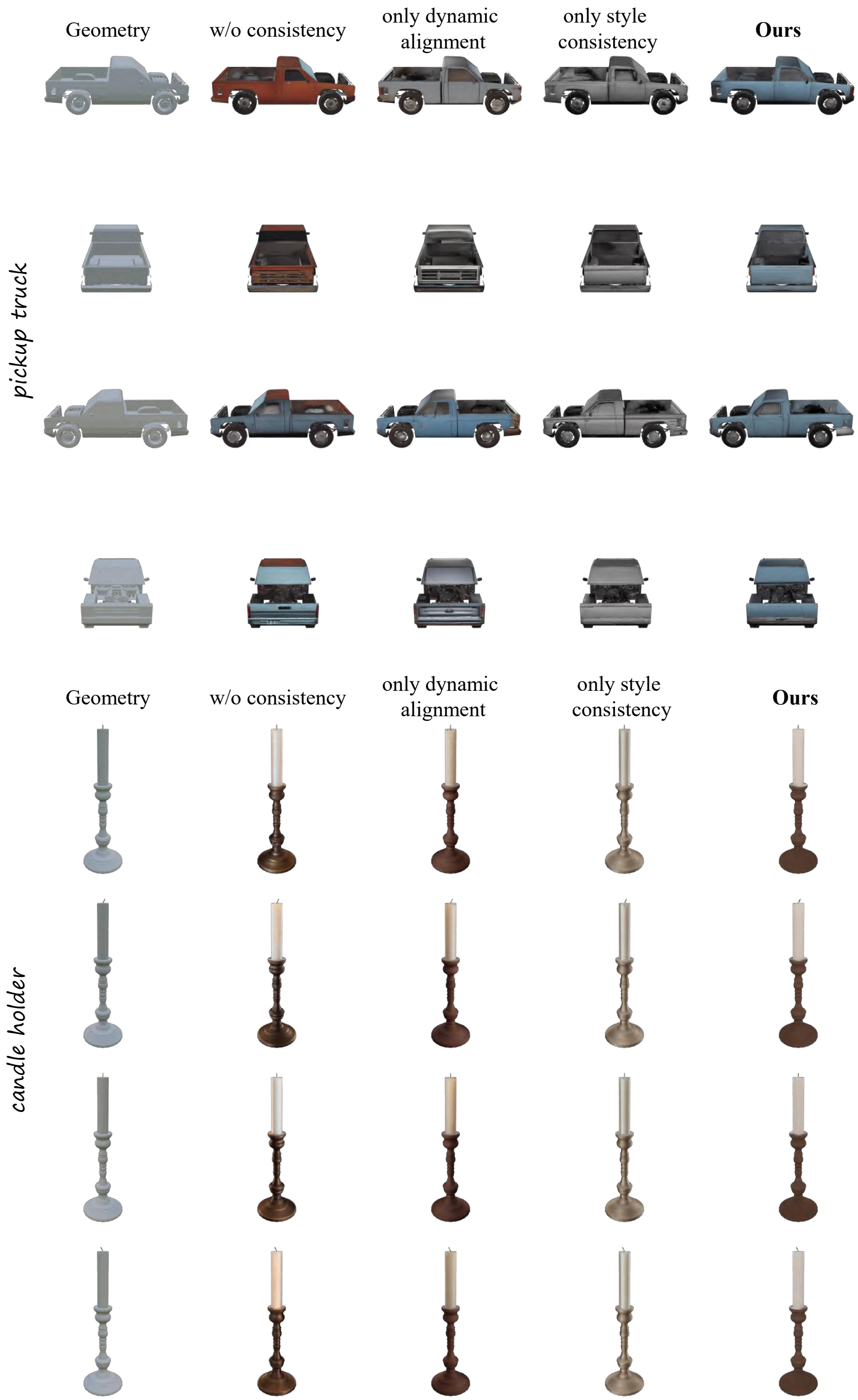

Figure 11. More ablation results I.

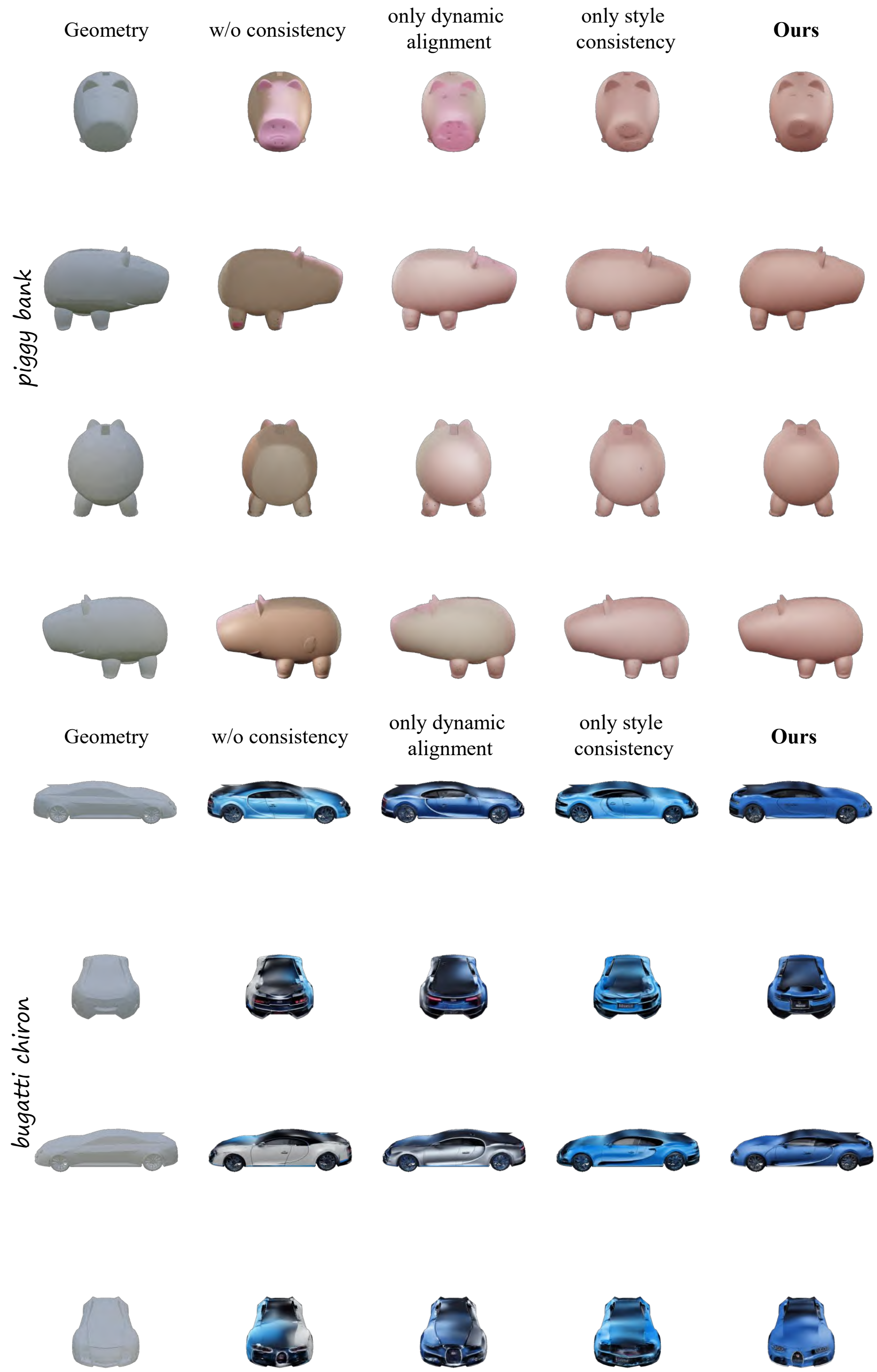

Figure 12. More ablation results II.