

MorphAny3D: Unleashing the Power of Structured Latent in 3D Morphing

Supplementary Material

This document includes the following supplementary sections:

- Initialization Details.
- Additional Experimental Details.
- Extended 3D Morphing Results.
- Additional Qualitative Experiments.
- Additional Application Details.
- Generalization Details.

1. Initialization Details

3D Inversion. For real-world assets, initial noisy latents f_{init} and image conditions c are obtained via 3D inversion, following the procedure in VoxHammer [6]. This involves two stages: Sparse Structure (SS) and Structured Latent (SLAT), which map a 3D asset to its corresponding initial latents $f_{\text{init-ss}}$ and $f_{\text{init-slat}}$. Further implementation details are provided in Sec. 3.2 of VoxHammer [6].

Initial Features Interpolation. The initial feature for frame n , f_{init}^n , is computed by spherical interpolation [12] with a deformation weight $\alpha^n \in [0, 1]$, ensuring $x^0 = x^{\text{src}} (\alpha^0 = 0)$ and $x^N = x^{\text{tgt}} (\alpha^N = 1)$. Specifically, we obtain the initial features in the SS stage of source and target assets via 3D inversion, denoted as $f_{\text{init-ss}}^{\text{src}}$ and $f_{\text{init-ss}}^{\text{tgt}}$. The initial feature in the SS stage for frame n is calculated as:

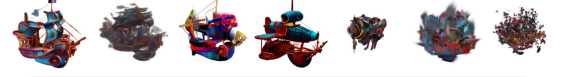
$$f_{\text{init-ss}}^n = \frac{\sin((1 - \alpha^n)\theta)}{\sin(\theta)} f_{\text{init-ss}}^{\text{src}} + \frac{\sin(\alpha^n\theta)}{\sin(\theta)} f_{\text{init-ss}}^{\text{tgt}}, \quad (1)$$

where $\theta = \arccos(\frac{f_{\text{init-ss}}^{\text{src}} \cdot f_{\text{init-ss}}^{\text{tgt}}}{\|f_{\text{init-ss}}^{\text{src}}\| \|f_{\text{init-ss}}^{\text{tgt}}\|})$. A similar approach is used for the SLAT stage, but requires finding corresponding sparse voxels based on Euclidean distance before performing spherical interpolation.

2. Additional Experimental Details

Implementation Details. Our method builds upon Image-to-3D Trellis [9], replacing all cross-attention modules with our proposed Morphing Cross-Attention (MCA) modules and self-attention modules with Temporal-Fused Self-Attention (TFSA) modules. After obtaining the structure of the n -th frame P^n , orientation correction is performed. All modules introduced require no training, allowing flexible and cost-effective application to other SLAT-based models. Both stages use a sampling step of 25 and a classifier-free guidance (CFG) guidance scale of 5.0.

Baselines Implementation Details. (1) 3DInterp and SLTAInterp: We first establish correspondence between the target and source objects using a well-known 3D shape



You are a professional reviewer specializing in digital art and 3D design. Multiple sets of 3D assets—depicting the same subject—have been generated using different methods (e.g., NeRF, 3D Gaussian Splatting, diffusion-based 3D generation, etc.).

Evaluate these results based on the following aesthetic criteria:

- (1) Overall harmony: balance, proportion, and structural coherence
- (2) Surface detail and texture: richness, realism, and material fidelity
- (3) Consistency of artistic style: unity and intentionality in visual language
- (4) Visual appeal: immediate aesthetic impact and emotional resonance

Identify the method that produces the most aesthetically pleasing asset, and justify your selection with specific observations tied to the criteria above. Focus exclusively on perceptual and artistic quality, not technical efficiency or computational cost.



The asset on the far left likely represents the output of the highest-quality. The subsequent assets (moving right) demonstrate a clear and rapid degradation of quality, which are likely artifacts or incomplete results from less stable methods:

- (1) The second asset is a blurry, amorphous blob, lacking any discernible detail or structure—a common failure state or low-iteration output.
 - (2) The third and fourth assets (the vehicles with the large wheels/engines) retain structural information but suffer from color noise, smearing, and a complete breakdown of material definition, suggesting a loss of texture fidelity.
 - (3) The final three assets are increasingly fragmented, showing a breakdown of the 3D representation into scattered points or colored clumps, demonstrating a total failure of overall harmony and consistency of artistic style.
- Therefore, the first asset on the far left is the clear winner because it is the only one that fully satisfies all four aesthetic criteria with a high degree of quality and completeness.



Figure 1. Examples of Aesthetics Score (AS) evaluation using vision-language models Gemini-2.5 [4] and ChatGPT-5 [7]. Given rendered images of morphed 3D assets, the models select the most aesthetically pleasing result across all methods.

correspondence method, DenseMatcher [13] (ICLR 2025, Spotlight). We then perform interpolation directly on the corresponding 3D-representation or SLAT-representation to generate the morphed 3D results. (2) DiffMorpher [12] and FreeMorph [2]: We use their officially released code to obtain 2D morphing results. Subsequently, we generate the final 3D assets from the morphed images using Image-to-3D Trellis [9]. (3) MorphFlow [8]: We first render multi-view images of the source and target 3D assets. Their official code is then used to generate the final morphed results.

Evaluation Metrics Details. (1) Fréchet Inception Distance (FID) [5]: For each morphed asset and its originals, we render 12 images uniformly sampled along the yaw axis and compute FID between the rendered distributions. (2) Perceptual Path Length (PPL) and Perceptual Distance Variance (PDV) [12]: For each source-target pair, we render six morphing videos (uniform yaw views) and compute PPL/PDV as in DiffMorpher [12]. (3) Aesthetics Scores (AS): Using Gemini-2.5 [4] and ChatGPT-5 [7], we prompt the models to select the most visually appealing morphed asset per pair (see Fig. 1). AS is the selection frequency (as a percentage). (4) User Preference (UP): We conducted a user study with 49 participants experienced in computer vision and graphics. They selected their preferred result

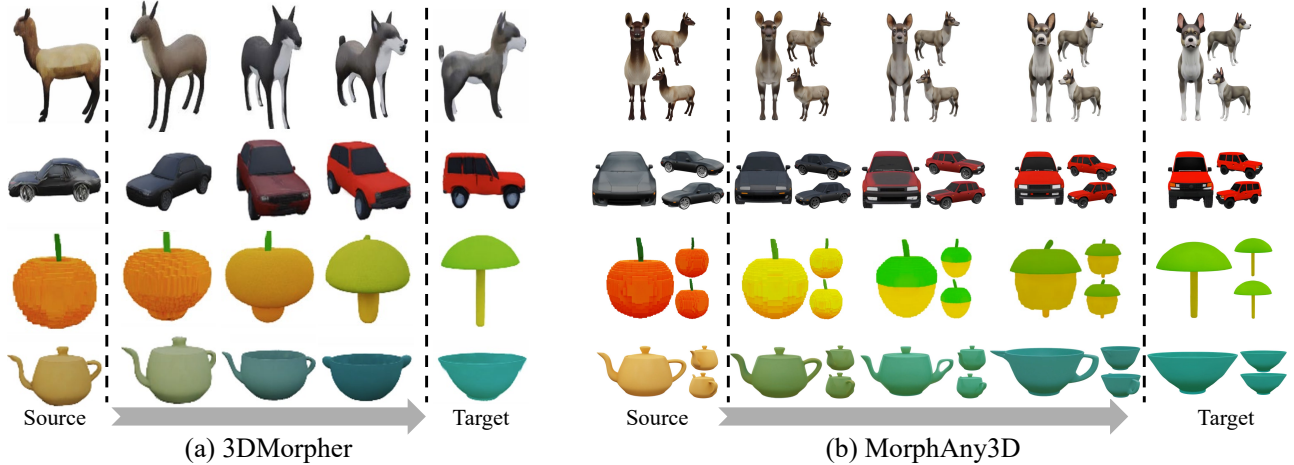


Figure 2. Qualitative comparison with 3DMorpher [10]. Results from 3DMorpher are reproduced directly from their published paper.

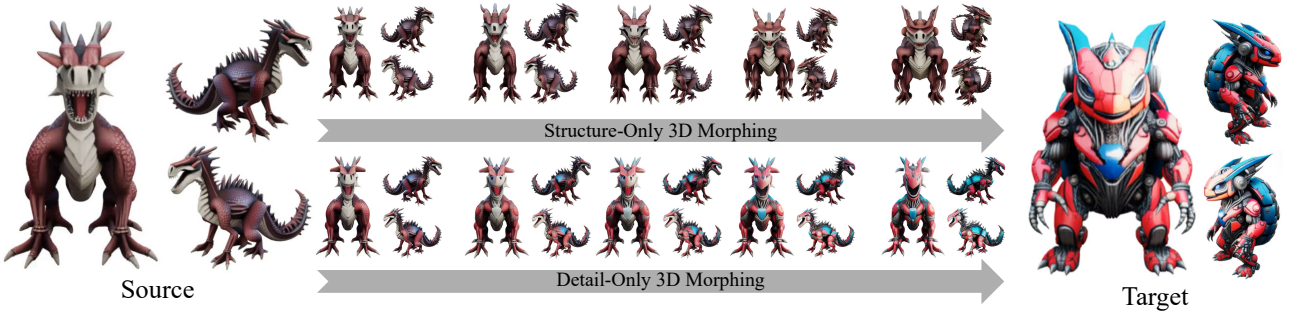


Figure 3. Disentangled 3D morphing. By applying our Morphing Cross-Attention (MCA) and Temporal-Fused Self-Attention (TFSA) modules exclusively in one stage (SS or SLAT), we independently control structural and detail deformation.

across 10 source-target pairs based on quality, smoothness, and realism (see Fig. 6). UP is reported as the selection percentage.

3. Extended 3D Morphing Results

Additional results are shown in Figs. 7 and 8. Our method produces high-quality, smooth morphs across diverse categories—including animals, vehicles, buildings, and humanoids. We visualize RGB renders from the 3DGS representation and normal maps from the mesh to highlight both appearance and geometric fidelity.

4. Additional Qualitative Experiments

Further comparisons appear in Figs. 9 and 10. Our method consistently outperforms baselines in morphing quality and temporal smoothness. Since 3DMorpher [10] has not released full code, we compare qualitatively using results reproduced from their paper (Fig. 2). We use the same source and target assets: cropped images from their figures are lifted to 3D via Image-to-3D Trellis [9]. As shown, our results exhibit superior detail and coherence. Moreover, 3DMorpher struggles with complex structures (e.g., Figs. 7 and 8) and relies solely on 3DGS, which lacks high-fidelity normals and compatibility with standard 3D software [1, 3].

5. Additional Application Details

Disentangled 3D Morphing. Thanks to Trellis’s decoupled SS and SLAT stages, our method supports disentangled morphing. As illustrated in Fig. 3:

- **Structure-only morphing:** Apply MCA/TFSA only in the SS stage to deform global shape while preserving local details.
- **Detail-only morphing:** Apply MCA/TFSA only in the SLAT stage to transfer texture and fine geometry without altering overall structure.

Dual-Target 3D Morphing. As shown in Fig. 4, distinct targets can be assigned to each stage: the SS stage adopts the source’s global shape from one target, while the SLAT stage incorporates local details from another target. This enables flexible composition of 3D assets, blending shapes and details from different sources.

3D Style Transfer. By conditioning the SLAT stage on a style reference while keeping the SS stage fixed to the source structure, our method performs 3D style transfer (Fig. 5), preserving shape while adopting the reference’s geometric and textural style.



Figure 4. Dual-target 3D morphing. Distinct target assets are assigned to the SS and SLAT stages, enabling flexible composition of global shape and local details from two different sources.

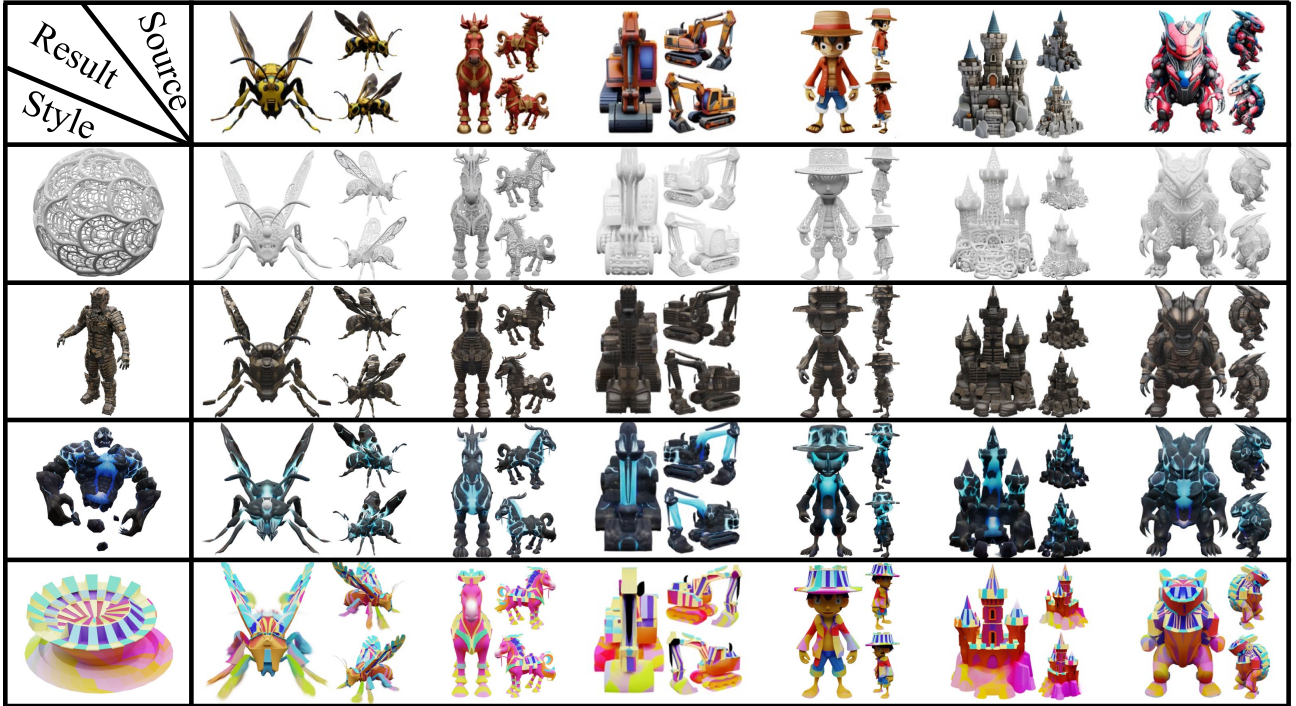


Figure 5. 3D style transfer. A style reference guides the SLAT stage while the SS stage retains the source structure, allowing texture and geometric style to be transferred without altering the underlying shape.

6. Generalization Details

ommend image-guided SLAT models when possible.

Hi3DGen [11]—a recent SLAT-based method for high-fidelity 3D generation—uses the same Trellis framework and image conditioning. Thus, our approach applies directly without retraining. Since Hi3DGen focuses on geometry (no texture), we report only normal maps.

For Text-to-3D Trellis [9], we replace image conditions in MCA with textual prompts. No other changes are needed, enabling zero-shot adaptation. However, textual conditions are more abstract than visual ones, leading to reduced deformation quality and temporal coherence. We therefore rec-

References

- [1] Inc. Autodesk. Autodesk maya - 3d animation and modeling software. <https://www.autodesk.com/products/maya>, 2024. 2
- [2] Yukang Cao, Chenyang Si, Jinghao Wang, and Ziwei Liu. Freemorph: Tuning-free generalized image morphing with diffusion model. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2025. 1
- [3] Blender Foundation. Blender - a 3d modelling and rendering software. <https://www.blender.org/>, 2024. 2
- [4] Google DeepMind. Gemini: A family of multimodal large language models, 2024. 1
- [5] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Adv. Neural Inf. Process. Syst.*, 2017. 1
- [6] Lin Li, Zehuan Huang, Haoran Feng, Gengxiong Zhuang, Rui Chen, Chunchao Guo, and Lu Sheng. Voxhammer: Training-free precise and coherent 3d editing in native 3d space. *arXiv preprint arXiv:2508.19247*, 2025. 1
- [7] OpenAI. ChatGPT: Optimizing language models for dialogue, 2023. 1
- [8] Chih-Jung Tsai, Cheng Sun, and Hwann-Tzong Chen. Multiview regenerative morphing with dual flows. In *Proc. Eur. Conf. Comput. Vis.*, 2022. 1
- [9] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2025. 1, 2, 3
- [10] Songlin Yang, Yushi Lan, Honghua Chen, and Xingang Pan. Textured 3d regenerative morphing with 3d diffusion prior. *arXiv preprint arXiv:2502.14316*, 2025. 2
- [11] Chongjie Ye, Yushuang Wu, Ziteng Lu, Jiahao Chang, Xiaoyang Guo, Jiaqing Zhou, Hao Zhao, and Xiaoguang Han. Hi3dgen: High-fidelity 3d geometry generation from images via normal bridging. In *Proc. IEEE Int. Conf. Comput. Vis.*, 2025. 3
- [12] Kaiwen Zhang, Yifan Zhou, Xudong Xu, Bo Dai, and Xingang Pan. Diffmorpher: Unleashing the capability of diffusion models for image morphing. In *Proc. IEEE Conf. Comput. Vis. Pattern Recogn.*, 2024. 1
- [13] Junzhe Zhu, Yuanchen Ju, Junyi Zhang, Muhan Wang, Zhecheng Yuan, Kaizhe Hu, and Huazhe Xu. Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo. *arXiv preprint arXiv:2412.05268*, 2024. 1

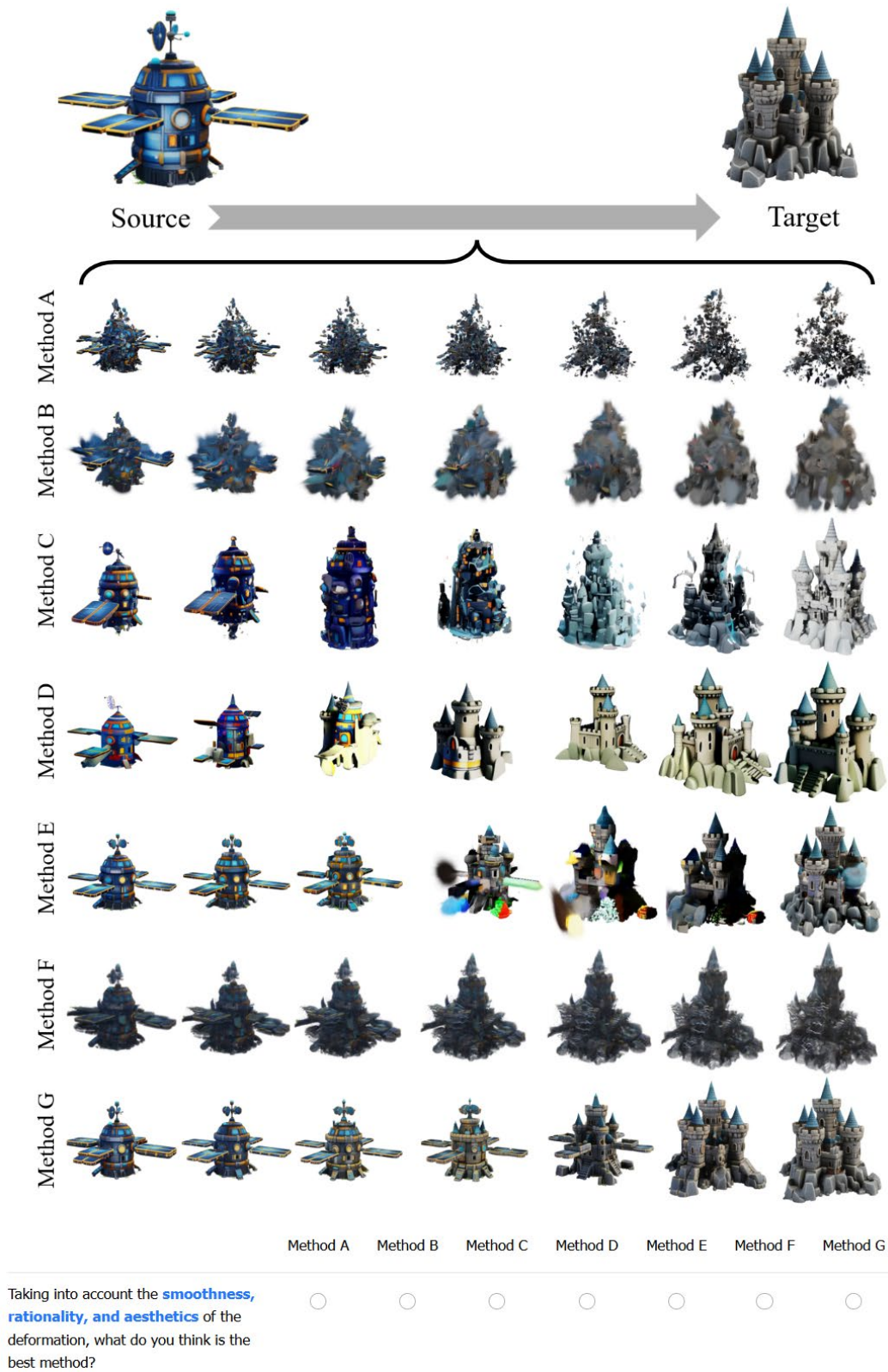


Figure 6. User Preference (UP) evaluation. Participants selected their preferred morphed 3D asset from all methods based on rendered images, judging quality, smoothness, and realism.



Figure 7. Extended 3D morphing results across diverse object categories (e.g., animals, vehicles, buildings). Our method produces high-fidelity and temporally smooth transitions.



Figure 8. Additional extended results demonstrating consistent morphing quality across varied and challenging object categories.



Figure 9. Qualitative comparison with baseline methods. Our approach achieves superior morphing quality and smoother transitions.

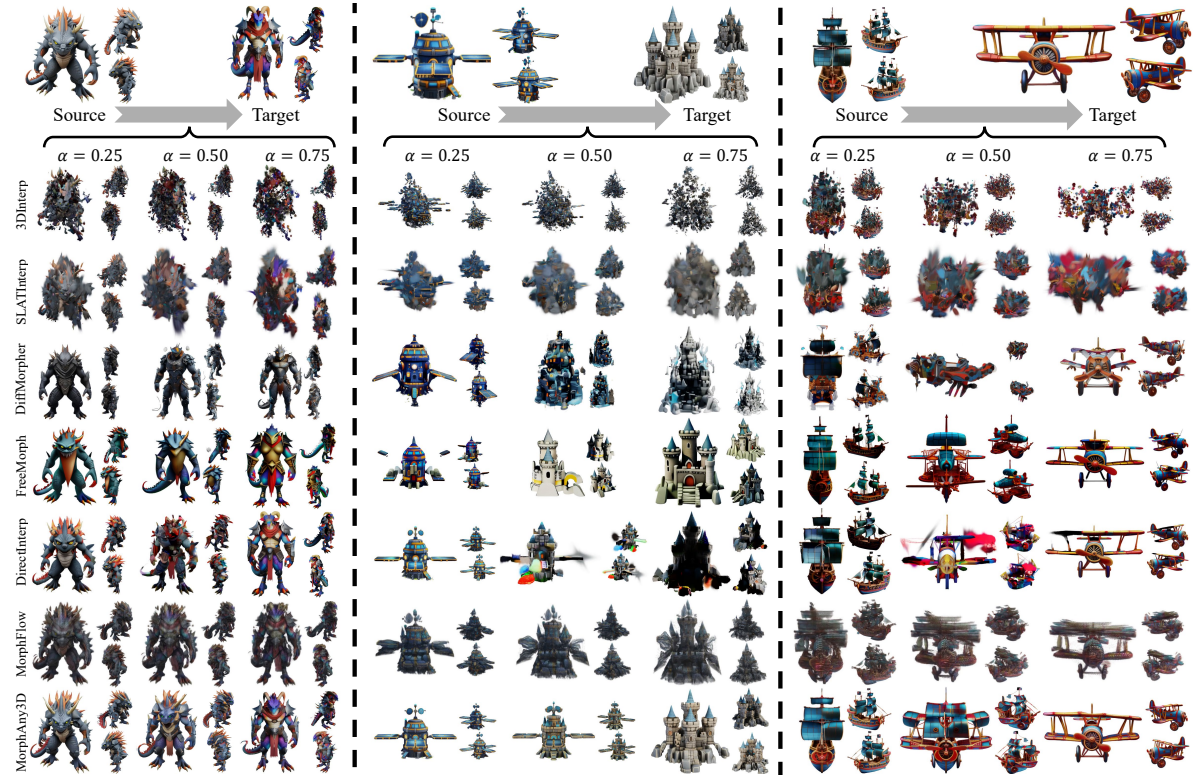


Figure 10. Further qualitative comparisons highlighting the advantages of our method over existing approaches in preserving geometry and ensuring temporal coherence.