Efficient Dynamic Scene Editing via 4D Gaussian-based Static-Dynamic Separation

Supplementary Material

7. Additional Qualitative Results

7.1. Results on Monocular Datasets

While 4D dynamic scene editing typically relies on multi-view video datasets to sufficiently capture spatio-temporal information, we evaluate our method on the DyCheck [16] and HyperNeRF [41] datasets to explore its potential applicability to monocular video inputs. For these monocular datasets, we cannot obtain edited multiview supervision images for editing canonical 3D Gaussians. Therefore, we skip Stage 1 (described in Sec. 4.2) and only apply Stage 2, the score-based temporal refinement (described in Sec. 4.3). Figure 9 presents a comparison between Instruct 4D-to-4D (*baseline*) and Instruct-4DGS (*ours*) on the DyCheck dataset, while Fig. 10 shows qualitative results of our method on the HyperNeRF dataset. Our Instruct-4DGS produces plausible dynamic scene editing results even on monocular datasets, and we expect the performance to further improve as techniques for reconstructing 4D Gaussians from monocular videos and editing with the SDS mechanism continue to advance.

7.2. Results with Varying Camera Poses

To further assess the spatial consistency and quality of our edited 4D Gaussian representations, we render the edited dynamic scenes from the DyNeRF [28] dataset under various camera poses. As shown in Fig. 11, the results produced by our Instruct-4DGS maintain plausible geometry and appearance across different viewpoints.

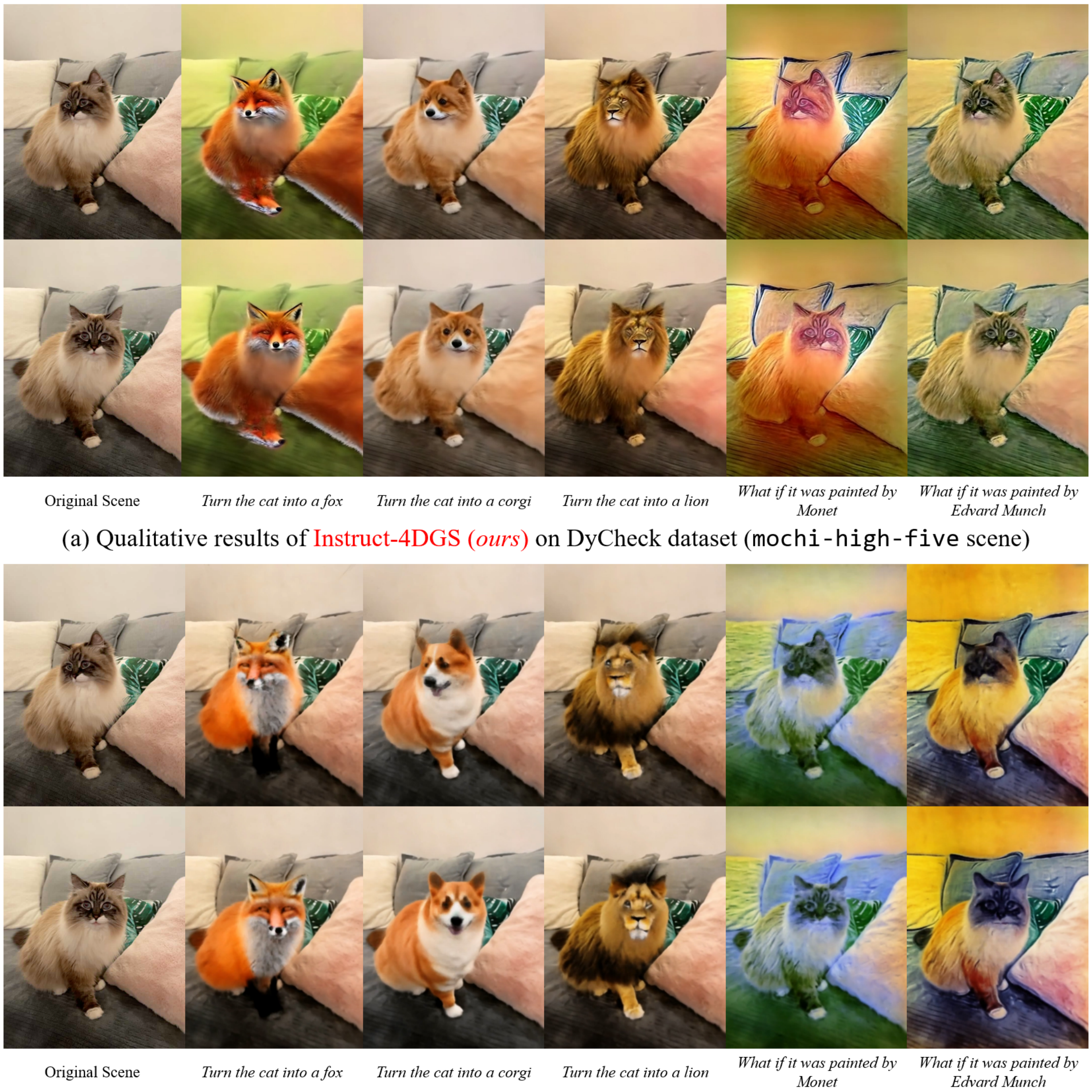
8. Full Set of Editing Instructions

Here, we provide the full set of editing instructions used for our dynamic scene editing experiments.

We used *“Make the person a statue”*, *“Make the person a marble Roman sculpture”*, and *“Make the person a wood sculpture”* for Tab. 1.

We used *“What if it was painted by {Makoto Shinkai, Henri Matisse, Utagawa Hiroshige, Van Gogh, Edvard Munch}?”*, *“Make it a Fauvism painting”*, *“Make the person a statue”*, *“Make the person a marble Roman sculpture”*, and *“Make the person a wood sculpture”* for Fig. 5.

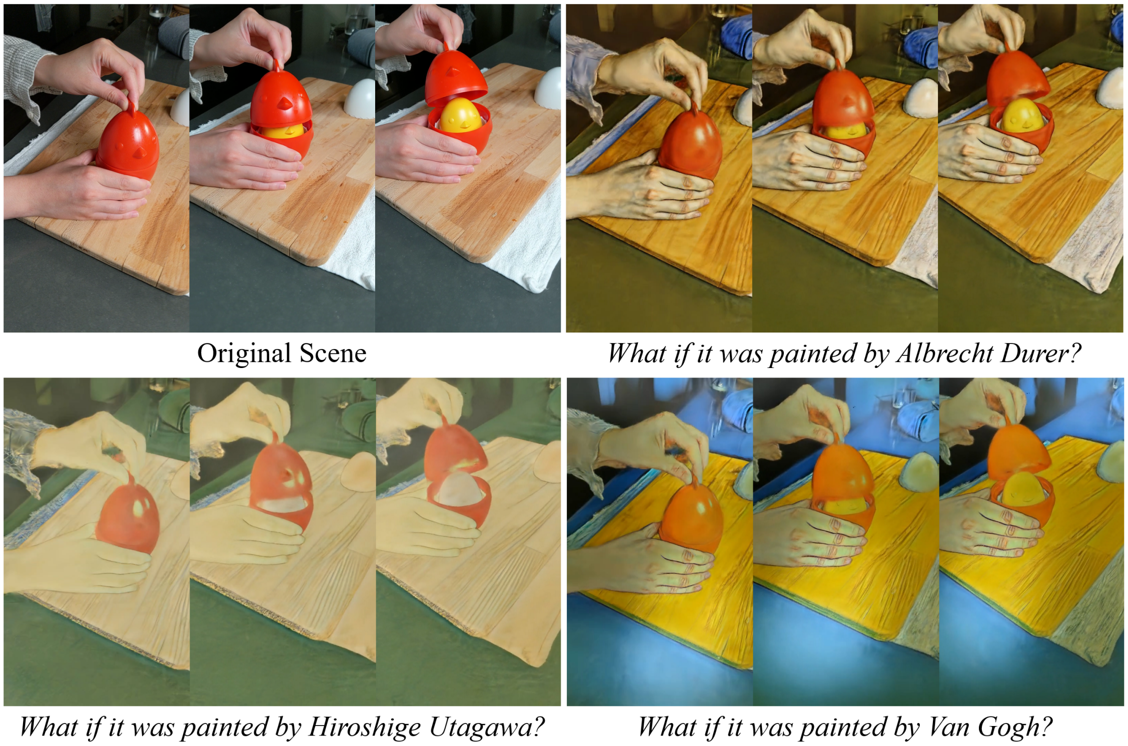
We used *“Make the person a marble Roman sculpture”*, *“What if it was painted by {Van Gogh, Edvard Munch}?”*, *“Make it a Fauvism painting”*, *“Make this a cozy wooden cabin bar with soft lighting and rustic decorations”*, *“Turn the man into a bronze sculpture”*, *“Add a beautiful sunset”*, *“Make it underwater”*, *“Give him a Victorian gentleman’s attire”*, *“Make it a Fauvism painting”*, and *“What if it was painted by Van Gogh?”* for Fig. 6


Original Scene
Turn the cat into a fox
Turn the cat into a corgi
Turn the cat into a lion
What if it was painted by Monet
What if it was painted by Edvard Munch
Original Scene
Turn the cat into a fox
Turn the cat into a corgi
Turn the cat into a lion
What if it was painted by Monet
What if it was painted by Edvard Munch

(a) Qualitative results of Instruct-4DGS (*ours*) on DyCheck dataset (`mochi-high-five` scene)

(b) Qualitative results of Instruct 4D-to-4D (*baseline*) on DyCheck dataset (`mochi-high-five` scene)

Figure 9. **Qualitative comparison of visual quality on the DyCheck [16] dataset (a *monocular* dataset)**: We compare our method, Instruct-4DGS (*ours*), with the baseline, Instruct 4D-to-4D [38] (*baseline*), on the *mochi-high-five* scene from the DyCheck dataset.



Original Scene

What if it was painted by Albrecht Durer?

What if it was painted by Hiroshige Utagawa?

What if it was painted by Van Gogh?

Figure 10. **Qualitative results of our Instruct-4DGS on the HyperNeRF [41] dataset (a *monocular* dataset)**: We evaluate our method on the *Interp_chickchicken* scene from the HyperNeRF dataset.

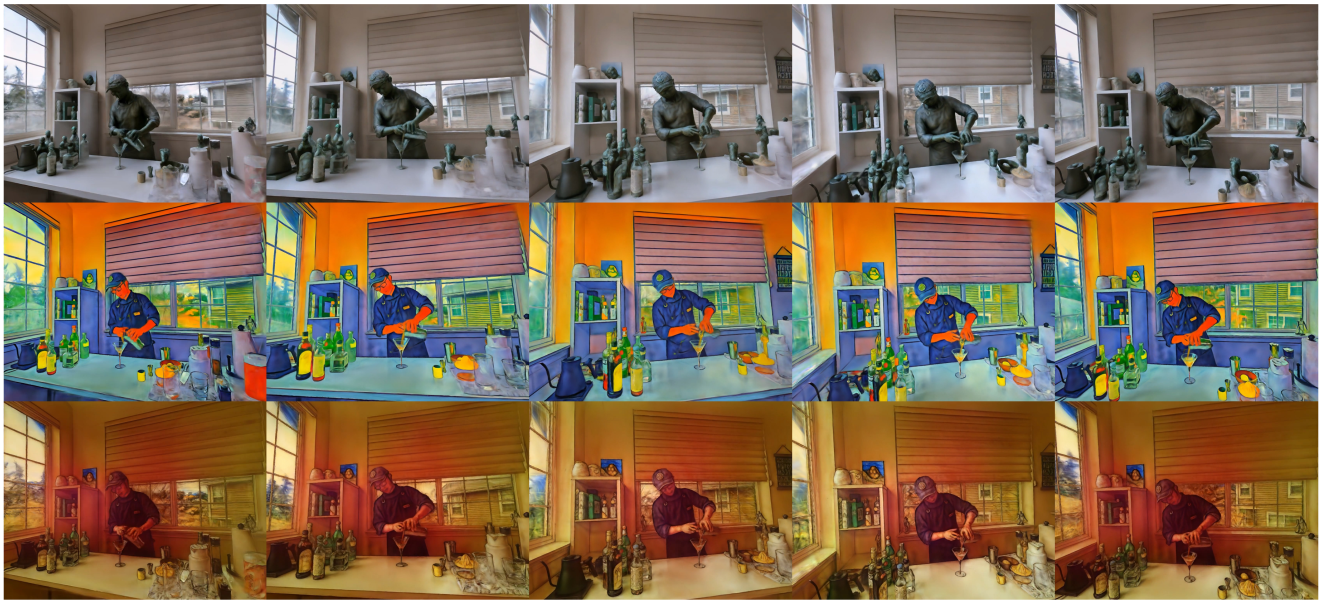


Figure 11. **Qualitative results of our Instruct-4DGS under various camera poses on the DyNeRF [28] dataset**: We render the edited dynamic scene from novel camera poses to evaluate the spatial consistency of our method. Our Instruct-4DGS produces view-consistent and geometrically plausible results.