GaussCtrl: Multi-View Consistent Text-Driven 3D Gaussian Splatting Editing
–*Supplementary Materials*–

Jing Wu*[1], Jia-Wang Bian*[2], Xinghui Li[1], Guangrun Wang[1], Ian Reid[2], Philip Torr[1], and Victor Adrian Prisacariu[1]

[1] University of Oxford
[2] Mohamed bin Zayed University of Artificial Intelligence
{jing.wu, philip.torr}@eng.ox.ac.uk, {jiawang.bian, ian.reid}@mbzuai.ac.ae, {xinghui, victor}@robots.ox.ac.uk, wanggrun@gmail.com

1 More Discussions

1.1 Depth Accuracy

Some may be concerned that the depth rendered from the 3DGS may be inaccurate due to issues like holes or complex geometries. However, our method is robust to this issue because ControlNet tolerates depth inaccuracy. As shown in [1], ControlNet performs well when using monocular depth, which is both inaccurate and multi-view inconsistent.

1.2 Attention Alignment Discussion

Intuitively, we consider that our cross-view attention aims to find the best parameters that enable the use of reference frames z_j as the basis "vector" to represent other frames z_i. In this way, it aligns the appearance and structure of the object and background of all the views to reference views, encouraging multi-view consistency. However, only performing cross-view attention would make each image lose its feature. When the difference between the sampled reference views and the current view is too large, it might exceedingly fit the reference views and degrade the editing results. Thus, we include self-attention to align the current view to itself, which makes the editing process more stable.

2 Qualitative Comparisons with Baseline Methods

We provide more qualitative comparisons between our method and baseline methods:

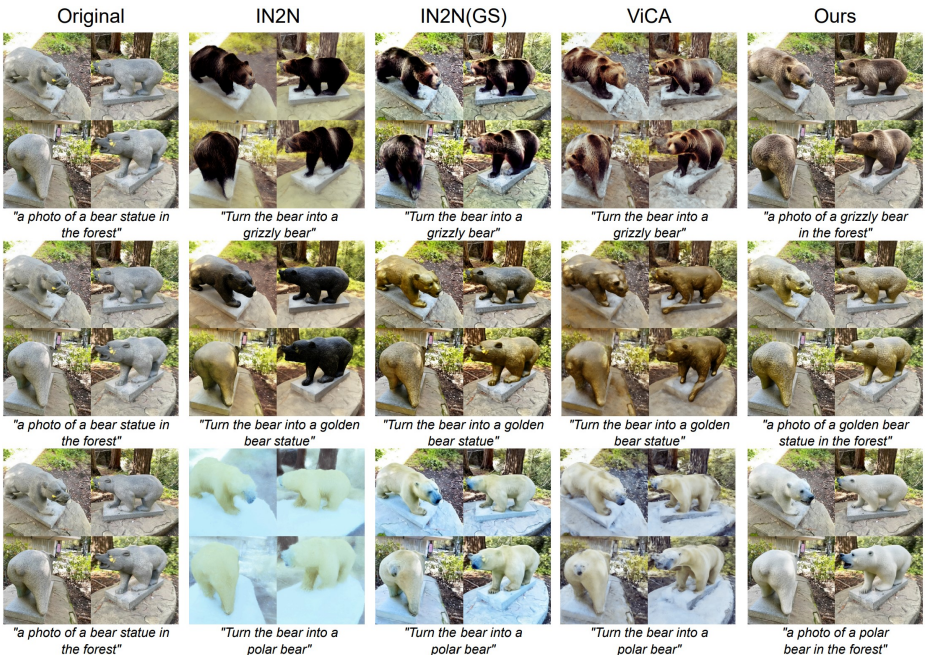

Fig. 1: Qualitative Results: 360-degree Scene (Bear Statue)

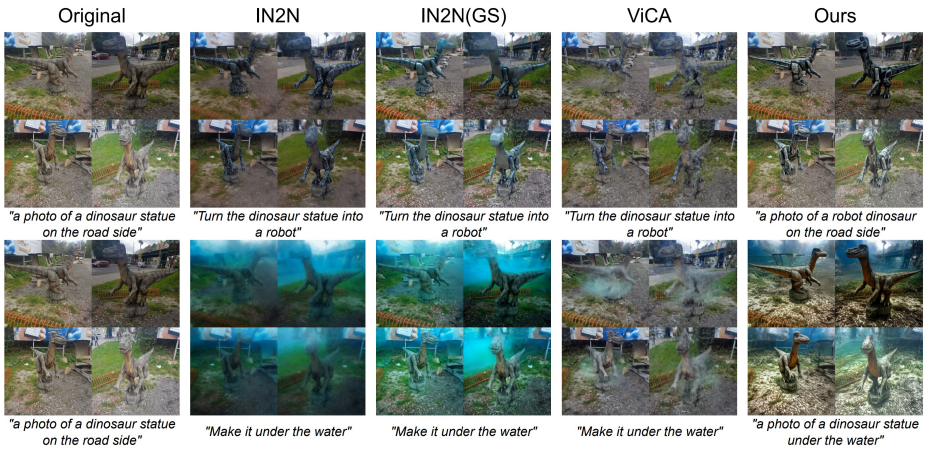

Fig. 2: Qualitative Results: 360-degree Scene (Dinosaur)

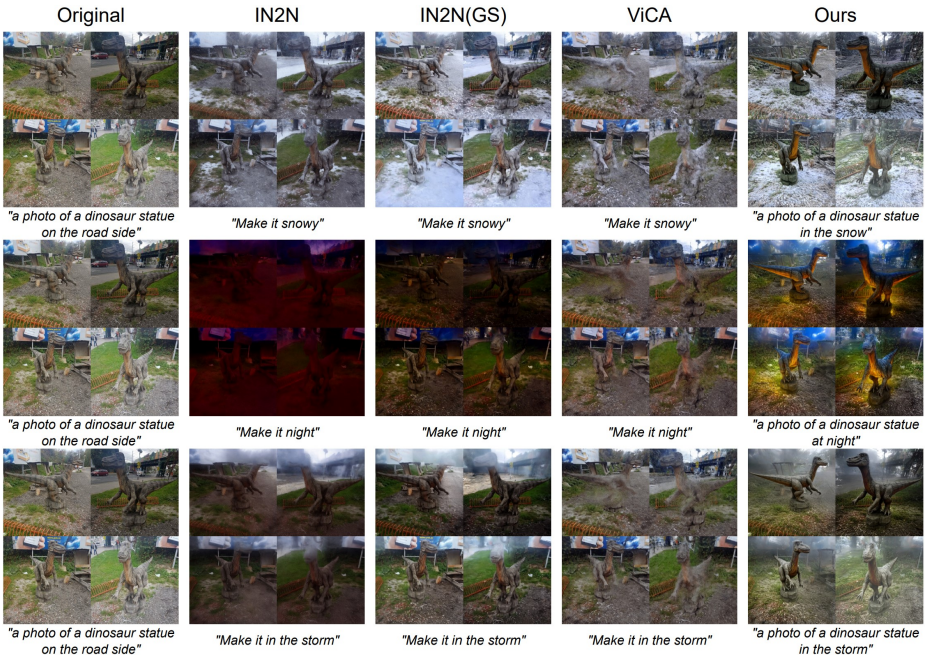

Fig. 3: Qualitative Results: 360-degree Scene (Dinosaur)

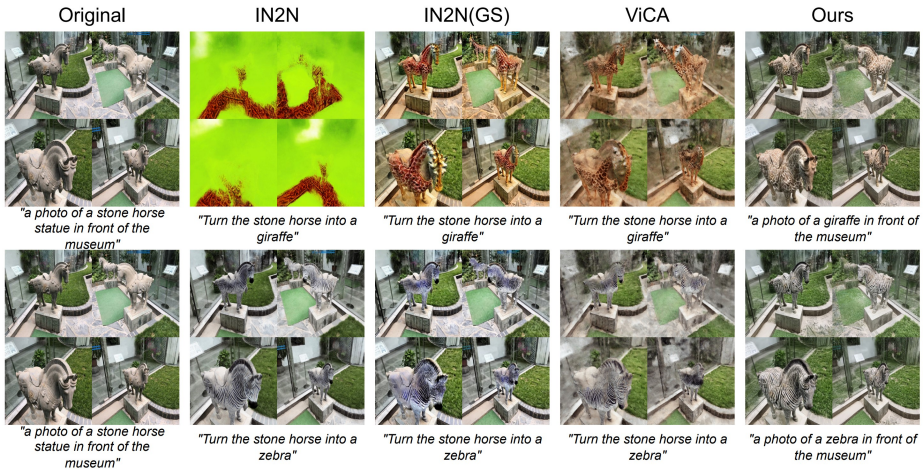

Fig. 4: Qualitative Results: 360-degree Scene (Stone Horse)

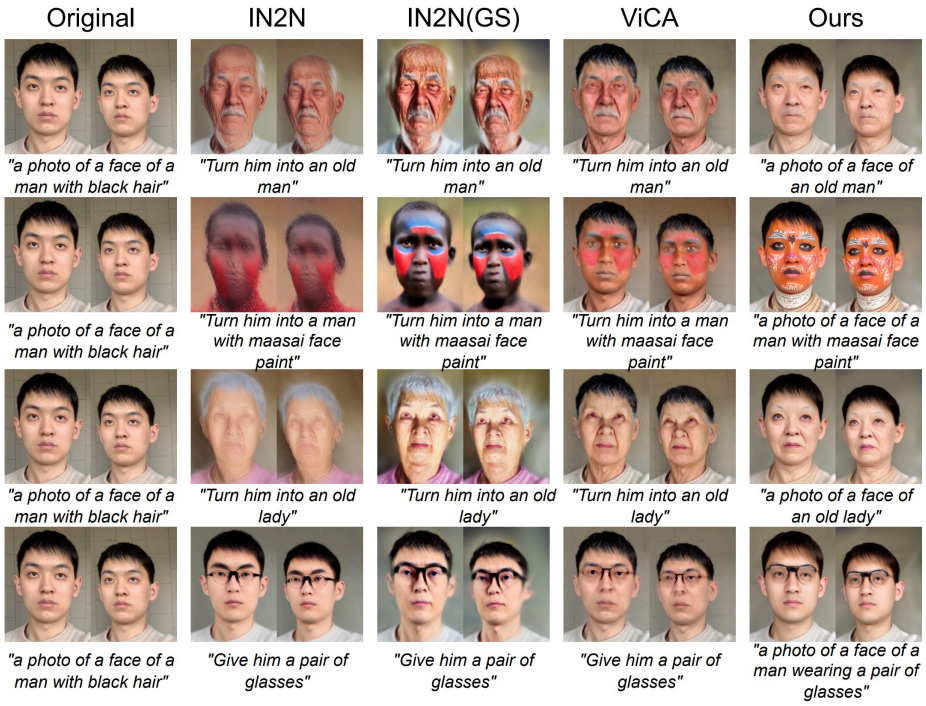

Fig. 5: Qualitative Results: Forward-facing Scene (Fangzhou)

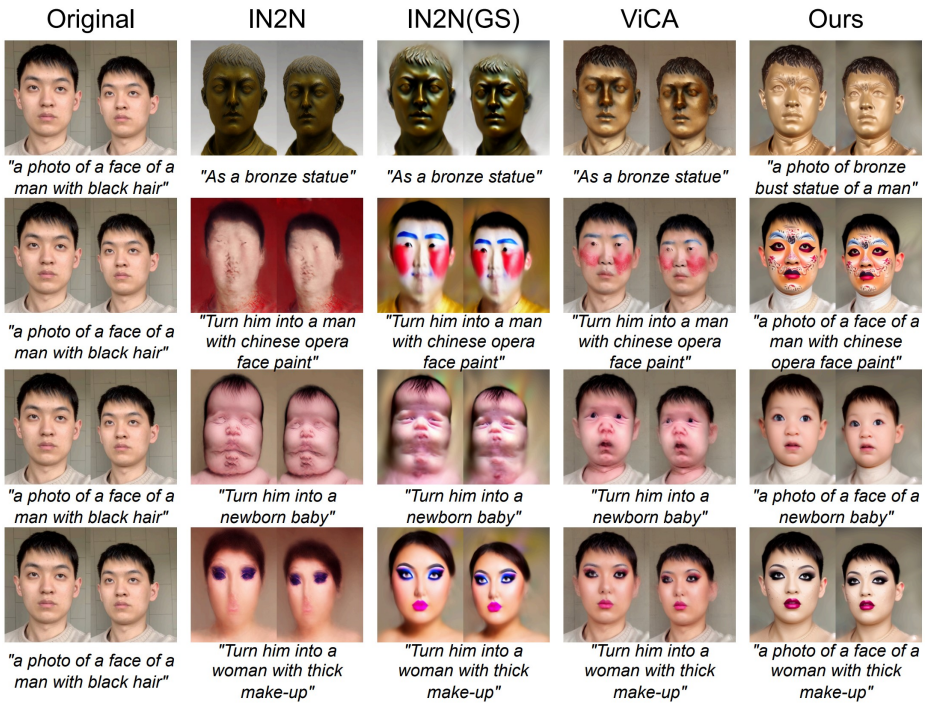

Fig. 6: Qualitative Results: Forward-facing Scene (Fangzhou)

Fig. 7: Qualitative Results: Forward-facing Scene (Face)

Fig. 8: Qualitative Results: Forward-facing Scene (Face)

References

1. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models (2023) 1