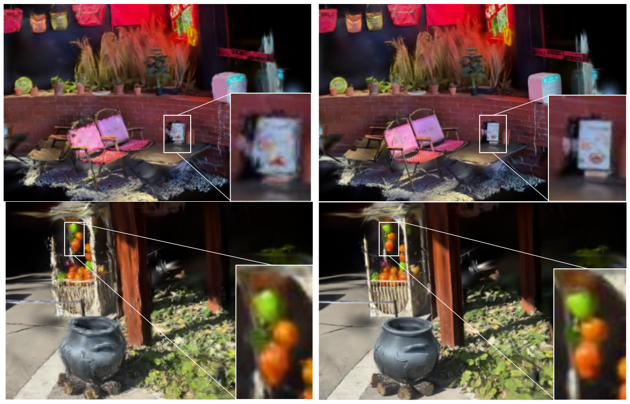
Appendix

In this supplementary material, we demonstrate the qualitative ablation study of our Can3Tok-based VAE model; more qualitative results; image-to-3DGS application and more discussions.



Without (left) and with (right) nearest voxel coordinates appending

Figure 1. Ablation study for with and without nearest voxel coordinate appending with each input 3DGS. The results indicate that appending structured volume coordinates to the unstructured input 3D Gaussians leads to better reconstruction.

1. Ablation Studies

We perform ablation study for the importance of each module of our method. In Tab. 2 of the main paper, we describe the overall quantitative comparison over different ablation studies. More specifically, we verify the performance by removing each of the following modules:
1) *w/o Learnable Query*: We remove the learnable canonical latent query and we replace the cross-attention block with self-attention. We observe that simply replacing it with self-attention fails to converge and is more likely to encounter out-of-memory issue, even with a batch size of one, although we are using a GPU with a large memory capacity (*e.g.,* 80GB). In fact, this highlights the importance of our Can3Tok module for its computational efficiency. Fig. 5 highlights that it was not possible to make a structured latent space to recover the original inputs without cross-attention with a low-dimensional learnable query.
2) *w/o normalization*: we do not apply the normalization of data to the entire 3DGS training dataset. Both Tab.2 and Fig. 6 highlight the severe scale inconsistency issue if a VAE model is trained on raw 3DGS input, which hinders scaling up training across thousands of scenes. A uniform data normalization strategy is essential to allow large-scale training and improve generalization.
3) *w/o data filtering*: we use raw 3DGS reconstruction results as a training set without data filtering. Fig. 8 implies that by suppressing the significant noise by data filtering, the models better learn the mapping between the latent and inputs in a way that preserves the local details.
4) *w/o voxel coordinate appending*: we turn off the dual positional embedding from 3DGS's position and its nearest voxel center. Instead, we append the positional embedding only from 3DGS's position. Fig. 7 show the effect of voxel coordinate appending where its to preserve the high-frequency local details in the reconstructed 3DGS.
5) *w/o voxel data enhancement*: we disable data enhancement during training.

2. More Results

In Fig. 9, we demonstrate more results from our Can3Tok with various test scenes.

3. Application: Image-to-3DGS

In this section, we showcase the application of our Can3Tok latent space modeling. Other than text-to-3DGS, our latent features can be used for image-to-3DGS generation. To this end, we use an image encoder (*e.g.,* [1]) that takes as input 2D images and outputs corresponding latents; and our pretrained Can3Tok decoder constructs the associated 3DGS scene. The pipeline is shown in Fig. 3. The objective of the encoder training is to minimize the L_2 error between the predicted latents $\mathbf{z}$ and "Ground-Truth 3D Gaussian latents" $\mathbf{z}_{GT}$, which are obtained by inputting 3D Gaussians into the Can3Tok encoder. The reason we use a regression objective instead of a diffusion objective is that almost all text-conditioned generations including text-to-image [3] and text-to-(3D object) [7] use diffusion objective due to their probabilistic nature. While image-to-3D is inherently more deterministic than text-to-3D, given that methods like Flash3D [4] and SplatterImage [5] both use regression objectives. Therefore, we follow the similar idea: regress $\mathbf{z}$ from an input image and then predict 3DGS from $\mathbf{z}$. We showcase some qualitative examples of this Image-to-3DGS applications in Fig. 2.

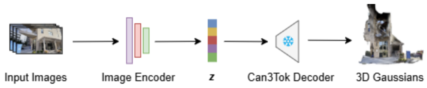


Figure 2. Illustration of image-to-3DGS architecture.

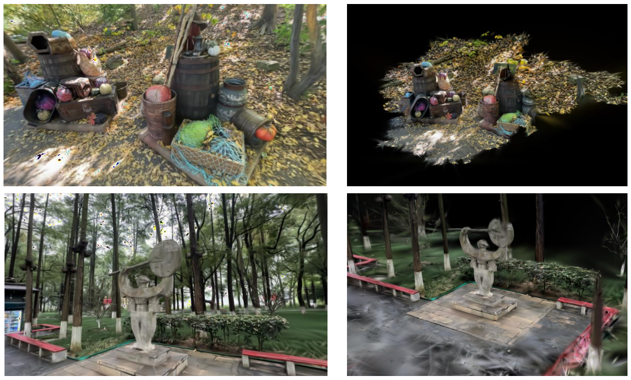


Figure 3. Qualitative examples of our image-to-3DGS applications.

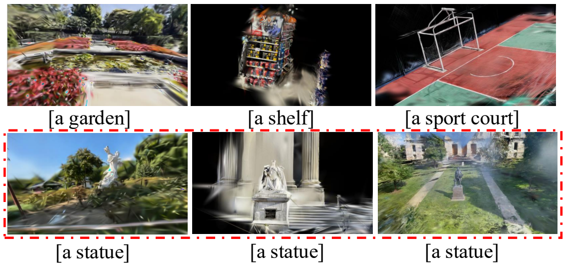


Figure 4. More generative results of our method with corresponding text conditions. Each prompt is intentionally brief and was not seen in the exact form during training to avoid bias toward any specific scene. The results highlight both inter-class and intra-class diversity, with the latter emphasized in red rectangles.

4. More Discussions

About speed, our Can3Tok VAE (1.1 s/iters) is 10 times faster than L3DG (11.3 s/iters). This is because our method accelerate self-attention steps by reducing the input dimension with a latent query, while 3D convolution step itself is slower than our self-attention even with Minkowski Engine. Moreover, 3D CNN requires non-bachify CPU-based voxel ID assignment for each Gaussian primitive, making 3D CNN even slower. We also evaluated our method on text-to-(3D scene) with FID metric, which is 28.32 calculated on rendered views randomly sampled over the unit sphere around 3D scenes, while PointTransformer achieves 153.76. Although we showcase 3DGS generation for general scenes as an application, the main focus of our paper is about 3D tokenization and latent modeling of scene-level 3D Gaussians. Accordingly, our emphasis is directed towards latent analysis similar to image-to-latent analysis [6] and the visualization of VAE reconstructions but on unseen inputs, similar to experimental results in Fig.4 of PerceiverIO [2].

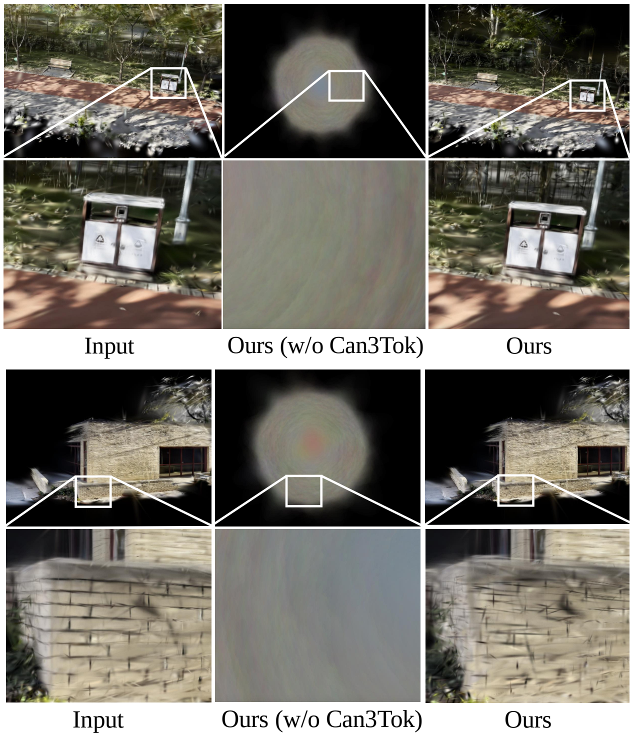


Figure 5. Qualitative comparisons of w/ and w/o Can3Tok.

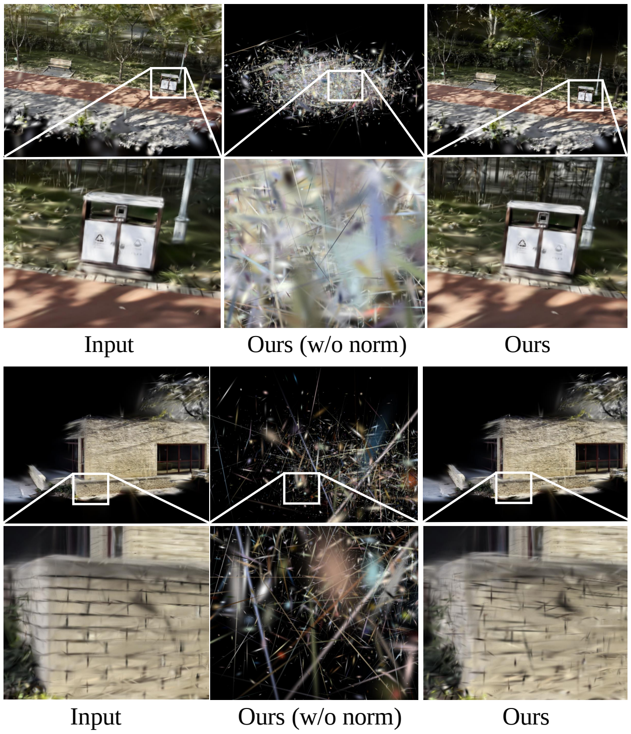


Figure 6. Qualitative comparisons w/ and w/o normalization.

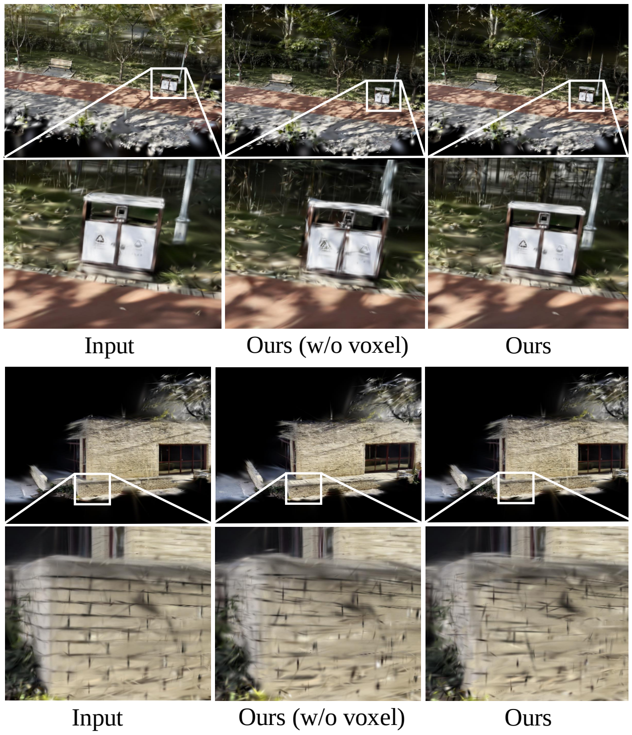


Figure 7. Qualitative comparisons of w/ and w/o voxel coordinate appending.

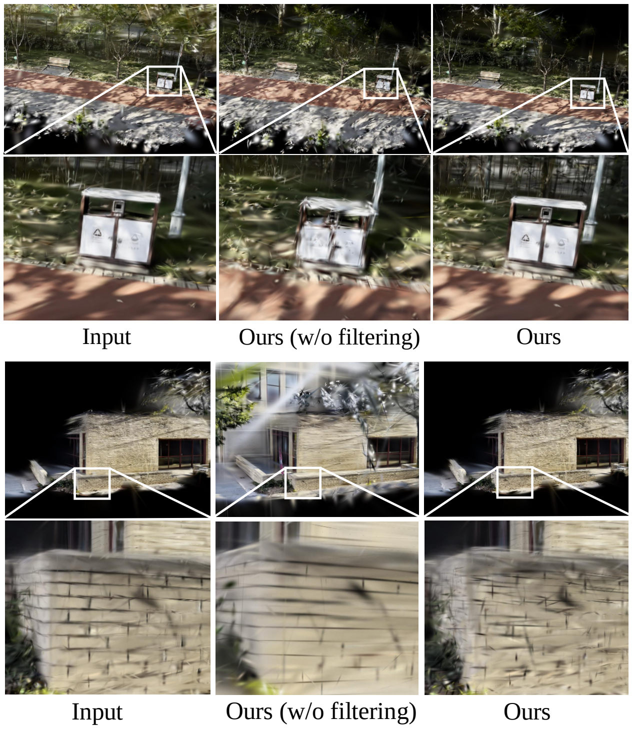


Figure 8. Qualitative comparisons of w/o semantic-aware 3DGS filtering.

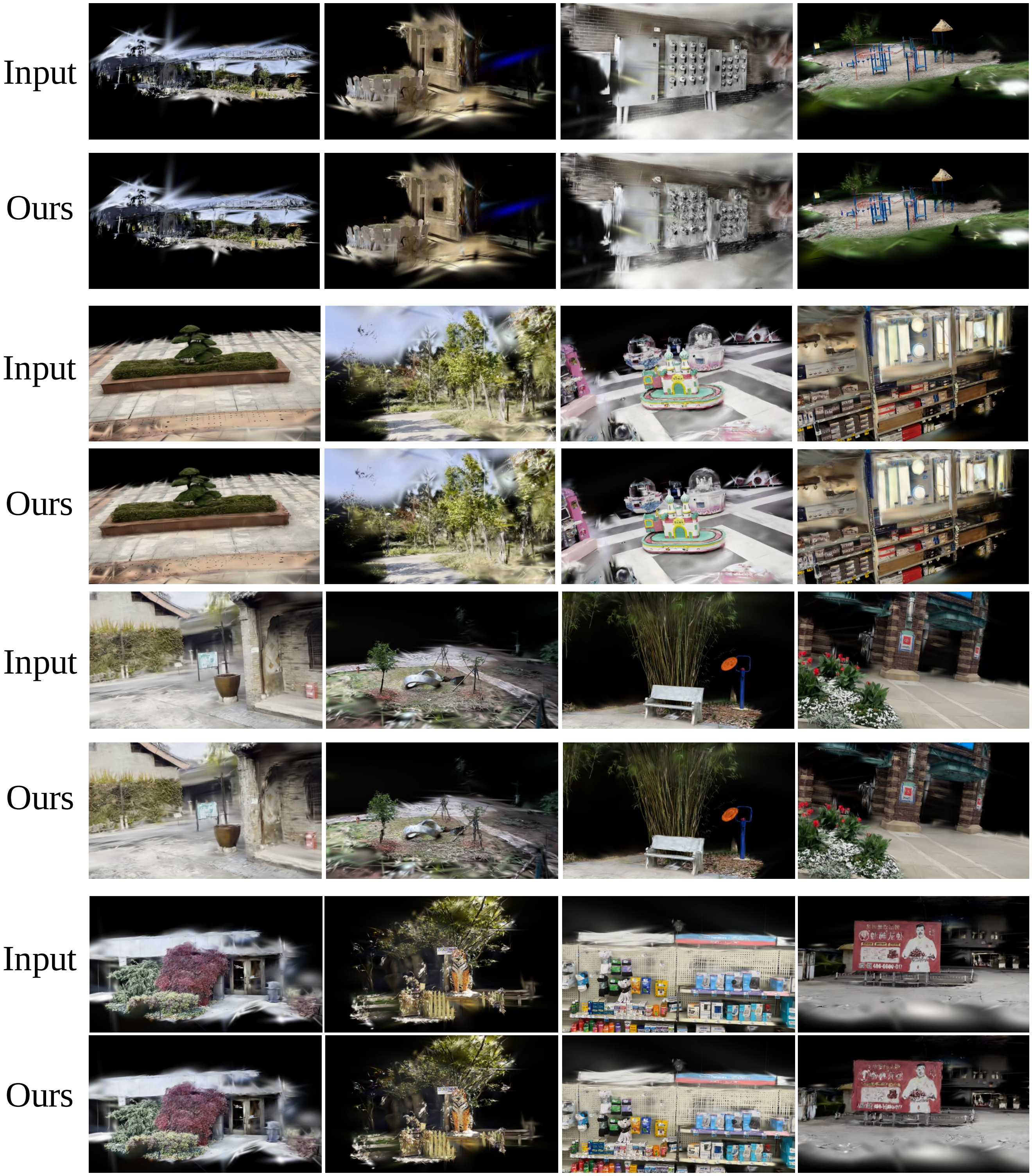


Figure 9. More qualitative results.

References

[1] Clément Chadebec, Louis Vincent, and Stephanie Allassonniere. Pythae: Unifying generative autoencoders in python - a benchmarking use case. In *Advances in Neural Information Processing Systems*, pages 21575–21589. Curran Associates, Inc., 2022. 1

[2] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 2

[3] 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. 1

[4] Stanislaw Szymanowicz, Eldar Insafutdinov, Chuanxia Zheng, Dylan Campbell, Joao F Henriques, Christian Rupprecht, and Andrea Vedaldi. Flash3d: Feed-forward generalisable 3d scene reconstruction from a single image. *arXiv preprint arXiv:2406.04343*, 2024. 1

[5] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10208–10217, 2024. 1

[6] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 2

[7] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, Bin Fu, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. *Advances in Neural Information Processing Systems*, 36, 2024. 1