

DiT360: High-Fidelity Panoramic Image Generation via Hybrid Training

Supplementary Material

6. Effect of Supervision on Polar Distortions

In this section, we further illustrate the effect of cube loss in addressing severe distortions around the polar regions. Figure 6 compares results generated from the same prompt without and with this supervision, showing that incorporating cube loss leads to clearer structures and fewer artifacts in the polar regions.

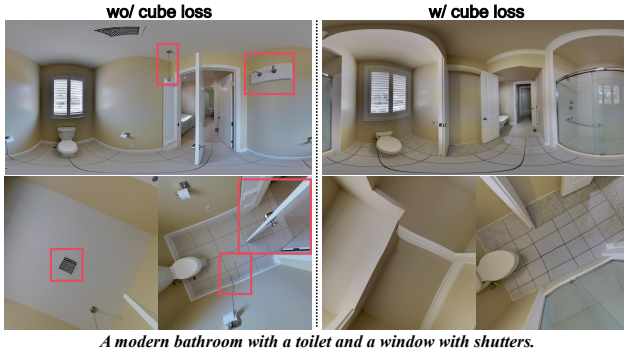


Figure 6. Qualitative comparison of generated panoramas and their top/bottom cube faces without (left) and with (right) cube loss. Red boxes mark regions where polar artifacts are significantly reduced when supervision is applied.

7. Inpainting and Outpainting

DiT360 demonstrates native inpainting and outpainting capabilities without requiring additional training, thereby establishing a unified framework for panoramic image generation, as illustrated in Fig. 7. Specifically, inspired by previous work [3], given a single perspective image, we first project it into the panoramic domain from a chosen viewpoint and derive the corresponding mask that specifies the regions requiring inpainting or outpainting. We then perform inversion on the projected image to obtain its initial noise representation while simultaneously extracting reference tokens without positional encodings. During subsequent inference, we adopt a token replacement strategy: tokens within the masked regions are replaced with those from the reference image, while retaining the positional encodings originally associated with the replaced tokens. This replacement scheme faithfully preserves subject details and maintains spatial consistency, effectively anchoring subject identity at the beginning of generation and naturally guiding the model toward coherent content completion. As a result, *DiT360* is able to produce consistent and semantically rich results in both inpainting and outpainting tasks. More results are provided in Fig. 8.

8. Experiment Settings

Implementation Details. We developed *DiT360* on top of Flux.1-dev [1], integrating LoRA [6] into the attention layers. The model was fine-tuned on 8 H20 GPUs using AdamW [7] with a learning rate of 2×10^{-5} for 20 epochs, a batch size per GPU of 1, and a gradient accumulation of 3, while only the LoRA modules were trainable. Our experiments revealed that the guidance scale plays a crucial role in convergence, with 1.0 yielding the most stable training. For inference, we set the guidance scale to 3.0 and employed 28 sampling steps.

Dataset. We adopt a hybrid training strategy that combines perspective and panoramic data. For the perspective branch, we curate 40k high-quality landscape images from the Internet, center-crop them to a 1:1 ratio, and project them onto random panoramic regions. For the panoramic branch, we follow PanFusion [20] and utilize Matterport3D [2], a large-scale RGB-D dataset comprising 10,800 panoramas across 90 building-scale scenes. To mitigate distortion, we refine the blurred polar regions and use 10k panoramas for training, consistent with prior work. For the validation benchmark, we follow previous work [14, 17, 20] and use the original Matterport3D [2] validation set without refinement.

Evaluation Metrics. Following prior work, we evaluate our method with a diverse set of complementary metrics. For realism, we adopt Fréchet Inception Distance (FID) [5] and its variants, including FID_{clip} for fair comparison by excluding blurred polar regions, and FID_{pole} and FID_{equ} following SMGD [14] to assess polar distortion and perspective projection quality. Since FID relies on an Inception network trained on perspective images and may not fully capture panoramic characteristics, we further employ Fréchet Auto-Encoder Distance (FAED) [11], a variant tailored for panoramas. For diversity, we report Inception Score (IS) [13], replacing the standard Inception-v3 [15] with a ResNet pretrained on Places365 [4, 21] to better reflect the scene-centric nature of our data. For text-image alignment, we compute CLIP Score (CS) [12], and for perceptual quality, we report Q-Align (QA) [18], BRISQUE [9], and NIQE [10], following HunyuanWorld [16]. Notably, since some methods [8, 16, 19] are trained on proprietary datasets, certain metrics such as FID [5] are provided only for reference, whereas the qualitative results in Fig. 9 more faithfully reflect the performance differences.

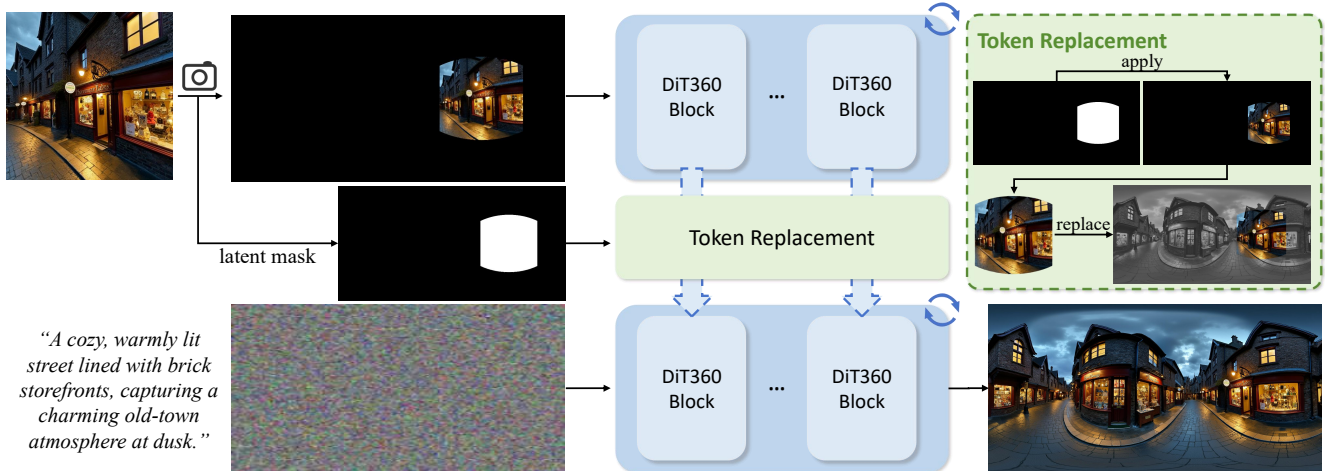


Figure 7. The inpainting and outpainting pipeline of DiT360 (outpainting shown). The input perspective image is first projected into the panoramic domain and obtain a task-specific mask (with white regions indicating masked areas). The model then performs inversion to derive the initial noise representation and extract reference tokens. During denoising, a token replacement mechanism fills masked regions with reference tokens while preserving positional structure, enabling coherent and semantically consistent panoramic completion.



Figure 8. More results on inpainting and outpainting.

9. Full Comparison

In this section, we present the complete qualitative comparison of text-to-panorama generation results. As shown in Fig. 9, our method demonstrates superior perceptual realism, producing sharper and more visually authentic panoramas. In addition, it achieves higher geometric fidelity by ef-

Table 5. User study results on text-to-panorama generation.

| Methods | TA \uparrow | BC \uparrow | Realism \uparrow | OQ \uparrow |
|--------------|---------------|---------------|--------------------|---------------|
| PanFusion | 21.7% | 19.6% | 2.1% | 0.3% |
| Matrix-3D | 24.1% | 27.5% | 23.7% | 5.1% |
| HunyuanWorld | 25.9% | 18.9% | 10.4% | 13.7% |
| Ours | 28.3% | 34.0% | 63.8% | 80.9% |

fectively handling distortions and preserving boundary continuity, whereas baseline methods often suffer from visible artifacts and structural inconsistencies.

10. User Study

To further evaluate human preference, we conducted a user study comparing our method with several representative baselines [8, 16, 20]. The study focused on four key aspects: text alignment (TA), boundary continuity (BC), realism, and overall quality (OQ). A total of 63 participants were asked to choose their preferred outputs from different methods on a test set consisting of 10 images. As shown in Tab. 5, our method received the highest preference across all metrics, clearly demonstrating its superior ability to generate realistic panoramic images with faithful alignment and coherent boundaries.

11. More Results

We present additional results in Figs. 10 and 11 to further illustrate the performance of DiT360 on panoramic image generation. These examples demonstrate that the model consistently produces high-quality, semantically coherent,



Figure 9. The full qualitative comparison on panorama generation. We highlight representative artifacts with red boxes.

and visually detailed completions across a variety of scenes.

12. Limitations and Future Work

Despite the strong performance of *DiT360* on panoramic image generation tasks, several limitations remain. The model’s effectiveness is constrained by the diversity and

scale of available datasets, leading to suboptimal results in certain scenarios, such as those containing high-resolution human faces or intricate scene details. Future work will focus on collecting larger and more diverse high-quality datasets to further enhance the model’s generative capabilities and image resolution. Additionally, leveraging syn-



Figure 10. More results on text-to-panorama generation.



Figure 11. More results on text-to-panorama generation.

thetic data to augment training samples can facilitate further advances in panoramic image generation. In the long term, extending the framework to three-dimensional scene generation and understanding represents a promising research direction.

References

- [1] Black Forest Labs. Flux. <https://github.com/black-forest-labs/flux>, 2024. Accessed: 2024-09-23. ¹
- [2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017. ¹
- [3] Haoran Feng, Zehuan Huang, Lin Li, Hairong Lv, and Lu Sheng. Personalize anything for free with diffusion transformer. In *arXiv*, 2025. ¹
- [4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *arXiv*, 2015. 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 *arXiv*, 2018. ¹
- [6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *arXiv*, 2021. ¹
- [7] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *arXiv*, 2019. ¹
- [8] Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Nahmias, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao, and Shiwei Li. Matrix3d: Large photogrammetry model all-in-one. In *arXiv*, 2025. ^{1, 2}
- [9] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. In *IEEE Trans. Image Process.*, 2012. ¹
- [10] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a “completely blind” image quality analyzer. In *IEEE Signal Process. Lett.*, 2013. ¹
- [11] Changgyoon Oh, Wonjune Cho, Daehee Park, Yujeong Chae, Lin Wang, and Kuk-Jin Yoon. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *arXiv*, 2021. ¹
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *arXiv*, 2021. ¹
- [13] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *arXiv*, 2016. ¹
- [14] Xiancheng Sun, Mai Xu, Shengxi Li, Senmao Ma, Xin Deng, Lai Jiang, and Gang Shen. Spherical manifold guided diffusion model for panoramic image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025. ¹
- [15] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *arXiv*, 2015. ¹
- [16] HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, Yihang Lian, Yulin Tsai, Lifu Wang, Sicong Liu, Puhua Jiang, Xianghui Yang, Dongyuan Guo, Yixuan Tang, Xinyue Mao, Jiaao Yu, Junlin Yu, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Chao Zhang, Yonghao Tan, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Minghui Chen, Zhan Li, Wangchen Qin, Lei Wang, Yifu Sun, Lin Niu, Xiang Yuan, Xiaofeng Yang, Yingping He, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Tian Liu, Peng Chen, Di Wang, Yuhong Liu, Linus, Jie Jiang, Tengfei Wang, and Chunchao Guo. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. In *arXiv*, 2025. ^{1, 2}
- [17] Chaoyang Wang, Xiangtai Li, Lu Qi, Xiaofan Lin, Jinbin Bai, Qianyu Zhou, and Yunhai Tong. Conditional panoramic image generation via masked autoregressive modeling. In *arXiv*, 2025. ¹
- [18] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *arXiv*, 2023. ¹
- [19] Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Yixuan Li, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. In *arXiv*, 2024. ¹
- [20] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *CVPR*, 2024. ^{1, 2}
- [21] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017. ¹