PanoFree: Tuning-Free Holistic Multi-view Image Generation with Cross-view Self-Guidance Supplementary Material

Aoming Liu[1,2*], Zhong Li[1†], Zhang Chen[1†], Nannan Li[2], Yi Xu[1], and Bryan A. Plummer[2]

[1] OPPO US Research Center, Palo Alto, CA 94303, USA
[2] Boston University, Boston, MA 02215, USA

A Detailed Method Illustrations

A.1 Panorama Generation Pipelines

In this part, we provide detailed illustrations of PanoFree's generation processes for Planar Panoramas, 360° Panoramas, and Full Spherical Panoramas.
Planar Panorama Generation is illustrated in Fig. 1. We use the Bidirectional Generation with Symmetric Guidance strategy to iteratively warp in two directions based on symmetric planar translation from the initial view located at the center of the planar panorama. Subsequently, we generate the image of the next view using inpainting.

Fig. 1: Detailed Illustration of Planar Panorama generation.

360° Panorama Generation is illustrated in Fig. 2. Similarly, the Bidirectional Generation with Symmetric Guidance strategy is employed. Generation starts from the initial view centering (pitch, yaw) = (0°,0°), then undergoes symmetric rotation in two directions around the yaw axis. Finally, the two generation paths converge at the merging view with (pitch, yaw) = (0°,180°). Inpainting is used to merge the two generation paths to ensure loop closure.

*Work partly completed during Aoming's internship at OPPO US Research Center.
†Corresponding Authors.

Fig. 2: Detailed Illustration of 360° Panorama generation.

Full Spherical Generation is illustrated in Fig. 3. We firstly generate a 360° panorama, then expand in the upward and downward directions. Finally, we use inpainting at the upper and lower poles to close up the entire spherical panorama. Specifically, with the generated 360° panorama, we firstly warp to (pitch, yaw) = (ϕ, 0°) and (pitch, yaw) = ($-\phi$, 0°) and inpaint the unknown areas to generate the initial views for the upward and downward expansions. Then, we apply PanoFree to expand the panorama's range in the pitch direction separately. Finally, we warp to (pitch, yaw) = (90°, 0°) and (pitch, yaw) = (-90°, 0°) and inpaint the unknown areas to close up the entire spherical panorama.

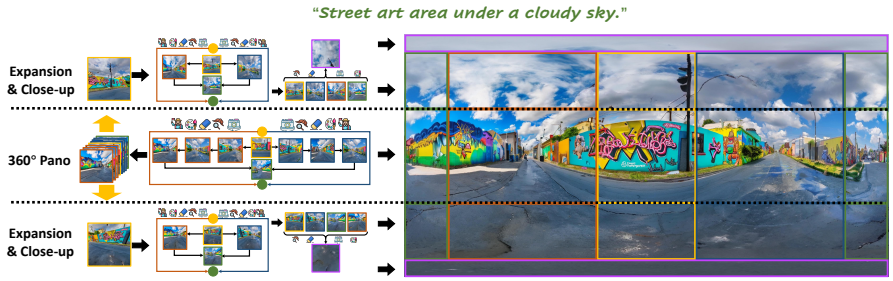


Fig. 3: Detailed Illustration of Full Spherical Panorama generation.

A.2 Accumulated Errors in Vanilla Sequential Generation

In this part, we provide detailed illustrations of the major accumulated errors that occur in the vanilla sequential generation process, and the risky erasing operations based on distance, edge color and smoothness.

Accumulated Inconsistency is illustrated in Fig. 4. As the vanilla sequential generation process solely condition the current view on the previous view, slight

style and content shifts that occur during every warping and inpainting step may accumulate. This accumulation can lead to significant differences between distant regions, thereby damaging the global consistency of the generated panorama.

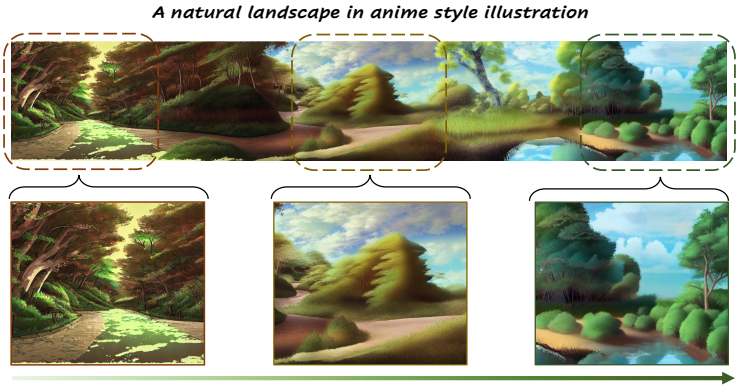


Fig. 4: Illustration of accumulated inconsistency.

Accumulated Artifacts are illustrated in Fig. 5. In the vanilla sequential generation process, normal content in the current view may become artifacts in the following view after several warping and inpainting steps. Moreover, these artifacts can propagate with sequential generation, leading to the generation of more severe artifacts in subsequent views. Those contents that lead to artifacts is called as artifact-inducing contents.

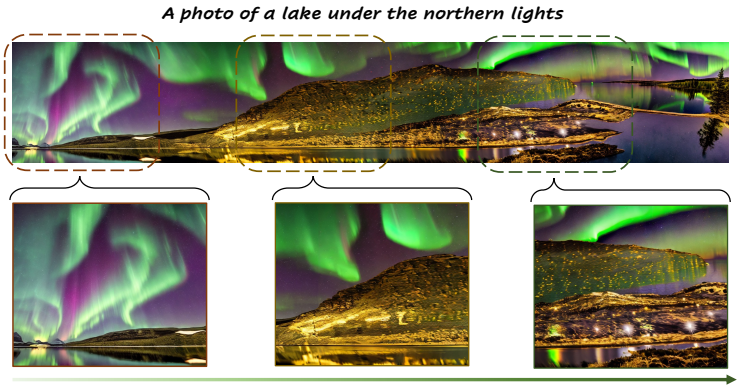


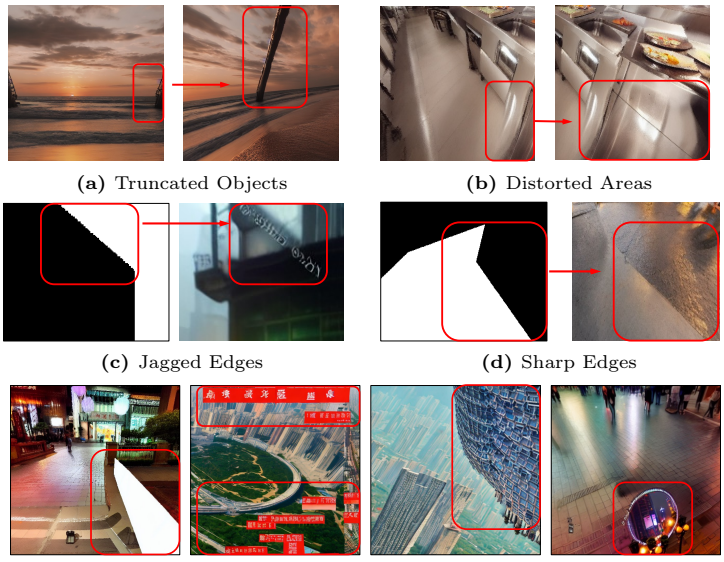
Fig. 5: Illustration of accumulated artifacts.

Artifact-inducing Contents are illustrated in Fig. 6. The major artifact-inducing contents and the artifacts they cause are summarized as follows:

– *Truncated Objects* are illustrated in Fig. 6a. We observed that pretrained T2I models often generate partial objects truncated by the edges of each view.

Those objects can be extended into unreasonable contents in the following views during the warping and inpainting process.

- *Distorted Areas* are illustrated in Fig. 6b. Regions heavily distorted during warping may appear blurred or exhibit strange shapes. After inpainting, these distorted regions may be extended, resulting in a large distorted area on the panorama. Those distorted areas often occur near the edges of each view.
- *Jagged Edges* are illustrated in Fig. 6c. Jagged edges on the inpainting masks often lead to jagged artifacts or letter-like artifacts.
- *Sharp Edges* are illustrated in Fig. 6d. Sharp edges on the inpainting masks often result in inconsistent connections between inpainted regions and other regions of the image, causing noticeable boundaries or color discrepancies.
- *Salient Areas* with abrupt colors or unevenness are illustrated in Fig. 6e. These salient areas may appear reasonable under the current view, but after warping and inpainting, they can easily accumulate into noticeable artifacts. On the other hand, abrupt colors or unevenness are already characteristic features of artifacts.



(a) Truncated Objects **(b)** Distorted Areas

(c) Jagged Edges **(d)** Sharp Edges

(e) Salient areas with abrupt colors or unevenness.

Fig. 6: Illustration of artifact-inducing contents.

Risky Area Erasing are briefly illustrated in Fig. 7, in which distance-based erasing, edge-based erasing, and color & smoothness-based erasing are progressively applied and black regions represent the areas that have been erased.
Hallucinations are illustrated in Fig. 8. Hallucinations often occur when generating full spherical panoramas. When we use the same text description to

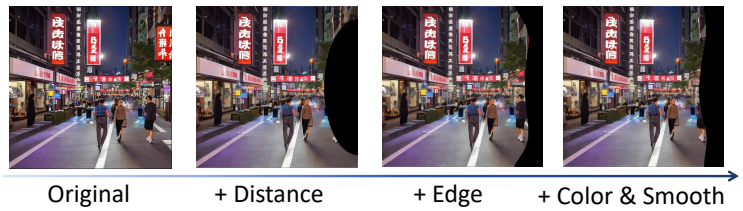


Fig. 7: Illustration of Risky Area Erasing.

generate all views on the spherical panorama, it may result in conflicts between the generated content's placement and the scene structure prior. For instance, another city on the ground or a mountain peak floating in the sky could be generated.

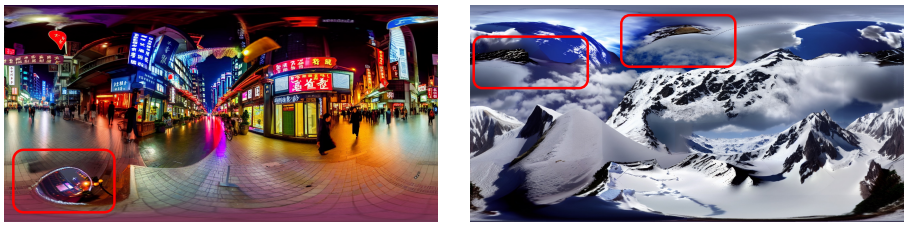


Fig. 8: Examples of Hallucinations.

B Ablation Study

We furthr conducted ablation studies to evaluate the impact of different components of PanoFree and guidance scale respectively.

Qualitative Ablation. Fig.9 showcases representative examples for qualitative ablation. Qualitative results support the prior observations in quantitative ablation, demonstrating progressive improvement in image coherence and quality with each integrated component.

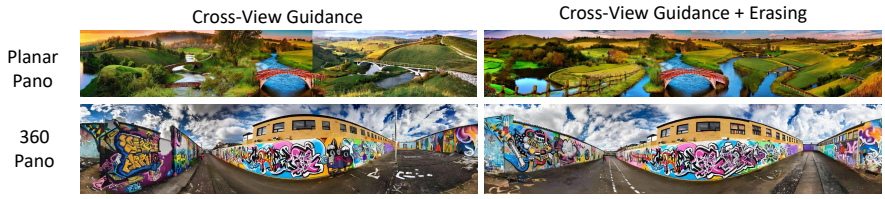


Fig. 9: Qualitative ablation of each component in PanoFree.

B.1 Ablation of Guidance Scale

As the guidance scale primarily affects full panorama generation, we perform a qualitative ablation in Fig. 10, showing that large guidance scales during expansion and close-up stages can cause artifacts, like the unusual structures in the upper part of the right image.

Fig. 10: Qualitative ablation of guidance scales.

C Additional Experiment Results and Comparisons

In this section, we present additional experimental results, including the Planar, 360°, and Full Spherical Panoramas generated by PanoFree. We categorize these results into indoor, street and city, as well as natural scenes, to comprehensively showcase PanoFree's generation capability. Furthermore, we provide further comparisons with baseline methods.

C.1 Planar Panorama Generation

Indoor Scene panoramas generated with PanoFree are shown in Fig. 11. We also provide additional comparisons with vanilla sequential generation (SG), MultiDiffusion (MD) [1], and SyncDiffusion (SYD) [4] in Fig. 12.
City and Street Scene panoramas generated with PanoFree are shown in Fig. 13. We also provide additional comparisons with vanilla sequential generation (SG), MultiDiffusion (MD) [1], and SyncDiffusion (SYD) [4] in Fig. 14.
Natural Scene panoramas generated with PanoFree are shown in Fig. 15. We also provide additional comparisons with vanilla sequential generation (SG), MultiDiffusion (MD) [1], and SyncDiffusion (SYD) [4] in Fig. 16.

We can observe that PanoFree is capable of generating panoramas with various scenes, styles and contents. In terms of image quality and global consistency, PanoFree significantly outperforms vanilla sequential generation and MultiDiffusion and it's comparable to SyncDiffusion.

C.2 360° Panorama Generation

Indoor Scene panoramas generated with PanoFree are shown in Fig. 17. We also provide additional comparisons with vanilla sequential generation (SG) and MVDiffusion (MVD) [5] in Fig. 18.

Fig. 11: Indoor scene planar panoramas generated by PanoFree.

City and Street Scene panoramas generated with PanoFree are shown in Fig. 19. We also provide additional comparisons with vanilla sequential generation (SG) and MVDiffusion (MVD) [5] in Fig. 20.
Natural Scene panoramas generated with PanoFree are shown in Fig. 21. We also provide additional comparisons with vanilla sequential generation (SG) and MVDiffusion (MVD) [5] in Fig. 22.

C.3 Full Spherical Generation

Indoor Scene panoramas generated with PanoFree are shown in Fig. 23.
Natural Scene panoramas generated with PanoFree are shown in Fig. 24.
Natural Scene panoramas generated with PanoFree are shown in Fig. 25.

C.4 Different Pre-trained T2I Model

Additionally, PanoFree is also highly flexible and can plug-and-play with different pre-trained T2I models. In Fig. 26, we illustrate this by applying Stable Diffusion v1 (SD1), Stable Diffusion v2 (SD2) and Stable Diffusion XL (SDXL) for PanoFree. The results show that PanoFree can work properly with different pre-trained T2I models.

D Panorama Diversity Comparison

In this section, we will provide more comparison results on panorama diversity and more detailed illustrations of the diversity issue of Joint Diffusion baselines [1, 4].

D.1 Planar Panorama Generation

For the planar panorama diversity comparison, we mainly compare PanoFree with SyncDiffusion [4]. For each text prompt, we select three different random seeds to generate three results. The comparison results are shown in Fig 27. We can observe that SyncDiffusion generates styles, contents, and scene structures that are very similar across different random seeds. In contrast, PanoFree can generate more diverse panoramas.

D.2 360° Panorama Generation

For the 360° panorama diversity comparison, we mainly compare PanoFree with MVDiffusion [5]. For each text prompt, we also select three different random seeds to generate three results. The comparison results are shown in Fig 28. We can see that MVDiffusion generates styles, contents, scene structures, etc., that are very similar across different random seeds. Even the diversity of MVDiffusion is poorer than SyncDiffusion. This is because MVDiffusion undergoes fine-tuning on top of Joint Diffusion design, which can bias the generation results towards the training dataset. In contrast, PanoFree can still generate more diverse 360 panoramas.

D.3 Generation with Rough Prompts

This diversity issue becomes particularly apparent when given some rough prompts. Therefore, we additionally generated a "rough set" consisting of 20 short and blurry prompts to exacerbate this issue. For each prompt, we used 20 different random seeds. Then, we calculated intra-LPIPS and cross-LPIPS based on the generated results from the rough set. By subtracting the intra-LPIPS from the cross-LPIPS, we can illustrate the trade-off between consistency and diversity. As shown in Table 1, we can observe that both MultiDiffusion and SyncDiffusion result in a significant decrease in diversity. Specifically, MultiDiffusion appears to perform an "equivalent exchange" between consistency and diversity.

Note that this loss of diversity appears to be "within the prompt". When provided with more detailed prompts, Joint Diffusion still has the ability to generate corresponding results. However, iteratively adjusting the prompt based on the output results also incurs a considerable additional time overhead.

Table 1: Results on Planar Panorama generation with rough prompts. Intra-LPIPS (10^{-2}) measures global consistency, Cross-LPIPS (10^{-2}) diversity, and CL-IL (Cross-LPIPS - Intra-LPIPS, (10^{-2})) the trade-off between consistency and diversity.

Method	Intra-LPIPS↓	Cross-LPIPS↑	CL-IL↑
SG	70.67	73.65	2.97
MD [1]	68.03	68.57	0.54
SYD [4]	64.32	67.48	3.16
PF (ours)	65.39	72.29	6.90

E Additional Experiment Details

E.1 Quantitative Evaluation Details

Reference Sets. For metrics that require a reference set, such as FID [3] and KID [2], we generate reference sets composed of perspective view images using the same stable diffusion model with identical prompts but different random seeds. We then crop an equal number of perspective view images from the panoramas generated by PanoFree and baseline methods to perform the calculations.

E.2 User Study Details

For planar panorama generation and 360° panorama generation, we conducted four user studies for each task to further evaluate the global consistency, image quality, prompt compatibility, and diversity of the generated panoramas, respectively. For the first three user studies, we follow the design of SyncDiffusion [4]. Participants were presented with panorama images generated by 2 methods and asked to measure their panorama consistency quality, prompt compatibility, and diversity (see supplementary for details). Then they are asked to choose one of them by answering the question: *"Which one appears a more consistent panorama image to you?"* **(Consistency)**, *"Which one is of higher quality?"* **(Quality)** and *"Which one best matches the shared caption?"* **(Prompt Compatibility)**.
Diversity. For the last user study, participants were presented with 2 groups of panoramas generated by the 2 methods in every question. Each group contains 3 panorama generated with same prompt and different random seeds. Then they are asked to choose one group by answering the question: *"Which group of panorama images is more diverse"* **(Diversity)**. We set 15 questions for each user study and collect responses from 5 Amazon MTurk workers for each question.
Baseline Methods. For the planar panorama generation task, we selected SyncDiffusion as the primary baseline. For the 360 panorama generation task, we chose MVDiffusion as the main comparative baseline.

E.3 Planar Panorama Generation

Task Configurations. In this paper, the resolution of the generated planar panoramas is 512x3072, while each view image has a resolution of 512x512. All

methods employed 50 diffusion steps.

Baseline and Configuration details are illustrated in the following.

- *Sequential Generation (SG)* refers to the vanilla iterative warping and inpainting process. We utilize 10 warping and inpainting steps to extend the initial view image into the desired panorama, with a translation step size of 256 pixels in the image space for each step.
- *MultiDiffusion (MD)* [1] adopts joint diffusion approach with multiple overlapping windows on the latent space. Each window has a separate diffusion process that are fused by averaging the latent features within the overlapping regions at every reverse diffusion step. We utilized default configurations from the official implementation, including a stride of 8 in the latent space.
- *SyncDiffusion (SYD)* [4] is another joint diffusion approach which achieves the state-of-the-art in Planar Panorama generation regarding global consistency. SyncDiffusion fuses multiple diffusion process and ensures global consistency by guiding the reverse diffusion process while adjusting the intermediate latent features at every step. We used default configurations from the official implementation, including a stride of 16 in the latent space, a weight of 20, and a weight decay with a rate of 0.95.

PanoFree Configurations. Similar to vanilla sequential generation, PanoFree also utilizes 10 warping and inpainting steps, 5 steps in each direction, to extend the initial view image into the desired panorama, with a translation step size of 256 pixels in the image space for each step. For SDEdit, we set $t_0 = 0.98$. We only estimate risk based on the distance to the initial view. At each step, we erase 30% of known areas based on the estimated risk.

E.4 360° Panorama Generation

Task Configurations. In this paper, the spherical surface is represented by a 2048x4096 2D image with equirectangular projection and we care about area with pitch $\in [-40, 40]$ for 360° panoramas. Each view image has a resolution of 512x512. All methods employed 50 diffusion steps.
Baselines. We have chosen 2 baselines for comparison. Except the *Vanilla Sequential Generation (SG)*, we also selected *MVDiffusion (MVD)* as a baseline. The baseline details and implementation details are available in the appendix.

- *Sequential Generation (SG)* still refers to the vanilla iterative warping and inpainting process. The only difference is that the current warping corresponds to optical geometric changes caused by rotation. We adopt a Field of View (FoV) of 80° and a yaw stride of 40°. 8 warping and inpainting steps were utilized.
- *MVDiffusion (MVD)* [5] also adopts the Joint Diffusion design, which fuses multiple diffusion processes together to generate consistency in multi-view images by incorporating correspondence-aware attention into a pretrained diffusion model. MVDiffusion requires panorama images for training, and we

chose to utilize the model weights provided by the authors. It is worth noting that although the model weights were trained on indoor scenes, MVDiffusion demonstrates impressive generalization ability and can generate outdoor data. We utilized the default configurations from the official implementation.

PanoFree Configurations. Similar to vanilla sequential generation, PanoFree also utilizes a Field of View (FoV) of 80° and a yaw stride of 40°. 7 warping and inpainting steps were utilized, including 3 symmetric steps in each direction, and a merging step. For SDEdit, we set $t_0 = 0.98$. We use $\mathbf{w} = [0.8, 0.2, 0, 0]$ to combine the risks $[\mathbf{r}^i, \mathbf{r}^e, \mathbf{r}^c, \mathbf{r}^s]$. At each step, we erase 5% of known areas based on the estimated risk.

E.5 Full Spherical Generation

Task Configurations. In this paper, the spherical surface is represented by a 2048x4096 2D image with equirectangular projection and we care about the whole spherical surface for full spherical panoramas. Each view image has a resolution of 512x512. All methods employed 50 diffusion steps.
PanoFree Configurations. The configurations for generating areas with pitch ∈ [-40°, 40°] are exactly the same as described for the 360° panorama. Then, we rotate the viewpoint upwards and downwards by 25° to generate areas with pitch ∈ [40°, 65°] and pitch ∈ [-40°, -65°]. During expansion, we use a Field of View (FoV) of 110° and a stride of 80°. Three warping and inpainting steps are required for expansion in both the upward and downward directions. Finally, using the upper and lower poles as centers, we use one warping and inpainting step each to generate areas with pitch ∈ [65°, 90°] and pitch ∈ [-65°, -90°]. For expansion stages, we set $t_0 = 0.90$. We use $\mathbf{w} = [0.6, 0.2, 0.1, 0.1]$ to combine the risks $[\mathbf{r}^i, \mathbf{r}^e, \mathbf{r}^c, \mathbf{r}^s]$. The guidance scale is set to 2.0, while the variance of noise is amplified by a factor of 1.05. At each expansion step, we erase 10% of known areas based on the estimated risk. At the final close-up stpes, we use a Field of View (FoV) of 90°. We set $t_0 = 0.90$. We use $\mathbf{w} = [0.6, 0.2, 0.1, 0.1]$. And we erase 20% of known areas based on the estimated risk. The guidance scale is set to 1.0, while the variance of noise is amplified by a factor of 1.1.

F Limitation and Failure Cases

In this section, we discuss about some limitations and failure cases of PanoFree.

F.1 Undesired Camera Pose

refers to the inconsistency between the underlying camera pose of the generated images and the camera pose we set. Undesired camera pose issue can lead to failure cases when generating 360 panoramas. As illustrated in Fig. 29a, this issue often results in ground deformation and may cause severe distortion, thereby making the generated panorama appear unreasonable.

Potential solutions include fine-tuning the T2I models with images having desired camera poses or incorporating the camera pose as an additional model conditioning input to control the generated images. However, these approaches require costly fine-tuning. In this paper, we narrow down this problem to the camera pose of the initial view: we find that as long as the underlying camera pose of the initial view image is relatively close to our set pose, there are fewer occurrences of undesired camera pose issues in the following views. Therefore, as illustrated in Fig 29b, we can mitigate this problem by leveraging open-source pre-trained camera pose estimation models to predict the camera pose of the initial view image. Specifically, we primarily care about the pitch and Field of View (FoV) of initial view image.

F.2 Biased Generation

Another issue is biased generation PanoFree may exhibit a bias towards ensuring local alignment with the prompt and scene priors, potentially leading to conflicts on a global scale. There are two common issues. Firstly, as shown in Fig 30, PanoFree may to generate duplicated semantic contents across the panorama. Secondly,, it can produce inconsistent scene characteristics in different parts of the image. For instance, in the left image of Fig. 31a,one section depict a winter landscape, while another section simultaneously presents spring-like features.

However, it's important to note that PanoFree's primary contribution lies in its tuning-free and efficient panorama generation approach, rather than completely eliminating these biases. On the other hand, biased generation remains a challenge in panorama generation tasks, affecting many methods including those trained on real panoramic data. For example, As shown in Fig 30, MVDiffusion [5] also exhibits issues with duplicated semantic content. Nevertheless, if data and computational costs are not constraints, PanoFree can be readily enhanced. For example, pretrained LLMs could be employed to generate denser prompts, potentially correcting biased generation, as demonstrated in Fig. 31b.

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation (2023)
2. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying mmd gans. ArXiv **abs/1801.01401** (2018), `https://api.semanticscholar.org/CorpusID:3531856`
3. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Neural Information Processing Systems (2017), `https://api.semanticscholar.org/CorpusID:326772`
4. Lee, Y., Kim, K., Kim, H., Sung, M.: Syncdiffusion: Coherent montage via synchronized joint diffusions. Advances in Neural Information Processing Systems **36** (2024)

5. Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. ArXiv **abs/2307.01097** (2023), https://api.semanticscholar.org/CorpusID:259316427

Fig. 12: Comparison: Indoor scene planar panoramas. PF is our PanoFree.

Fig. 13: City and street scene planar panoramas generated by PanoFree.

Fig. 14: Comparison: City and street scene planar panoramas. PF is our PanoFree.

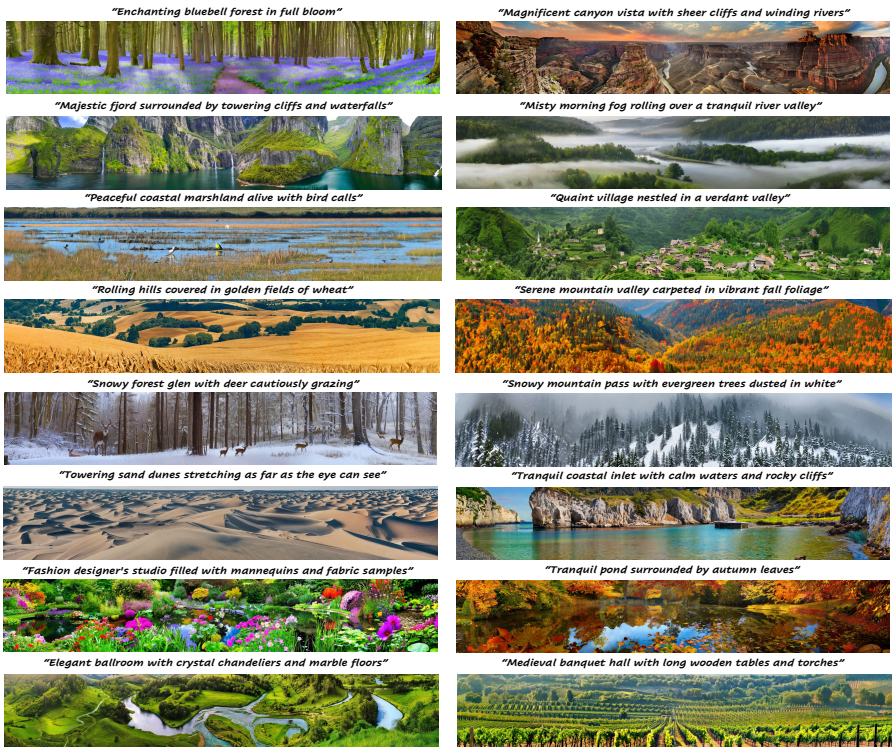


Fig. 15: Natural scene planar panoramas generated by PanoFree.

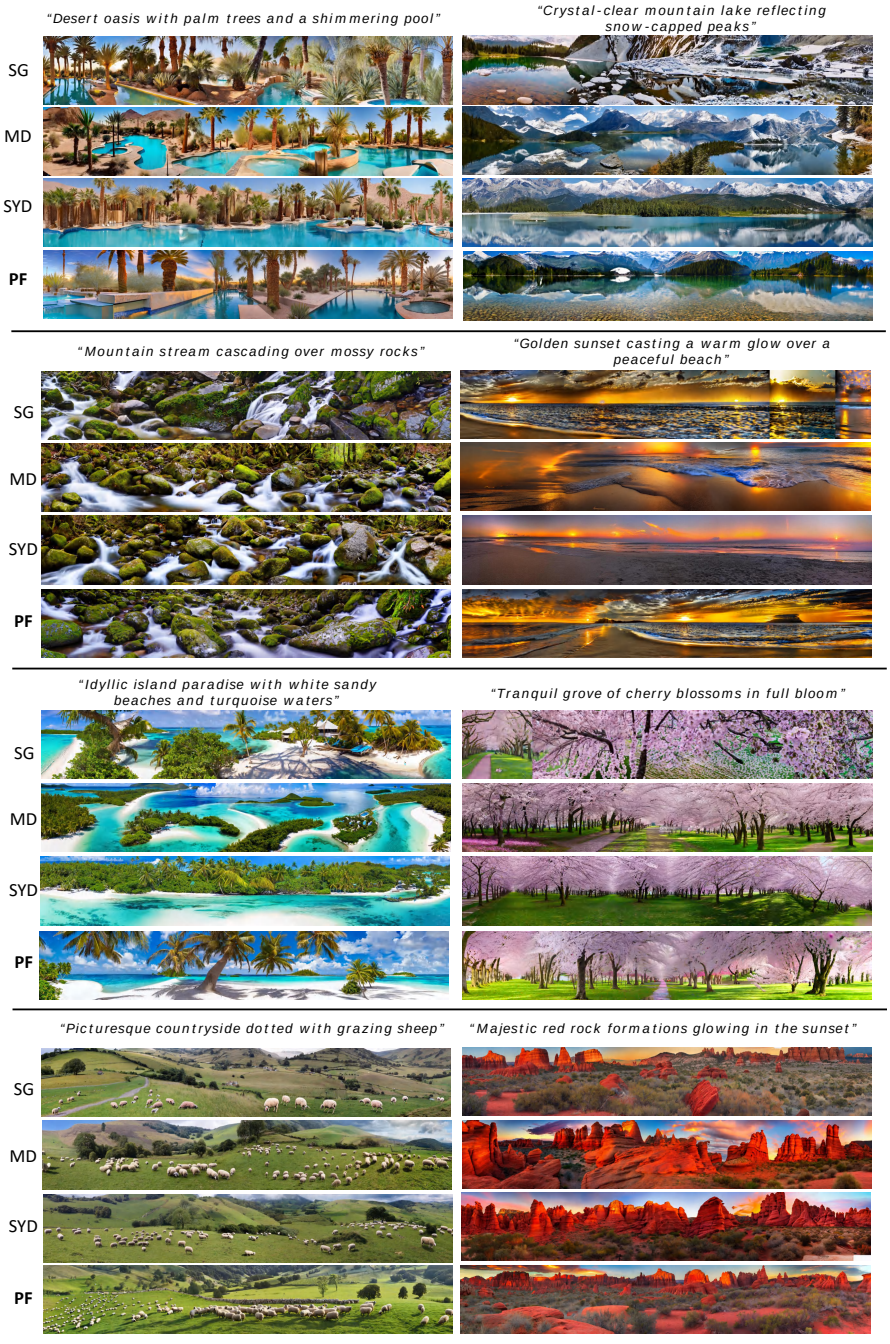


Fig. 16: Comparison: Natural scene planar panoramas. PF is our PanoFree.

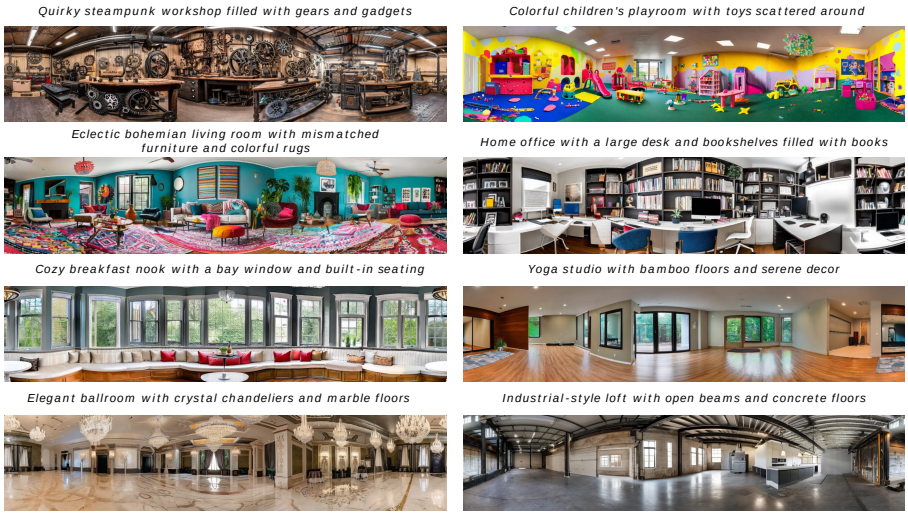


Fig. 17: Indoor scene 360° panoramas generated by PanoFree.

Fig. 18: Comparison: Indoor scene 360° panoramas. PF is our PanoFree.

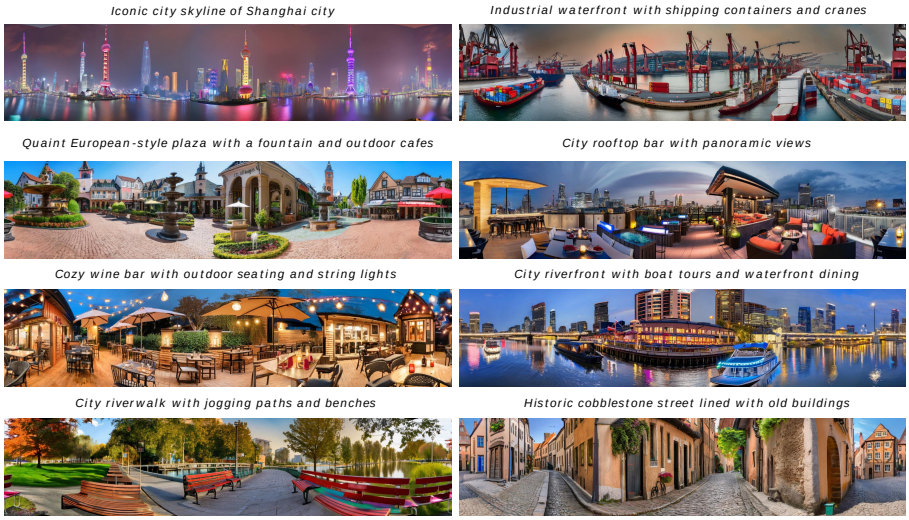


Fig. 19: City and Street scene 360° panoramas generated by PanoFree.

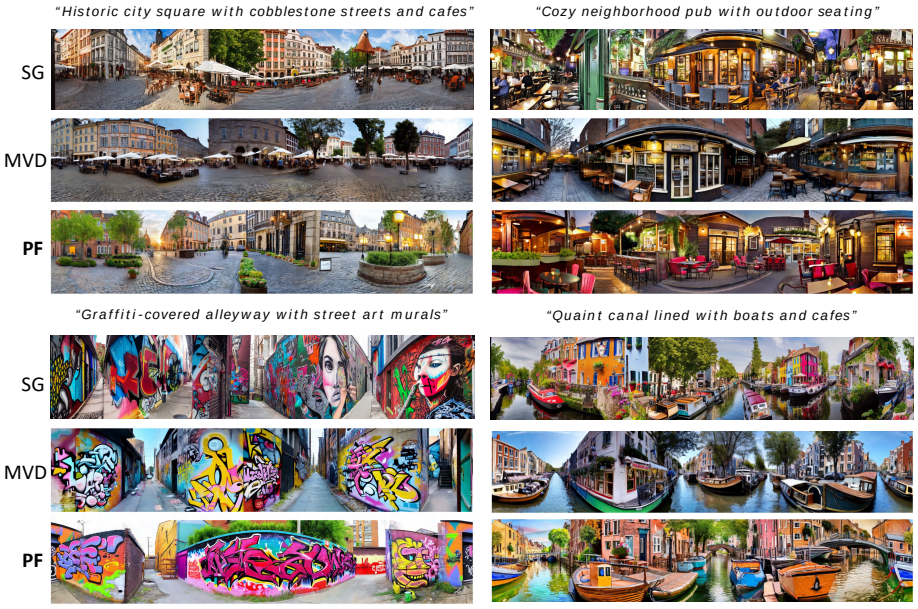


Fig. 20: Comparison: City and Street scene 360° panoramas. PF is our PanoFree.

Fig. 21: Natural scene 360° panoramas generated by PanoFree.

Fig. 22: Comparison: Natural scene 360° panoramas. PF is our PanoFree.

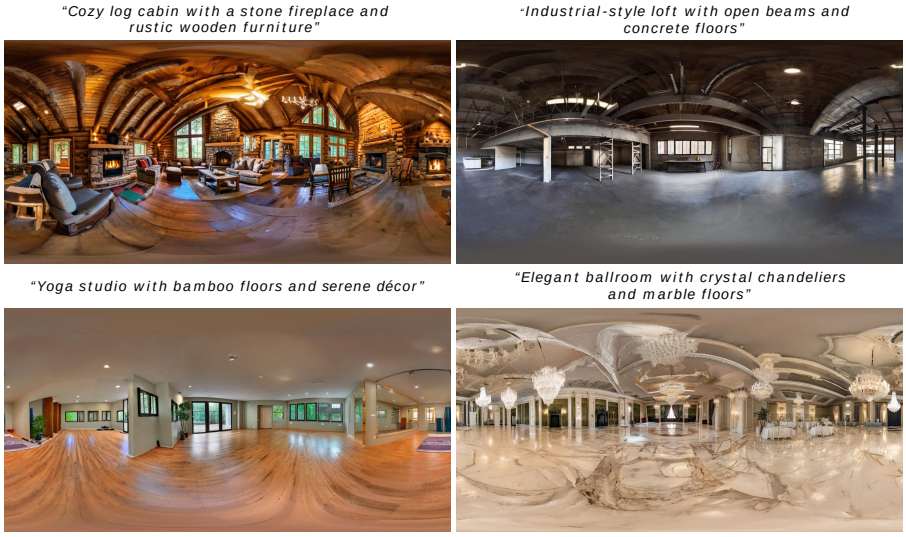


Fig. 23: Indoor scene full spherical panoramas generated by PanoFree.

Fig. 24: City and street scene full spherical panoramas generated by PanoFree.

Fig. 25: Natural scene full spherical panoramas generated by PanoFree.

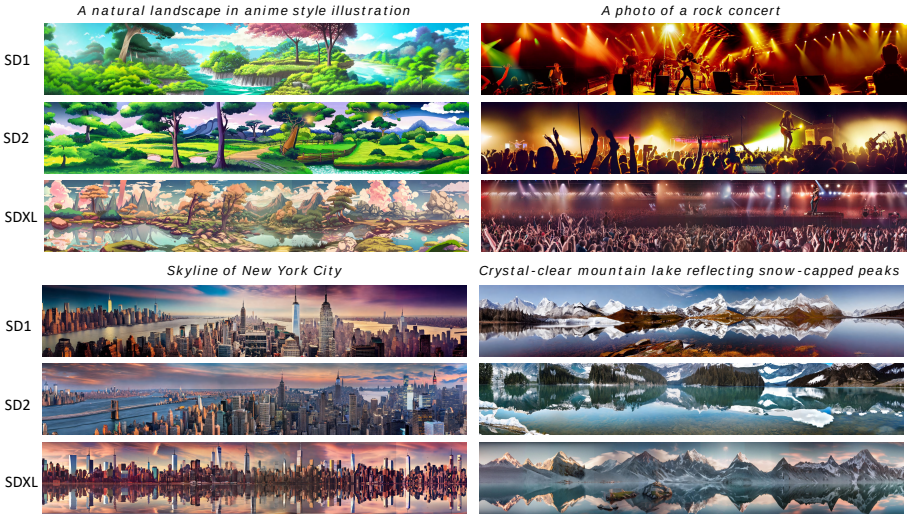


Fig. 26: Panoramas generated by PanoFree using different pre-trained T2I models.

Fig. 27: Planar panorama diversity illustration and comparison. PF is our PanoFree.

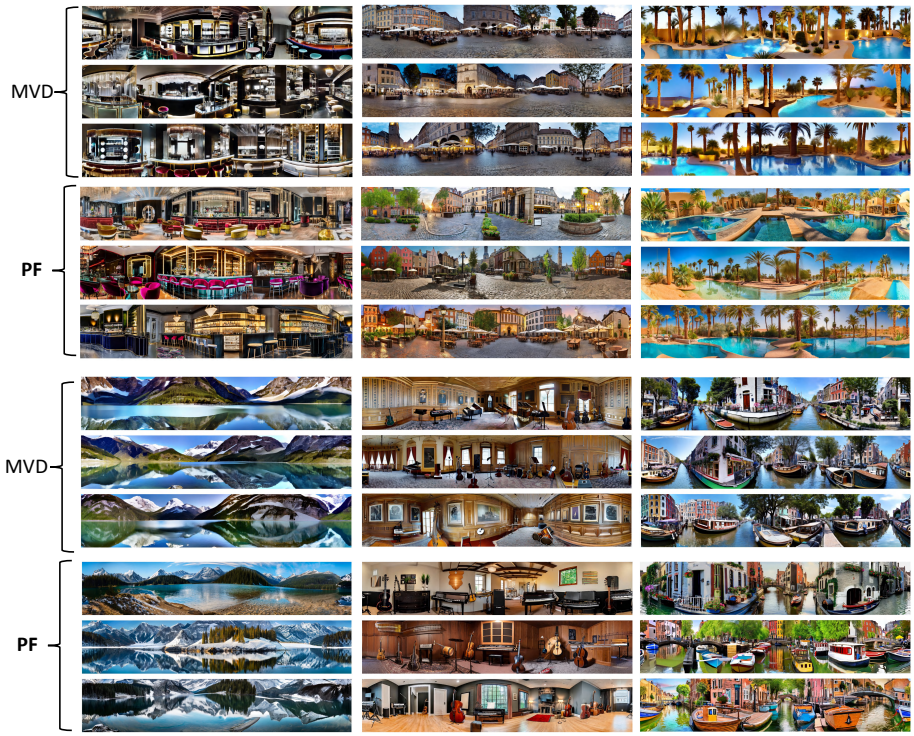
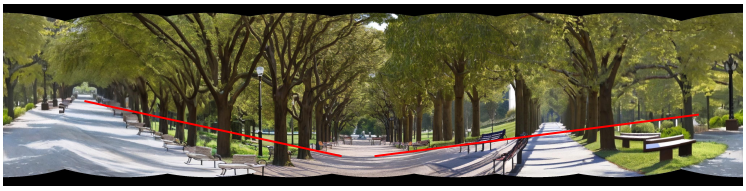


Fig. 28: Planar panorama diversity illustration and comparison. PF is our PanoFree.



(a) Undesired camera poses

(b) Estimate the camera pose of initial view

Fig. 29: Results with the undesired camera poses and the estimated initial camera pose.

Fig. 30: Duplicated semantic contents generated with MVDiffusion [5] and PanoFree.

(a) Biased Generation.

(b) Correction with dense prompt.

Fig. 31: Correcting biased generation with denser prompts.