

# What Makes for Text to 360-degree Panorama Generation with Stable Diffusion?

Jinhong Ni<sup>1\*</sup>   Chang-Bin Zhang<sup>2</sup>   Qiang Zhang<sup>3,4</sup>   Jing Zhang<sup>1</sup>

<sup>1</sup>Australian National University   <sup>2</sup>The University of Hong Kong

<sup>3</sup>Beijing Innovation Center of Humanoid Robotics

<sup>4</sup>Hong Kong University of Science and Technology (Guangzhou)

{jinhong.ni, jing.zhang}@anu.edu.au   cbzhang@connect.hku.hk   jony.zhang@x-humanoid.com

## Abstract

*Recent prosperity of text-to-image diffusion models, e.g. Stable Diffusion, has stimulated research to adapt them to 360-degree panorama generation. Prior work has demonstrated the feasibility of using conventional low-rank adaptation techniques on pre-trained diffusion models to generate panoramic images. However, the substantial domain gap between perspective and panoramic images raises questions about the underlying mechanisms enabling this empirical success. We hypothesize and examine that the trainable counterparts exhibit distinct behaviors when fine-tuned on panoramic data, and such an adaptation conceals some intrinsic mechanism to leverage the prior knowledge within the pre-trained diffusion models. Our analysis reveals the following: 1) the query and key matrices in the attention modules are responsible for common information that can be shared between the panoramic and perspective domains, thus are less relevant to panorama generation; and 2) the value and output weight matrices specialize in adapting pre-trained knowledge to the panoramic domain, playing a more critical role during fine-tuning for panorama generation. We empirically verify these insights by introducing a simple framework called UniPano, with the objective of establishing an elegant baseline for future research. UniPano not only outperforms existing methods but also significantly reduces memory usage and training time compared to prior dual-branch approaches, making it scalable for end-to-end panorama generation with higher resolution. The code will be released<sup>1</sup>.*

## 1. Introduction

Creating 360-degree panoramic images has gained substantial attention due to its significant potential [21, 53]. Despite the considerable advancement in text-to-image synthesis re-

cently [33–35], generating panoramas from text prompts remains challenging from the following aspect. Panoramic images encompass the entire surrounding view with a 360-degree horizontal and 180-degree vertical field of view, typically represented using equirectangular projection geometry. This results in distinctive features such as a 2 : 1 aspect ratio and spherical distortion, setting them apart from standard square perspective images. On top of this, due to the high cost of capturing panoramic images in practice, the panoramic datasets are often relatively scarce, e.g. Matterport3D [5] contains 10,800 panoramic images. The lack of data complicates the training of generative models, as conventional perspective diffusion models [35] generally require billions of text-image pairs for training [38].

To mitigate data scarcity, the typical strategy is to fine-tune pre-trained diffusion models for downstream applications [18, 37, 57]. However, as stated in [17], the fundamental structural differences between panoramic and perspective images intuitively suggest that the embedded perspective knowledge within the pre-trained diffusion models may not be readily transferable. Aligning with this intuition, prior work [17, 19, 44] has proposed generating multiple perspective images according to predefined camera poses and stitching them into a panorama. Contrary to the aforementioned intuition, another line of work [56] has demonstrated that fine-tuning pre-trained diffusion models on limited panoramic data using conventional low-rank adaptation (LoRA) [15] still yields effective text-to-panorama generation results. This empirical success suggests the presence of some intrinsic mechanism that enables LoRA to effectively leverage prior knowledge from the pre-trained perspective diffusion models, thereby circumventing the structural differences. This motivates us to explore the following question: *What exactly makes for fine-tuning Stable Diffusion for text-to-panorama generation?*

We base our analysis on the LoRA fine-tuning paradigm to study the behaviors and ideally functionalities of all trainable counterparts, particularly their impact on panorama generation, with the ultimate goal of elucidating the mecha-

<sup>\*</sup>Work partially done at The University of Hong Kong.

<sup>1</sup><https://github.com/jinhong-ni/UniPano>

*“Amidst the ruins of an ancient civilization, deciphering hieroglyphics that tell the story of a lost world.”*



*“a living room with a fireplace.”*



*“a home with pool and patio.”*



Figure 1. Our UniPano can synthesize realistic 360-degree panoramic images by fine-tuning Stable Diffusion. (Top)  $1024 \times 2048$  panoramic images generated by UniPano. (Bottom)  $512 \times 1024$  panoramic images generated by UniPano.

nism that thrives in adapting perspective diffusion models for panorama generation. Our launching point is to isolate the trainable components within LoRA fine-tuning (*i.e.*,  $W_{\{q,k,v,o\}}$ , *cf.* Fig. 2) and examine their relevance for learning panoramic structures. Subsequently, we identify the underlying behaviors of each trainable component when they are tuned jointly. We draw two major empirical findings (*cf.* Sec. 3.2 for details):

- $W_{\{q,k\}}$  in the attention blocks fail to learn the panoramic structures when they are trained in isolation, whereas  $W_{\{v,o\}}$  both succeed in capturing such information.
- When  $W_{\{q,k,v,o\}}$  are jointly trained,  $W_{\{v,o\}}$  are responsible for learning panoramic-specific information (*i.e.*, equirectangular structure), whereas  $W_{\{q,k\}}$  learn shared knowledge across panoramic and perspective domains that are irrelevant to the panoramic structure.

Our analysis reveals the following: After fine-tuning with panoramic images, we discover that the query and key

within the cross-attention blocks capture less panoramic-specific information, namely, they function to ‘preserve’ or ‘enhance’ the pre-trained perspective knowledge; In contrast, the value and output weight matrices are responsible for adapting such perspective information into the panoramic domain. Based on the analysis, we believe that fine-tuning the query and key matrices is less relevant to panorama generation, whereas the representational capability of the value and output matrices should be emphasized. This yields our straightforward yet efficacious uni-branch solution, dubbed *UniPano*, targeting to serve as a simple baseline to foster future research. UniPano achieves state-of-the-art results on  $512 \times 1024$  text-to-panorama generation while requiring notably less memory and training time compared to the current SoTA [56]. Thanks to this computational efficiency, UniPano can be scaled to generate panoramic images with even higher resolution in an end-to-end manner, as shown in Fig. 1.



## 2. Related Work

**Diffusion Models.** The recent breakthrough in diffusion models [14, 40, 41] has accelerated the inference process [23, 42, 58, 59] and significantly boosted the generation quality [9, 30]. The prosperity of large-scale pre-trained diffusion models [33–35] has prompted various applications, including text-to-3D generation [8, 31], personalized customization [11, 37], image inpainting [26, 51], depth estimation [18], perception [6, 52], *etc.* The core of most of these works is to exploit pre-trained text-to-image diffusion models as a priori thus circumventing the data scarcity which is common in downstream applications. Such an adaptation is usually achieved by parameter-efficient fine-tuning techniques such as low-rank adaptation (LoRA) [15], or via distillation [27]. This paper targets the former approach and attempts to demystify what makes for panorama generation by fine-tuning pre-trained diffusion models with LoRA.

**Panorama Generation.** Existing works can be roughly divided into two categories, namely panorama outpainting and text-to-panorama generation. The former approach [3, 17, 25, 28, 46, 48] aims to complete a panoramic image based on a partial input image, exemplified by CubeDiff [17] which proposes to jointly generate six faces of cubemap for panorama generation. Aligning with the prosperity of text-conditioned generation as in the perspective domain, text-to-panorama generation [4, 19, 22, 44, 47, 54–56] has gained attention recently. Among these works, [4, 19, 44, 55] generate a sequence of consistent perspective images and stitch them into a panorama. A separate branch of works fine-tunes the pre-trained text-to-image diffusion models to generate an equirectangular panoramic image in an end-to-end manner [47, 54, 56]. StitchDiffusion [47] fine-tunes pre-trained diffusion models with techniques ensuring panoramic continuity. DiffPano [54] includes multi-view panoramic awareness into the framework. PanoFree [22] stands apart from the aforementioned methods by employing a tuning-free approach to generate panoramas. PanFusion [56] introduces a dual-branch framework by simultaneously generating perspective and panoramic images and ensuring consistency through a cross-branch attention mechanism. Of particular importance to our study, [56] shows that LoRA fine-tuning on Stable Diffusion reports reasonable performance, and our work aims to investigate and elucidate the underlying factors contributing to this empirical success. As a side product of our analysis, we present an efficient and effective uni-branch panorama generation framework, serving as a baseline method for future research.

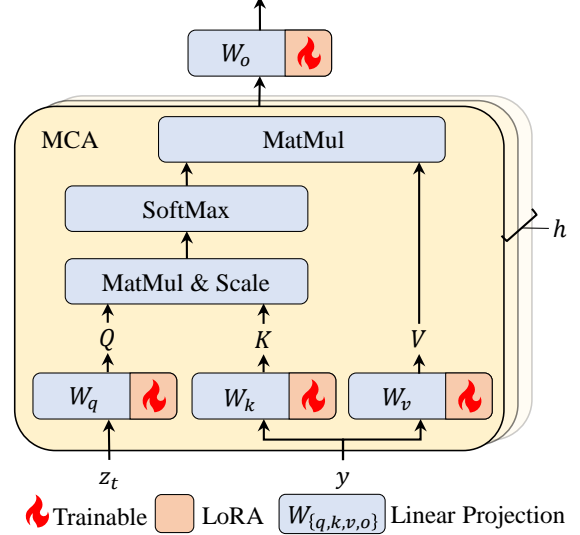


Figure 2. **Illustration of fine-tuning cross-attention blocks within diffusion models with low-rank adaptation (LoRA).** MCA, MatMul,  $h$  denotes the multi-head cross attention, matrix multiplication, and the number of attention heads respectively.

## 3. What Makes for Panorama Generation?

### 3.1. Preliminary

Diffusion models [14, 40, 41] generate images by iteratively transforming the noise sampled from the prior distribution into the target data distribution, where each sampling step involves predicting the noise from the input noisy image. We defer details on diffusion models to the supplementary materials. Of particular relevance to our study, the conditioning in diffusion models is often accomplished by cross-attention. Formally, given an input latent  $z_t$  and the corresponding condition  $y$ , the attention computes:

$$\text{MHCA}(z_t, y)W_o,$$

where MHCA denotes the multi-head cross-attention, and  $W_o$  represents the output weight; for notational simplicity, we write MHCA in the single-head form

$$\text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right)V,$$

with head channel dimension  $d_h$ ,  $Q = z_t W_q$ ,  $K = y W_k$ , and  $V = y W_v$ , where  $W_{\{q,k,v,o\}}$ <sup>2</sup> are the set of trainable weights. The community has shown that training these attention modules within text-to-image diffusion models using parameter efficient fine-tuning techniques (*e.g.* LoRA [15]) suffices to surrogate fine-tuning the entire models, exemplified by [11, 37]. We illustrate fine-tuning diffusion models with LoRA in Fig. 2.

<sup>2</sup>Throughout the paper, we denote  $W_{\{q,k,v,o\}}$  as the set of trainable weights  $\{W_q, W_k, W_v, W_o\}$ .



Figure 3. **Qualitative comparison for training  $W_{\{q,k,v,o\}}$  in isolation separately.** Training  $W_q$  or  $W_k$  in isolation fails to capture the spherical structure, as in (a) and (b); whereas training  $W_v$  or  $W_o$  in isolation successfully captures the spherical distortion of the panoramic images, as in (c) and (d). All visualizations are generated with the text prompt “a kitchen with stainless steel appliances”.

### 3.2. Motivation and Insights

360° panoramic images capture a complete spherical view of the surroundings, involving a full 360-degree horizontal field of view and a vertical field of view of 180 degrees, typically stored in equirectangular format. These unique characteristics in view structures make legitimate panoramic images fundamentally different from the perspective ones, which at first glance, implies that perspective-related knowledge within the pre-trained diffusion models may not be immediately relevant. Contrary to this intuition, previous works [56] have demonstrated the feasibility of directly fine-tuning pre-trained perspective diffusion models (e.g., Stable Diffusion) with LoRA (cf., Sec. 3.1) for text-to-panorama generation, given a relatively scarce set of panoramic data. Evinced by this empirical success, we speculate that such an adaptation to the panoramic domain has to conceal some intrinsic mechanism to leverage perspective knowledge within the pre-trained diffusion models.

To reveal such a mechanism and elucidate what makes such perspective-based diffusion adaptation succeed in panorama generation, we start by decomposing the trainable components and training them in isolation to identify the behaviors – and ideally the functionalities – of each counterpart. Specifically, we fine-tune  $W_{\{q,k,v,o\}}$  (cf., Fig. 2) separately with LoRA for panorama generation, and showcase the comparison both qualitatively in Fig. 3 and quanti-

	Panorama		20 Views	8 Views
	FAED↓	FID↓	FID↓	FID↓
$W_q$	10.86	81.09	30.66	27.44
$W_k$	13.27	67.63	27.01	24.47
$W_v$	<b>8.66</b>	<b>52.60</b>	<b>17.01</b>	<b>19.35</b>
$W_o$	<b>9.38</b>	<b>52.17</b>	<b>20.32</b>	<b>20.24</b>

Table 1. **Quantitative comparison for training  $W_{\{q,k,v,o\}}$  in isolation separately.** Training only  $W_v$  or  $W_o$  reports considerably better FAED and FID than  $W_q$  or  $W_k$ . Details of reported metrics are in Sec. 4.1.

tatively in Tab. 1. We highlight the following.

**Observation 3.1.** As shown in Fig. 3 (a) and (b), it is evident that training  $W_q$  or  $W_k$  in isolation notably fails to capture the spherical structure within panoramic images, while both  $W_v$  and  $W_o$  are capable of achieving such goals, as in Fig. 3 (c) and (d). The quantitative results in Tab. 1 also align with the qualitative observation, as training  $W_v$  or  $W_o$  in isolation leads to considerably better FAED and FID than  $W_q$  or  $W_k$ .

**Conclusion 3.2.** The four trainable components  $W_{\{q,k,v,o\}}$  within the cross-attention modules exhibit varying abilities to learn the spherical structures for successful adaptation to panorama generation. In particular,  $W_q$  and  $W_k$  fail to capture such distortions even when fitted purely on



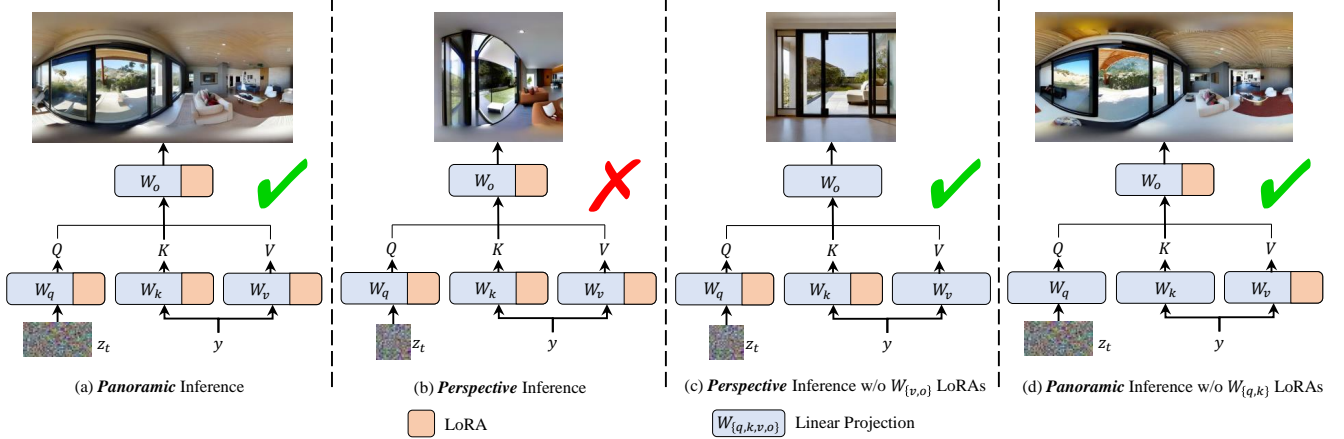


Figure 4. **Illustration of roles of  $W_{\{q,k,v,o\}}$  when jointly fine-tuned.** We first fine-tune  $W_{\{q,k,v,o\}}$  jointly with LoRA on panoramic data, then optionally deactivate some LoRAs for inference with different purposes. (a) the panoramic image generated by fine-tuning  $W_{\{q,k,v,o\}}$  jointly with LoRA, where we simplify the LoRA architecture (cf. Fig. 2); (b) diffusion models with panorama fine-tuned LoRAs can only generate distorted, panoramic-like images, are thus no longer capable of generating perspective images; (c) by excluding  $W_v$  and  $W_o$  LoRAs, diffusion models fine-tuned on panoramic data recover the ability to generate perspective images; (d) excluding  $W_q$  and  $W_k$  LoRAs does not affect the model’s ability to generate panoramic images.

panoramic data, whereas  $W_v$  and  $W_o$  both successfully adapt the pre-trained diffusion model to panorama generation when trained in isolation.

**Remark 3.3.** Conclusion 3.2 suggests that  $W_{\{q,k\}}$  have limited capability to be adapted for panorama generation. As such, excluding them during fine-tuning shall not impact the model’s capability to learn the panoramic structure.

Our subsequent step is to investigate how each component contributes to adaptation when fine-tuned collectively. Knowing that  $W_q$  and  $W_k$  are unable to capture the spherical structure characteristic of panoramic images, we hypothesize that their role is limited to learning common knowledge across both the perspective and panoramic domains. To validate this argument, we design the following experiment: we first fine-tune  $W_{\{q,k,v,o\}}$  with LoRA jointly on panoramic data, then optionally deactivate some LoRAs during inference to examine their relevance to panorama generation. The illustrations are provided in Fig. 4, from which we draw the observation and conclusion below.

**Observation 3.4.** After fine-tuning the diffusion models with panoramic data (Fig. 4 (a)), the model when attempted to generate perspective images with all LoRAs applied can only generate panoramic-like images, as shown in Fig. 4 (b). If the LoRAs associated with  $W_{\{v,o\}}$  are excluded during inference, the models with only the remaining  $W_{\{q,k\}}$  LoRAs successfully recover the ability to generate valid perspective images, evidenced in Fig. 4 (c). On top of this, excluding  $W_{\{q,k\}}$  LoRAs does not affect the model’s ability to generate panoramic images, as shown in Fig. 4 (d).

**Conclusion 3.5.** When all weights  $W_{\{q,k,v,o\}}$  are trained jointly, LoRAs associated with  $W_q$  and  $W_k$  learn shared

knowledge across panoramic and perspective domains that are irrelevant to panoramic structures, whereas LoRAs associated with  $W_v$  and  $W_o$  are responsible for learning the spherical structures of the panoramic images.

**Remark 3.6.** Conclusion 3.5 differentiates the trainable components based on their roles. In particular,  $W_{\{q,k\}}$  even when trained on panoramic images do not apply any spherical distortion to the generated images, such a finding also aligns with Conclusion 3.2 as they also struggle to learn spherical information when trained isolatedly. Contrarily, as  $W_{\{v,o\}}$  are responsible for learning the panoramic structure, their capacity needs to be emphasized more during fine-tuning.

To this end, we have explicated what exactly makes for panorama generation with pre-trained diffusion models. Specifically, after fine-tuning with panoramic images, we find that  $W_q$  and  $W_k$  within the cross-attention modules do not capture the spherical distortion within the panorama at all. They act as if they were tuned using perspective images, and their roles are likely to be ‘preserving’ or ‘refining’ the pre-trained perspective knowledge. In stark contrast,  $W_v$  and  $W_o$  are responsible for adapting the information – captured both in  $W_{\{q,k\}}$  and within the pre-trained model itself – into the panoramic domain. Their roles are thus learning the spherical structure of the panorama, which is much more instrumental to the task.

### 3.3. UniPano

As a byproduct of our insights, we present a memory-efficient uni-branch fine-tuning framework for adapting pre-trained text-to-image diffusion models to panorama generation, dubbed *UniPano*. The core idea of UniPano is based on Remarks 3.3 and 3.6: in essence, it freezes  $W_{\{q,k\}}$  as

	Panorama		20 Views	8 Views
	FAED↓	FID↓	FID↓	FID↓
Pano Only [56]	7.90	50.40	20.10	20.56
DA [50]	9.78	<b>46.41</b>	<b>16.56</b>	<b>18.77</b>
SE [16]	8.72	50.67	17.71	19.61
LoRA ( $r = 8$ ) [15]	8.34	<u>48.58</u>	<u>16.94</u>	19.42
LA [12]	<u>7.65</u>	50.39	18.24	<u>19.41</u>
MoE [10, 49]	<b>7.21</b>	48.83	19.50	20.05

Table 2. **Comparison across several designs for increasing the capacity of  $W_o$ .** Based on the panorama branch only baseline (Pano Only), we compare several plausible attempts to increase the capacity of the output layer ( $W_o$ ). Details of reported metrics are in Sec. 4.1.

they are associated with non-parametric-specific information and emphasizes  $W_{\{v,o\}}$  as they are critical to capture equirectangular distortion within the panoramas. We highlight that the purpose of UniPano is to empirically verify our insights in Sec. 3.2, and with the aim to provide a simple yet effective baseline for future research.

**Design Choices of  $W_o$ .** We propose to increase the representational capacity of the corresponding modules to enhance model’s learning ability for the spherical structure of panoramic images. We differentiate  $W_v$  and  $W_o$  with the intuition that  $W_v$  is associated with head-wise projection while  $W_o$  interacts directly with the entire set of representations. Due to this reason, although both components are capable of learning the spherical structure, we choose to increase the capacity of  $W_o$  because of its directness. We adopt and compare several common strategies to enhance the representational capability:

- *Larger LoRA ranks* (LoRA  $r = 8$ ) [15] is the most straightforward way to increase capability. This approach simply doubles the LoRA ranks of  $W_o$  from 4 to 8.
- *Local Window Attention* (LA) [12, 32] constrains the receptive field of the attention operation to the neighboring pixels. We insert such an attention block before each  $W_o$  LoRA.
- *Deformable Attention* (DA) [50] introduces a learnable offset and computes the attention based on the sampled features. We attempt to insert a deformable attention block prior to each  $W_o$  LoRA.
- *Squeeze and Excitation* (SE) [16] adaptively recalibrates channel-wise feature responses to strengthen the representational power. We similarly insert a SE block before each  $W_o$  LoRA.
- *Mixture of Experts* (MoE) [10, 49] involves computing weighted sum over several expert networks, where the weights are learned via a routing network. We replace each  $W_o$  LoRA with a MoE module. We adopt the same MoE architecture as [10], except that each expert network

is a LoRA, similar to the design in [49]. We apply the same auxiliary loss as [10] for routing load balancing.

The comparison is detailed in Tab. 2, where we also provide a reference to the panorama branch only (Pano Only) baseline. To ensure a fair comparison with this baseline, we also fine-tune  $W_{\{q,k\}}$  LoRAs in all compared strategies. We simply opt to use MoE to enhance the capacity of  $W_o$  because of its superiority on FAED, while we highlight that such a choice may not be optimal and encourage future work to investigate further.

## 4. Experiments

### 4.1. Experimental Setup

**Dataset.** Matterport3D dataset [5] is a scene understanding dataset with 10,800 panoramic images. We use the same captions as [56], which are generated by BLIP-2 [20] with a prompt of “a 360 - degree view of”. We adopt the same data split as [44, 56], containing 9,820 and 1,092 pairs for training and evaluation respectively.

**Implementation Details.** To facilitate a fair comparison, we strictly follow [44, 56] to train our model using AdamW optimizer [24] with a batch size of 4 and a learning rate of  $2 \times 10^{-4}$  for 10 epochs, with identical cosine annealing learning rate scheduler. Following [56], we base our model on Stable Diffusion 2 base version.

**Evaluation Metrics.** We follow previous works to evaluate the generated panoramic images in the panorama [7, 56] and perspective [44] domain.

- *Panorama.* Following [7, 56], we report Fréchet Inception Distance (FID) and Inception Score (IS) to measure the quality and realism of the generated panoramas. In addition, we report the CLIP Score (CS) to evaluate text-image consistency. Since both FID and IS are based on InceptionNet [43] which is trained using perspective images only, we follow [56] to report a panoramic-customized metric Fréchet Auto-Encoder Distance (FAED) [28] for panorama evaluation.
- *Perspective.* We follow [56] to randomly sample 20 perspective views to simulate practical navigation on panoramas, and these views are evaluated based on FID and IS. Following [44], we also report FID, IS, and CS on 8 horizontally evenly spaced views.

As stated in [56], IS evaluates the diversity of objects within the generated image, as such, lower IS does not necessarily reflect the quality and realism of images in case models do not tend to generate unexpected objects. Similarly, as in [56], higher CS may be due to the repetition of objects to strengthen text-image alignment. On top of these, we also note that the 20 randomly sampled views may capture the top or the bottom of the panoramas, which are often



Methods	Peak Mem. <sup>†</sup> (GB)	Dur. <sup>†</sup> (hrs)	Panorama				20 Views		Horizontal 8 Views		
			FAED↓	FID↓	IS↑	CS↑	FID↓	IS↑	FID↓	IS↑	CS↑
SD+LoRA [15, 35, 56]	31.69	2.26	7.19	51.69	<u>4.40</u>	<b>28.83</b>	19.32	<u>6.90</u>	20.68	6.48	24.77
MVDiffusion [44]	26.66 (-15.9%)	9.86	-	-	-	-	-	-	25.27	<b>6.90</b>	<b>26.34</b>
Pano Only [56]	31.81 (+0.4%)	2.33	7.90	<u>50.40</u>	<b>4.54</b>	<u>28.67</u>	20.10	<b>7.06</b>	20.56	6.37	24.85
PanFusion [56]	60.12 (+89.7%)	6.61	<u>6.04</u>	<b>46.47</b>	4.36	28.58	<b>17.04</b>	6.85	<u>19.88</u>	<u>6.50</u>	<u>24.98</u>
UniPano (Ours)	32.59 (+2.8%)	3.43	<b>5.90</b>	<b>46.47</b>	4.16	28.37	<u>17.09</u>	6.74	<b>17.74</b>	6.00	24.82

Table 3. **Comparison between SoTA methods on 512×1024 panorama generation.** We quantitatively evaluate the panorama images based on Fréchet Auto-Encoder Distance (FAED) Fréchet Inception Distance (FID), Inception Score (IS), and CLIP Score (CS). We follow [56] to randomly sample 20 views from a panoramic image and [44] to horizontally sample 8 evenly spaced views to evaluate the quality of cropped perspective images. We report the peak allocated GPU memory (Peak Mem.) and time duration (Dur.) for 10-epoch training. All evaluated results are based on Stable Diffusion 2 base. †: results are reproduced with FP32 precision.



Figure 5. **Selected qualitative comparisons between UniPano (Ours) and PanFusion.** We show the generated panoramic image for each text prompt and 4 randomly sampled horizontal perspective views below. We highlight notable artifacts such as non-perspective lines with red boxes. More qualitative comparisons can be found in the supplementary material.

blurred even on real panoramic images, impacting the evaluation quality. *For the above reasons, we emphasize FAED and FID while caring horizontal FID more than 20-view FID among all evaluation metrics.*

## 4.2. Main Results

**Compared Methods.** We compare our uni-branch approach with several baseline methods.

- *MVDiffusion* [44] trains a multi-view diffusion model to simultaneously generate 8 horizontal views, which can be stitched into a panorama.
- *SD+LoRA* [15, 35] is the baseline method which fine-tunes the Stable Diffusion model [35] with LoRA [15] on panoramic images.
- *Pano Only* [56] is another baseline method introduced in [56] which additionally includes circular padding on top of SD+LoRA to ensure loop consistency.
- *PanFusion* [56] is the SoTA solution to date, which adopts a dual-branch approach and adds a cross-attention module between panoramic and perspective branches to en-

sure consistency.

**Quantitative Comparison.** We present the quantitative comparison in Tab. 3. Our UniPano achieves state-of-the-art FAED and horizontal FID, while being on par with PanFusion on FID and 20-view FID. We highlight that UniPano introduces minimal computational overhead, with an additional 2.8% allocated GPU memory and about 1 additional hour of training compared to SD+LoRA baseline. In comparison, training PanFusion [56] almost double the allocated GPU memory (+89.7%) and almost triple the time required for training compared to SD+LoRA baseline.

**Qualitative Comparison.** We showcase the qualitative comparison in Fig. 5. PanFusion may sometimes generate panoramic images with invalid equirectangular projection, evidenced by the notable artifacts (curvy and panoramic-like lines) in the regions of the perspective views highlighted with red boxes. With the boosted capacity of the panoramic-specific modules, UniPano faithfully generates

$W_{\{q,k\}}$	$W_v$	$W_o$	Panorama		20 Views	8 Views
			FAED↓	FID↓	FID↓	FID↓
LoRA	LoRA	LoRA	7.90	50.40	20.10	20.56
LoRA	LoRA	MoE	<u>7.21</u>	48.83	19.50	20.05
❄	LoRA	LoRA	7.99	<u>48.62</u>	<u>19.27</u>	<u>19.30</u>
❄	LoRA	MoE	<b>5.90</b>	<b>46.47</b>	<b>17.09</b>	<b>17.74</b>

Table 4. **Ablation study on fine-tuning strategies.** We compare several fine-tuning strategies for weights  $W_{\{q,k,v,o\}}$  in attention modules, with various combinations of freezing (❄), LoRA fine-tuning [15], and MoE [39, 49].

$n$	$k$	Panorama		20 Views	8 Views
		FAED↓	FID↓	FID↓	FID↓
2	2	6.75	51.01	<u>18.27</u>	19.71
4	2	<b>5.90</b>	<b>46.47</b>	<b>17.09</b>	<b>17.74</b>
8	2	<u>6.17</u>	<u>47.10</u>	20.19	<u>19.51</u>
8	4	7.31	47.61	22.05	21.12

Table 5. **Ablation study on the mixture of experts in  $W_o$ .** We experiment with different hyperparameters for the mixture of experts (MoE) by adjusting the number of experts  $n$ , and selecting top- $k$  experts with the highest weighting.

panoramic images that follow equirectangular projection in the illustrated cases. More qualitative comparisons are deferred to the supplementary material.

### 4.3. Ablation Study

**Different fine-tuning strategies.** We compare several different fine-tuning strategies for trainable weights  $W_{\{q,k,v,o\}}$ , in Tab. 4, with different combinations of the following: freezing, fine-tuning with LoRA [15], and with mixture of experts (MoE) [10, 39, 49]. The panorama branch only baseline (highlighted in gray) fine-tunes all trainable components with LoRA and is thereby considered as the baseline. Switching to MoE for fine-tuning  $W_o$  contributes to a notable improvement in FAED and FID metrics, demonstrating the benefits of increasing the capacity for panoramic-specific components. Based on the baseline, we further experiment with keeping  $W_{\{q,k\}}$  frozen throughout fine-tuning, resulting in improved FID scores while the FAED remained comparable to the baseline. This result is in line with our analysis that  $W_{\{q,k\}}$  relate to non-parametric-specific information. Merging these two strategies – *i.e.*, freezing  $W_{\{q,k\}}$  and fine-tuning  $W_o$  with MoE – yields our state-of-the-art UniPano.

**Different settings for mixture of experts.** We additionally ablate different sets of hyperparameters for the mixture of experts (MoE) in Tab. 5, namely the number of experts  $n$  and the number of top- $k$  selected experts per token. Our first observation is that unlike many other MoE appli-



Figure 6. **Failure cases of UniPano.** Similar to PanFusion [56], UniPano sometimes generates scenes with invalid layouts, such as rooms without entrances.

cations [10, 39, 49], the performance in our context saturates for a relatively small amount of experts ( $n = 4$ ), evidenced by the notable deterioration in FID-related metrics when scaling  $n$  from 4 to 8. This is likely due to the relative simplicity of fine-tuning a pre-trained diffusion model, in comparison to the typical use cases of MoE *i.e.* training the entire model from scratch. Additionally, aligning with the previous work [10, 39, 49], we find that sparsity is critical, as increasing the number of used experts per token  $k$  from 2 to 4 reduces the FAED and FID metrics rapidly.

### 4.4. Scaling to Higher Resolution

Panoramic images store the entire 360-degree surrounding scenes within one equirectangular image, cropping perspective views from a panoramic image up to  $512 \times 1024$  thus still leading to somewhat low-resolution images. This motivates the importance of scaling panorama generation to higher resolution, which is typically achieved with a separate super-resolution stage [7]. As a direct benefit of lowering the memory burden of PanFusion, UniPano can be readily scaled for higher-resolution panorama generation in an end-to-end manner. As Stable Diffusion 2 base is optimized for generating images up to  $512 \times 512$ , directly adapting it to generate higher-resolution images leads to suboptimal results. We thus adopt the state-of-the-art Stable Diffusion 3 [9] which natively supports  $1024 \times 1024$  image generation. We show UniPano based on Stable Diffusion 3 can generate realistic  $1024 \times 2048$  panoramic images, and is robust to out-of-distribution prompts and extremely complex prompts, in Fig. 1. More high-resolution results and experimental details are deferred to the supplementary material.

## 5. Conclusion

We have elucidated the underlying mechanism that facilitates low-rank adaptation of pre-trained perspective diffusion models to panorama generation. Particularly, our analysis reveals that the query and key matrices ( $W_{\{q,k\}}$ ) learn common semantic information that can be shared between the panoramic and perspective domains, whereas the value and output matrices ( $W_{\{v,o\}}$ ) specialize in capturing the equirectangular structure of panoramic images. Based on these insights, we propose *UniPano*, which outperforms and reduces the memory and training burden compared to the previous dual-branch approach.



**Limitations.** The primary focus of this paper is to investigate the underlying behaviors of the trainable components within LoRAs when adapting pre-trained perspective diffusion models to panorama generation. While our UniPano reports the state-of-the-art results, given the abundance of hyperparameters, the performance is still possibly far from optimal. Additionally, similar to the drawbacks of PanFusion [56], we find that UniPano sometimes generates scenes with invalid layouts, such as rooms without entrances, as shown in Fig. 6.

## References

- [1] Chief architect 360° panorama renderings. <https://www.chiefarchitect.com/products/360-panorama-viewer>. Accessed: 2025-03-03. 11
- [2] Online 360° panorama viewer vr. <https://renderstuff.com/tools/360-panorama-web-viewer>. Accessed: 2025-03-03. 11
- [3] Naofumi Akimoto, Yuhi Matsuo, and Yoshimitsu Aoki. Diverse plausible 360-degree image outpainting for efficient 3deg background creation. In *CVPR*, 2022. 3
- [4] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. In *ICML*, 2023. 3
- [5] 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, 6, 11
- [6] Shoufa Chen, Peize Sun, Yibing Song, and Ping Luo. Diffusiondet: Diffusion model for object detection. In *ICCV*, 2023. 3
- [7] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *TOG*, 2022. 6, 8
- [8] Zilong Chen, Feng Wang, Yikai Wang, and Huaping Liu. Text-to-3d using gaussian splatting. In *CVPR*, 2024. 3
- [9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 3, 8
- [10] Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Scaling diffusion transformers to 16 billion parameters. *arXiv preprint arXiv:2407.11633*, 2024. 6, 8
- [11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3
- [12] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *CVPR*, 2023. 6
- [13] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 11
- [14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3
- [15] 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 *ICLR*, 2022. 1, 3, 6, 7, 8, 11
- [16] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 6
- [17] Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. Cubediff: Repurposing diffusion-based image models for panorama generation. In *ICLR*, 2025. 1, 3
- [18] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 1, 3
- [19] Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. In *NeurIPS*, 2023. 1, 3
- [20] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 6
- [21] Renjie Li, Panwang Pan, Bangbang Yang, Dejia Xu, Shijie Zhou, Xuanyang Zhang, Zeming Li, Achuta Kadambi, Zhangyang Wang, Zhengzhong Tu, et al. 4k4dgen: Panoramic 4d generation at 4k resolution. *arXiv preprint arXiv:2406.13527*, 2024. 1
- [22] Aoming Liu, Zhong Li, Zhang Chen, Nannan Li, Yi Xu, and Bryan A Plummer. Panofree: Tuning-free holistic multi-view image generation with cross-view self-guidance. In *ECCV*, 2024. 3
- [23] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023. 3
- [24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6
- [25] Zhuqiang Lu, Kun Hu, Chaoyue Wang, Lei Bai, and Zhiyong Wang. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. In *AAAI*, 2024. 3
- [26] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 3
- [27] Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *CVPR*, 2023. 3
- [28] Changgyoon Oh, Wonjune Cho, Yujeong Chae, Daehee Park, Lin Wang, and Kuk-Jin Yoon. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *ECCV*, 2022. 3, 6
- [29] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 11
- [30] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *ICLR*, 2024. 3

- [31] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 3
- [32] Jack Rae and Ali Razavi. Do transformers need deep long-range memory? In *ACL*, 2020. 6
- [33] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 1, 3
- [34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.
- [35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 7, 11
- [36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 11
- [37] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1, 3
- [38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 1
- [39] Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017. 8
- [40] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2020. 3, 11
- [41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021. 3
- [42] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *ICML*, 2023. 3
- [43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 6
- [44] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *NeurIPS*, 2023. 1, 3, 6, 7, 11
- [45] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, William Berman, Yiyi Xu, Steven Liu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>, 2022. 11
- [46] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *ECCV*, 2022. 3
- [47] Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *WACV*, 2024. 3
- [48] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. Panodiffusion: 360-degree panorama outpainting via diffusion. In *ICLR*, 2024. 3
- [49] Xun Wu, Shaohan Huang, and Furu Wei. Mixture of lora experts. In *ICLR*, 2024. 6, 8
- [50] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *CVPR*, 2022. 6
- [51] Shaoan Xie, Zhifei Zhang, Zhe Lin, Tobias Hinz, and Kun Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *CVPR*, 2023. 3
- [52] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 3
- [53] Bangbang Yang, Wenqi Dong, Lin Ma, Wenbo Hu, Xiao Liu, Zhaopeng Cui, and Yuewen Ma. Dreamspace: Dreaming your room space with text-driven panoramic texture propagation. In *IEEE VR*, 2024. 1
- [54] Weicai Ye, Chenhao Ji, Zheng Chen, Junyao Gao, Xiaoshui Huang, Song-Hai Zhang, Wanli Ouyang, Tong He, Cairong Zhao, and Guofeng Zhang. Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. In *NeurIPS*, 2024. 3
- [55] Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 2023. 3
- [56] 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, 3, 4, 6, 7, 8, 9, 11, 12, 13
- [57] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 1
- [58] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. In *NeurIPS*, 2023. 3
- [59] Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. In *NeurIPS*, 2023. 3



## A. Preliminary on Diffusion Models

For completeness sake, we provide preliminary on diffusion models, particularly latent diffusion models, below.

Diffusion models involve iteratively transforming the noise into the target data. The sampling step thus requires learning a time-conditioned noise (or equivalently the score function) prediction network  $\epsilon_\theta$ , often in the form of U-Net [36] or transformers [29]. In practice, to optimize efficiency and performance, the denoising process is generally performed in the latent space of a pre-trained encoder  $\mathcal{E}$ , which leads to the training objective:

$$\min_{\theta} \mathbb{E}_{t \sim \mathcal{U}(0, T), (x, y) \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(0, \text{Id})} [\|\epsilon_\theta(z_t, t, y) - \epsilon\|^2],$$

where  $\mathcal{U}$  is the uniform distribution,  $p_{\text{data}}$  denotes the data distribution for which each sample contains an input image  $x$  and an input condition  $y$  (which is text in our context),  $z_t = \alpha(t)\mathcal{E}(x) + \beta(t)\epsilon$  is the noisy latent at a timestep  $t$  with  $\alpha(t)$  and  $\beta(t)$  defining the diffusion trajectory, and  $T$  is the largest timestep such that  $z_T \sim \mathcal{N}(0, \text{Id})$ . During sampling, a random noise  $z_T$  is first drawn from the prior distribution, and gradually denoised to the clean latent  $z_0$  by the learned denoising network  $\epsilon_\theta$  following a pre-defined noise schedule. The clean latent is finally converted into the image space using the pre-trained decoder  $\mathcal{D}$ .

## B. Experimental Details

We provide more details on the experimental setup of  $512 \times 1024$  panorama generation below.

**Implementation Details.** Our implementation is based on Stable Diffusion from `diffusers` [45]. In addition to the implementation details listed in the main article, we strictly follow MVDiffusion [44] and PanFusion [56] using the DDIM sampler [40] with 50 sampling steps and classifier-free guidance scale [13] of 9 for inference.

**Compared Methods.** We provide more details on the compared methods and the reported results in Tab. 3 below.

- MVDiffusion [44] trains diffusion models with multi-view awareness, which generate 8 horizontal perspective views simultaneously. These images can then be stitched into a panorama, however, we note that the panoramas are incomplete due to the missing top and bottom regions. The reported results are directly taken from [56], where the only difference with the original MVDiffusion paper is to downsample to  $256 \times 256$  for evaluation to match the resolution of the ground truth images.
- SD+LoRA [15, 35] directly fine-tunes Stable Diffusion with LoRA [15] on panoramic images, which is a standard technique for adapting pre-trained diffusion models

for downstream tasks. The reported metrics are directly taken from [56].

- Pano Only [56] is a baseline method proposed in [56] which adopts circular padding on top of SD+LoRA to ensure loop consistency. The reported metrics are again taken from [56].
- PanFusion [56] is the state-of-the-art solution to date and our most important baseline method. It adopts a dual-branch approach, consisting of a panoramic and a perspective branch. It proposes an equirectangular-perspective projection attention module to establish a correspondence between these two branches to ensure consistency. The reported metrics are directly taken from the original PanFusion paper [56].

## C. Additional Qualitative Comparisons

In addition to the qualitative results in Sec. 4.2, we showcase more qualitative comparisons in Figs. 7 and 8. We randomly sample 4 horizontal perspective views below each generated panoramic image. One may also use panorama viewer (e.g. [1, 2]) to freely navigate the panoramas.

## D. Higher-resolution Panorama Generation

### D.1. Implementation Details

The setup for our higher-resolution generation experiments besides the base model is identical to Sec. 4.1. As the current SoTA PanFusion is not capable of generating  $1024 \times 2048$  panoramic images, we emphasize that our experiments serve primarily as illustrations rather than comparisons with current baseline models. Another special note is that since Stable Diffusion 3 is based on transformer architectures, for which circular padding cannot be trivially applied and thus has been left out for our implementation.

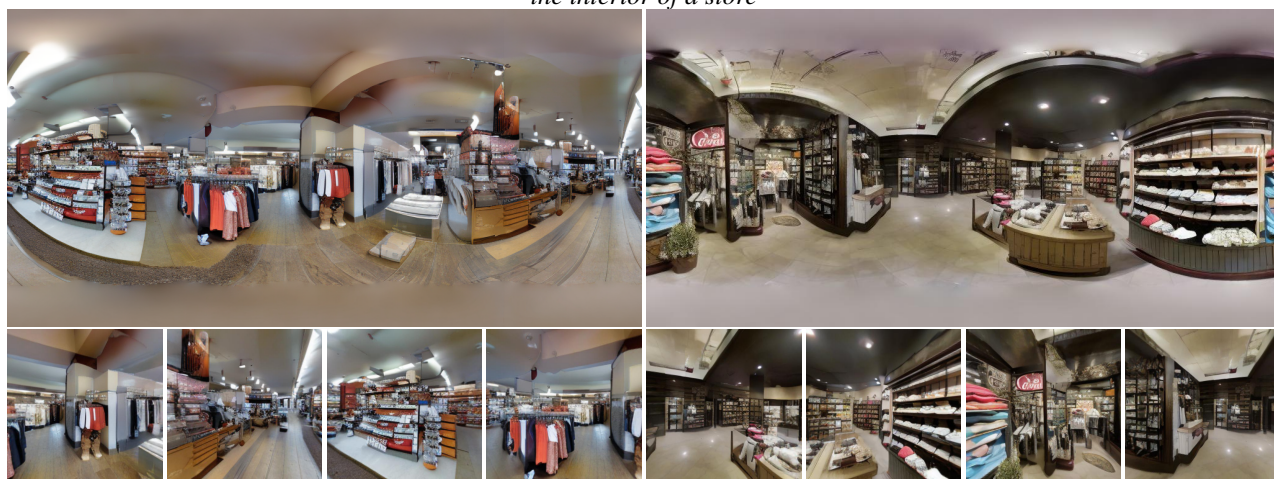
### D.2. Additional High-resolution Results

We present the qualitative results for scaling UniPano to generate  $1024 \times 2048$  panoramic images. We provide more qualitative results in Fig. 9. To illustrate the power of implementing UniPano on a more powerful base model, we showcase the results with out-of-distribution text prompts in Figs. 10 to 12 and with extremely long and complex text prompts in Fig. 13. We refer the reader to the semantic class distribution of Matterport3D in [5, Fig 5] for the definition of in- and out-of-distribution.

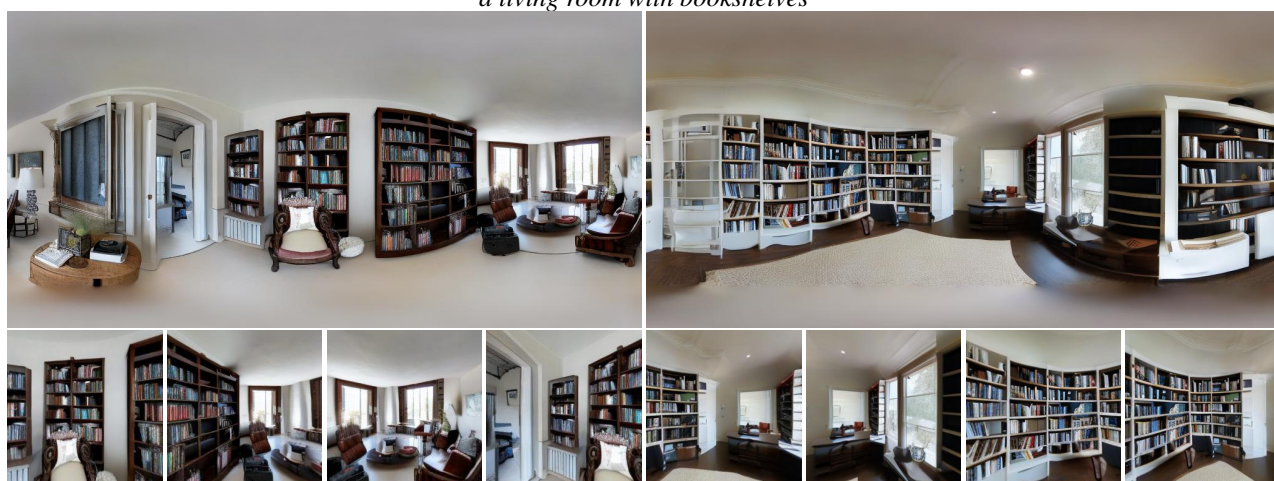
*“a bedroom with a ceiling fan”*



*“the interior of a store”*



*“a living room with bookshelves”*



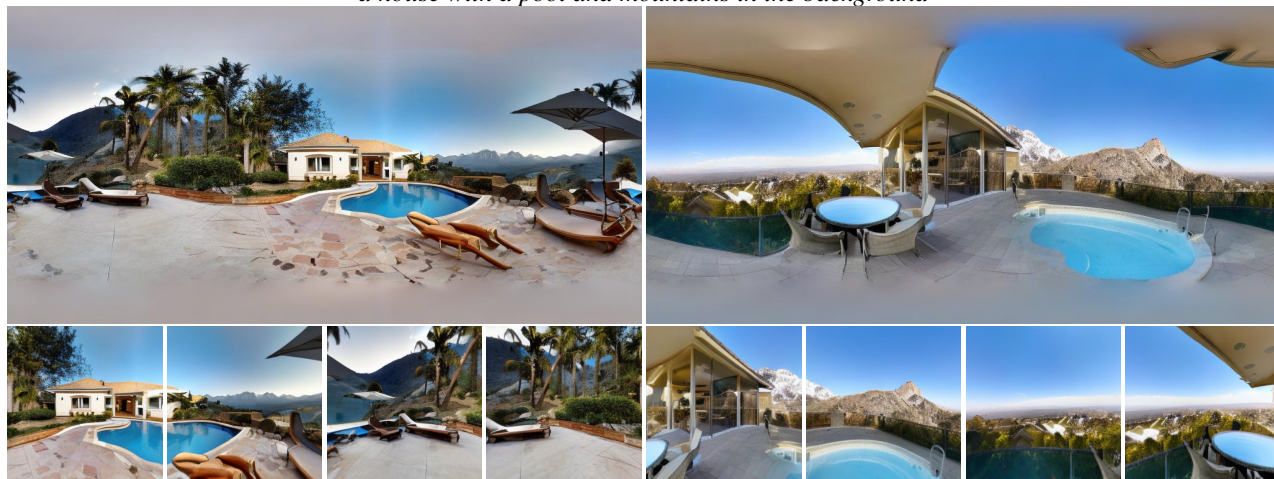
UniPano (Ours)

PanFusion [56]

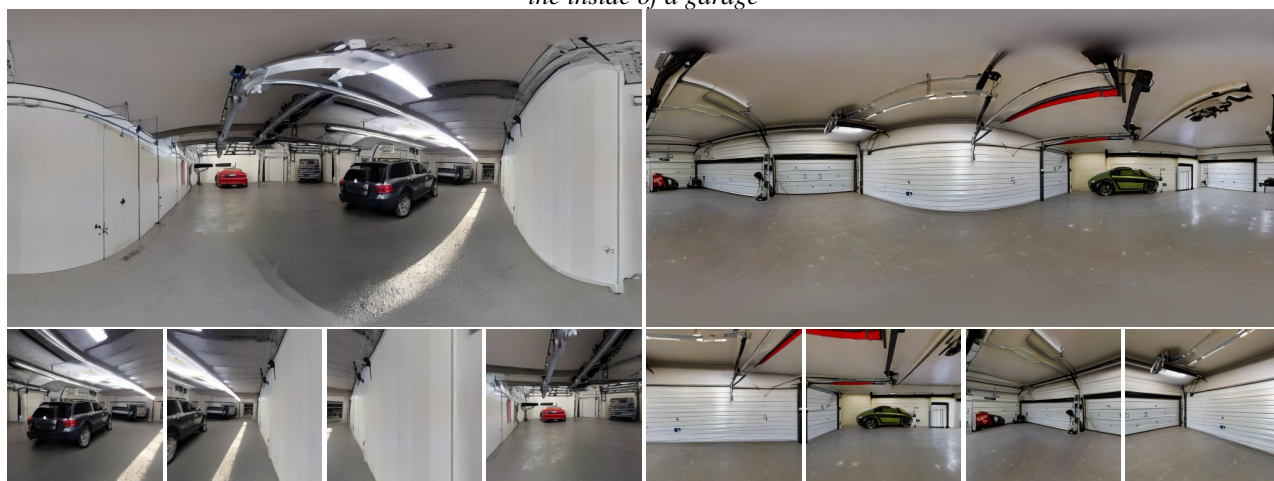
Figure 7. Additional qualitative comparisons.



*“a house with a pool and mountains in the background”*



*“the inside of a garage”*



*“a hallway in a luxury home”*



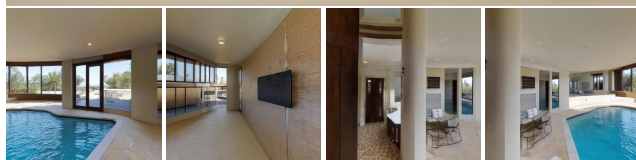
UniPano (Ours)

PanFusion [56]

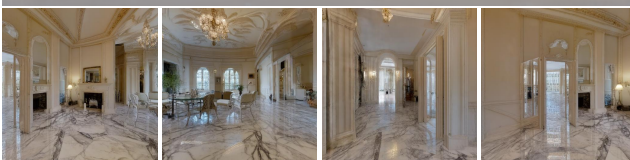
Figure 8. Additional qualitative comparisons.



*“a room with a swimming pool”*



*“a room with marble floors”*



*“the inside of a home”*



*“a garage with a car in it”*



Figure 9. Additional high-resolution ( $1024 \times 2024$ ) results. Note that all results are generated using UniPano based on Stable Diffusion 3.



*“Exploring the historic streets of Prague, with its charming architecture, cobblestone alleys, and medieval ambiance.”*



*“A traditional Italian trattoria, where locals gather for hearty meals, laughter, and the warmth of shared conversation.”*



Figure 10. Additional high-resolution results for out-of-distribution prompts.



*“Alpine village, snow-covered rooftops, nestled between majestic peaks—a picture-perfect scene of winter tranquility.”*



*“Exploring an abandoned underwater city, where sunken buildings are now home to schools of bioluminescent fish.”*

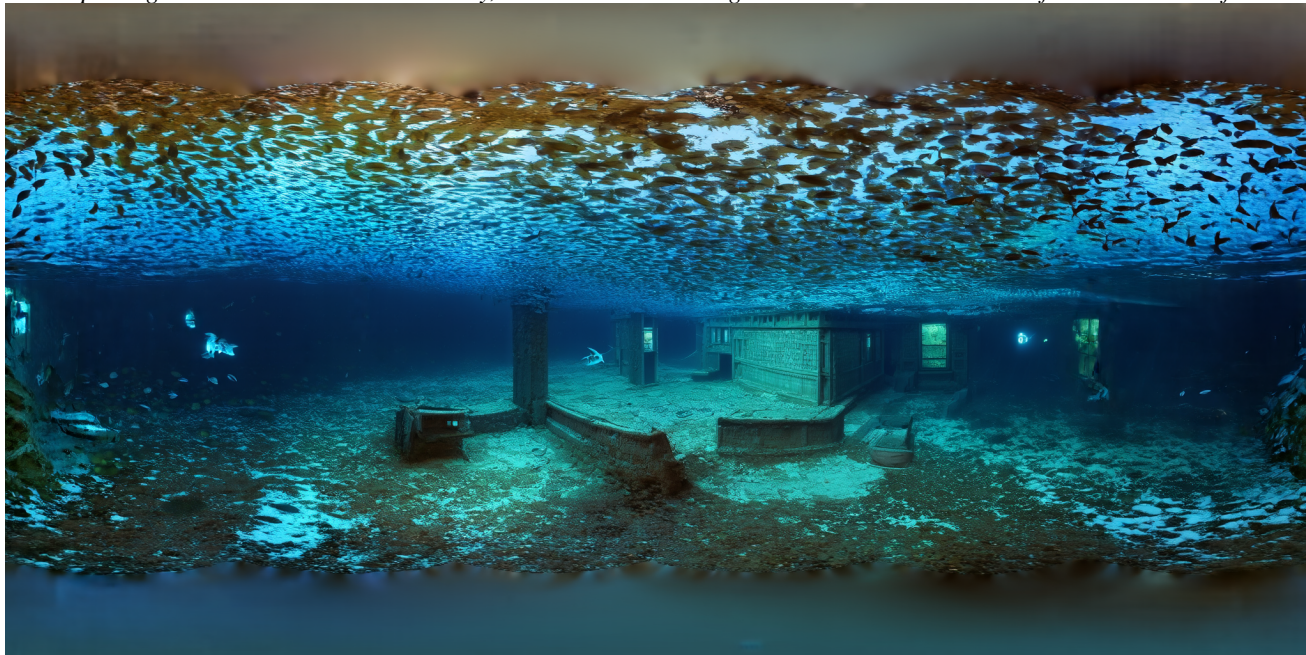
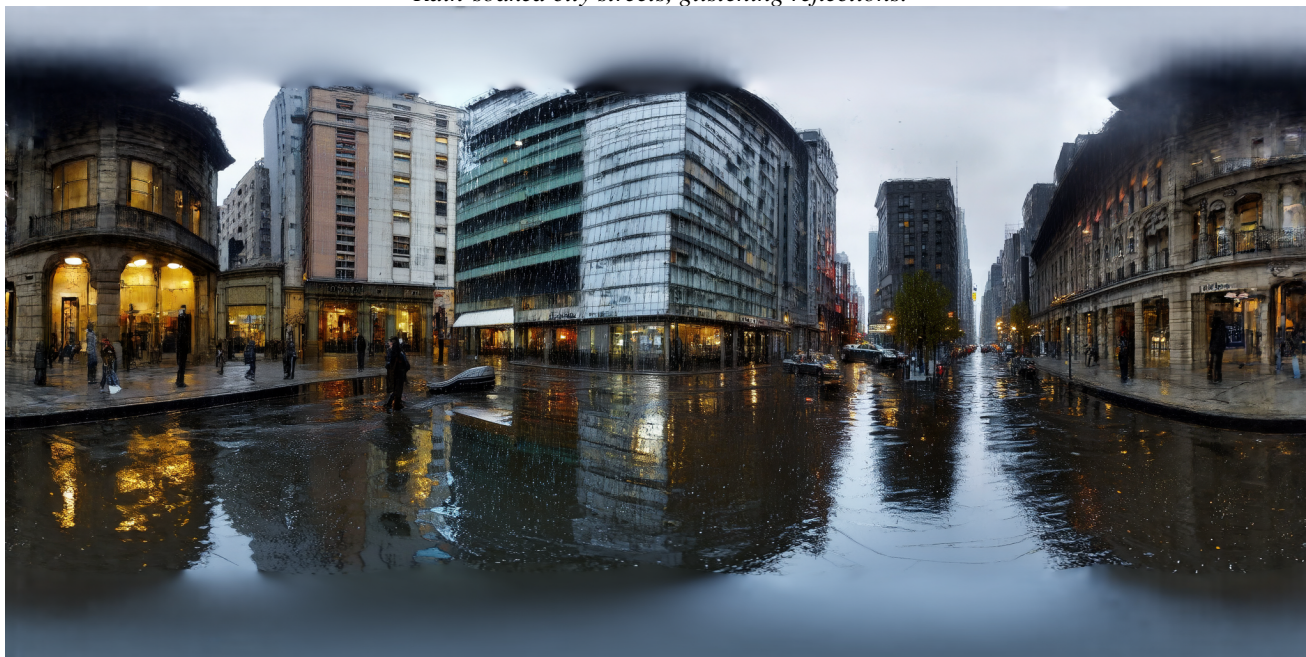


Figure 11. Additional high-resolution results for out-of-distribution prompts.



*“Rain-soaked city streets, glistening reflections.”*



*“Cobblestone alley, historic architecture bathed in soft morning light.”*



Figure 12. Additional high-resolution results for out-of-distribution prompts.



*“On a distant planet’s surface, towering crystalline structures rise against an alien sky. The landscape is surreal, with bioluminescent flora casting an otherworldly glow. Strange creatures move gracefully through the phosphorescent mist, creating an ethereal scene that defies earthly imagination.”*



*“Amidst the bustling energy of a busy market, vendors peddle their wares with animated fervor. A kaleidoscope of colors, from fresh produce to woven textiles, creates a vibrant tapestry. The air is thick with the mingling scents of spices, street food, and the lively chatter of buyers and sellers.”*



Figure 13. Additional high-resolution results with complex prompts.