What Makes for Text to 360-degree Panorama Generation with Stable Diffusion?
– Supplementary Material –

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 ϵ_θ, often in the form of U-Net [34] or transformers [27]. 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:

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

where $\mathcal{U}$ is the uniform distribution, p_{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 $z_T \sim \mathcal{N}(0, \mathrm{Id})$. 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 ϵ_θ 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 512×1024 panorama generation below.

Implementation Details. Our implementation is based on Stable Diffusion from `diffusers` [43]. In addition to the implementation details listed in the main article, we strictly follow MVDiffusion [42] and PanFusion [52] using the DDIM sampler [38] with 50 sampling steps and classifier-free guidance scale [12] of 9 for inference.

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

- MVDiffusion [42] 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 [52], where the only difference with the original MVDiffusion paper is to downsample to 256×256 for evaluation to match the resolution of the ground truth images.
- SD+LoRA [14, 33] directly fine-tunes Stable Diffusion with LoRA [14] on panoramic images, which is a standard technique for adapting pre-trained diffusion models for downstream tasks. The reported metrics are directly taken from [52].
- Pano Only [52] is a baseline method proposed in [52] which adopts circular padding on top of SD+LoRA to ensure loop consistency. The reported metrics are again taken from [52].
- PanFusion [52] 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 [52].

Methods	Win Rate (↑)
PanFusion	0.47
UniPano (**Ours**)	0.53

Table 6. **User Study**.

C. Additional Qualitative Comparisons

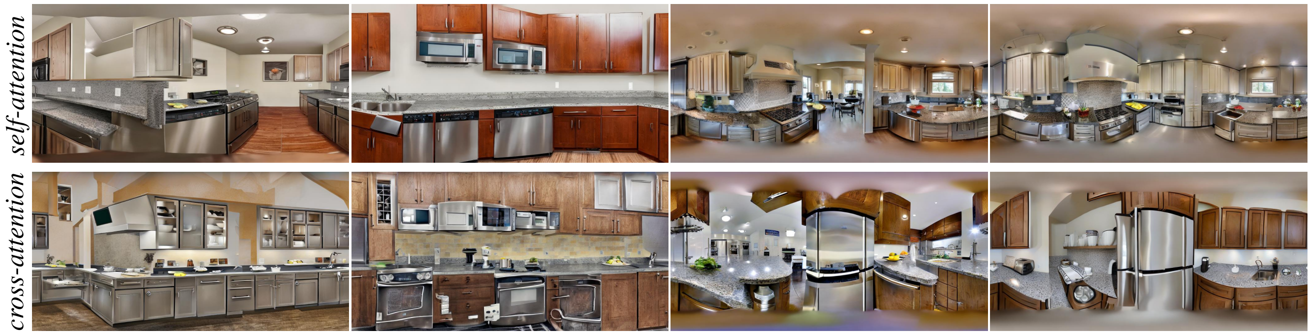
In addition to the qualitative results in Sec. 4.2, we showcase more qualitative comparisons in Figs. 9 and 10. We randomly sample 4 horizontal perspective views below each generated panoramic image. One may also use panorama viewer (*e.g*. [1, 2]) to freely navigate the panoramas.

D. User Study

We generate 25 text prompts with `ChatGPT`, and ask 12 human annotators to select their preferred one from panoramas generated by PanFusion and UniPano for each text prompt. We report the win rate (successful trials / total trials) in Tab. 6. We highlight that UniPano is much more efficient in terms of GPU memory and training time than PanFusion, while achieving on-par and even better performance.

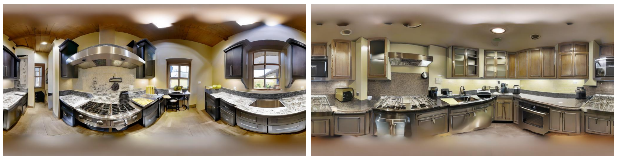
E. Additional Ablation Studies

Quantitative results of Fig. 4. We first present the quantitative results of Fig. 4 in Tab. 7. We evaluate the quality of panoramic and perspective inference by FAED and FID respectively. Since all LoRAs are activated during training, disabling some LoRAs during inference inevitably lowers



Training W_q in isolation Training W_k in isolation Training W_v in isolation Training W_o in isolation

Figure 7. **Qualitative comparison for training $W_{\{q,k,v,o\}}$ in isolation for self- and cross-attention**. Similar to Fig. 3, training W_q or W_k in isolation fails to capture the spherical structure in the case of either self- or cross-attention; training W_v or W_o in isolation manages to capture the distortion.



Training W_q in isolation (FFT) Training W_k in isolation (FFT)

Figure 8. **Qualitative comparison for training $W_{\{q,k\}}$ in isolation with full fine-tuning (FFT)**. FFT on $W_{\{q,k\}}$ successfully generates panoramic images.

	FAED↓	Perspective FID↓
Pano Only	7.90	86.50
- Disabling $W_{\{v,o\}}$	-	68.72
- Disabling $W_{\{q,k\}}$	9.69	-

Table 7. **Quantitative evaluation of roles of $W_{\{q,k,v,o\}}$ when jointly fine-tuned**. We evaluate the ability to generate panoramas and perspective images by FAED and FID respectively.

			Panorama		20 Views	8 Views
$W_{\{q,k\}}$	W_v	W_o	FAED↓	FID↓	FID↓	FID↓
LoRA	LoRA	LoRA	7.90	50.40	20.10	20.56
MoE	❄	❄	7.77	67.15	25.90	22.21
❄	LoRA	MoE	**5.90**	**46.47**	**17.09**	**17.74**

Table 8. **Additional Ablation Study**. Solely relying on $W_{\{q,k\}}$ with increased capacity (blue row) leads to deteriorated metrics.

the quality. We confirm that disabling $W_{\{q,k\}}$ still generates panorama, and disabling $W_{\{v,o\}}$ generates perspective images, as shown in Fig. 4.

Additional experiments on MoE. We then present an additional ablation study on MoE, where we solely increase the capacity of $W_{\{q,k\}}$ using MoE. The results are in Tab. 8. Solely relying on enhanced $W_{\{q,k\}}$ leads to significantly deteriorated FIDs compared to the Pano Only baseline.

	Panorama		20 Views	8 Views
	FAED↓	FID↓	FID↓	FID↓
W_q	14.84	82.17	36.81	37.29
W_k	14.77	81.03	37.50	42.57
W_v	13.48	62.34	31.12	33.42
W_o	**8.67**	**61.92**	**26.91**	**32.32**

Table 9. **Quantitative comparison for training $W_{\{q,k,v,o\}}$ in isolation on DiT**. We observe similar patterns as in Tab. 1, where training $W_{\{v,o\}}$ leads to notable improved metrics compared to $W_{\{q,k\}}$

	Panorama		20 Views	8 Views
	FAED↓	FID↓	FID↓	FID↓
W_q	9.77	48.51	19.87	20.20
W_k	9.45	50.68	19.77	19.17
W_v	**6.28**	47.84	**17.55**	**19.09**
W_o	8.46	**46.63**	17.59	19.33

Table 10. **Quantitative comparison for training $W_{\{q,k,v,o\}}$ in isolation with full fine-tuning**. Training $W_{\{v,o\}}$ still generates panoramas of higher quality than $W_{\{q,k\}}$, but the gap is considerably closer compared to Tab. 1.

Generalization of insights. We conduct additional experiments of isolating attention matrices – similar to Tab. 1 – on DiT-based Stable Diffusion 3, in Tab. 9. These results affirm that our insights generalize to other architectures.

Disentangling self- and cross-attention. We further disentangle the attention mechanisms and consider them separately for analyzing their behaviors. We conduct experiments similar to Fig. 3 on cross-attention and self-attention only in Fig. 7. It is obvious that our insights hold for both cases.

Full fine-tuning. Lastly, we conduct experiments of isolating attention matrices with full fine-tuning (FFT) rather

Methods	Condition	Training Size	FAED↓
PanFusion	Text	11K	71.7
UniPano (**Ours**)	Text	11K	**46.8**
CubeDiff	Image	48K	22.0
CubeDiff	Image+Text	48K	**18.4**
UniPano (**Ours**)	Text	48K	36.9

Table 11. **Quantitative comparison of SoTA methods on LAVAL Indoor**. We highlight the difference in the conditioning scheme and the size of the training data. UniPano, with the same setting, outperforms PanFusion by a significant margin.

than LoRA fine-tuning. As shown in Fig. 8 and Tab. 10, FFT on $W_{\{q,k\}}$ successfully generates panoramic images, despite lower quality compared to $W_{\{v,o\}}$. This implies that $W_{\{q,k\}}$ can also contribute to panorama generation given sufficient capacity, while $W_{\{v,o\}}$ exhibit better efficacy.

F. Additional Experiments on LAVAL Indoor

We conduct additional experiments for comparison on LAVAL Indoor with CubeDiff [16] and PanFusion [52]. We highlight that CubeDiff requires mandatory image conditioning, whereas PanFusion and our method are purely text-based. Aside from this different problem setting, to facilitate fairer comparisons, we additionally train UniPano on the same datasets consisting of around 48K panoramas. We train UniPano for 10 epochs and have observed improvements by using a larger dataset. The results are shown in Tab. 11.

G. Higher-resolution Panorama Generation

G.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 × 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.

G.2. Quantitative Comparison

We provide the quantitative comparison in Tab. 12. We follow [52] to train an autoencoder on 1024 × 2048 panoramic images for 60 epochs, and its latent space is used to compute FAED. As the dual-branch approach PanFusion [52] requires over 80 GB of GPU memory to train, we are not able to reproduce their results on 1024 × 2048 panoramas and thus have excluded it from our comparison. UniPano outperforms the baseline of Stable Diffusion fine-tuned with LoRA (SD+LoRA) on important evaluation metrics, including FAED, 20-view, and horizontal 8-view FID. The gap between the 20-view and horizontal 8-view FID is likely due to excluding circular padding, resulting in notable artifacts in the lapping regions.

	Panorama		20 Views	8 Views
	FAED↓	FID↓	FID↓	FID↓
SD+LoRA [14, 33]	25.80	**55.03**	22.48	28.78
UniPano (**Ours**)	**24.47**	55.47	**21.66**	**28.36**

Table 12. **Quantitative comparisons on 1024×2048 panorama generation**. The dual-branch approach PanFusion [52] requires more than 80 GB GPU memory to train on this resolution, thus is infeasible to train even with the finest GPUs to date.

G.3. Additional High-resolution Results

We present the qualitative results for scaling UniPano to generate 1024 × 2048 panoramic images. We provide more qualitative results in Fig. 11. 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. 12 to 17 and with extremely long and complex text prompts in Figs. 18 and 19. We refer the reader to the semantic class distribution of Matterport3D in [5, Fig 5] for the definition of in- and out-of-distribution.

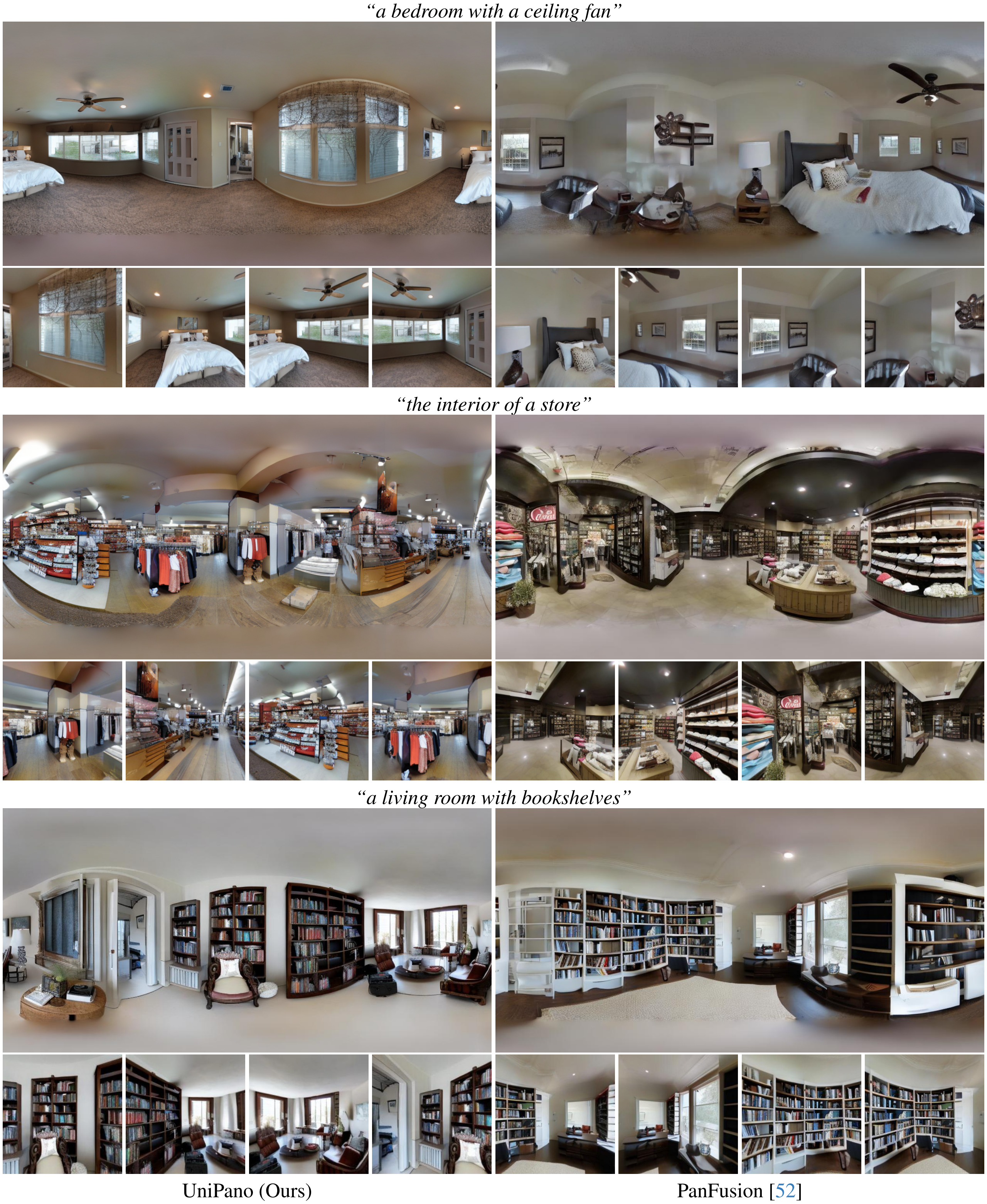

"a bedroom with a ceiling fan"
"the interior of a store"
"a living room with bookshelves"
UniPano (Ours)
PanFusion [52]

Figure 9. Additional qualitative comparisons.

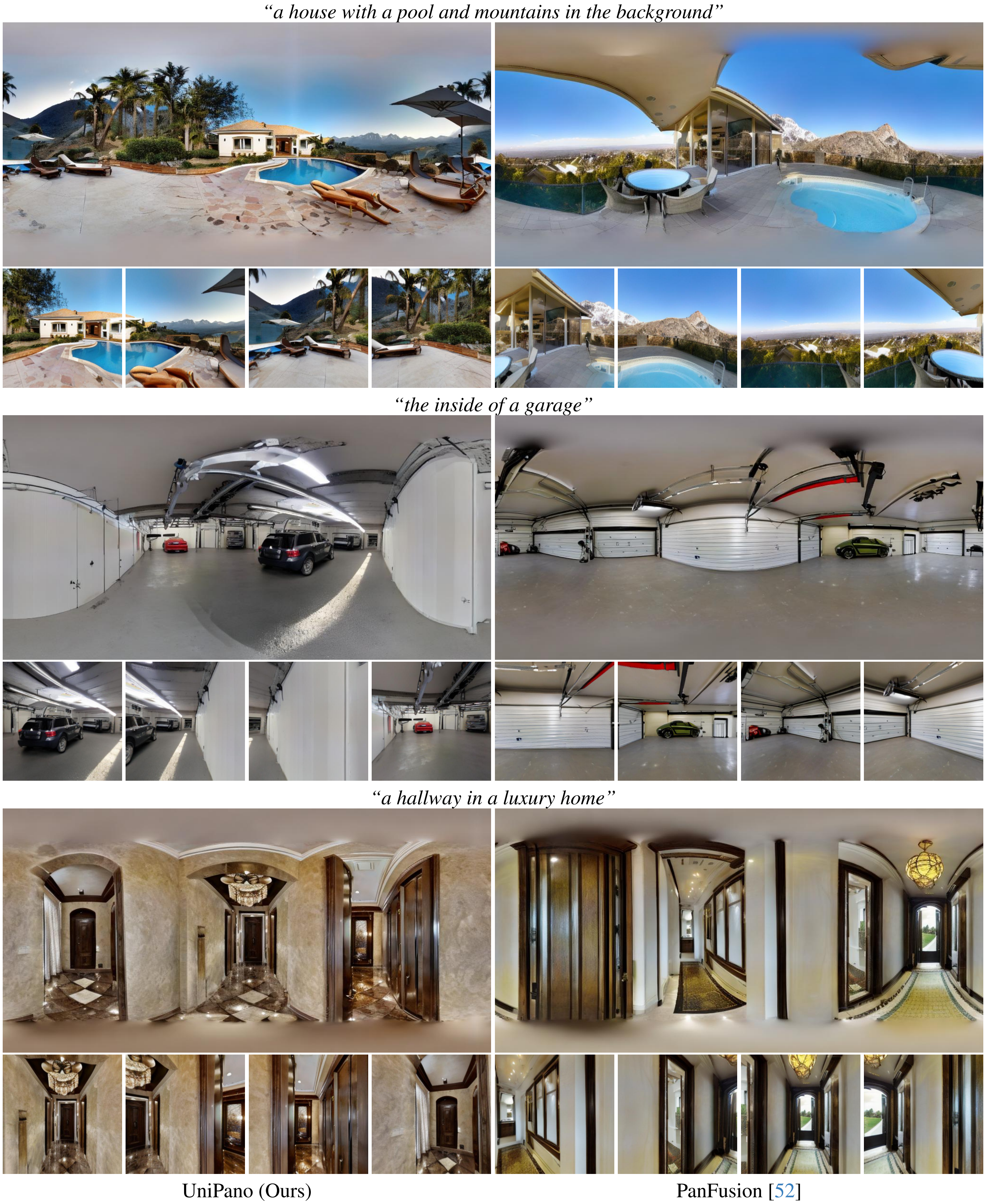
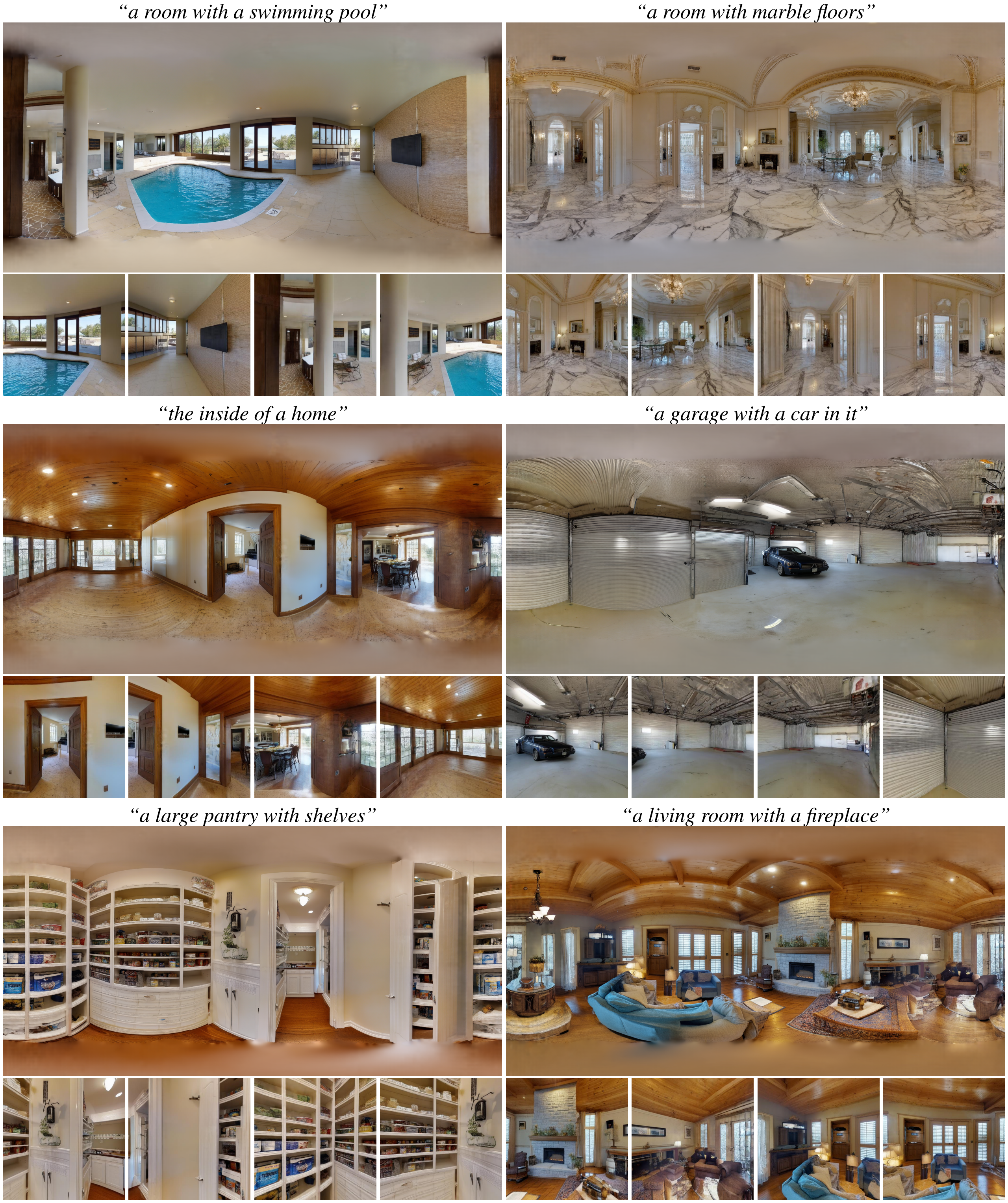

“a house with a pool and mountains in the background”
“the inside of a garage”
“a hallway in a luxury home”
UniPano (Ours)
PanFusion [52]

Figure 10. Additional qualitative comparisons.



“a room with a swimming pool”

“a room with marble floors”

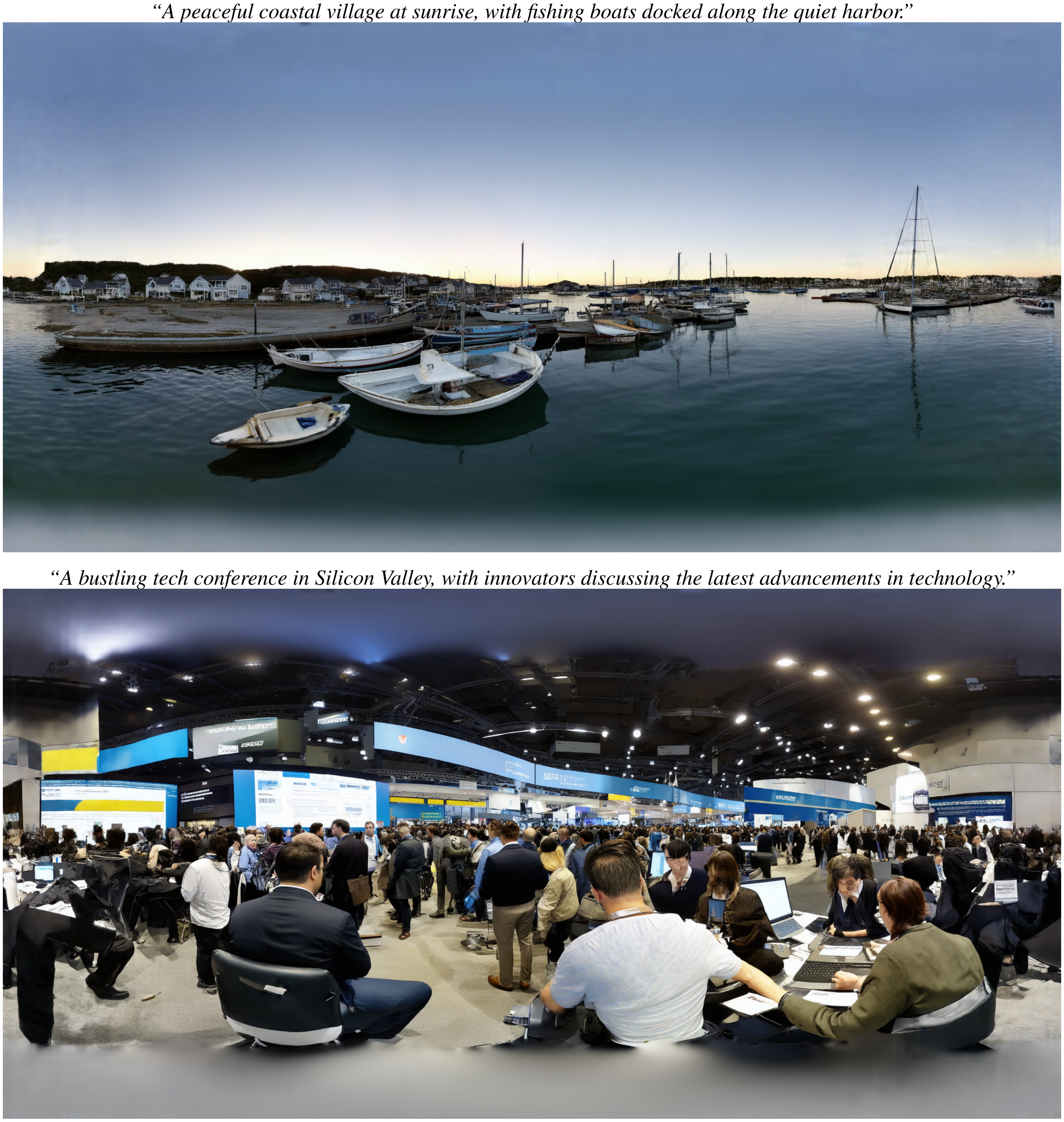
“the inside of a home”

“a garage with a car in it”

“a large pantry with shelves”

“a living room with a fireplace”

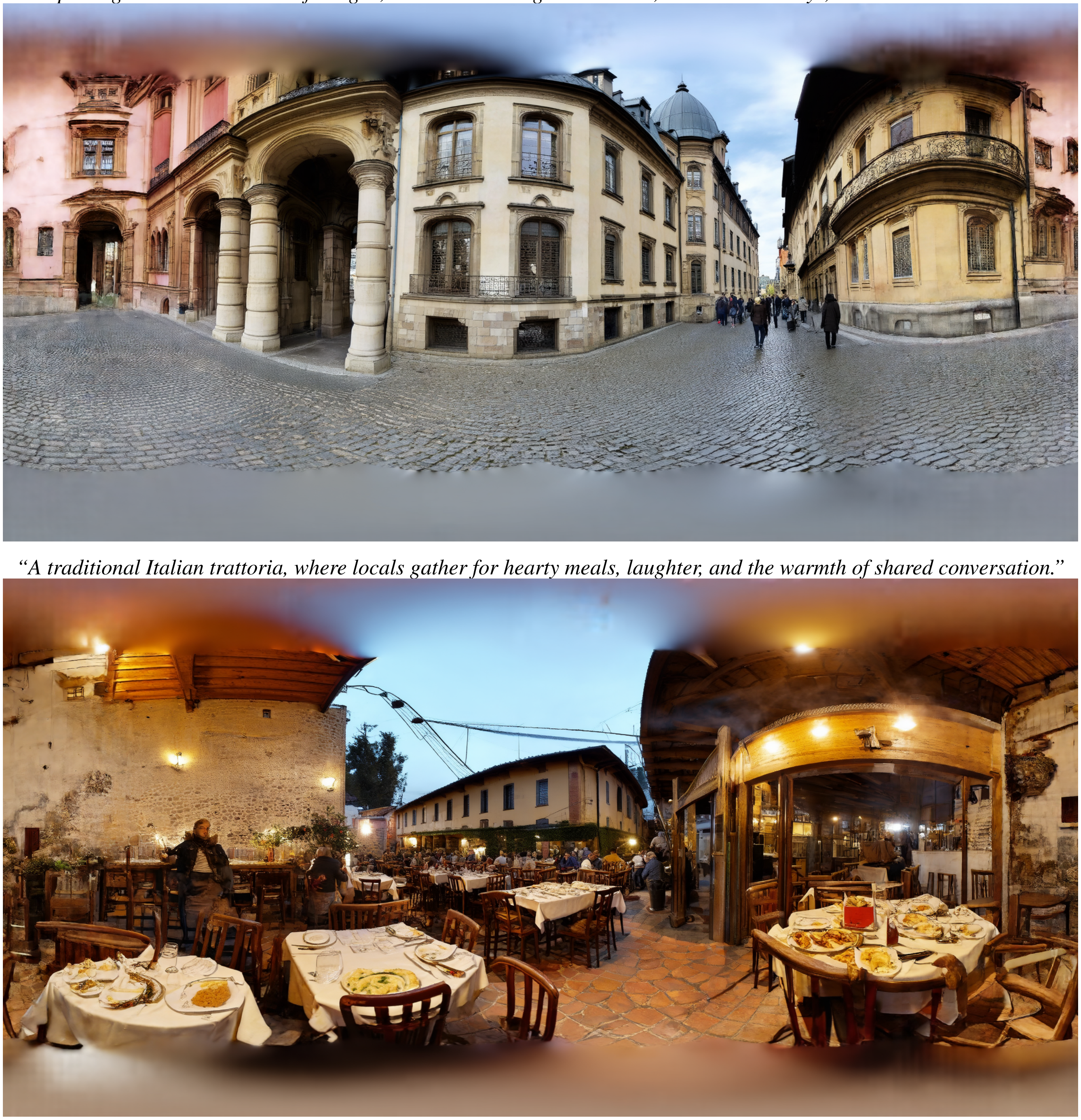
Figure 11. Additional high-resolution (1024 × 2024) results. Note that all results are generated using UniPano based on Stable Diffusion 3.



"A peaceful coastal village at sunrise, with fishing boats docked along the quiet harbor."

"A bustling tech conference in Silicon Valley, with innovators discussing the latest advancements in technology."

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

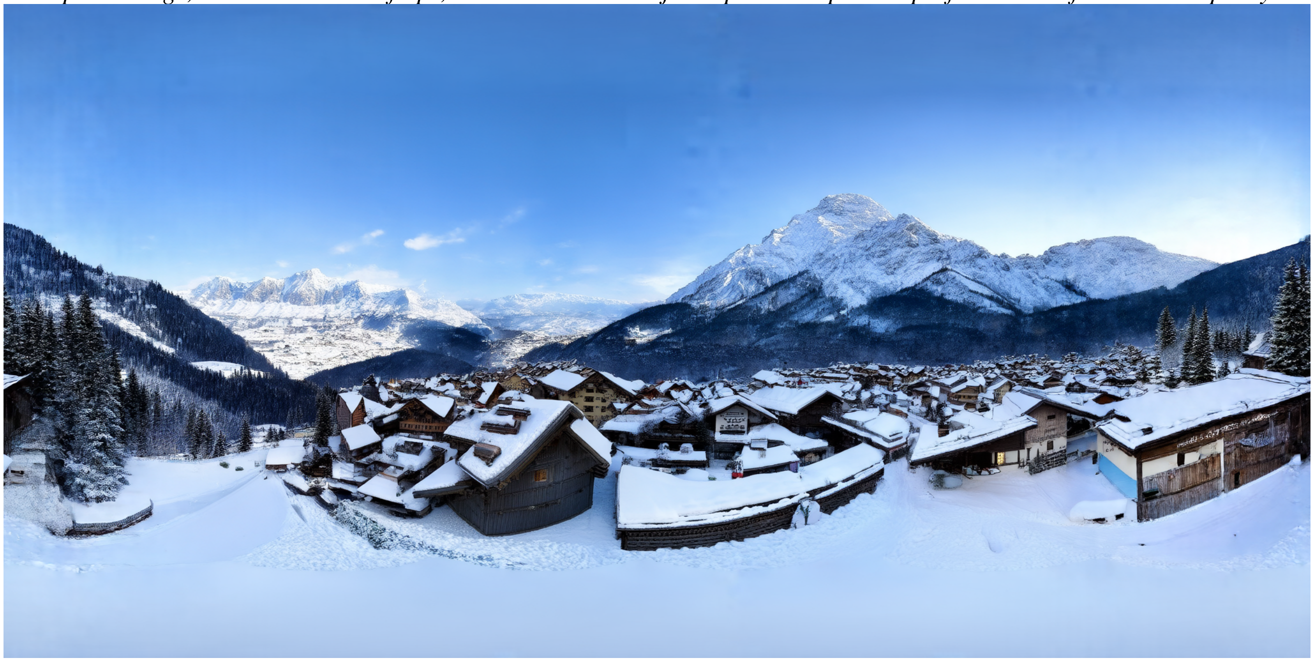


“Exploring the historic streets of Prague, with its charming architecture, cobblestone alleys, and medieval ambiance.”

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

Figure 13. 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."

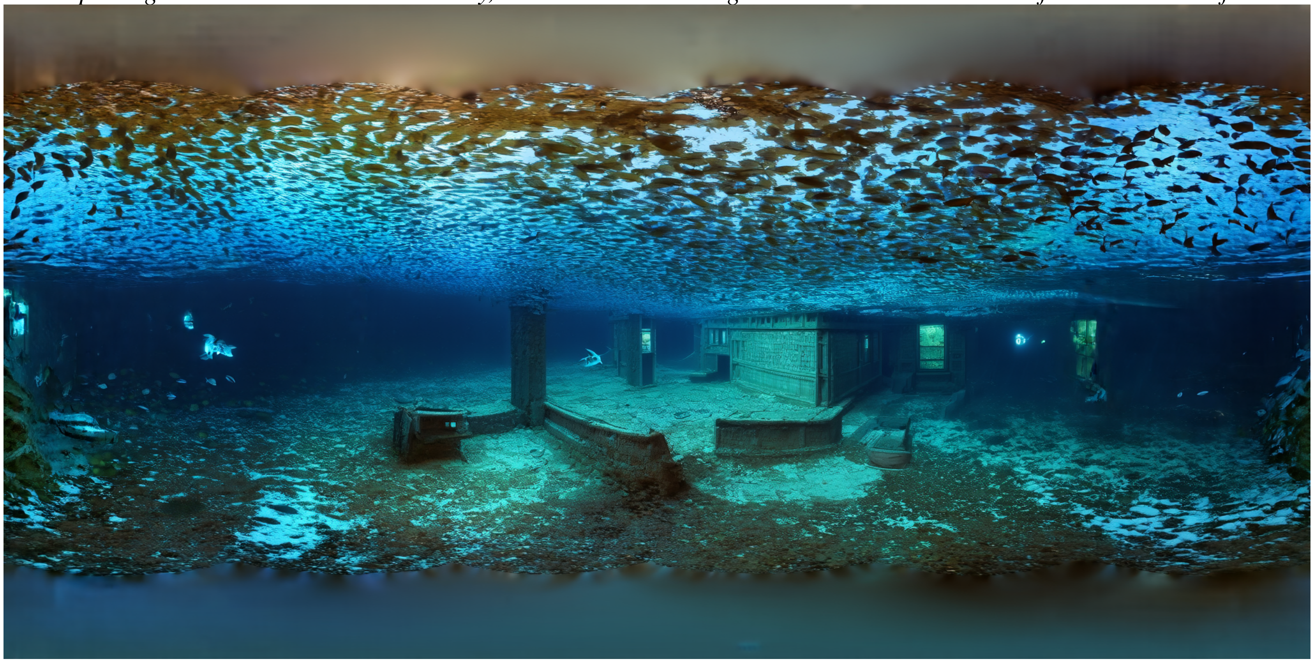
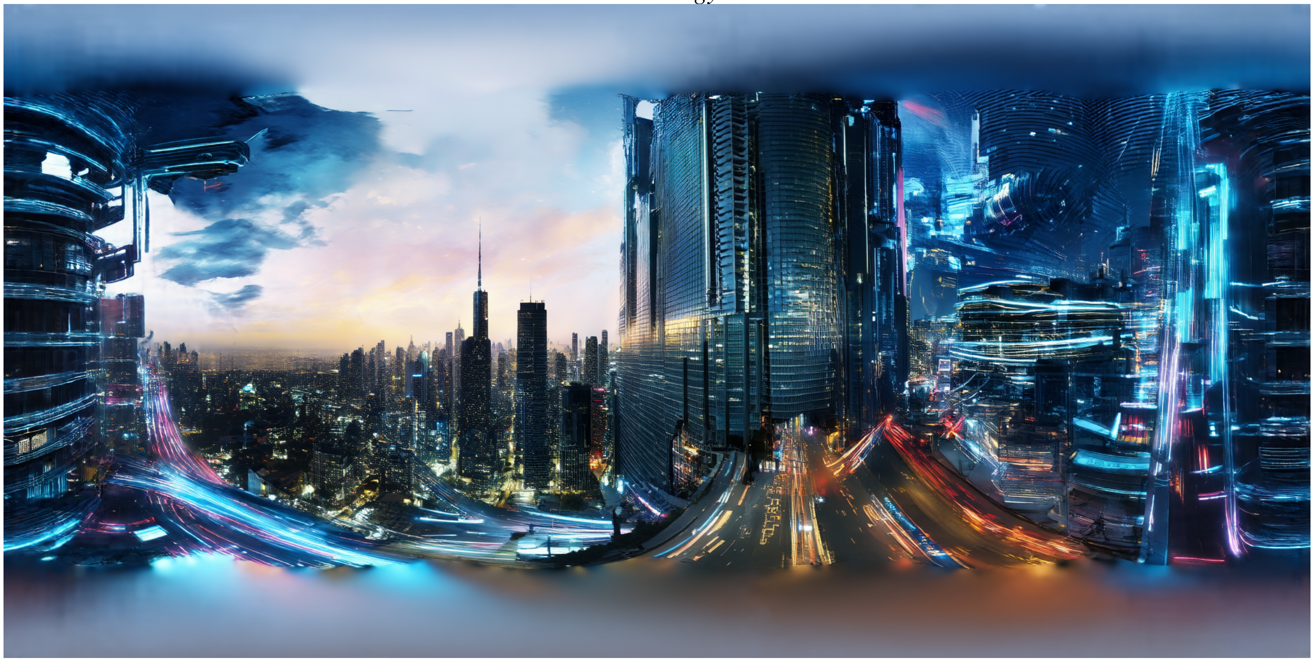


Figure 14. Additional high-resolution results for out-of-distribution prompts.

“Futuristic cityscape, skyscrapers piercing the sky, neon lights painting the streets—an electric dreamscape alive with urban energy.”



“On the surface of a distant planet, a landscape of alien rock formations and swirling, multicolored gases.”

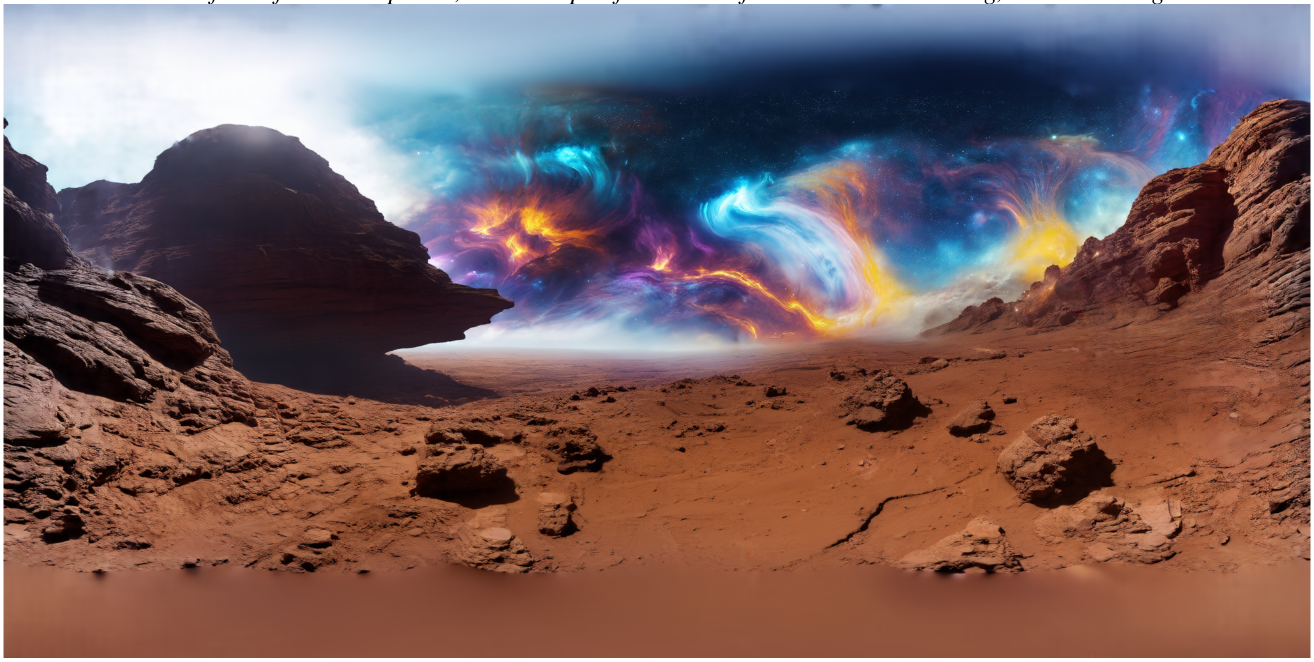
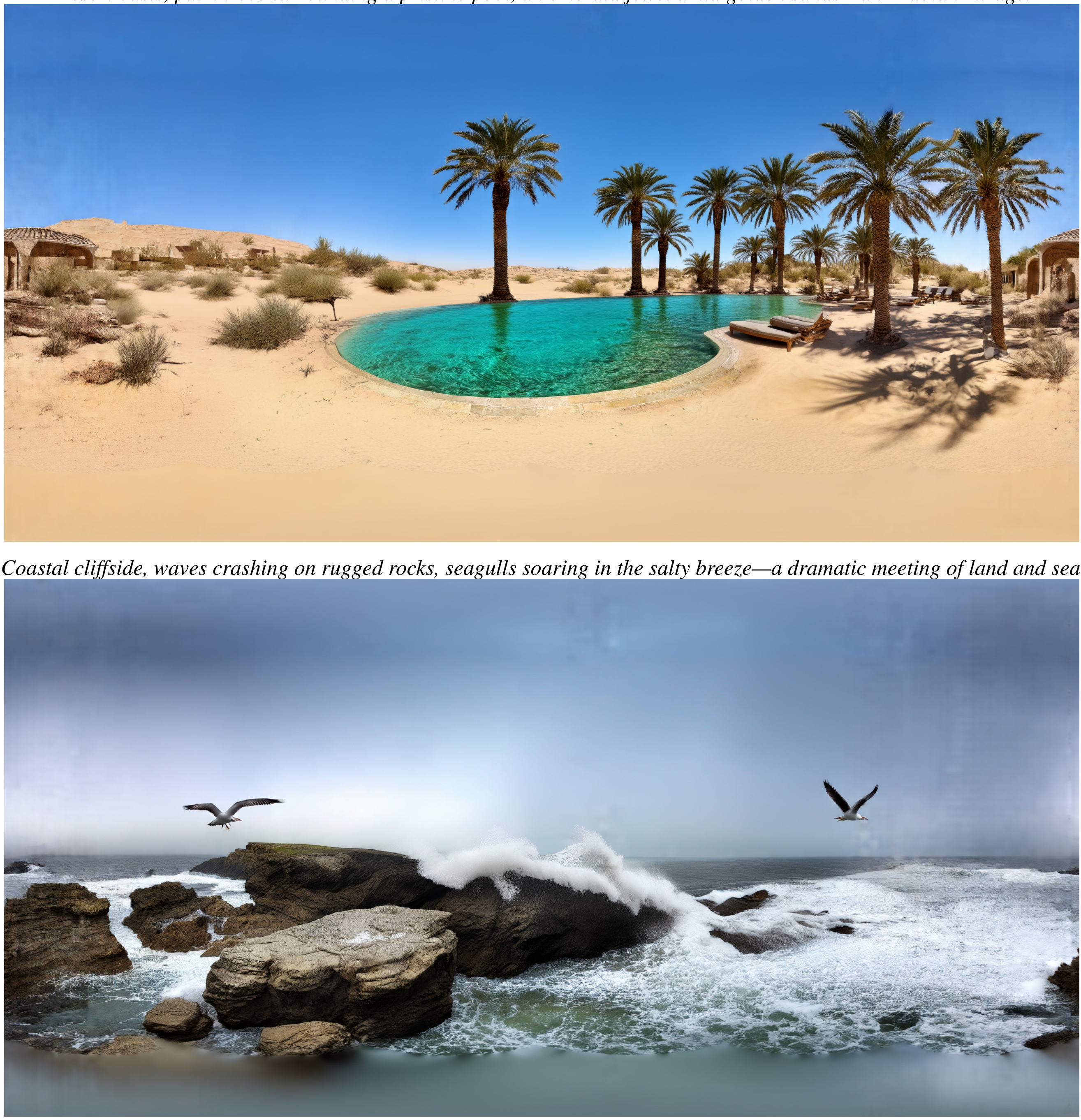


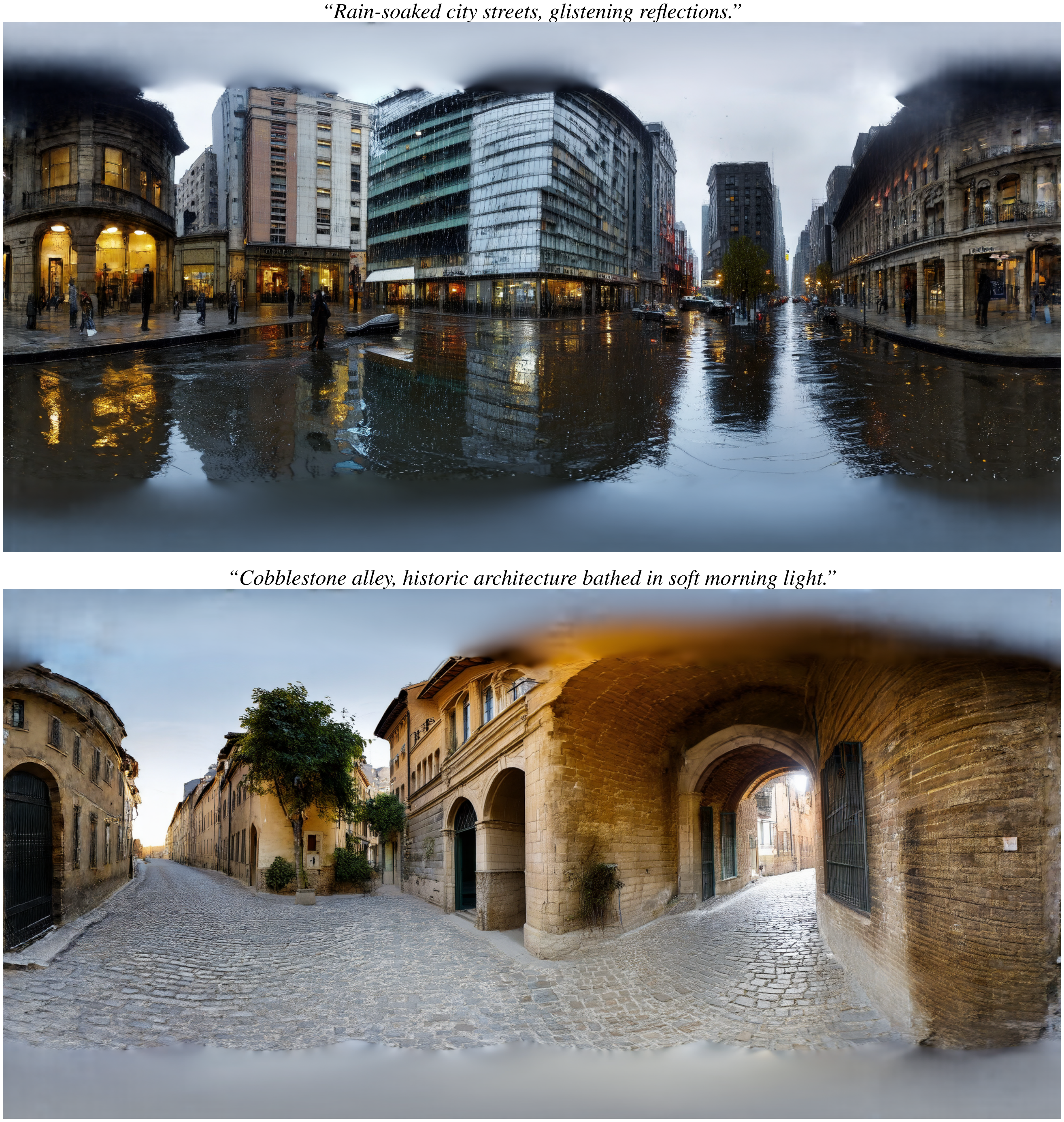
Figure 15. Additional high-resolution results for out-of-distribution prompts.



“Desert oasis, palm trees surrounding a pristine pool, an emerald jewel amid golden sands—an Arabian mirage.”

“Coastal cliffside, waves crashing on rugged rocks, seagulls soaring in the salty breeze—a dramatic meeting of land and sea.”

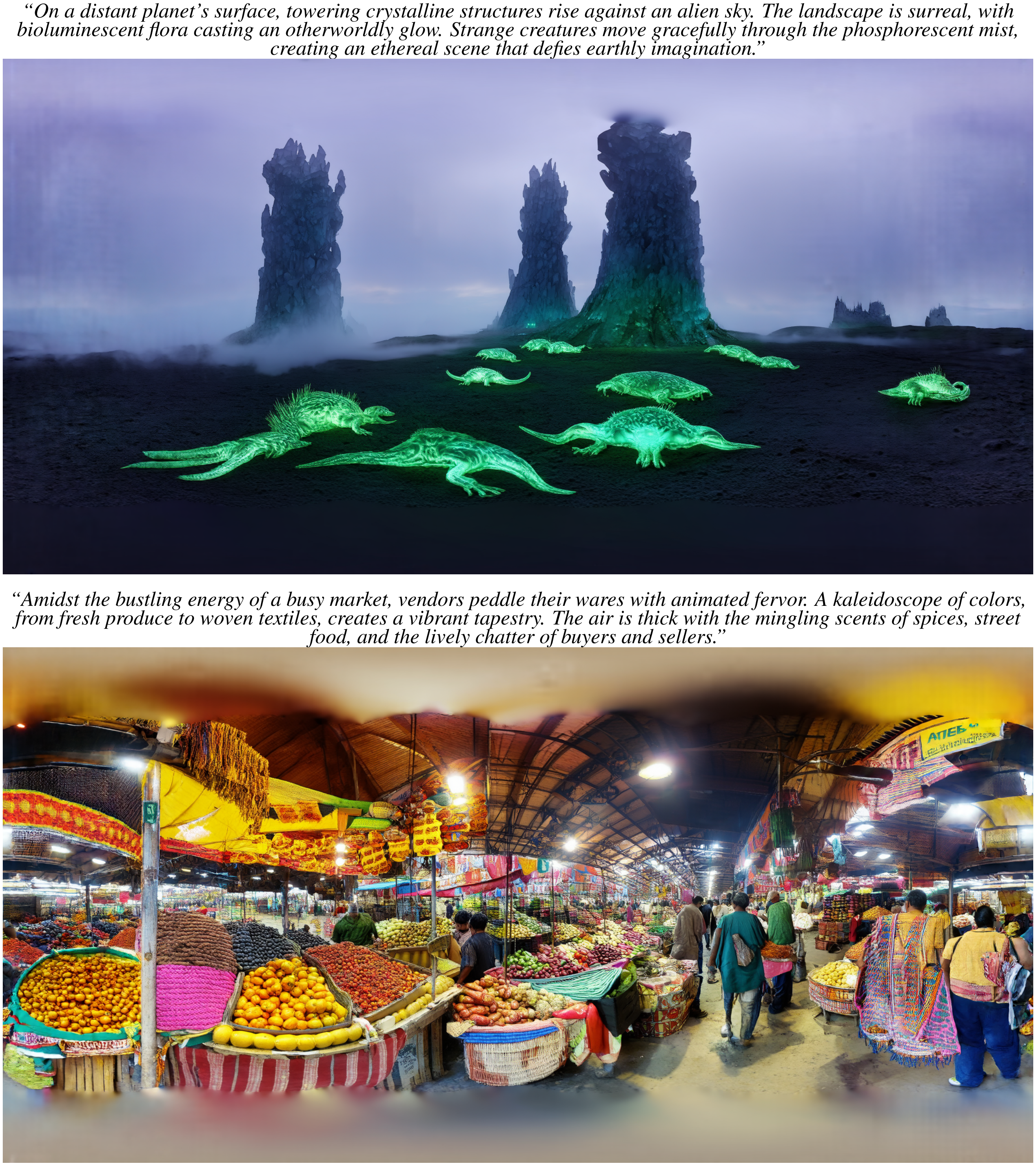
Figure 16. Additional high-resolution results for out-of-distribution prompts.



“Rain-soaked city streets, glistening reflections.”

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

Figure 17. Additional high-resolution results for out-of-distribution prompts.

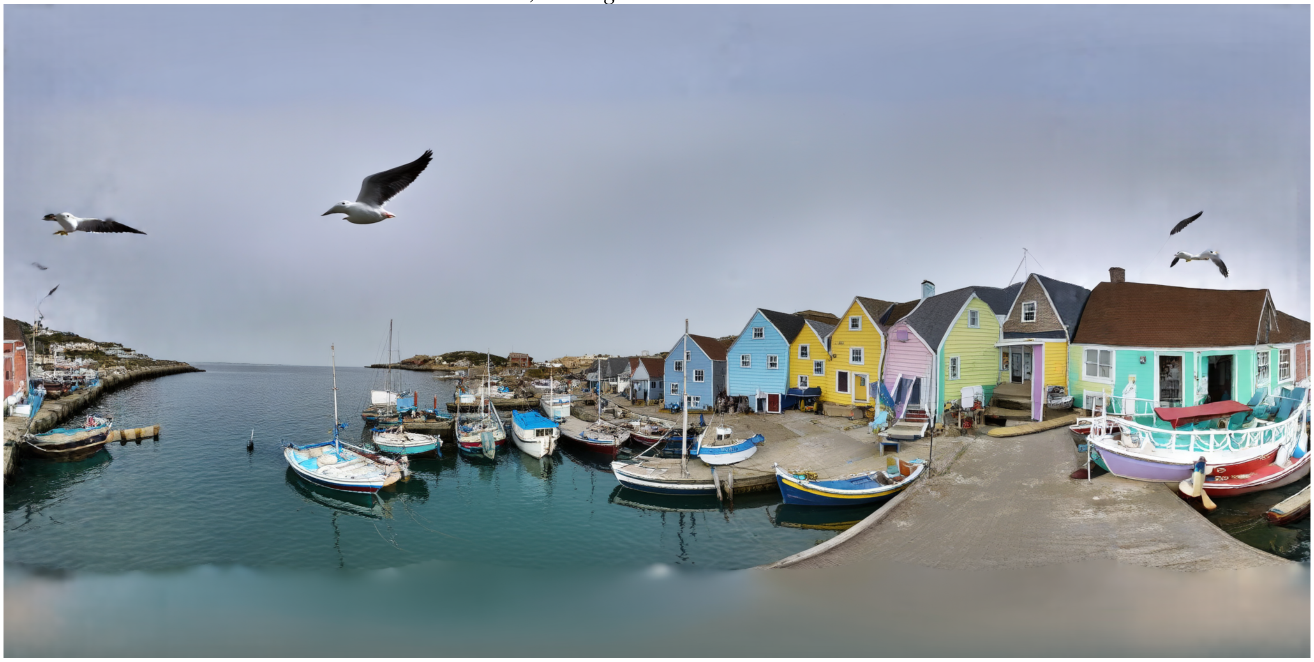


“On a distant planet’s surface, towering crystalline structures rise against an alien sky. The landscape is surreal, with bioluminescent flora casting an otherworldly glow. Strange creatures move gracefully through the phosphorescent mist, creating an ethereal scene that defies earthly imagination.”

“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 18. Additional high-resolution results with complex prompts.

"In a quaint coastal village, weathered cottages line the shore, their pastel hues blending with the colors of the sea. Fishing boats bob gently in the harbor, and the air is tinged with the scent of saltwater and freshly caught fish. Seagulls circle overhead, adding to the maritime chorus."



"Nestled in a canyon, a pueblo village stands against the red earth. Adobe homes with turquoise accents blend seamlessly into the landscape. The sun sets, casting a warm glow on the cliffs, and the rhythmic beats of traditional drums resonate in the stillness, echoing ancient tales."

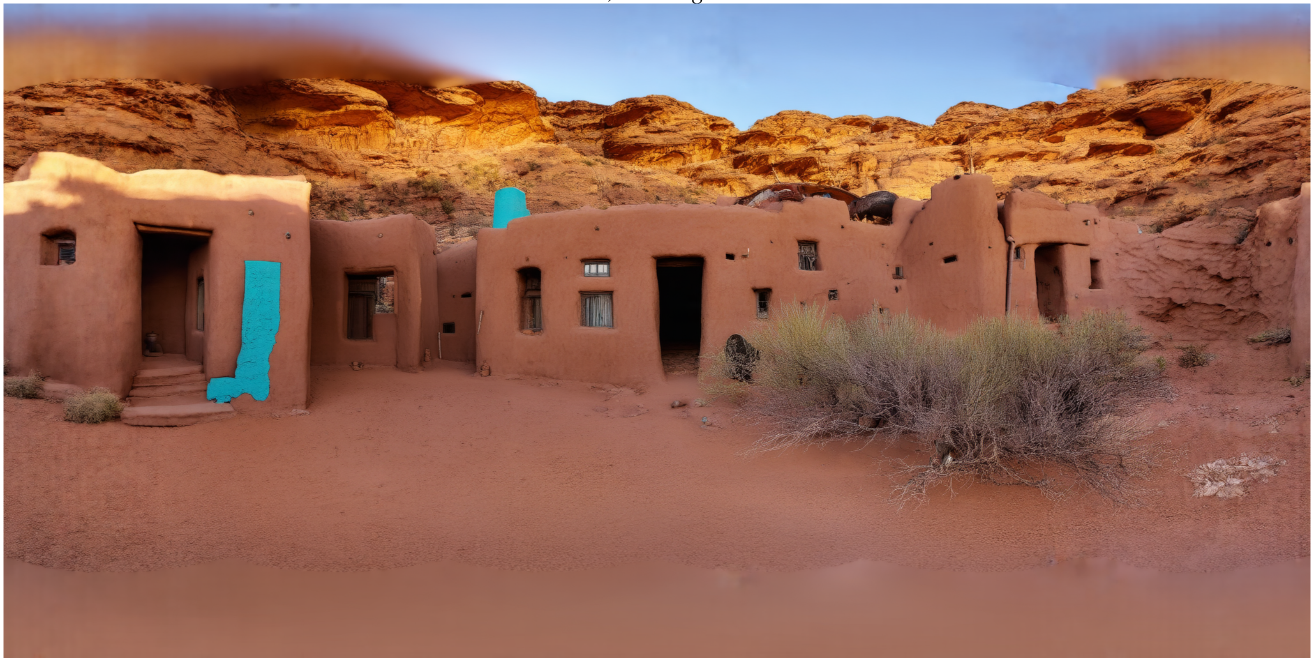


Figure 19. Additional high-resolution results with complex prompts.