OmniX: From Unified Panoramic Generation and Perception To Graphics-Ready 3D Scenes

Yukun Huang[1], Jiwen Yu[1,2], Yanning Zhou[3], Jianan Wang[4], Xintao Wang[2], Pengfei Wan[2], and Xihui Liu[1]

[1]University of Hong Kong [2]Kuaishou Technology
[3]Tencent [4]Astribot

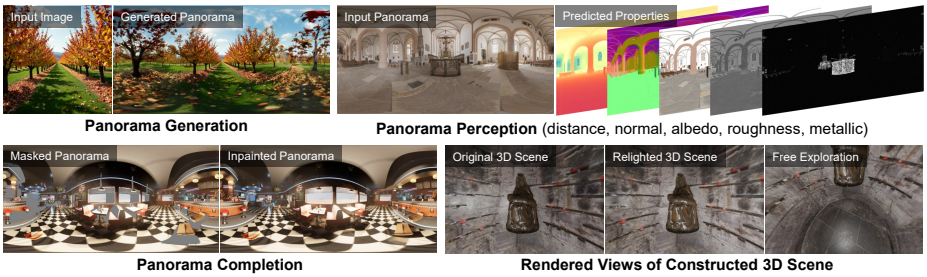

Fig. 1: We present **OmniX**, a versatile and unified framework for panorama generation, perception, and completion. This framework enables the automatic construction of 3D scenes with support for physically based rendering (PBR) from input images.

Abstract. There are two prevalent ways for automatic 3D scene construction: procedural generation and 2D lifting. Among these, panorama-based 2D lifting has emerged as a promising technique, leveraging powerful 2D generative priors to produce immersive, realistic, and diverse 3D environments. In this work, we advance this technique to generate graphics-ready 3D scenes suitable for physically based rendering (PBR), relighting, and simulation. Our key insight is to repurpose 2D generative models for panorama perception of geometry, textures, and PBR materials. Unlike existing 2D lifting approaches that emphasize appearance generation and neglect the perception of intrinsic properties, we present OmniX, a versatile and unified framework for panorama generation, perception, and completion. Built upon cross-modal adapter structure and cyclic spatial operators, OmniX effectively repurposes pre-trained 2D flow matching priors for joint modeling of multimodal, seamless equirectangular representations. Furthermore, we construct a large-scale synthetic panorama dataset comprising high-quality multimodal panoramas from diverse indoor and outdoor scenes. Extensive experiments demonstrate the effectiveness and generality of OmniX as a unified framework for panorama generation and perception across geometry, lighting, and semantics, enabling graphics-ready 3D scene generation and opening new possibilities for immersive and physically realistic virtual world creation. Project page is available at `https://yukun-huang.github.io/OmniX/`.

1 Introduction

Digitizing the 3D world we live in is a technological endeavor that is both imaginative and valuable. Digital replication [6, 17] of 3D scene allows us humans to obtain entertainment and interactive experiences that are difficult to obtain in daily life, or enables near-zero-cost simulation learning for intelligent agents or robots. However, constructing complex 3D scenes requires significant effort and time from artists and engineers, which limits the scale of 3D scene data and hinders the development of native 3D scene generative models.

To automatically build 3D scenes while circumventing data shortages, the community has leveraged large visual language foundation models trained on large-scale text, image, and video data. Based on these powerful models, two typical approaches emerge: procedural generation [10, 39] and 2D lifting [38, 54]. While procedural generation relies on retrieving objects from a 3D asset library to build the scene, 2D lifting methods directly repurpose 2D generative priors for 3D scene generation, achieving diverse and high-quality results. Recent works [16, 27, 52, 60] further introduces panoramic representations, which serve as a bridge between 2D and 3D, greatly improving the cross-view consistency of generated 3D scenes. However, these works emphasize appearance generation rather than intrinsic perception, generally using off-the-shelf depth estimation models to extract scene geometry without textures and PBR materials. This hinders the integration of generated 3D scenes into modern graphics pipelines.

In this paper, we present **OmniX**, a versatile framework that repurposes pre-trained 2D flow matching models for panorama generation, perception, and completion. To this end, we introduce a unified formulation for panoramic vision tasks, extending the 2D generative paradigm to image-to-panorama generation, panorama-to-X perception, and their mask-guided generalization.

Specifically, to enhance multimodal panorama modeling, we introduce two key technical designs. First, we develop *circular synchronization* to mitigate the prevalent issue of discontinuous seams in equirectangular panorama generation. In contrast to prior approaches [9, 49] that directly manipulate latent features, our method provides a fundamental solution by enforcing circular translation equivariance in the model's spatial operators. This structural mechanism ensures circular consistency without compromising the generative capacity of pre-trained models. Second, we investigate cross-modal adapter architectures for handling diverse input modalities and introduce an effective and flexible design, *modality-specific adapters*. These adapters integrate multiple LoRAs [15], allowing modality-specific adaptation while leveraging pre-trained 2D generative priors, thereby improving performance on cross-modal vision tasks.

In addition, we construct a synthetic panorama dataset, **PanoX**, covering indoor and outdoor scenes and various visual modalities such as distance, normal, albedo, roughness, and metallic. This dataset addresses the shortage of high-quality panorama data with dense geometry and material annotations.

Our main contributions are as follows:

- We present OmniX, a unified framework that repurposes pre-trained 2D flow matching models for panorama generation, perception, and completion.

- To enhance multimodal panorama modeling, we introduce two key designs: circular synchronization, which ensures seamless equirectangular panoramas, and modality-specific adapters, which allow effective integration of diverse cross-modal inputs while leveraging pre-trained 2D generative priors.
- We introduce PanoX, a synthetic panorama dataset covering both indoor and outdoor scenes. PanoX addresses the scarcity of high-quality panoramic data with dense geometric/material annotations.
- Extensive experiments demonstrate the effectiveness of OmniX in panorama generation and perception. Moreover, our framework enables the automatic construction of immersive, PBR-ready 3D scenes from images.

2 Related Work

2.1 Inverse Rendering

Inverse rendering [3] aims to estimate intrinsic scene properties such as geometry, materials, and lighting from images. With the rapid progress of generative models, particularly diffusion models, researchers have explored their potential for inverse rendering [21, 30, 31, 56]. IntrinsiX [21] generates high-quality PBR maps from text prompts using a diffusion process, supporting precise material and lighting editing. DiffusionRenderer [31] leverages video diffusion models for joint inverse and forward rendering, combining G-buffer estimation with photo-realistic image generation through co-training on synthetic and real data.

Panoramic images capture a wider field of view and provide more comprehensive scene information, making them versatile for various applications. Yet, inverse rendering with panoramas remains underexplored. PhyIR [29] recovers geometry, complex SVBRDFs, and spatially-coherent illumination from a panoramic indoor image using an enhanced SVBRDF model and a physics-based in-network rendering layer to handle complex materials like glossy, metal, and mirror surfaces. However, it is limited to indoor scenes, while we leverage 2D generative priors to generalize across both indoor and outdoor environments.

2.2 3D Scene Generation

Existing 3D scene generation methods can be mainly divided into procedural generation [10, 33, 34, 39] and 2D lifting [7, 16, 44, 53, 55].

Procedural generation creates 3D scenes based on predefined rules or constraints. These methods are scalable and widely used in domains such as gaming, urban planning, and architecture, but often lack diversity and realism due to their rule-based nature. Representative works include CityEngine [34], which uses grammar-based rules for urban layouts, and InfiniGen [39], which integrates terrain, material, and creature generators to produce diverse environments.

2D lifting methods bridge 2D inputs and 3D representations. Image-based approaches reconstruct 3D scenes from single or sequential images using outpainting or depth estimation, with works like ImmerseGAN [7] and MVDiffusion [44] generating panoramas for scene synthesis. Video-based methods leverage temporal information to ensure coherent dynamic scenes, exemplified by

VividDream [25] and 4Real [53]. These methods emphasize appearance generation, relying on off-the-shelf depth estimators for geometry while neglecting intrinsic properties such as albedos, normals, and PBR materials.

3 Method

3.1 Overview

We first construct a multimodal synthetic panorama dataset, PanoX (Sec. 3.2), which provides high-quality supervision for panorama understanding. Building upon this dataset, we introduce OmniX, a versatile and unified framework that repurposes pre-trained 2D flow matching models [8, 32] for panorama generation, perception, and completion (Sec. 3.3). Finally, We demonstrate OmniX's potential in PBR-ready 3D scene generation (Sec. 3.4).

3.2 PanoX: A Synthetic Panorama Dataset with Dense Annotations

Omnidirectional visual perception is crucial for visual understanding and spatial intelligence. To effectively learn perception across a wide field of view (FoV), large-scale panorama datasets with dense annotations are necessary. While several narrow FoV image datasets [28, 40, 61] offer rich geometry and material annotations, there remains a scarcity of panorama datasets equipped with dense annotations within the research community.

To this end, we introduce PanoX, a multimodal synthetic panorama dataset with dense geometry and material annotations. Given the challenges of collecting real-world panorama data and the high cost of manual annotation, we leverage synthetic 3D scene assets and Unreal Engine 5 to generate pixel-aligned multimodal panorama data. A preview of PanoX is shown in Figure 2.

Specifically, PanoX comprises eight large-scale 3D environments, including five indoor and three outdoor scenes such as stores, warehouses, and wilderness areas. Each scene is rendered into RGB panoramas, along with corresponding distance, world normal, albedo, roughness, and metallic. We also provide text descriptions corresponding to the panoramic images, which are extracted using Florence 2 [50]. The entire dataset contains more than 10,000 instances, corresponding to 60,000 panoramic images of different modalities. We split the

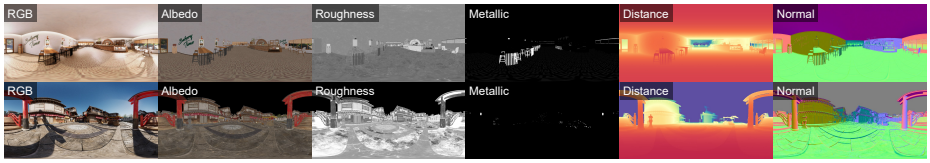

Fig. 2: A preview of the proposed PanoX dataset, providing high-quality panoramic rendered images with rich pixel-aligned annotations, including albedo, roughness, metallic, distance, and normal, across both indoor and outdoor scenes.

Table 1: Comparison between PanoX and existing synthetic scene datasets.

Dataset	Geometry	Material	Include Panorama	Indoor	Outdoor
InteriorNet [26]	✓	✓	✓†	✓	×
Structured3D [59]	✓	×	✓	✓	×
Hypersim [40]	✓	×	×	✓	×
InteriorVerse [61]	✓	✓	×	✓	×
FutureHouse‡ [29]	✓	✓	✓	✓	×
MatrixCity [28]	✓	✓	×	×	✓
PanoX (Ours)	✓	✓	✓	✓	✓

† InteriorNet only contains panoramic RGB images without dense annotations.
‡ FutureHouse is no longer publicly available.

samples from the first six scenes into training, validation, and test segments in an 8:1:1 ratio, obtaining PanoX-Train, PanoX-Val, PanoX-Test. The remaining two scenes are grouped as out-of-domain test set (*i.e.*, PanoX-OutDomain) for generalization evaluation.

We compare the proposed PanoX dataset with existing image datasets rendered from synthetic 3D scenes, as shown in Table 1. To the best of our knowledge, the proposed PanoX is the first panorama dataset covering both indoor and outdoor scenes with dense geometry and material annotations.

3.3 OmniX: A Unified Framework for Panorama Generation, Perception, and Completion

OmniX is a versatile framework for unified panorama perception and generation, built on the pre-trained 2D flow matching model [24]. To this end, we provide a general formulation for unified generation and perception.

Unified formulation. Typically, a flow matching-based image generator f_θ is trained to predict the velocity vector $\mathbf{v}$ from latent representation $\mathbf{z}_0$ to latent representation $\mathbf{z}_1$, given a textual prompt y and the current timestep t:

$$\mathbf{v}_t = f_\theta(\mathbf{z}_t, y, t), \tag{1}$$

The predicted target $\hat{z}_1$ can be obtained by solving the following ordinary differential equation (ODE):

$$\hat{\mathbf{z}}_1 = \mathbf{z}_0 + \int_0^1 \mathbf{v}_t \, dt = \mathbf{z}_0 + \int_0^1 f_\theta(\mathbf{z}_t, y, t) dt. \tag{2}$$

Our goal is to expand this image generation paradigm into a unified panorama generation, perception, and completion framework. To this end, we generalize the model f_θ to take multiple condition inputs:

$$\hat{\mathbf{z}}_1 = \mathbf{z}_0 + \int_0^1 f_\theta(\mathbf{z}_t, \mathbf{c}^0, \mathbf{c}^1, ..., y, t) dt, \tag{3}$$

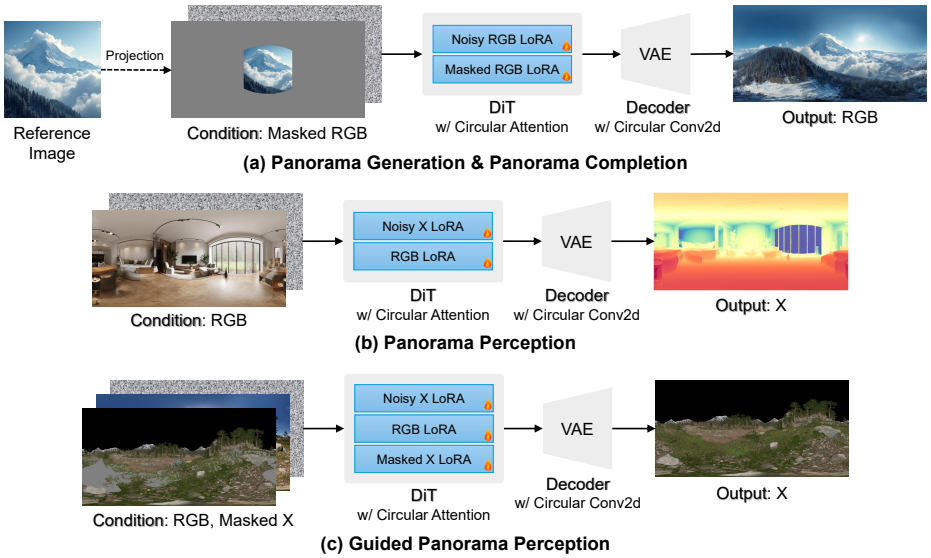

Fig. 3: OmniX is a versatile and unified framework that repurposes pre-trained 2D flow-matching models for panorama generation, perception, and completion. Built upon circular synchronization operators and modality-specific LoRAs, OmniX flexibly adapts to different panorama tasks and produces seamless results.

where $\{\mathbf{c}^i | i = 0, 1, ...\}$ denotes the conditioning inputs, each spatially aligned with $\mathbf{z}_t$. The modality and number of the conditioning inputs depend on the target task. Figure 3 shows three representative target tasks, each corresponding to a different configuration of conditioning inputs. Detailed descriptions of these task settings are provided in the supplementary material.

Circular synchronization for seamless generation. Circular blending [9] is a commonly used technique [45, 52] for achieving seamless panorama generation with latent diffusion models [41], which weights and blends the leftmost and rightmost edges of latent features to ensure closed loops. While effective, such latent-level operations may introduce unwanted feature distortions near the boundary. We argue that seam discontinuities fundamentally originate from the lack of circular translation equivariance in spatial operators (*e.g.*, attentions and convolutions). To address this issue, we propose Circular Synchronization, a training-free approach that adapts spatial operators for seamless generation without fine-tuning the pre-trained model parameters.

The core idea of circular synchronization is to adapt all spatial operators to be circular translation-equivariant, as illustrated in Figure 4. For convolutions, this adaptation is straightforward, as it only requires replacing the conventional padding strategy with circular padding [47]. For attentions equipped with RoPE [43], the adaptation involves two steps: token padding and attention masking. Specifically, token padding applies circular padding to the left and right boundaries of key and value tokens (excluding query tokens), with the padding

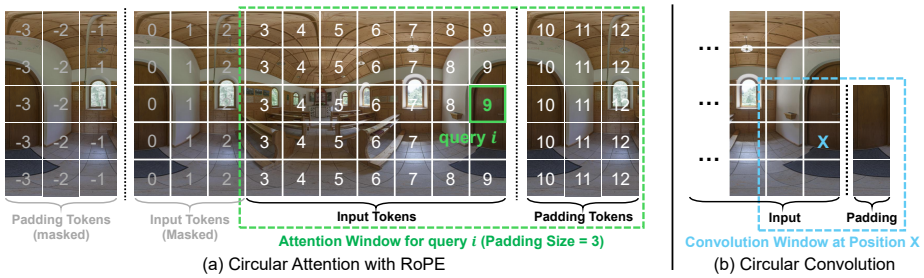

Fig. 4: Circular Synchronization adapts the receptive window of spatial operators to enforce circular translational equivariance, including: (a) Circular Attention with RoPE [43], where the numbers in the image patches indicate horizontal position IDs, and (b) Circular Convolution (also referred to as circular padding).

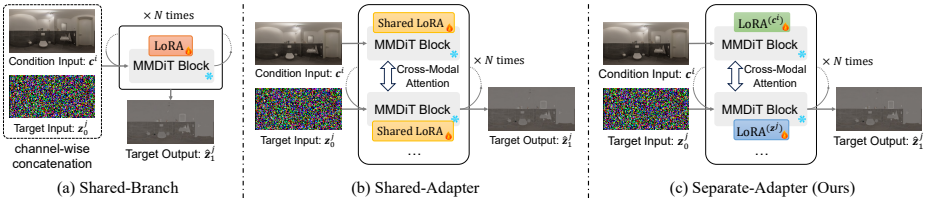

Fig. 5: Different cross-modal adapter structures for multiple condition inputs $\{\mathbf{c}^i \mid i = 0, 1, ...\}$ and multiple target outputs $\{\hat{\mathbf{z}}_1^j \mid j = 0, 1, ...\}$. Specifically, (a) *Shared-Branch* concatenates different inputs along the channel dimension; (b) *Shared-Adapter* performs token-wise concatenation across multiple inputs; and (c) *Separate-Adapter* builds upon *Shared-Adapter* by assigning modality-specific weights to each input type.

size controlled by a padding-size hyperparameter. Attention masking constrains the attention window to ensure a circular and spatially uniform receptive field for each query token. By enforcing circular synchronization across all convolutional and attention operations in the pre-trained DiT [35] and VAE, our method enables seamless circular image generation, as shown in Figure 10.

For multi-view methods [18,42,44], boundary consistency is typically achieved through overlapping fields of view (FoVs), where adjacent views provide mutual constraints for alignment. Such strategies require additional view aggregation mechanisms and computational overhead, while treating panoramic consistency as an external constraint. In contrast, our Circular Synchronization directly adapts spatial operators to be circular translation-equivariant, making boundary consistency an intrinsic property of the generation process. This design preserves the pretrained generative priors without modifying the underlying model parameters and produces more coherent boundary transitions.

Modality-specific adapters for cross-modal generation. We explore multiple ways to adapt the pre-trained DiT for cross-modal 2D inputs, as shown in Figure 5. Specifically, depending on how branches and adapters are shared, these methods can be divided into: Shared-Branch, Shared-Adapter, and Separate-

Adapter. Given multiple 2D inputs, Shared-Branch and Shared-Adapter perform channel-wise and token-wise concatenations, respectively. Building upon token-wise concatenation, Separate-Adapter further assigns different LoRAs [15] to different types of inputs. Note that all 2D inputs and outputs are spatially aligned and share the same 2D positional encoding.

Empirical results indicate that the Separate-Adapter design not only achieves superior performance in cross-modal panorama perception tasks (as reported in Table 5) but also allows flexible expansion to new input modalities with minimal impact on the model's weight distribution. Therefore, OmniX employs this modality-specific adapter design.

Optimization. Built upon modality-specific adapters, multiple LoRAs [15] are trained to exploit a pre-trained 2D flow matching model for feature extraction from conditional inputs and predicting velocity vectors for target outputs. While both the condition $\mathbf{c}$ and the target $\mathbf{z}_t$ are input to DiT, only the target output is used to compute the flow matching loss [32]:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{z}_1,\mathbf{z}_0} \|\mathbf{v} - f_\theta(\mathbf{z}_t, \mathbf{c}, t)\|^2, \tag{4}$$

where the velocity vector $\mathbf{v} = \mathbf{z}_1 - \mathbf{z}_0$. Note that this objective can be generalized to **M**ultiple condition **I**nputs $\{\mathbf{c}^i \mid i = 0, 1, ...\}$ and **M**ultiple target **O**utputs $\{\mathbf{z}_1^j \mid j = 0, 1, ...\}$, yielding a MIMO version of flow matching loss:

$$\mathcal{L}_{\text{mimo}} = \mathbb{E}_{t,\mathbf{z}_1^j,\mathbf{z}_0} \|\mathbf{v} - f_\theta(\mathbf{z}_t, \mathbf{c}^0, \mathbf{c}^1, ..., t)\|^2. \tag{5}$$

Note that circular synchronization is a structural constraint and does not need to be performed during training, which avoids additional computational costs.

3.4 Application: PBR-Ready 3D Scene Generation

The OmniX framework opens up the possibility of automatically constructing PBR-ready 3D scenes from images or panoramas. Specifically, the entire pipeline consists of three stages: (a) multimodal panorama generation, (b) scene reconstruction, and (c) iterative completion.

Multimodal panorama generation. OmniX offers a general solution for image-to-panorama generation and RGB-to-X panorama perception. We train multiple adapters to repurpose the pre-trained flow matching model for these tasks. By switching adapters of different tasks, we can achieve a generative chain of "image → panorama → panorama with intrinsic properties".

Scene reconstruction. Given a panoramic distance map, since the ray direction corresponding to each pixel is known, the pixels can be projected into 3D space as vertices of a 3D mesh. The connectivity of these vertices can be further determined based on pixel neighbors and relative distances. Once the 3D mesh of the scene is obtained, the panoramic maps of other modalities (*i.e.*, albedo, normal, roughness, and metallic) can be assigned to each triangle face via spherical UV unwrapping, resulting in a PBR-ready scene-level 3D asset.

Iterative completion. A panoramic view provides only an omnidirectional observation from a fixed viewpoint, and thus the reconstructed scene does not

Fig. 6: Occlusion-aware masked data construction pipeline. Based on the panoramic distance map and a randomly sampled 3D displacement, we can estimate the occluded regions by ray intersection. These regions are used as masks to construct masked data for training the panorama completion models.

support free exploration. Iterative scene completion [52, 54] is therefore essential for constructing explorable and potentially large-scale 3D scenes. To this end, we augment the OmniX adapters with mask inputs and fine-tune them for completion and guided perception, resulting in OmniX-Fill. Specifically, to simulate occlusion-induced scene holes, we propose a depth-based sampling strategy to generate occlusion-aware masks, as illustrated in Figure 6. Through the interaction between OmniX-Fill and the graphics engine, we are able to generate new regions while preserving previously reconstructed content.

4 Experiment

4.1 Implementation Details

Our method is implemented in PyTorch, trained and evaluated on four NVIDIA L40S GPUs. We use the same optimization settings for all tasks. Specifically, we adopt an AdamW optimizer with learning rate of 1e-4 for training. No learning rate decay strategy is employed. All panoramic images are resized to a resolution of 512×1024 for training, using the batch size of 1 for each GPU.

We trained 12 adapter models based on Flux.1-dev [24] for different panoramic vision tasks, including: image-to-panorama generation, panorama perception (distance, normal, albedo, roughness, metallic), and their corresponding masked versions. Each of the adapter model consists of two or more LoRAs, depending on the number of input modalities.

4.2 Results on Panorama Generation

For panorama generation, we consider a perspective-to-panorama setting [18], in which text–image pairs are used as conditioning inputs.

Datasets. For training, we combine multiple panorama datasets and publicly available sources, including Pano360 [20], Structured3D [59], Poly-Haven [37], Humus [36], and our proposed PanoX. For evaluation, we report results on the Laval Indoor [11] and SUN360 [51] datasets following CubeDiff [18]. The text prompts for all panoramic images are extracted by Florence 2 [50].

Evaluation metrics. To evaluate visual quality, we follow CubeDiff [18] and report FID [14], KID [4], CLIP-FID [23], and FAED [57]. FID, KID, and

Table 2: Quantitative evaluation of OmniX on panorama generation compared to existing image-to-panorama generation methods: OmniDreamer [1], PanoDiffusion [49], Diffusion360 [9], and CubeDiff [18].

Method	Laval Indoor					SUN360				
	FID↓	KID ($\times 10^2$)↓	CLIP-FID↓	FAED↓	CS↑	FID↓	KID ($\times 10^2$)↓	CLIP-FID↓	FAED↓	CS↑
OmniDreamer	71.0	5.17	23.9	19.2	-	92.3	8.89	51.7	30.4	-
PanoDiffusion	58.6	4.08	26.6	106.8	-	52.9	3.51	28.9	98.0	-
Diffusion360	33.1	2.07	16.9	23.7	26.38	45.4	3.73	18.5	12.6	22.89
CubeDiff	<u>9.5</u>	**0.32**	<u>3.2</u>	<u>18.4</u>	<u>27.02</u>	**25.5**	<u>1.33</u>	<u>8.1</u>	**7.6**	<u>25.00</u>
OmniX (Ours)	**7.4**	<u>0.41</u>	**2.5**	**5.2**	**28.47**	<u>39.4</u>	**0.66**	**6.5**	<u>8.7</u>	**27.54**

Fig. 7: Qualitative evaluation of OmniX on panorama generation compared to two image-to-panorama methods: Diffusion360 [9] and CubeDiff [18].

CLIP-FID are computed on perspective crops from generated ERP panoramas, while FAED is evaluated directly on ERP outputs. We additionally report CLIP Score (CS) [13] to measure text alignment.

Quantitative evaluation. As shown in Table 2, OmniX achieves state-of-the-art performance across both the Laval Indoor and SUN360 datasets. On Laval Indoor, it obtains the best results on FID (7.4), CLIP-FID (2.5), FAED (5.2), and CLIP Score (28.47), while remaining competitive on KID, outperforming previous methods in both visual fidelity and text alignment. On SUN360, OmniX continues to deliver strong results, achieving the best KID (0.66), CLIP-FID (6.5), and CLIP Score (27.54), while maintaining competitive performance on FID and FAED. Compared to strong baselines like CubeDiff, OmniX demonstrates more consistent improvements across metrics, highlighting its ability to produce visually realistic and semantically aligned panoramic images.

Qualitative evaluation. As illustrated in Figure 7, OmniX outperforms previous perspective image-to-panorama methods, including Diffusion360 [9] and CubeDiff [18], in terms of panorama quality and scene consistency. Diffusion360 suffers from noticeable visual artifacts and inconsistent scene layouts, while CubeDiff alleviates some appearance artifacts but still produces structurally inconsistent panoramas, particularly for large-scale scene layouts requiring long-range spatial reasoning. In contrast, OmniX generates panoramas with improved structural fidelity, seamless global consistency, and realistic scene composition,

Table 3: Quantitative results of OmniX on panoramic intrinsic estimation compared to five competing methods: RGB↔X [56], MGNet [61], IDArb [30], IID [22], and DiffusionRenderer [31].

Method	PanoX-OutDomain						Structured3D	
	Albedo		Roughness		Metallic		Albedo	
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
RGB↔X	6.35	0.591	8.18	0.628	4.38	0.720	<u>14.40</u>	<u>0.350</u>
MGNet	7.93	0.583	10.22	0.625	6.37	0.656	13.21	0.496
IDArb	9.42	0.562	9.57	0.603	4.30	0.554	10.87	0.505
IID	10.25	0.640	10.09	0.631	7.89	0.726	10.80	0.539
DiffusionRenderer	<u>10.91</u>	<u>0.556</u>	<u>10.45</u>	<u>0.591</u>	<u>14.45</u>	<u>0.425</u>	9.12	0.396
OmniX (Ours)	**17.76**	**0.344**	**16.21**	**0.398**	**18.87**	**0.254**	**20.35**	**0.174**

while preserving fine-grained local details and coherent geometry across the entire 360° field of view.

4.3 Results on Panorama Perception

We divide panorama perception into intrinsic decomposition (albedo, roughness, metallic) and geometry estimation (distance, normal), and present both qualitative and quantitative results compared to competing methods.

Datasets. We use the standard splits of both PanoX and Structured3D for training and evaluation. During training, each batch is sampled from these two data sources with equal probability. Since Structured3D does not include PBR materials, only PanoX is used for roughness and metallic models.

Evaluation metrics. For visual perception tasks with ground truths, we adopt a variety of quantitative metrics for different modalities. Specifically, we use PSNR (Peak Signal-to-Noise Ratio) and LPIPS [58] as metrics for albedo, roughness, and metallic. For Euclidean distances, we use four commonly used metrics: AbsRel, δ-1.25, MAE, and RMSE, following the implementation in [5]. For surface normals, We measure the pixel-wise angular error with ground truth and report the mean, median, and the percentage of pixels with an error below 5° and 30° following [2]. Note that there may be invalid values in the ground truths (*e.g.*, pixels at infinite distance), we exclude these invalid values when calculating the metrics.

Panoramic intrinsic decomposition. We compare our OmniX with five state-of-the-art intrinsic decomposition methods: RGB↔X [56], MGNet [61], IDArb [30], IID [22], DiffusionRenderer [31]. Note that DiffusionRenderer is a video-based inverse rendering method, so we render each panorama into multiple frames to fit its input. The quantitative results are reported in Table 3. Our method achieves consistent state-of-the-art performance on the prediction of three intrinsic properties: albedo, roughness, and metallic. A qualitative comparison is shown in Figure 8 to illustrate the prediction results.

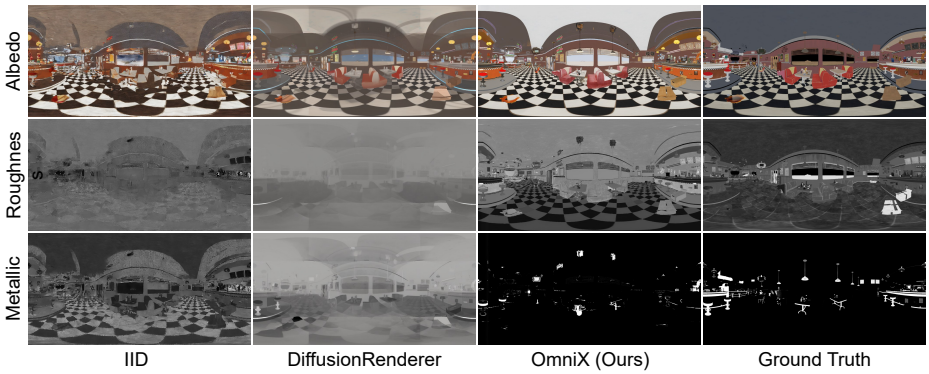

Fig. 8: Qualitative evaluation of OmniX on panoramic intrinsic estimation compared to two competing methods: IID [22] and DiffusionRenderer [31].

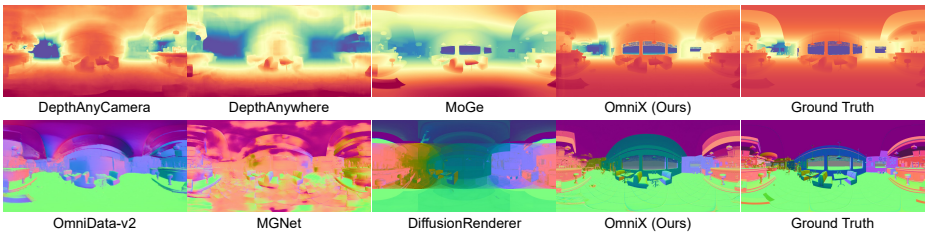

Fig. 9: Qualitative evaluation of OmniX on panoramic geometry estimation compared to state-of-the-art geometry estimation methods: DepthAnyCamera [12], DepthAnywhere [46], OmniData-v2 [19], MGNet [61], DiffusionRenderer [31], and MoGe [48]. Our method shows a notable advantage in capturing fine image details.

Panoramic geometry estimation. We compare our OmniX with two panoramic geometry estimation methods: DepthAnyCamera [12], DepthAnywhere [46], and four narrow-FoV geometry estimation methods: OmniData-v2 [19], MGNet [61], DiffusionRenderer [31], and MoGe [48]. The quantitative results are reported in Table 4, where we achieve the highest normal estimation accuracy and and the second highest depth estimation accuracy. Note that MoGe [48] integrate 21 large-scale datasets for training, while we use much less data to achieve competitive performance. We further provide a qualitative comparison in Figure 9 to illustrate the prediction results.

In-the-wild panorama perception. We empirically find that the proposed OmniX demonstrates superior generalization performance and is able to achieve satisfactory prediction results on in-the-wild images from the Internet. These results are presented in the supplementary material.

Table 4: Quantitative results of OmniX on panoramic geometry estimation compared to six state-of-the-art competing methods: DiffusionRenderer [31], MGNet [61], DepthAnywhere [46], OmniData-v2 [19], DepthAnyCamera [12], and MoGe [48]. For fair comparison, we use PanoX-OutDomain as the evaluation set to ensure all methods are evaluated in unseen scenarios. Note that MoGe uses far more depth annotations than ours (9.0M vs. 0.087M). Moreover, MoGe requires multi-view inference and stitching for panorama inputs, which is both complex and inefficient.

Type	Method	Distance				Normal			
		AbsRel↓	δ-1.25↑	MAE↓	RMSE↓	Mean↓	Median↓	5°↑	30°↑
Depth-Only	DepthAnywhere	0.345	0.392	1.80	9.59	/	/	/	/
	DepthAnyCamera	0.199	0.680	1.93	7.86	/	/	/	/
	MoGe	**0.106**	**0.898**	**1.04**	**5.35**	/	/	/	/
Geometry	DiffusionRenderer	0.709	0.246	2.55	16.10	97.19	89.62	0.001	0.023
	MGNet	0.433	0.396	3.97	11.32	<u>79.96</u>	<u>82.84</u>	0.019	<u>0.269</u>
	OmniData-v2	0.342	0.440	1.94	10.76	85.22	100.60	<u>0.150</u>	0.245
	OmniX (Ours)	<u>0.158</u>	<u>0.787</u>	<u>1.68</u>	<u>6.83</u>	**27.14**	**14.88**	**0.155**	**0.663**

Table 5: Ablation Studies. We analyze the impact of 2D generative pre-trained weights and different adapter architectures on panoramic perception performance. Both datasets PanoX-Test and PanoX-OutDomain are used as the evaluation set to comprehensively cover both in-domain and out-domain scenarios.

Method	Albedo		Roughness		Distance			
	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓	δ-1.25↑	AbsRel↓	RMSE↓	MAE↓
Shared-Branch	15.29	0.650	11.73	0.667	0.464	0.386	8.565	2.122
Shared-Adapter	20.46	0.305	16.92	0.363	0.689	0.219	6.346	1.363
Separate-Adapter (from scratch)	17.74	0.534	14.83	0.510	0.382	0.461	8.771	2.729
Separate-Adapter (Ours)	**21.68**	**0.260**	**18.16**	**0.329**	**0.808**	**0.154**	**4.755**	**1.110**

4.4 Ablation Analysis and Discussion

We conduct ablation studies to evaluate the key components of OmniX, with additional analysis and discussion provided in the supplementary material.

Cross-modal adapter structures. We investigate the effect of different adapter structures on panorama perception performance, as reported in Table 5. Among the evaluated variants, the Separate-Adapter design adopted in OmniX achieves the best performance. This improvement stems from its ability to enable modality-specific adaptation while effectively leveraging pre-trained 2D generative priors. By decoupling adaptations across modalities without disrupting the original weight distribution, the proposed structure preserves the strengths of the pre-trained model while enhancing cross-modal learning capacity.

Leveraging 2D generative priors. We compare training from scratch with initialization from 2D generative pre-trained weights to evaluate their impact on panorama perception performance, as shown in Table 5. Initializing with 2D generative priors leads to consistent and substantial performance improvements, even for modalities with significantly different pixel-value distributions. These

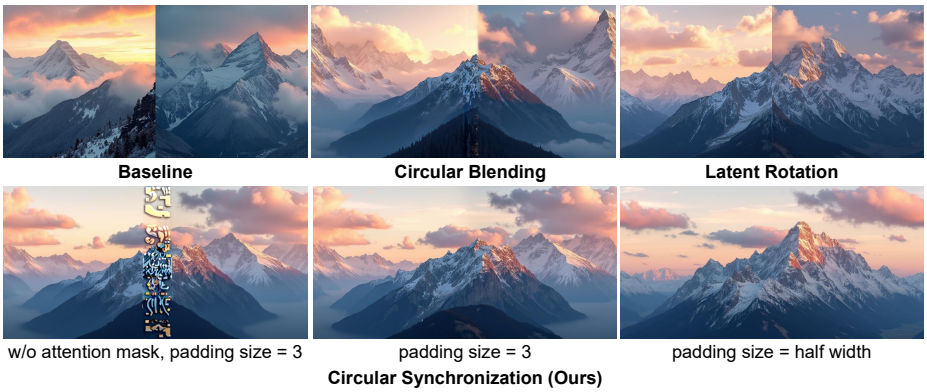

Fig. 10: Qualitative analysis of circular synchronization. The displayed images are generated by Flux.1-dev [24] and horizontally rolled by half their width to show the continuity at the seams. Compared to circular blending [9] and latent rotation [49], our method achieves near-perfect seamless image generation without any fine-tuning.

results highlight the strong transferability and effectiveness of 2D generative priors for panorama perception tasks.

Circular synchronization. We compare circular synchronization with baseline (Flux.1-dev [24]), circular blending [9], and latent rotation [49], as shown in Figure 10. To better visualize seam continuity, the generated images are horizontally rolled by half of their width. The baseline exhibits clear discontinuities at the boundary, while circular blending and latent rotation partially alleviate seam artifacts but still introduce visible structural inconsistencies or texture distortions near the boundary. In contrast, circular synchronization produces nearly seamless transitions across the horizontal wrap-around. This result demonstrates that enforcing circular translation equivariance at the structural level effectively guarantees global consistency along the seam while preserving the generative fidelity of pre-trained 2D flow matching models.

4.5 Applications

OmniX enables automatic production of PBR-ready 3D scenes from images or panoramas. To evaluate the practicality of these generated 3D scenes, we import them into Blender and implement various graphics workflows, including free exploration, PBR-based relighting, and physical simulation, as shown in Figure 11. Specifically, for free exploration, we move the camera forward to render a novel panoramic view. For PBR-based relighting, we add a point light source and animate its horizontal movement in a circular path around the scene's center. For physical simulation, we introduce an elastic ball into the scene, assigning it an initial horizontal velocity to enable dynamic interactions within the environment. Demonstration videos are provided in the supplementary material.

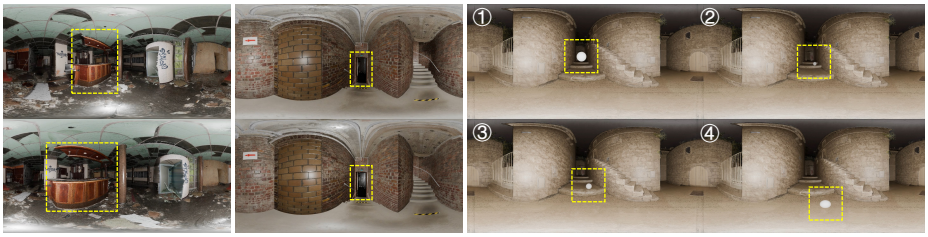

Fig. 11: Demonstrations of the PBR-ready 3D scenes constructed by OmniX, ready for free exploration, PBR-based relighting, and physical simulation.

5 Conclusion and Limitations

In this work, we introduce OmniX, a versatile framework that repurposes pre-trained 2D flow-matching models for panorama generation, perception, and completion. We develop a unified formulation that casts both visual perception (RGB→X) and visual completion (masked X→X) into a 2D generative paradigm, and propose an efficient, lightweight cross-modal adapter to capture diverse task-specific knowledge. In addition, we construct a synthetic panorama dataset, PanoX, which spans indoor and outdoor environments and multiple visual modalities. PanoX serves as a comprehensive benchmark for panorama perception, addressing the scarcity of panorama data with dense geometric and material annotations. Extensive experiments demonstrate the effectiveness of our approach across panorama generation, perception, and completion tasks. Our approach further enables the automatic creation of immersive, PBR-ready 3D scenes that integrate seamlessly with standard 3D workflows.

Limitations. Our method is built on top of pre-trained 2D flow matching models and thus inherits their shortcomings such as slow training and inference efficiency. In addition, OmniX's prediction of Euclidean distance is still not accurate enough, resulting in bumpy reconstructed 3D surfaces, which affects the subsequent PBR rendering effect. We also empirically observe that OmniX-Pano2Metallic, used for metallic prediction, performs poorly in generalization. This is partly due to the scarcity of panoramic PBR material data for training. Furthermore, the significant differences between neural rendering (*i.e.*, 2D generative modeling) and PBR rendering may indicate that pre-trained 2D image priors have limited benefits for PBR material estimation.

Acknowledgment

The research work described in this paper was conducted in the JC STEM Lab of Autonomous Intelligent Systems funded by The Hong Kong Jockey Club Charities Trust.

References

1. Akimoto, N., Matsuo, Y., Aoki, Y.: Diverse Plausible 360-Degree Image Outpainting for Efficient 3DCG Background Creation. In: CVPR. pp. 11441–11450 (2022)
2. Bae, G., Davison, A.J.: Rethinking Inductive Biases for Surface Normal Estimation. In: CVPR. pp. 9535–9545 (2024)
3. Barrow, H., Tenenbaum, J., Hanson, A., Riseman, E.: Recovering Intrinsic Scene Characteristics From Images. Comput. Vis. Syst **2**(3-26), 2 (1978)
4. Bińkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: ICLR (2018)
5. Cheng, X., Wang, P., Yang, R.: Depth Estimation via Affinity Learned with Convolutional Spatial Propagation Network. In: ECCV. pp. 103–119 (2018)
6. Dai, T., Wong, J., Jiang, Y., Wang, C., Gokmen, C., Zhang, R., Wu, J., Fei-Fei, L.: Automated Creation of Digital Cousins for Robust Policy Learning. In: CoRL (2024)
7. Dastjerdi, M.R.K., Hold-Geoffroy, Y., Eisenmann, J., Khodadadeh, S., Lalonde, J.F.: Guided Co-Modulated GAN for 360° Field of View Extrapolation. In: 3DV. pp. 475–485 (2022)
8. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. In: ICLR (2024)
9. Feng, M., Liu, J., Cui, M., Xie, X.: Diffusion360: Seamless 360 Degree Panoramic Image Generation based on Diffusion Models. arXiv preprint arXiv:2311.13141 (2023)
10. Feng, W., Zhu, W., Fu, T.j., Jampani, V., Akula, A., He, X., Basu, S., Wang, X.E., Wang, W.Y.: LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. NeurIPS **36**, 18225–18250 (2023)
11. Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.F.: Learning to Predict Indoor Illumination from a Single Image. arXiv preprint arXiv:1704.00090 (2017)
12. Guo, Y., Garg, S., Miangoleh, S.M.H., Huang, X., Ren, L.: Depth Any Camera: Zero-Shot Metric Depth Estimation from Any Camera. In: CVPR. pp. 26996–27006 (2025)
13. Hessel, J., Holtzman, A., Forbes, M., Le Bras, R., Choi, Y.: CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In: EMNLP. pp. 7514–7528 (2021)
14. 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. NeurIPS **30** (2017)
15. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: LoRA: Low-Rank Adaptation of Large Language Models. In: ICLR (2022)
16. Huang, Y., Zhou, Y., Wang, J., Huang, K., Liu, X.: DreamCube: RGB-D Panorama Generation via Multi-plane Synchronization. In: ICCV. pp. 24922–24932 (2025)
17. Huang, Z., Wu, X., Zhong, F., Zhao, H., Nießner, M., Lasenby, J.: LiteReality: Graphic-Ready 3D Scene Reconstruction from RGB-D Scans. In: NeurIPS (2025)
18. Kalischek, N., Oechsle, M., Manhardt, F., Henzler, P., Schindler, K., Tombari, F.: Cubediff: Repurposing diffusion-based image models for panorama generation (2025), `https://arxiv.org/abs/2501.17162`
19. Kar, O.F., Yeo, T., Atanov, A., Zamir, A.: 3D Common Corruptions and Data Augmentation. In: CVPR. pp. 18963–18974 (2022)

20. Kocabas, M., Huang, C.H.P., Tesch, J., Müller, L., Hilliges, O., Black, M.J.: SPEC: Seeing People in the Wild With an Estimated Camera. In: ICCV. pp. 11035–11045 (2021)
21. Kocsis, P., Höllein, L., Nießner, M.: IntrinsiX: High-Quality PBR Generation using Image Priors. In: NeurIPS (2025)
22. Kocsis, P., Sitzmann, V., Nießner, M.: Intrinsic Image Diffusion for Indoor Single-view Material Estimation. In: CVPR. pp. 5198–5208 (2024)
23. Kynkäänniemi, T., Karras, T., Aittala, M., Aila, T., Lehtinen, J.: The Role of ImageNet Classes in Fréchet Inception Distance. In: ICLR (2023)
24. Labs, B.F.: Flux.1-dev. `https://huggingface.co/black-forest-labs/FLUX.1-dev` (2025), accessed: 2025-01-19
25. Lee, Y.C., Chen, Y.T., Wang, A., Liao, T.H., Feng, B.Y., Huang, J.B.: VividDream: Generating 3D Scene with Ambient Dynamics. arXiv preprint arXiv:2405.20334 (2024)
26. Li, W., Saeedi, S., McCormac, J., Clark, R., Tzoumanikas, D., Ye, Q., Huang, Y., Tang, R., Leutenegger, S.: InteriorNet: Mega-scale Multi-sensor Photo-realistic Indoor Scenes Dataset. In: BMVC (2018)
27. Li, W., Cai, F., Mi, Y., Yang, Z., Zuo, W., Wang, X., Fan, X.: SceneDreamer360: Text-Driven 3D-Consistent Scene Generation with Panoramic Gaussian Splatting. arXiv preprint arXiv:2408.13711 (2024)
28. Li, Y., Jiang, L., Xu, L., Xiangli, Y., Wang, Z., Lin, D., Dai, B.: MatrixCity: A Large-scale City Dataset for City-scale Neural Rendering and Beyond. In: ICCV. pp. 3205–3215 (2023)
29. Li, Z., Wang, L., Huang, X., Pan, C., Yang, J.: PhyIR: Physics-Based Inverse Rendering for Panoramic Indoor Images. In: CVPR. pp. 12713–12723 (2022)
30. Li, Z., Wu, T., Tan, J., Zhang, M., Wang, J., Lin, D.: IDArb: Intrinsic Decomposition for Arbitrary Number of Input Views and Illuminations. In: ICLR (2025)
31. Liang, R., Gojcic, Z., Ling, H., Munkberg, J., Hasselgren, J., Lin, C.H., Gao, J., Keller, A., Vijaykumar, N., Fidler, S., et al.: Diffusion Renderer: Neural Inverse and Forward Rendering with Video Diffusion Models. In: CVPR. pp. 26069–26080 (2025)
32. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow Matching for Generative Modeling. In: ICLR (2023)
33. Musgrave, F.K., Kolb, C.E., Mace, R.S.: The Synthesis and Rendering of Eroded Fractal Terrains. ACM Siggraph Computer Graphics **23**(3), 41–50 (1989)
34. Parish, Y.I., Müller, P.: Procedural Modeling of Cities. In: Annual conference on Computer Graphics and Interactive Techniques. pp. 301–308 (2001)
35. Peebles, W., Xie, S.: Scalable Diffusion Models with Transformers. In: ICCV. pp. 4195–4205 (2023)
36. Persson, E.: Texture from Humus. https://www.humus.name/index.php?page=Textures (accessed 02/2025)
37. polyhaven.com: HDRIs. https://polyhaven.com/hdris (accessed 02/2025)
38. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: DreamFusion: Text-to-3D using 2D Diffusion. In: International Conference on Learning Representations (2022)
39. Raistrick, A., Lipson, L., Ma, Z., Mei, L., Wang, M., Zuo, Y., Kayan, K., Wen, H., Han, B., Wang, Y., et al.: Infinite Photorealistic Worlds Using Procedural Generation. In: CVPR. pp. 12630–12641 (2023)
40. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In: ICCV. pp. 10912–10922 (2021)

41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-Resolution Image Synthesis with Latent Diffusion Models. In: CVPR. pp. 10684–10695 (2022)
42. Schwarz, K., Rozumny, D., Bulò, S.R., Porzi, L., Kontschieder, P.: A Recipe for Generating 3D Worlds From a Single Image. In: ICCV. pp. 3520–3530 (2025)
43. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: RoFormer: Enhanced Transformer with Rotary Position Embedding. Neurocomputing **568**, 127063 (2024)
44. Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion. In: NeurIPS (2023)
45. Team, H., Wang, Z., Liu, Y., Wu, J., Gu, Z., Wang, H., Zuo, X., Huang, T., Li, W., Zhang, S., et al.: HunyuanWorld 1.0: Generating Immersive, Explorable, and Interactive 3D Worlds from Words or Pixels. arXiv preprint arXiv:2507.21809 (2025)
46. Wang, N.H.A., Liu, Y.L.: Depth Anywhere: Enhancing 360 Monocular Depth Estimation via Perspective Distillation and Unlabeled Data Augmentation. NeurIPS **37**, 127739–127764 (2024)
47. Wang, Q., Li, W., Mou, C., Cheng, X., Zhang, J.: 360DVD: Controllable Panorama Video Generation with 360-Degree Video Diffusion Model. In: CVPR. pp. 6913–6923 (2024)
48. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: MoGe: Unlocking Accurate Monocular Geometry Estimation for Open-Domain Images with Optimal Training Supervision. In: CVPR. pp. 5261–5271 (2025)
49. Wu, T., Zheng, C., Cham, T.J.: PanoDiffusion: 360-degree Panorama Outpainting via Diffusion. In: ICLR (2024)
50. Xiao, B., Wu, H., Xu, W., Dai, X., Hu, H., Lu, Y., Zeng, M., Liu, C., Yuan, L.: Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In: CVPR. pp. 4818–4829 (2024)
51. Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing Scene Viewpoint using Panoramic Place Representation. In: CVPR. pp. 2695–2702 (2012)
52. Yang, S., Tan, J., Zhang, M., Wu, T., Wetzstein, G., Liu, Z., Lin, D.: LayerPano3D: Layered 3D Panorama for Hyper-Immersive Scene Generation. In: SIGGRAPH. pp. 1–10 (2025)
53. Yu, H., Wang, C., Zhuang, P., Menapace, W., Siarohin, A., Cao, J., Jeni, L., Tulyakov, S., Lee, H.Y.: 4Real: Towards Photorealistic 4D Scene Generation via Video Diffusion Models. NeurIPS **37**, 45256–45280 (2024)
54. Yu, H.X., Duan, H., Herrmann, C., Freeman, W.T., Wu, J.: WonderWorld: Interactive 3D Scene Generation from a Single Image. In: CVPR. pp. 5916–5926 (2025)
55. Yu, H.X., Duan, H., Hur, J., Sargent, K., Rubinstein, M., Freeman, W.T., Cole, F., Sun, D., Snavely, N., Wu, J., et al.: WonderJourney: Going from Anywhere to Everywhere. In: CVPR. pp. 6658–6667 (2024)
56. Zeng, Z., Deschaintre, V., Georgiev, I., Hold-Geoffroy, Y., Hu, Y., Luan, F., Yan, L.Q., Hašan, M.: RGB↔X: Image Decomposition and Synthesis Using Material- and Lighting-aware Diffusion Models. In: ACM SIGGRAPH (2024)
57. Zhang, C., Wu, Q., Gambardella, C.C., Huang, X., Phung, D., Ouyang, W., Cai, J.: Taming Stable Diffusion for Text to 360 Panorama Image Generation. In: CVPR. pp. 6347–6357 (2024)
58. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In: CVPR (2018)

59. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3D: A Large Photo-Realistic Dataset for Structured 3D Modeling. In: ECCV. pp. 519–535. Springer (2020)
60. Zhou, S., Fan, Z., Xu, D., Chang, H., Chari, P., Bharadwaj, T., You, S., Wang, Z., Kadambi, A.: DreamScene360: Unconstrained Text-to-3D Scene Generation with Panoramic Gaussian Splatting. In: ECCV. pp. 324–342. Springer (2024)
61. Zhu, J., Luan, F., Huo, Y., Lin, Z., Zhong, Z., Xi, D., Wang, R., Bao, H., Zheng, J., Tang, R.: Learning-Based Inverse Rendering of Complex Indoor Scenes with Differentiable Monte Carlo Raytracing. In: ACM SIGGRAPH Asia (2022)