

# Omni<sup>2</sup>: Unifying Omnidirectional Image Generation and Editing in an Omni Model

Liu Yang  
SJTU  
Shanghai, China  
ylyl.yl@sjtu.edu.cn

Huiyu Duan\*  
SJTU  
Shanghai, China  
huiyuduan@sjtu.edu.cn

Yucheng Zhu  
SJTU  
Shanghai, China  
zyc420@sjtu.edu.cn

Xiaohong Liu  
SJTU  
Shanghai, China  
xiaohongliu@sjtu.edu.cn

Lu Liu  
SJTU  
Shanghai, China  
letteliu@sjtu.edu.cn

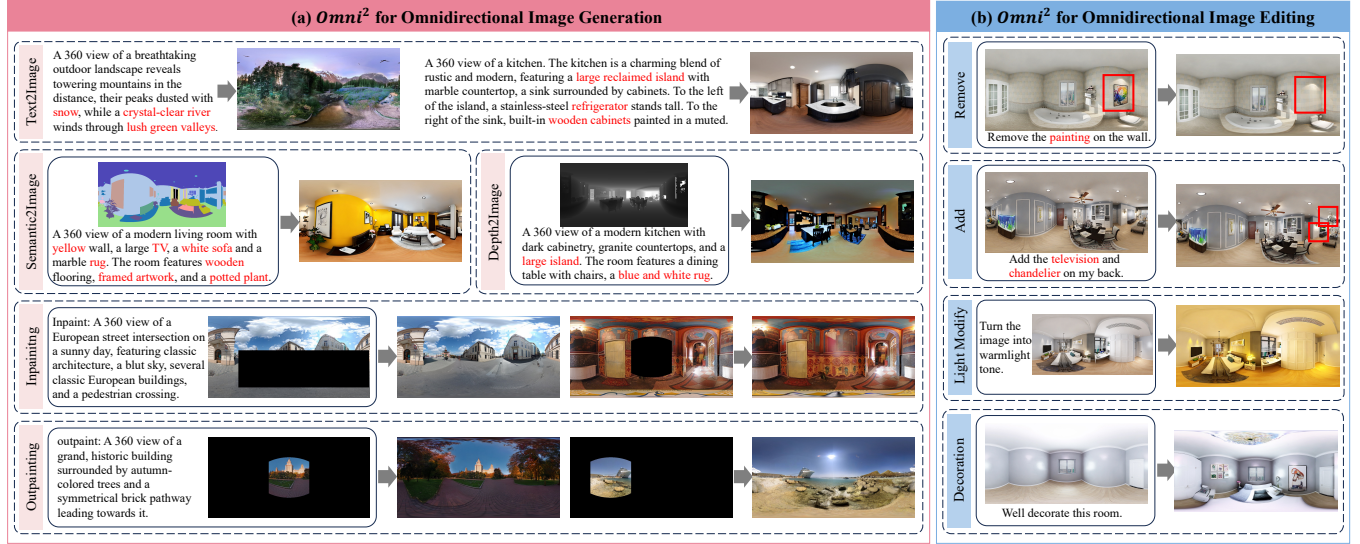
Zitong Xu  
SJTU  
Shanghai, China  
xuzitong@sjtu.edu.cn

Guangji Ma  
UESTC  
Shanghai, China  
guangjima0806@gmail.com

Xiongkuo Min\*  
SJTU  
Shanghai, China  
minxiongkuo@sjtu.edu.cn

Guangtao Zhai\*  
SJTU  
Shanghai, China  
zhaiguangtao@sjtu.edu.cn

Patrick Le Callet  
Université de Nantes  
Nantes, France  
patrick.lecallet@univ-nantes.fr



**Figure 1: We propose the first omni model for omnidirectional image generation and editing, termed *Omni*<sup>2</sup>. *Omni*<sup>2</sup> is capable of handling both omnidirectional image generation and editing with various input conditions, demonstrating strong potential across diverse tasks, as demonstrated in (a) and (b).**

## Abstract

360° omnidirectional images (ODIs) have gained considerable attention recently, and are widely used in various virtual reality (VR) and augmented reality (AR) applications. However, capturing such

\* Corresponding authors.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MM '25, Dublin, Ireland

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.  
ACM ISBN 979-8-4007-2035-2/2025/10  
<https://doi.org/10.1145/3746027.3755405>

images is expensive and requires specialized equipment, making ODI synthesis increasingly important. While common 2D image generation and editing methods are rapidly advancing, these models struggle to deliver satisfactory results when generating or editing ODIs due to the unique format and broad 360° Field-of-View (FoV) of ODIs. To bridge this gap, we construct *Any2Omni*, the first comprehensive ODI generation-editing dataset comprises 60,000+ training data covering diverse input conditions and up to 9 ODI generation and editing tasks. Built upon *Any2Omni*, we propose an *Omni* model for *Omni*-directional image generation and editing (*Omni*<sup>2</sup>), with the capability of handling various ODI generation and editing tasks under diverse input conditions using one model. Extensive experiments demonstrate the superiority and effectiveness of the proposed *Omni*<sup>2</sup> model for both the ODI generation and

editing tasks. Both the Any2Omni dataset and the Omni<sup>2</sup> model are publicly available at: <https://github.com/IntMeGroup/Omni2>.

## CCS Concepts

• **Computing methodologies** → **Computer vision; Virtual reality.**

## Keywords

Omnidirectional Image Generation, Omnidirectional Image Editing, Generative Models, Virtual Reality

### ACM Reference Format:

Liu Yang, Huiyu Duan\*, Yucheng Zhu, Xiaohong Liu, Lu Liu, Zitong Xu, Guangji Ma, Xionghuo Min\*, Guangtao Zhai\*, and Patrick Le Callet. 2025. Omni<sup>2</sup>: Unifying Omnidirectional Image Generation and Editing in an Omni Model. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*, October 27–31, 2025, Dublin, Ireland. ACM, New York, NY, USA, 19 pages. <https://doi.org/10.1145/3746027.3755405>

## 1 Introduction

With the rapid advancement of virtual reality (VR) technology, 360° omnidirectional images (ODIs) have gained increasing attention. However, capturing ODIs requires expensive and specialized hardwares, making ODI synthesis a crucial task. While 2D image generation techniques have gradually matured with the rapid advancement of AIGC [7, 10, 24, 32, 38], ODI generation remains underexplored. Previous works, such as MVDiffusion [37], can only generate ODIs with a limited vertical Field-of-View (FoV) of  $360^\circ \times 90^\circ$ , restricting its applicability in real-world scenarios. Though some other methods [8, 51, 53] are capable of generating full  $360^\circ \times 180^\circ$  ODIs, they primarily focus on text-driven ODI generation, while ignoring other input conditions such as narrow viewport images, semantic maps, depth maps, etc.

High-quality 2D image editing datasets [5, 26, 34, 48, 54] have recently driven the advancement of image editing models [13, 54]. In the context of ODIs, editing is of great importance for enhancing quality of experience in immersive environment. However, omnidirectional image editing remains unexplored. Unlike 2D images, ODIs are stored in the warped equirectangular projection (ERP) format, thus making common 2D image editing algorithms cannot be directly applied. An alternative approach is applying 2D editing models on the designated single view of an ODI after viewport splitting, but it is both time-consuming and inefficient, and may generate inconsistent views. Moreover, due to the unique depth and spatial characteristics of omnidirectional images, conventional 2D editing models struggle to understand the spatial relationships between viewports or even within individual ODI views, making the split-and-edit approach impractical. Therefore, dedicated ODI editing dataset and method are necessary to advance research in this domain.

Built upon the rapid progress in 2D image generation and editing, integrating these tasks into a unified framework has become increasingly popular [39, 45, 47]. In this paper, we aim to unify ODI generation and editing into an efficient model. To this end, we first introduce **Any2Omni**, the first comprehensive dataset for omnidirectional image generation and editing tasks. As shown in Table 1, our dataset integrates multiple input modalities for ODI generation tasks. For the newly defined omnidirectional image editing

tasks, we start by proposing a simple yet effective pipeline capable of generating high-quality, object-level indoor editing samples. Additionally, we introduce two scene-level ODI editing tasks utilizing existing ODI datasets [11, 49, 50, 55]. Overall, our Any2Omni dataset comprises over **60,000** training samples, covering **9** categories of omnidirectional image generation and editing tasks with various input conditions.

Based on Any2Omni, we further introduce the first omni model for omnidirectional image generation and editing, termed **Omni<sup>2</sup>**. Omni<sup>2</sup> adopts a simple yet effective Transformer-based framework to support  $360^\circ \times 180^\circ$  high-quality omnidirectional image synthesis under a variety of multimodal input conditions. In contrast to existing diffusion-based ODI generation methods, which incorporate additional attention blocks for multi-view consistency [37, 53], we introduce a novel approach by executing viewport-based bidirectional attention within a unified Transformer. Our model demonstrates superior performance across a wide range of ODI generation and editing tasks with various input conditions, as shown in Fig. 1.

The main highlights of this work include:

- We unify the ODI generation and editing tasks. To the best of our knowledge, our study is the first work to achieve multimodal input-based ODI generation for various tasks, and the first to explore ODI editing.
- We construct **Any2Omni**, the first comprehensive dataset for omnidirectional image generation and editing that contains over 60,000 training samples covering 9 ODI generation and editing tasks with various input conditions.
- We propose **Omni<sup>2</sup>**, the first omni model for omnidirectional image generation and editing. Omni<sup>2</sup> is capable of processing various input conditions and producing high-quality, viewport-consistent ODIs.
- Extensive experimental results demonstrate that the proposed Omni<sup>2</sup> model exhibits state-of-the-art performance on ODI generation tasks, and shows great potential on ODI editing tasks.

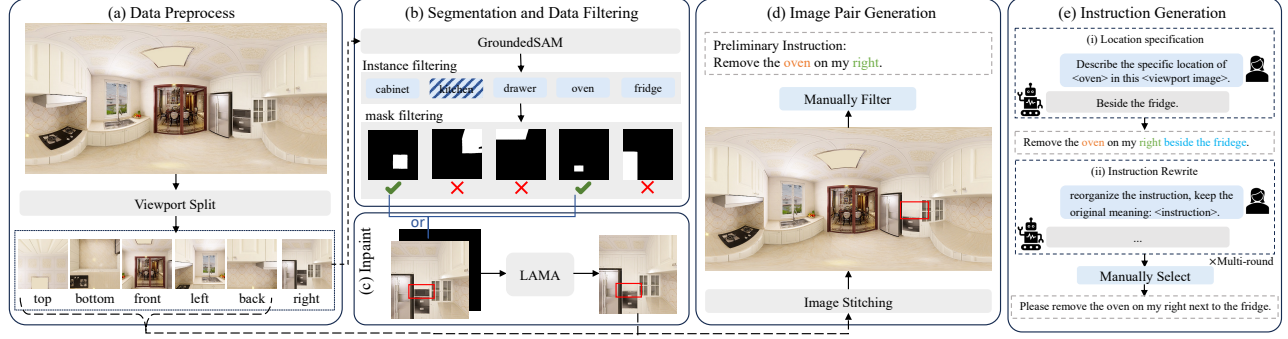
## 2 Related Work

### 2.1 Omnidirectional Image Generation

Although 2D image generation has made significant progress recently, omnidirectional image generation remains notably limited. Prior works, such as MultiDiffusion [4], employ pretrained diffusion model to generate long images from text input. However, these images are not composed of stitched multi-view images thus are not aligned with the true projection process of ODIs. MVDiffusion [37] solves this problem by synchronously generating eight overlapping views using pretrained stable diffusion model [32] and introduces Correspondence-Aware Attention (CAA) module to ensure viewport consistency. However, it only generates the central perspectives, omitting the top and bottom views, limiting its applicability in real-world scenarios. Other methods like DiffPano [51] and PanFusion [53] are able to generate omnidirectional images with the full FoV, yet they are still limited to specific input conditions, and lack support for generation tasks with other input conditions, such as depth-to-image generation and outpainting, etc.







**Figure 2: We present a simple yet effective pipeline for constructing high-quality object-level indoor omnidirectional image editing dataset. Our pipeline is mainly consisted of two parts, i.e., image pair generation and instruction refinement. (a) The ODI input undergoes viewport splitting to generate six perspective images. (b) Viewport image is processed through a segmentation model for instance-level segmentation and class labeling, followed by dual-stage filtering for quality control. (c) A selected instance is removed via inpainting [36]. (d) The edited viewport image is seamlessly stitched with other perspective images to form edited ODI. (e) InternVL2-5 [9] is deployed to refine editing instructions, adding positional details and enhancing linguistic diversity.**

Inst-Inpaint [52] leverages existing dataset [19] to obtain scene graphs within the image, which significantly simplifies the task. However, due to the immaturity of current ODI generation methods and limited sources of ODI-related datasets, we are forced to build the omnidirectional image editing dataset from scratch. In response, we develop a pipeline that can generate high-quality editing image pairs and detailed editing instructions from a single, unconditioned indoor ODI input, as illustrated in Fig. 2.

Compared to outdoor omnidirectional images, indoor ODIs are often more diverse and contain easily segmentable objects. As an initial attempt at omnidirectional image editing, we select Structured3D [55], a large-scale indoor ODI dataset, to generate object-level removal/addition image pairs. Inspired by [52], we constructed our editing image pairs through a pipeline that integrates segmentation and inpainting models to generate precise edited image pairs. However, unlike Inst-Inpaint, which relies on object categories provided by the scene graphs in the GQA dataset [19] for instance segmentation, our task requires an algorithm capable of both detecting instances and outputting corresponding instance categories. To accomplish this, we choose GroundedSAM [31] due to its ability to effectively detect and classify instances in a single pass. Directly applying segmentation on the ERP formatted ODIs yields imprecise results. Furthermore, given the richness of objects in indoor scenes, we propose to generate **editing pairs with predefined viewports**. Specifically, we first split the ODIs into six viewpoints: front, back, left, right, top, and bottom. Then, segmentation is applied on each individual viewport, as shown in Fig. 2(a).

We perform dual-stage post-processing after the segmentation process to ensure data quality, as illustrated in Fig. 2(b). First, we apply instance filtering to remove instances that are irrelevant for our object-level editing purpose, such as “floor”, “cityscape”, and “bedroom”. Since the inpainting operation is applied to each individual viewport, instances that span across multiple viewports may experience disruptions in scene consistency when inpainting is applied to each viewport separately. To address this, we implement an automatic instance filtering method, where instances with segmentation masks located at the image edges are discarded. This

process effectively filters out approximately 60% of the segmentation results.

With the selected instance classes and corresponding segmentation masks, we adopt LAMA [36] to inpaint the instances, thereby generating object-removal outputs, shown in Fig. 2(c)-(d). The outputs are manually selected to ensure image quality. Inspired by [13, 52], object-addition pairs are generated by swapping the input and output images. By executing multiple rounds of the pipeline on the same perspective image and stitching together images edited from various viewports, we further obtain a set of **multi-object editing image pairs**.

Preliminary editing instructions are formed using the detected instance class and the viewport information in the format as “Remove the <instance class> on my <viewport>.” In certain editing cases, the relative location of the object should be pointed out. To address this, we input both the original image and segmented instance class into Internvl2-5 [9], instructing it to generate relative position descriptions, as shown in Fig. 2(e). With these relative position descriptions, we create more detailed editing instructions: “Remove the <instance class> on my <viewport> beside the <reference instance>.” For multi-object editing, the instructions are formed as: “Add the <instance class1> beside the <reference instance1> on my my <viewport1> and <instance class2> on my <viewport2>.” To ensure data diversity, we further utilize Internvl2-5 [9] to refine the generated editing instructions, thereby creating a range of diverse expressions that contribute to improving the model’s robustness. The instruction generation process is carried out over multiple iterations, during which we manually select the most appropriate instructions.

**3.2.2 Scene-level ODI Editing.** The ability to manipulate entire scenes is a defining characteristic that enhances immersion and interactivity in omnidirectional images within virtual reality applications. To this end, we introduce two novel scene-level indoor editing tasks, namely light modify and indoor decoration, designed to enable comprehensive editing of ODIs. We collect the image pairs from Structured3D and standardize the tasks into the input-output pair format, as detailed in the supplementary.



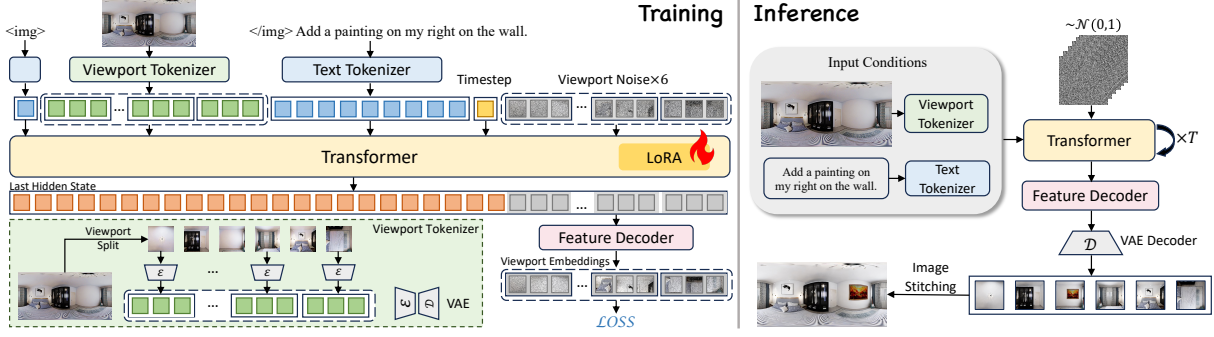


Figure 3: Overview of proposed Omni<sup>2</sup>. Input prompt and Image are tokenized through text tokenizer and viewport tokenizer separately before feeding into a simple yet effective transformer for viewport image generation. During inference, the generated viewports are seamlessly integrated to reconstruct a high-quality omnidirectional image.

## 4 Proposed Method

In this section, we introduce our *unified* model, Omni<sup>2</sup>, towards unifying omnidirectional image generation and editing from various condition inputs in free form using one model.

### 4.1 Overall Architecture

The overall Architecture of Omni<sup>2</sup> is presented in Fig. 3, which adopts a pretrained Transformer as the denoising network. The model takes arbitrarily interleaved text and images in free form as input conditions, which are then tokenized and fed into the Transformer for denoising, along with timestep tokens and viewport noise. The last hidden state is then passed through a feature decoder to obtain the final viewport embeddings. During inference, the viewport embeddings are subsequently fed into a VAE decoder to generate viewport images, which are then stitched together to form a seamless ODI output.

### 4.2 Model Design

**4.2.1 Input Embedding.** We propose generating separate viewports instead of treating the entire ODI as a single entity, in order to ensure viewport consistency and align with the capture process of ODIs. To address this, the noise input is defined on a viewport basis for the viewport-based diffusion process, as shown in Fig. 3.

Input image condition is also viewport tokenized to align with the output. Given an ODI as an input image, we employ a viewport tokenizer to convert the image into viewport-based tokens. Specifically, the input ODI is divided into six overlapping viewports, each of which is then transformed into viewport-based latent representations using a frozen VAE encoder. These representations are then flattened into a sequence of visual tokens with the patch size set to 2 following [28]. During training, the viewport noise is formulated as:

$$z_t^i = tz^i + (1-t)\epsilon^i \quad (1)$$

where  $z^i$  is the viewport-based latent representation of the ground truth  $\{x_i\}_{i=1}^6$ ,  $t$  is the diffusion timestep and  $\epsilon^i \sim \mathcal{N}(0, I)$  is the Gaussian noise. During the inference stage, each viewport noise is formulated as a random sampled Gaussian noise, as shown in Fig. 3. The text prompt and timestep are also tokenized and fed into the pretrained transformer to facilitate the denoising process.

**4.2.2 Attention Mechanism for Viewport Consistency.** In omnidirectional image generation, maintaining viewport consistency is

especially crucial. Previous works often adopt diffusion model for ODI generation, with external attention modules designed to ensure viewport consistency [37, 53]. However, these approaches introduce additional training parameters and lack holistic control across viewports, as viewports are generated separately [37]. In contrast, we propose to leverage and modify the attention mechanism within the Transformer architecture itself to achieve multi-view consistency. Specifically, we adopt the Transformer from Phi-3 [1] as the denoising network. We believe that viewports should be treated equally during the generation process to maintain viewport consistency. To this end, we propose a novel approach to apply bidirectional attention mechanism within the viewport sequences while maintain the causal attention mechanism elsewhere. This approach is not only simpler but also effectively preserves viewport consistency. Furthermore, since viewports are generated using the shared model, it allows for greater overall control over the image, ensuring more coherent results across all generated views.

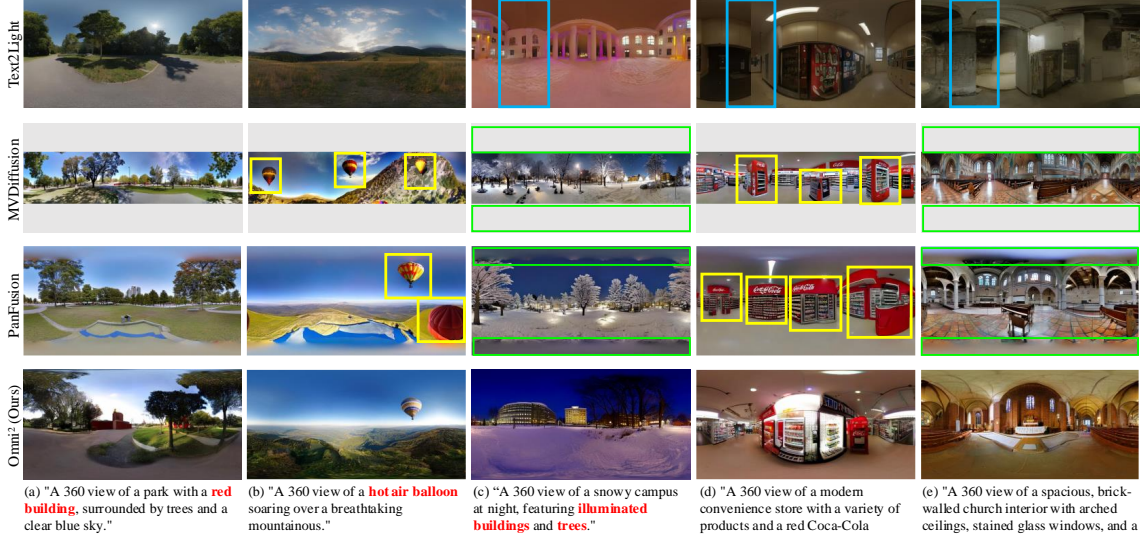
**4.2.3 Omnidirectional Image Generation.** The final hidden state from the Transformer is passed into a feature decoder to map the language space features into latent representations for each viewport. We utilize the final layer from [28] to accomplish this process. The VAE decoders then decode the generated viewport representations into predicted viewports, which are subsequently stitched together to form a seamless omnidirectional image.

### 4.3 Training

**4.3.1 LoRA-based Finetuning.** 2D generation models possess excellent prior knowledge of image generation and strong text comprehension capabilities. Due to the limited scale of omnidirectional images, we use the VAE from SDXL [29] and initialize the Transformer with pretrained weights from [47] to take advantage of the excellent generation ability on 2D images. During training, only the Transformer is finetuned using LoRA [17] to improve training efficiency and retain generalization capability as much as possible.

**4.3.2 Loss Function.** Flow matching [25] is employed to train the model. We extend the loss from single-view to multi-view. For each training step, we randomly sample a timestep  $t$  for all the viewport images  $\{x_i\}_{i=1}^N$ . The loss function is defined as:

$$\mathcal{L} = \mathbb{E} \left[ \sum_{i=1}^N \|(z^i - \epsilon^i) - v_\theta(z_t^i, t, c)\|^2 \right] \quad (2)$$



**Figure 4: Text to ODI comparison between Text2Light [8], MVDiffusion [37], PanFusion [53] and ours. We highlight the left-right inconsistency, repeating objects in different views and top-bottom blurriness/missing with corresponding color boxes. Objects that are missing in some baselines but present in our method are **bolded and highlighted in red** in the prompts.**

**Table 2: Comparison with state-of-the-art methods on text to omnidirectional image task. \*Since MVDiffusion cannot generate full-FOV ODIs, we evaluate its performance here for reference purposes.**

Methods	FAED↓	FID↓	IS↑	CS↑	Inference Time(s)↓
Text2Light [8]	2.70	91.05	4.90	0.7007	135.36
MVDiffusion* [37]	3.06	92.59	6.64	0.5367	252.08
PanFusion [53]	2.54	80.66	7.36	0.8463	67.83
SD+LoRA [32]	2.30	57.97	7.41	0.8540	31.25
Omni <sup>2</sup> (Ours)	<b>2.25</b>	<b>47.32</b>	<b>7.62</b>	<b>0.8887</b>	<b>22.55</b>

**Table 3: User study of text to ODIs.**

Methods	Image Quality↑	Image-Text Consistency↑	Omni-Scene Consistency↑
Text2Light [8]	2.17	2.28	2.78
PanFusion [53]	3.33	3.94	4.44
Omni <sup>2</sup> (Ours)	<b>4.06</b>	<b>4.72</b>	<b>4.83</b>

where  $z^i = \{\mathcal{E}(x_i)\}_{i=1}^N$  represents the latent embeddings of the  $i$ -th viewport image,  $\epsilon^i \sim \mathcal{N}(0, 1)$  is the Gaussian noise,  $z_t^i$  is the noised latent for the  $i$ -th viewport as defined in Eq. 1, and  $c$  denotes the embed condition.

For tasks like light modify and object-level editing, different regions should be adjusted with varying intensities. Specifically, for light modify, distinct areas require different levels of adjustment, while in object-level editing, only a small portion of the regions needs to be modified, thus we make slight modifications to the loss function for these tasks to enable the model to learn the appropriate intensities and to prevent it from simply copying the input image as the output, as discussed in [47]. This was achieved by introducing a weighted loss, where regions that differ significantly from the input image are assigned higher weights. However, although the decoration task belongs to editing, since the change is made to the overall image, applying weighted loss leads to unsatisfactory results. Therefore, we use the original loss for the decoration task.

## 5 Experiment

### 5.1 Implementation Details

**5.1.1 Training Settings.** We train the Omni<sup>2</sup> model on our Any2Omni dataset using an **all in one** training strategy, *i.e.*, all nine tasks, including both generation and editing, are trained simultaneously. The image resolution is set to  $512 \times 1024$  and the viewport FoV is set to  $110^\circ$  with resolution of  $256 \times 256$ . The total number of training step is 10k. The training is performed using the AdamW optimizer with a batch size of 16 and a learning rate of  $1e^{-3}$ , utilizing 2 A6000 GPUs.

**5.1.2 Inference Settings.** The sampling steps  $T$  is set to 50. The guidance scale and image guidance scale are adjusted according to the specific tasks. More implementation details are provided in the supplementary material.

### 5.2 Text to Image

**5.2.1 Qualitative Results.** We compare the text2ODI performance of proposed Omni<sup>2</sup> with other 3 models trained on their respective datasets. The qualitative comparison results are presented in Fig. 4. Images generated by MVDiffusion [37] are padded gray in the top and bottom since MVDiffusion can only generate images with  $90^\circ$  vertical FoV. The images are rotated by  $90^\circ$  to better visualize left-right consistency. As can be observed, ODIs generated by Text2Light [8] exhibit poor left-right consistency and fail to capture detailed object specified in the text prompt, such as the missing “red building” in (a) and the “hot air balloon” in (b). Despite the top-bottom view missing problem, MVDiffusion tends to generate repeating objects across different viewports since the ODIs are stitched from 8 viewports generated by SD [32] blocks, separately, resulting in a loss of overall image coherence. ODIs generated by PanFusion [53] suffer from top-bottom blurriness and also exhibit insensitivity to specific details in the text prompt, *e.g.*, missing illuminated buildings in (c). In contrast, our model is capable of generating ODIs with clear  $360^\circ \times 180^\circ$  FoV, while preserving fine details from the text prompt, demonstrating superior text2ODI performance for both indoor and outdoor ODI generation.

**5.2.2 Quantitative Results.** Table 2 presents the quantitative results of text2ODI task. We also LoRA-finetuned Stable Diffusion [32] to



Figure 5: Outpainting comparison between SIG-SS [14], OmniDreamer [2], PanoDiffusion [44], PanoDiff [40], and ours.

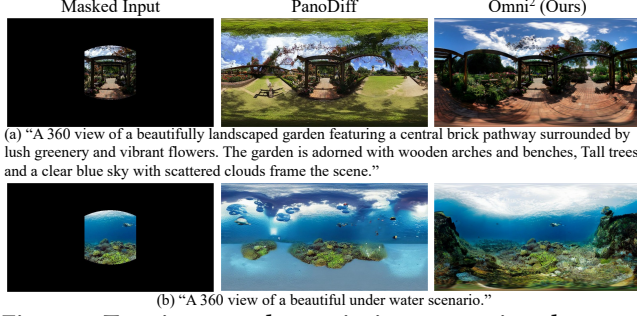


Figure 6: Text-instructed outpainting comparison between PanoDiff [40] and ours.

Table 4: Comparison with state-of-the-art methods on outpainting task.

Task Methods	Image Input			Text-Image Input			
	FAED↓	FID↓	IS↑	FAED↓	FID↓	IS↑	CS↑
SIG-SS [14]	1.08	66.67	5.04	N/A	N/A	N/A	N/A
OmniDreamer [2]	1.83	73.80	5.15	N/A	N/A	N/A	N/A
PanoDiffusion [44]	1.54	127.30	4.19	N/A	N/A	N/A	N/A
PanoDiff [40]	2.21	61.03	6.30	1.22	45.54	6.78	0.8535
Omni <sup>2</sup> (Ours)	<b>1.00</b>	<b>44.13</b>	<b>6.86</b>	<b>0.99</b>	<b>37.40</b>	<b>6.93</b>	<b>0.8620</b>

provide more comparative data. We use Fréchet Auto-Encoder Distance (FAED) [27], Fréchet Inception Distance (FID) [16], Inception Score (IS) [33] and Clip Score (CS) [15] to compare the quality of ODIs generated by Omni<sup>2</sup> with that by other methods. We also report inference time on a single A6000 for efficiency comparison. It can be observed that our model outperforms others in terms of all the evaluation metrics. Moreover, our method has a substantial advantage in inference time due to the attention mechanism using kv-cache.

**5.2.3 User Study.** To better quantize the performance of different methods, we collect 108 text prompts and recruit 20 volunteers to rate the generated ODIs from three perspectives: image quality, image-text consistency and omni-scene consistency. Experimental results presented in Table 3 show that our proposed Omni<sup>2</sup> exhibits an overall superior performance, demonstrating its strong capability of ODI generation.

### 5.3 Multi-modal to Image

We compare the performance of our model with the state-of-the-art methods on the ODI outpainting task. For image-based outpainting, we report FID and IS scores for comparison. For text-guided outpainting, we add the CS metric to evaluate the correspondence between image and text. The results are presented in Table 4. Our model achieves state-of-the-art performance on ODI outpainting tasks, both with and without text prompts.

The qualitative results of image conditioned outpainting are presented in Fig. 5. We only provide the outpainting results from center viewport here, outpainting results from diverse input masks are presented in the supplementary. It should be noted that since we do not know the specific data split for these models, images in our test set could be in their training set. As can be seen in the figure, outpainting results generated by SIG-SS [14] and OmniDreamer [2] suffer from noticeable boundary inconsistencies, while PanoDiffusion [44] tends to produce images with evident artifacts. Although PanoDiff [40] demonstrates relatively better visual quality in terms of boundary consistency, the outpainting content appears less semantically reasonable. In contrast, our model significantly outperforms these baselines by generating semantically meaningful and visually coherent images with superior boundary consistency and enhanced overall visual fidelity.

We also compare the performance of proposed Omni<sup>2</sup> with PanoDiff on text-guided ODI outpainting, qualitative results are shown in Fig. 6. PanoDiff struggles to generate semantically plausible results, e.g., grass in the sky in Fig. 6 (a). In addition, artifacts and meaningless objects appear in the generated ODIs. In contrast, our proposed Omni<sup>2</sup> is able to generate meaningful content that is consistent with the text prompt.

Due to the limited number of dedicated methods for ODI inpainting, semantic2image, and depth2image, the qualitative results of these tasks are presented in the supplementary material.

### 5.4 Omnidirectional Image Editing

Since there is no existing ODI editing method, we compare our method with existing 2D image editing methods in two aspects: directly applying existing editing methods on ODI, and editing the designated viewport separately after viewport splitting.

We first compare our method with directly applying 2D editing on ODIs. The results are shown in Fig. 7. As illustrated, HQ-Edit [20] generates images that are completely unrelated to the source image, while InstructPix2Pix [5] fails to understand the instructions and produces incorrect images, such as the wall being mistakenly painted yellow in Fig. 7(b). Although MagicBrush [54] seems to partially understand some of the editing instructions and makes adjustments to the source image, it struggles with understanding the direction of the edits, which is critical for immersive VR editing. Furthermore, all these methods fail to truly understand the object within the image, as can be clearly seen in fig. 7(a), possibly due to the fact that 2D editing models are not trained to understand the warped ERP format.

We also compare our method with applying 2D editing method on designated views. The results for the edited viewport are presented in Fig. 8. We split the view such that the targeted object is centered in the viewport image. As can be observed, both Instruct-Pix2Pix and HQ-Edit fail to comprehend the instructions while





Figure 7: Visual comparison of directly applying existing 2D editing methods [5, 20, 54] on ODIs versus our method.

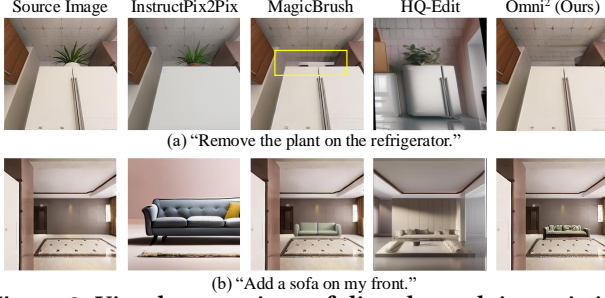


Figure 8: Visual comparison of directly applying existing 2D editing methods [5, 20, 54] on designated views of ODIs versus our method. Only the designated view is presented.

MagicBrush introduces noticeable artifacts, as highlighted in the yellow box. Additionally, the added object lacks the level of detail produced by our method. More importantly, this split-and-edit approach is not applicable to real-world scenarios, as it overlooks the global coherence of the entire 360° image and the complex interactions between different regions of the scene.

The experimental results clearly demonstrate the ineffectiveness of applying 2D editing methods on ODIs, which underscores the importance of developing methods tailored to the unique demands of omnidirectional image editing.

## 5.5 Ablation Study

In this section, we conduct ablation studies to validate the utility of the core modules in Omni².

**5.5.1 Bidirectional Attention Mechanism.** We adopt bidirectional attention within the viewport sequence to maintain viewport consistency. Fig. 9-Top shows the text2image result with causal attention. As can be seen, great distortion occurs, and there are clear boundaries for overlapping viewports. The result demonstrates the effectiveness of the proposed viewport-based bidirectional attention.

**5.5.2 Loss Function.** We modify the loss function so that the model learns to modify dedicated regions within the input image while keeping other parts unchanged. For the decoration task, we use the origin loss for better results. Ablation studies are conducted to demonstrate the effectiveness of the modified loss function, with results presented in Fig. 9-Middle and -Bottom. For object-level image editing, the model tends to simply copying the input as output without weighted loss. For the decoration task, since modifications are made on the whole image, applying weighted loss leads to heavy distortion in generated image.

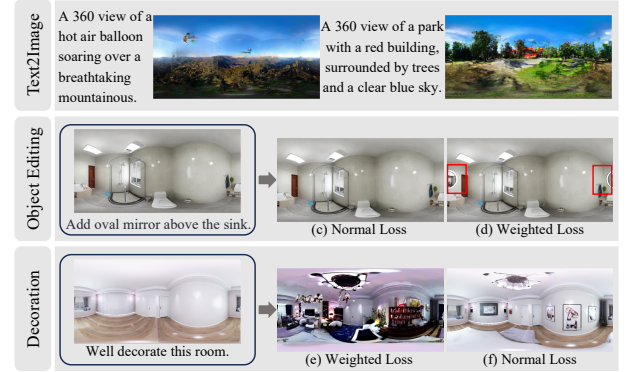


Figure 9: Visual results of ablation studies. Top: ablation on viewport-based bidirectional attention; Middle: ablation on weighed loss for object-level editing tasks; Bottom: ablation on normal loss for decoration task.

Table 5: Ablation study on LoRA rank, with T2I performance results reported.

LoRA rank $r$	FID↓	IS↑	CS↑
8	48.68	6.96	0.8873
32	47.63	7.34	0.8853
16 (Ours)	<b>47.32</b>	<b>7.62</b>	<b>0.8887</b>

**5.5.3 LoRA Rank.** The LoRA rank  $r$  is set to 16 in the paper, we also compare the influence of different rank and report the quantitative results of text2image task in Table 5. As shown in the table, using  $r = 16$  generally performs better than using  $r = 8$  and  $r = 32$ , which further validates the effect of LoRA finetuning.

## 6 Conclusion

In this paper, we aim to unify omnidirectional image generation and editing tasks. Specifically, we first construct Any2Omni, the first comprehensive dataset containing 60,000+ data encompassing various ODI generation and editing tasks. Any2Omni contains the first comprehensive multi-task ODI generation subset with diverse input conditions and the first ODI editing subset featuring both object-level and scene-level editing tasks. Based on the dataset, we propose Omni², an omni model for omnidirectional image generation and editing via a Transformer architecture with viewport-based bidirectional attention mechanism, which is able to process multi-modal input conditions and generate high-quality ODIs across various tasks. Extensive experiments demonstrate that our proposed model achieves state-of-the-art performance on various ODI generation tasks and exhibits strong potential for ODI editing tasks.

## 7 Acknowledgments

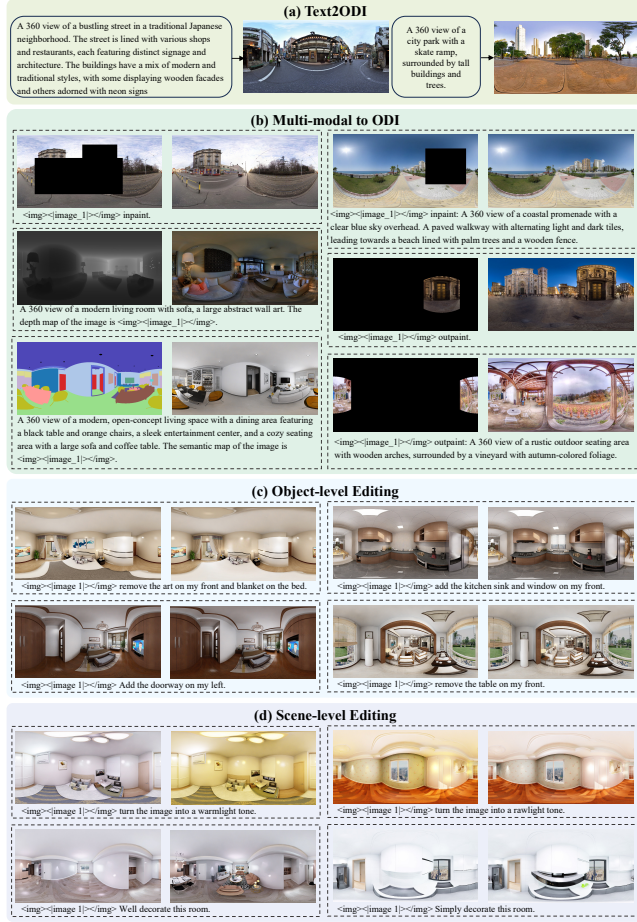
This work was supported in part by the National Natural Science Foundation of China under Grants 62401365, 62225112, 62271312, 62132006, U24A20220, and in part by the China Postdoctoral Science Foundation under Grant Number BX20250411, 2025M773473, and in part by STCSM under Grant 22DZ2229005.

## References

- [1] Marah Abidin, Jyoti Aneja, Hany Awadallah, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219* (2024).
- [2] Naofumi Akimoto, Yuh Matsuo, and Yoshimitsu Aoki. 2022. Diverse Plausible 360-Degree Image Outpainting for Efficient 3DCG Background Creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11441–11450.
- [3] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiro Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. 2021. Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 3727–3737.
- [4] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. [n. d.]. Multidiffusion: Fusing diffusion paths for controlled image generation. ([n. d.]).
- [5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18392–18402.
- [6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158* (2017).
- [7] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. 2023. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704* (2023).
- [8] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. 2022. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.
- [9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 24185–24198.
- [10] Huiyu Duan, Qiang Hu, Jiarui Wang, Liu Yang, Zitong Xu, Lu Liu, Xiongkuo Min, Chunlei Cai, Tianxiao Ye, Xiaoyun Zhang, et al. 2025. Finevq: Fine-grained user generated content video quality assessment. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*. 3206–3217.
- [11] Huiyu Duan, Xiongkuo Min, Yucheng Zhu, Guangtao Zhai, Xiaokang Yang, and Patrick Le Callet. 2022. Confusing image quality assessment: Toward better augmented reality experience. *IEEE Transactions on Image Processing (TIP)* 31 (2022), 7206–7221.
- [12] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to Predict Indoor Illumination from a Single Image. *ACM Transactions on Graphics (SIGGRAPH Asia)* 9, 4 (2017).
- [13] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. 2024. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12709–12720.
- [14] Takayuki Hara, Yusuke Mukuta, and Tatsuya Harada. 2023. Spherical Image Generation From a Few Normal-Field-of-View Images by Considering Scene Symmetry. *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)* 45, 05 (May 2023), 6339–6353. doi:10.1109/TPAMI.2022.3215933
- [15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).
- [16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems (NIPS)* 30 (2017).
- [17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *Proceedings of The International Conference on Learning Representations (ICLR)* 1, 2 (2022), 3.
- [18] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. 2024. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8362–8371.
- [19] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6700–6709.
- [20] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990* (2024).
- [21] Masum Shah Junayed, Arezoo Sadehghzadeh, Md Baharul Islam, Lai-Kuan Wong, and Tarkan Aydin. 2022. HiMODE: A hybrid monocular omnidirectional depth estimation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5212–5221.
- [22] Black Forest Labs. 2024. FLUX. <https://github.com/black-forest-labs/flux>.
- [23] Yuyan Li, Zhixin Yan, Ye Duan, and Liu Ren. 2021. Panodepth: A two-stage approach for monocular omnidirectional depth estimation. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 648–658.
- [24] Lu Liu, Huiyu Duan, Qiang Hu, Liu Yang, Chunlei Cai, Tianxiao Ye, Huayu Liu, Xiaoyun Zhang, and Guangtao Zhai. 2024. F-Bench: Rethinking Human Preference Evaluation Metrics for Benchmarking Face Generation, Customization, and Restoration. *arXiv preprint arXiv:2412.13155* (2024).
- [25] Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003* (2022).
- [26] Xiongkuo Min, Huiyu Duan, Wei Sun, Yucheng Zhu, and Guangtao Zhai. 2024. Perceptual video quality assessment: A survey. *Science China Information Sciences* 67, 11 (2024), 211301.
- [27] Changgyoon Oh, Wonjune Cho, Yujeong Chae, Daehee Park, Lin Wang, and Kuk-Jin Yoon. 2022. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 352–371.
- [28] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4195–4205.
- [29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).
- [30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.
- [31] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159* (2024).
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.
- [33] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in Neural Information Processing Systems (NIPS)* 29 (2016).
- [34] Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. 2024. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8871–8879.
- [35] Peter Sushko, Ayana Bharadwaj, Zhi Yang Lim, Vasily Ilin, Ben Caffee, Dongping Chen, Mohammadreza Salehi, Cheng-Yu Hsieh, and Ranjay Krishna. 2025. REALEDIT: Reddit Edits As a Large-scale Empirical Dataset for Image Transformations. *arXiv preprint arXiv:2502.03629* (2025).
- [36] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2149–2159.
- [37] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. 2023. MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion. *arXiv:2307.01097 [cs.CV]* <https://arxiv.org/abs/2307.01097>
- [38] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. 2024. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in Neural Information Processing Systems (NIPS)* 37 (2024), 84839–84865.
- [39] Xueyun Tian, Wei Li, Bingbing Xu, Yige Yuan, Yuanzhuo Wang, and Huawei Shen. 2025. Mige: A unified framework for multimodal instruction-based image generation and editing. *arXiv preprint arXiv:2502.21291* (2025).
- [40] Jionghao Wang, Ziyu Chen, Jun Ling, Rong Xie, and Li Song. 2023. 360-Degree Panorama Generation from Few Unregistered NFOV Images. In *Proceedings of the*

- 31st ACM International Conference on Multimedia (ACM MM). ACM, 6811–6821. doi:10.1145/3581783.3612508
- [41] Qian Wang, Weiqi Li, Chong Mou, Xinhua Cheng, and Jian Zhang. 2024. 360dvd: Controllable panorama video generation with 360-degree video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6913–6923.
  - [42] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18359–18369.
  - [43] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhui Chen. 2024. OmniEdit: Building Image Editing Generalist Models Through Specialist Supervision. In *Proceedings of The Thirteenth International Conference on Learning Representations (ICLR)*.
  - [44] Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. 2024. PanoDiffusion: 360-degree Panorama Outpainting via Diffusion. arXiv:2307.03177 [cs.CV] <https://arxiv.org/abs/2307.03177>
  - [45] Bin Xia, Yuechen Zhang, Jingyao Li, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. 2024. DreamOmni: Unified Image Generation and Editing. *arXiv preprint arXiv:2412.17098* (2024).
  - [46] Jianxiong Xiao, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2012. Recognizing scene viewpoint using panoramic place representation. In *2012 IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2695–2702.
  - [47] Shitao Xiao, Yuezhe Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Chaofan Li, Shutong Wang, Tiejun Huang, and Zheng Liu. 2024. Omnigen: Unified image generation. *arXiv preprint arXiv:2409.11340* (2024).
  - [48] Zitong Xu, Huiyu Duan, Bingnan Liu, Guangji Ma, Jiarui Wang, Liu Yang, Shiqi Gao, Xiaoyu Wang, Jia Wang, Xiongkuo Min, et al. 2025. LMM4Edit: Benchmarking and Evaluating Multimodal Image Editing with LMMs. *arXiv preprint arXiv:2507.16193* (2025).
  - [49] Liu Yang, Huiyu Duan, Long Teng, Yucheng Zhu, Xiaohong Liu, Menghan Hu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. 2024. Aigcoiqa2024: Perceptual quality assessment of ai generated omnidirectional images. In *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1239–1245.
  - [50] Liu Yang, Huiyu Duan, Jiarui Wang, Jing Liu, Menghan Hu, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. 2025. Quality Assessment and Distortion-aware Saliency Prediction for AI-Generated Omnidirectional Images. *arXiv preprint arXiv:2506.21925* (2025).
  - [51] Weicai Ye, Chenhao Ji, Zheng Chen, Junyao Gao, Xiaoshui Huang, Song-Hai Zhang, Wanli Ouyang, Tong He, Cairong Zhao, and Guofeng Zhang. 2024. Diff-Pano: Scalable and Consistent Text to Panorama Generation with Spherical Epipolar-Aware Diffusion. *arXiv preprint arXiv:2410.24203* (2024).
  - [52] Ahmet Burak Yildirim, Vedat Baday, Erkut Erdem, Aykut Erdem, and Aysegül Dundar. 2023. Inst-inpaint: Instructing to remove objects with diffusion models. *arXiv preprint arXiv:2304.03246* (2023).
  - [53] Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. 2024. Taming stable diffusion for text to 360 panorama image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6347–6357.
  - [54] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems (NIPS)* 36 (2023), 31428–31449.
  - [55] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. 2020. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 519–535.





**Figure 10: Examples of Any2Omni dataset. The input of all tasks are organized into an interleaved image-text sequence format.**

## A More Details of Any2Omni Construction

Fig. 10 presents some examples of Any2Omni dataset. Detailed composition and task description of Any2Omni is presented below.

### A.1 Text to Image

Input for this subset of data is text only. Although prior text2ODI works utilize captioned ODI datasets for training [37, 53], some are captioned only for individual views rather than the entire ODI [37], and there is no standardized annotation format for full ODIs [51, 53]. We aim to construct a comprehensive text-to-ODI dataset with both short and detailed text descriptions for diverse ODI generation.

Leveraging the powerful multi-modal understanding capabilities of vision-language models (VLMs), we employ InternVL2-5 [9] to generate text descriptions for 20,000 ODIs from the SUN360 dataset [46]. SUN360 encompasses a diverse range of indoor and outdoor scenes, making it well-suited for text-to-ODI training both in terms of scene diversity and dataset scale. We adopt a dual-strategy approach for generating diverse text descriptions: for half of the images, we input length-unconstrained prompts to generate detailed text descriptions, while for the other half, we restrict

prompts to concise the output text of no more than 20 words. This approach ensures that our dataset supports both generating high-quality ODIs from brief prompts and producing highly-detailed text-aligned ODIs based on fine-grained text descriptions.

### A.2 Multi-modal to Image

We integrate existing ODI generation tasks and propose new ones utilizing existing ODI data to create diverse multi-modal to ODI training examples. Examples of this subset is presented in Fig. 10(b). Trained on these data, our Omni<sup>2</sup> model is capable of supporting a wide range of task types with multi-modal inputs, enabling more versatile and comprehensive ODI generation capability.

**A.2.1 Inpainting.** We select image from SUN360 [46] dataset as the ground truth and generate masks with random shapes including rectangular, irregular and single-view masks, to create the inpainting dataset. The inpainting task is further categorized into text-guided inpainting, *i.e.*, the masked image and GT description generated by Internvl2-5 are used as input, and non-text-guided inpainting, where only the masked image is used.

**A.2.2 Outpainting.** We select images from the SUN360 dataset and extract partial viewpoints as input for outpainting. Similar to inpainting, the outpainting task is also divided into text-guided and non-text-guided categories, depending on whether a description is provided alongside the input image.

**A.2.3 Depth to Image.** ODIs are highly relevant to depth information, and some existing works have explored depth estimation for ODIs [3, 21, 23]. However, few works have been done to generate ODIs with the aid of 360° depth information. To bridge this gap, we introduce a Depth2Image task for ODI and construct a dedicated subset using existing ODI depth estimation datasets. Specifically, we select images and their corresponding depth maps from the Pano3D dataset [3] and generate scene descriptions using Internvl2-5.

**A.2.4 Semantic to Image.** Structured3D dataset [55] offers a diverse collection of indoor ODIs and their corresponding semantic segmentation maps. Similarly to 2D images, we define a semantic2image task for omnidirectional images. We utilize Internvl2-5 to generate descriptions for the ODIs and incorporate both these descriptions and the semantic segmentation maps as inputs to train the model that produces semantically coherent omnidirectional images.

### A.3 Object-level ODI Editing

We utilize the pipeline proposed in the main paper to construct the first object-level ODI editing datasets, including single-object removal/addition and multi-object removal/addition. Notably, the simple pipeline we propose supports the generation of large-scale object-level ODI editing datasets. In this work, to ensure balanced training data for each task, we generate a total of over 12,000 editing data entries across thousands of scenes. Given the richness of semantic information in indoor ODI environments, the object-level ODI editing can be further categorized into: single object removal/add, multi-object removal/add in one view and multi-object removal/add across different views. Some examples are presented in Fig. 10(c).

**Table 7: Comparison with more baselines on text to omnidirectional image task.**

Methods	FAED↓	FID↓	IS↑	CS↑
MVDiffusion (retrained) [37]	3.11	69.21	5.58	0.564
SD+LoRA [53]	2.30	57.97	7.41	0.854
Flux+LoRA [32]	2.53	53.87	<b>7.87</b>	0.873
Omni <sup>2</sup> (Ours)	<b>2.25</b>	<b>47.32</b>	7.62	<b>0.888</b>

**Table 6: Inference settings for different tasks.**

Tasks	Guidance Scale	Image Guidance Scale
Text2ODI	2.5	N/A
Inpainting	2.5	1.8
Outpainting	2.5	1.8
Depth2Image	2.0	1.8
Semantic2Image	2.0	1.8
Object-level Editing	3.0	1.8
Light-modify	3.0	1.8
Decoration	3.5	1.8

#### A.4 Scene-level ODI Editing

Apart from using synthetic data pipeline for generating object-level editing, we propose two insightful scene-level indoor editing tasks utilizing data in Structured3D dataset [55] as shown in Fig. 10(d). Structured3D is a large photo-realistic dataset with high-quality rendered images of various types, thus can be well applied for image editing tasks.

**A.4.1 Light-Modify.** Indoor lighting modification plays a crucial role in virtual reality experiences, and has broad application in real life. Structured3D dataset utilizes industry-leading rendering engines to simulate photo-realistic indoor scenes under different lighting conditions. We select data from this dataset and generate corresponding instructions to build a dataset for indoor light-modify tasks.

**A.4.2 Decoration.** Structured3D generates different configurations (full, simple and empty) of the same room by removing some or all furniture. Building upon this dataset, we further define an interesting and meaningful task, termed indoor decoration. We utilize the data from Structured3D under different configurations and generate instructions to construct a dataset for this task.

## B More Details of Experiment Settings

### B.1 Training and Inference Details

Omni<sup>2</sup> is able to process interleaved texts and images as input conditions. This is achieved by **joint training** with data from multiple tasks. We train the model with a resolution of  $256 \times 256$  and FoV of  $100^\circ$  for perspective images for 10k steps, with a batch size of 16, utilizing two A6000 GPUs. Guidance scale is set as the strength of the text guidance, the larger the guidance, the more similar the generated image will be to the prompt. Image guidance scale indicates the guidance strength of the input image, where larger image guidance leads to generating images closer to the input image condition. During inference, guidance scale and image guidance scale are set based on different tasks to achieve better visual results, as presented in Table 6.

## C More Details of Comparison Experiment

### C.1 Text2ODI

#### C.1.1 Baseline Models.

- Text2Light [8] adopts a two-stage approach that first generates a low-resolution ODI based on the input text, and then expands it to ultra-high resolution. The model provides two sets of checkpoints for indoor and outdoor ODI generation, during inference, we adopt these two checkpoints based on the input prompt.
- MVDiffusion [37] generates 8 viewports using pretrained SD [32] blocks separately and fine-tunes the inserted CAA block for multi-view consistency. We adopt the pretrained weights for comparison purposes. During training, each viewport is assigned a viewport-specific prompt. However, during inference, the model lacks the ability to process the prompt for the entire ODI; instead, it simply copies the prompt for each viewport, leading to noticeable object repetition across different views. Since the generated image lacks top and bottom views, the quantitative results may not be completely reliable and are presented in the main paper for reference only.
- PanFusion [53] is a text-to-ODI model designed to reduce distortion caused by projecting perspective images onto an ODI canvas, while also offering global layout guidance. However, the generated images still suffer from quality issues, such as repeated objects appearing across different views, which significantly affects the visual coherence and overall aesthetic quality. Additionally, noticeable blurriness is observed at the top and bottom regions of the generated ODIs, as discussed in the main paper.

**C.1.2 More baseline comparison.** We further adapt MVDiffusion [37] to generate six viewports as our model and retrain it on our database and LoRA-finetuned Stable Diffusion [32] and Flux [22] on Any2Omni T2I dataset for more baseline comparison. The results are reported in Table 7. As reported in the table, MVDiffusion shows great overfitting after retraining. Our model attains an overall state-of-the-art performance.

**C.1.3 User Study.** User study is conducted to a human preference perspective to provide additional insights for comparing text2ODI methods. MVDiffusion is excluded from the comparison as it is unable to generate full Field-of-View (FoV) ODIs. We select 108 prompts, covering both indoor and outdoor scenarios for generating ODIs. Examples of the prompts and generated ODIs are presented in Fig. 11 and Fig. 12 to serve as more T2ODI comparison results. The images are rotated by  $90^\circ$  to show left-right consistency.

Participants are instructed to rate the generated ODIs on a scale from 1 to 5, in increments of 1, from the following three perspectives:

**Image Quality:** The overall quality of the generated ODI, including both low-level aspects (e.g., color, brightness, etc.) and high-level attributes (e.g., authenticity, aesthetic appeal and scene coherence). (1) Poor: The ODI contains severe distortions, with noticeable noise, unnatural colors, and numerous artifacts, all of which significantly impact viewing comfort and visual coherence. (2) Bad: The overall image quality is low, with visible noise and artifacts. While distortions are present, they are less disruptive compared to the “Poor”

level. (3) Fair: There are slight distortions within the generated ODI, the low-level quality is acceptable but there are artifacts and unrealistic scenes that affect the aesthetic quality. (4) Good: The overall image quality is high, with minimal distortions and natural colors. However, the scene may still contain elements that appear somewhat unrealistic, slightly affecting authenticity. (5) Excellent: No distortions occur in the generated ODI, the color is natural and colorful, the image is smooth and the scene is coherent and closely resembles real-world environments.

*Image-Text Consistency:* The degree to which the generated image aligns with the content of the text prompt. (1) Poor: The scene is totally irrelevant to the text descriptions. (2) Bad: The scene is of low relevance to the text descriptions, while the general background may loosely match, the specific details mentioned in the prompt are missing or incorrect. (3) Fair: The generated image captures some key elements of the text description, but several important details are either inaccurate, missing, or only partially reflected. (4) Good: Most of the content described in the prompt is accurately represented in the generated image, though a few minor details may be inconsistent or underrepresented. (5) Excellent: The generated image aligns very closely with the text prompt, with both global structure and fine-grained details accurately depicted.

*Omni-Scene Consistency.* The degree of consistency across the scene, including left-right alignment and overall scene coherence. (1) Poor: The left and right regions are completely inconsistent, depicting different or unrelated scene content. (2) Bad: The left and right views are misaligned, though they attempt to depict similar content. Clear discontinuities are present. (3) Fair: Left-right consistency is generally acceptable, but inconsistencies remain across adjacent viewports, affecting the perception of a coherent 360° scene. (4) Good: The overall scene consistency is ok with only minor mismatches or discontinuities in certain regions. (5) The scene is fully consistent across all directions, with seamless transitions between viewports and no perceptible misalignments.

The mean opinion scores (MOS) for each model, averaged across all participants, are reported in the main paper for each evaluation criterion.

## C.2 Outpainting

We provide more outpainting results from diverse viewport input masks to better show the performance of our proposed Omni<sup>2</sup>. The results without text guidance are provided in Fig. 13-Top. Our model shows state-of-the-art performance on image-input outpainting tasks.

Fig. 13-Bottom presents the results of proposed Omni<sup>2</sup> and PanoDiff [40] from text-image input. PanoDiff is an ODI outpainting model trained on SUN360 dataset, which may have overlap with our testing set. As can be seen from the figure, outpainting results of PanoDiff suffer from semantic inconsistencies within the scenes and noticeable floating artifacts especially in the top and bottom viewports. Moreover, the model demonstrates limited sensitivity to the detailed content described in the text prompt.

## D More visualization results of Omni<sup>2</sup>

### D.1 Inpainting

As there are limited ODI inpainting methods, we only present qualitative results in Fig. 14 for visualization. Fig. 14-Top presents ODI inpaint with a single masked image as input, different mask shapes are adopted. Our Omni<sup>2</sup> is capable of generating plausible content within the masked regions while maintaining overall scene consistency. The results of ODI inpainting with both text guidance and masked image are presented in Fig. 14-Bottom presents inpainting results guided by both the masked image and a text prompt. As illustrated, Omni<sup>2</sup> excels at incorporating textual guidance to produce seamless, text-aligned ODIs.

### D.2 Semantic2Image

The input conditions for this task are semantic maps and the corresponding scene descriptions. Our proposed Omni<sup>2</sup> demonstrates superior performance on these newly introduced tasks, generating images that exhibit high alignment with the text descriptions based on the semantic maps. The qualitative results are presented in Fig. 15-Top.

### D.3 Depth2Image

The input conditions for this task are depth maps and the corresponding scene descriptions. Qualitative results of this task is presented in Fig. 15-Bottom. The performance on this task is currently limited by the quality of the training dataset, which can be enhanced in future work.

### D.4 ODI Editing

As discussed in the main paper, we are the first to explore editing tasks specifically for omnidirectional images, where existing 2D image editing methods are not directly applicable. In this section, we provide more qualitative of Omni<sup>2</sup> on ODI editing tasks, including both object-level editing and scene-level editing. The results are shown in Fig. 16. Omni<sup>2</sup> demonstrate strong capabilities across these proposed editing tasks.



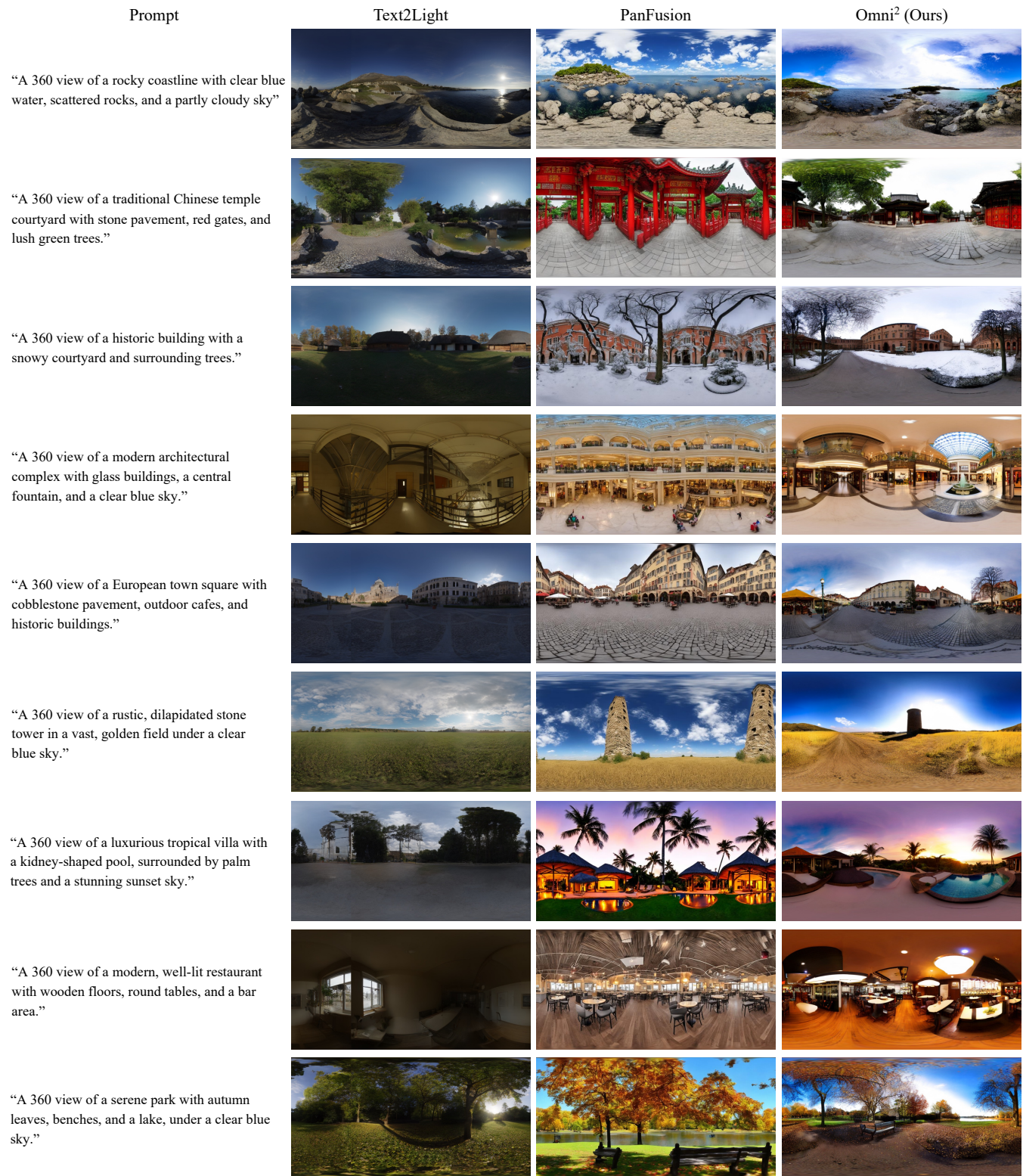


Figure 11: Examples of generated ODIs for user study.

























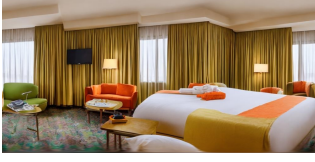




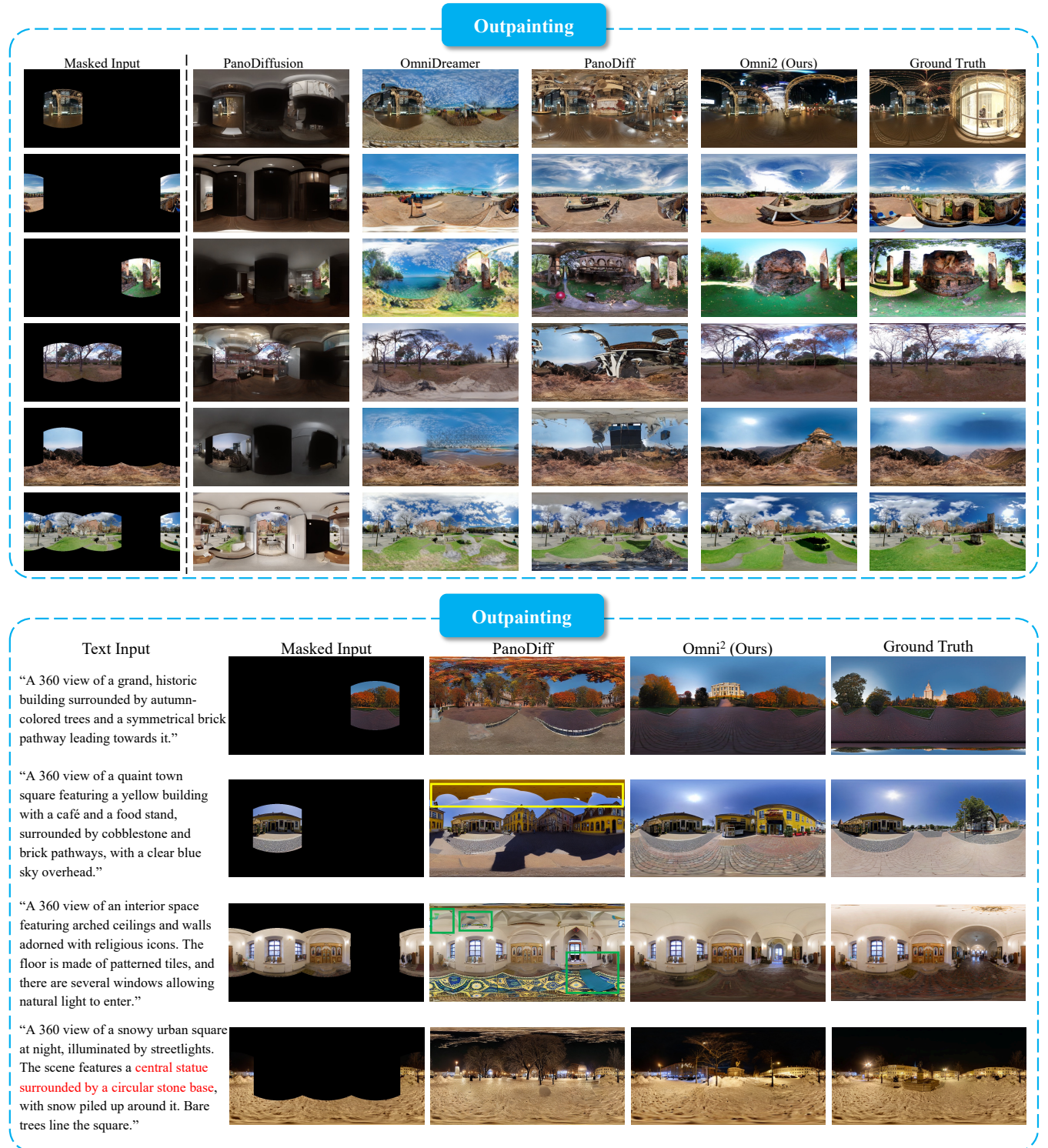
Prompt	Text2Light	PanFusion	Omni <sup>2</sup> (Ours)
“A 360 view of a historic church interior with arched ceilings, wooden pews, and a checkered floor.”			
“A 360 view of a luxurious, sunlit atrium with elegant furniture, a central fountain, and a grand staircase.”			
“A 360 view of a grand, illuminated building at night, surrounded by cobblestone plazas and flanked by trees. And the lighting highlights the building's intricate details against the dark sky.”			
“A 360 view of a well-lit, modern convenience store with a variety of products neatly arranged on shelves and in refrigerated sections. The store features bright lighting and a clean, organized layout. The center has a display of snacks and a red Coca-Cola vending machine.”			
“A 360 view of a ferry deck with a yellow and white painted floor, featuring a large white bridge structure in the center. The sky is overcast with a hint of sunset on the horizon.”			
“A 360 view of a charming courtyard surrounded by white-washed buildings adorned with lush greenery and colorful flowers. Potted plants and hanging baskets adding vibrant touches.”			
“A 360 view of an abandoned, graffiti-covered building with large windows and a curved ceiling. There are scattered leaves and debris. The walls are adorned with colorful murals and graffiti, including a prominent red and black design near the center.”			
“A 360 view of hotel room featuring a large bed with a green bedspread and white pillows, positioned near a balcony with colorful curtains. The room has a green sofa, a television, and a small table with a lamp.”			
“A 360 view of a picturesque canal scene in Venice, Italy, featuring a wooden bridge leading to a vibrant area with colorful buildings. The architecture includes a striking orange building on the left and a series of yellow and white buildings on the right. The sky is clear with a hint of sunset.”			

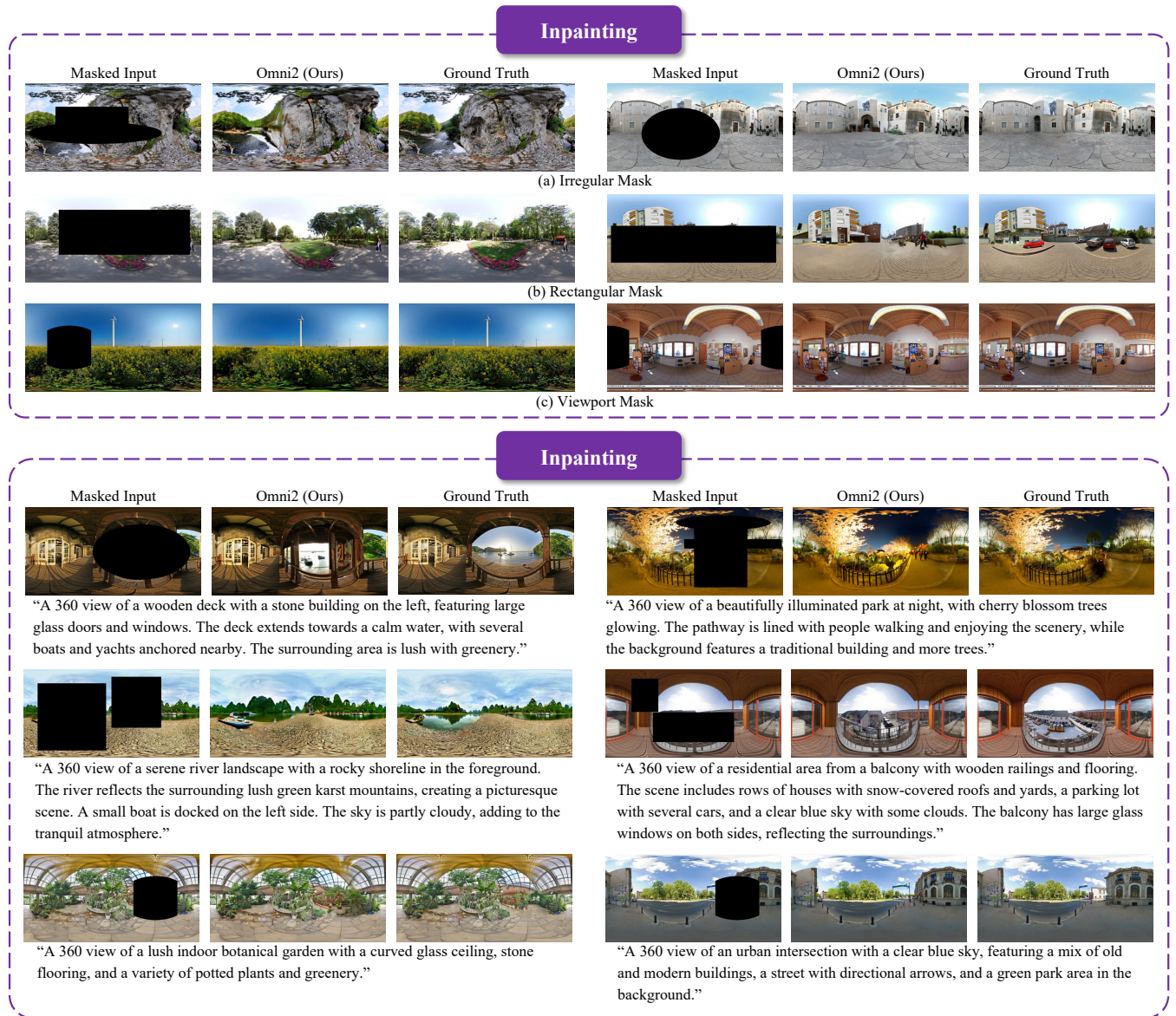
Figure 12: Examples of generated ODIs for user study.





**Figure 13: More comparison results of outpainting.** Top: outpainting comparison between PanoDiffusion [44], OmniDreamer [2], PanoDiff [40], and ours from various masked input images. Bottom: Outpainting comparison between PanoDiff and ours from various masked input images. We highlight the **semantic inconsistencies** and **folating artifact** with corresponding color boxes. Key objects that are missing in PanoDiff but present in our method are highlighted in **red** in the prompts.





**Figure 14: Qualitative results of ODI inpainting with Omni<sup>2</sup>. Top: ODI inpainting using a single masked image as input. Bottom: ODI inpainting guided by both a masked image and text descriptions.**



Figure 15: Qualitative results of Semantic2Image (top) and Depth2Image (bottom). Key words in the text prompt are marked in red.



Figure 16: Qualitative results of object-level (top two rows) and scene-level (bottom two rows) editing tasks.