ENHANCING IN-CONTEXT PANORAMIC GENERATION VIA GEOMETRIC-AWARE PRETRAINING

Anonymous authors
Paper under double-blind review

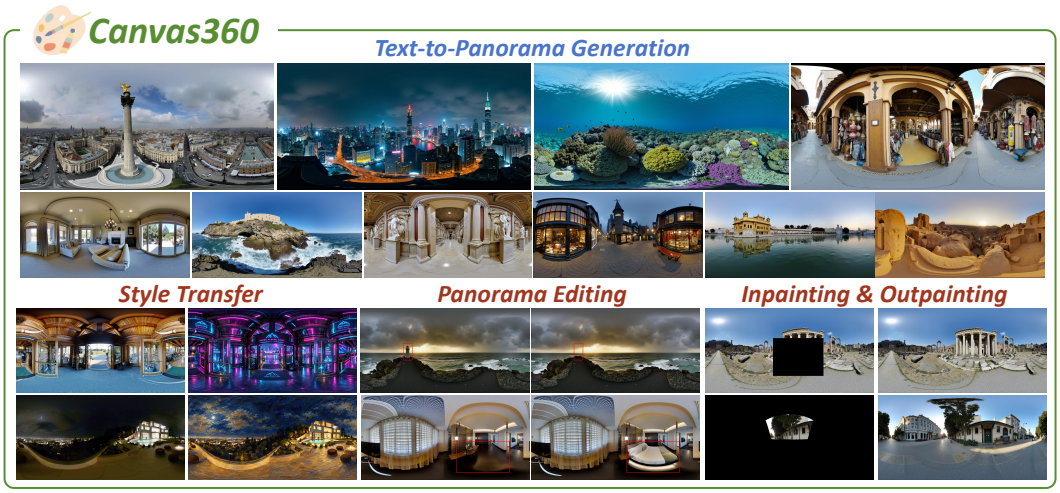

Figure 1: *Canvas360* transfers an RGB–depth panoramic prior to four in-context tasks. The examples show inpainting, outpainting, editing, and style transfer that preserve scene layout under equirectangular projection. Additional results appear in Sec. G.

ABSTRACT

Panoramic image generation must preserve scene structure across a full spherical field of view, but horizontal wrap-around and latitude-dependent distortion make scene completion and editing difficult. We present *Canvas360*, a two-stage framework that learns a geometry-aware panoramic prior with parallel RGB–depth pretraining and transfers it to a unified RGB-only generator for four in-context tasks, namely style transfer, inpainting, outpainting, and editing. The pretraining stage combines pseudo-depth supervision, modality-specific representations, and panorama-aware boundary handling to capture scene structure while respecting equirectangular adjacency. To support this two-stage design, we construct *Canvas360Dataset* with annotated RGB–depth panoramas and task-specific context–target pairs for prior learning and downstream adaptation. On text-to-panorama generation, *Canvas360* reduces panorama-specific distributional distance by 49.4% relative to the strongest baseline and outperforms the evaluated baselines in semantic alignment, perceptual quality, and boundary consistency. Completion and editing evaluations further show improvements in perceptual similarity, panoramic fidelity, and reconstruction accuracy. These results indicate that geometry-aware pretraining provides a transferable prior for panoramic generation.

1 INTRODUCTION

Panoramic image generation supports the creation and modification of immersive scenes. Text-to-panorama models synthesize environments from language descriptions (Ye et al., 2024; Zhang et al., 2024; Xie, 2025; Sun et al., 2025; Ni et al., 2025; Li & Bansal, 2023), while in-context generators transform a reference image according to a text prompt (Brooks et al., 2023; Black Forest Labs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

et al., 2025; Liu et al., 2025). These operations include style transfer, inpainting, outpainting, and object manipulation (Suvorov et al., 2022; Google, 2025; ByteDance Seed, 2026; Wu et al., 2025b; Deng et al., 2025). For panoramas, every change must remain compatible with a full spherical scene. An inserted object must follow the local projection, and a completed region must connect to visible content across the horizontal boundary. Local visual plausibility is therefore insufficient. The generator must preserve scene layout and continuity at the same time.

The equirectangular projection (ERP) makes this requirement explicit. It maps a sphere to a rectangular grid whose left and right edges are adjacent, with distortion increasing toward the poles. A plausible perspective-image edit can therefore introduce a seam or incompatible object geometry in ERP. Existing panoramic editors address the spherical domain through cube-map processing (Yang et al., 2025) or spherical positional embeddings (Zhong et al., 2025), together with task-specific image pairs. These representations encode spatial relationships, but the training signal must also teach the generator which surfaces and objects occupy the scene. Learning scene structure and respecting ERP adjacency are complementary requirements. This motivates learning an explicitly supervised panoramic scene prior that can support different transformations, while handling the horizontal boundary within the model.

Depth provides a suitable scene-structure signal and has supported geometric conditioning and supervision in image, video, and scene generation (Huang et al., 2025a; Bai et al., 2025b; Bhat et al., 2024; Zhang et al., 2023a; Yu et al., 2025b). Panoramic radial depth measures distance from the camera center along each viewing direction, providing a geometric target over the full spherical field of view. It supplies supervision for surfaces and object boundaries while preserving pixel correspondence. Joint RGB–depth learning must still distinguish modalities with different visual statistics, and the resulting prior must transfer to tasks that provide RGB context alone. These requirements call for shared scene learning with modality-specific targets, followed by adaptation that removes the need to generate depth.

We present *Canvas360*, a two-stage framework that learns a geometry-aware panoramic prior with parallel RGB–depth pretraining and transfers it to a unified RGB-only generator for four in-context tasks, namely style transfer, inpainting, outpainting, and editing. During pretraining, a shared transformer jointly generates RGB panoramas and depth maps under pseudo-depth supervision. Positional separation and similarity regularization distinguish the modalities, while velocity circular padding encodes horizontal adjacency in the predicted field and its supervision. Task adaptation then removes depth generation and uses a shared RGB context-and-prompt interface. All four tasks follow the same target flow-matching objective, allowing the learned panoramic prior to guide downstream generation without depth at inference.

This training strategy requires both geometric supervision and task pairs that preserve panoramic structure. We therefore construct *Canvas360Dataset* from 100K indoor and outdoor panoramas drawn from existing resources (Chang et al., 2017; Feng et al., 2025), web images, and generated scenes. Captions and pseudo-depth support prior learning, while 900K context–target pairs provide supervision for downstream adaptation. Completion pairs retain the original panorama as the target, reversed removal pairs preserve source geometry for object addition, and structural filtering rejects inconsistent style-transfer pairs. These construction choices maintain the connection between learning scene structure and preserving it during task adaptation.

Experiments evaluate the prior in text-to-panorama generation and its transfer to all four in-context tasks. *Canvas360* reduces panorama-specific FAED by 49.4% and horizontal boundary error by 35.1% relative to the next-best baseline on each metric. It also leads the evaluated methods in semantic alignment and perceptual quality, while the in-context results show gains in perceptual similarity, panoramic fidelity, and reconstruction accuracy. Component ablations clarify the roles of depth supervision, modality separation, and boundary treatment, and matched completion fine-tuning evaluates transfer of the full pretrained prior. Our contributions are as follows.

- We introduce *Canvas360*, which transfers a geometry-aware panoramic prior to four in-context tasks through one RGB-only model and a shared target flow-matching objective.
- We develop parallel RGB–depth pretraining with positional separation, velocity circular padding, and similarity regularization to supervise scene structure and horizontal continuity.
- We construct *Canvas360Dataset*, comprising 100K annotated RGB–depth panoramas and 900K paired examples for four in-context tasks.

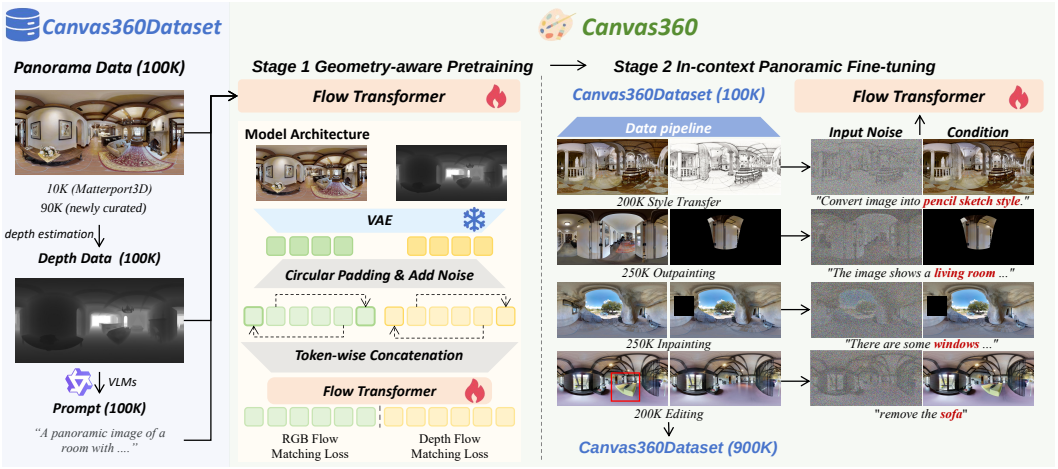

Figure 2: The two-stage design transfers a geometry-aware panoramic prior to four RGB-only tasks. Stage one learns the prior through joint RGB–depth generation on 100K annotated panoramas. Stage two adapts it to style transfer, inpainting, outpainting, and editing using 900K context–target pairs. Both stages concatenate image tokens within a shared transformer, while stage two uses only RGB inputs and outputs.

2 Related Work

Image Generation Priors. Diffusion models synthesize images by learning to reverse a noising process (Dhariwal & Nichol, 2021; Nichol et al., 2022; Saharia et al., 2022; Ramesh et al., 2022). Latent diffusion reduces the cost of high-resolution synthesis through a compressed image representation (Rombach et al., 2022; Podell et al., 2023), while transformer backbones model global interactions among image tokens with explicit positional information (Peebles & Xie, 2023; Vaswani et al., 2017; Black Forest Labs, 2024; Esser et al., 2024; Yu et al., 2025c; Ma et al., 2024). Flow matching learns a continuous-time velocity field connecting noise and data and underlies several recent image generators (Black Forest Labs, 2024; Esser et al., 2024; Lipman et al., 2022; Liu et al., 2022). We build on these priors and add supervision for panoramic scene geometry and horizontal continuity.

Panoramic and In-context Generation. In-context image generation conditions synthesis on a reference image and text. Existing methods address style transfer (Zhang et al., 2023c), inpainting (Black Forest Labs, 2026; Suvorov et al., 2022), outpainting (Cheng et al., 2022), editing (Brooks et al., 2023; Liu et al., 2025; ByteDance Seed, 2026), and object manipulation (Deng et al., 2025).

Panoramic generation must maintain consistency across a full spherical field of view. View-based scene synthesis (Fang et al., 2023; Höllein et al., 2023; Yu et al., 2023; Song et al., 2023; Shi et al., 2023; Tang et al., 2023; Lu et al., 2025) and coordinated diffusion over image regions (Bar-Tal et al., 2023; Lee et al., 2023; Zhang et al., 2023b) provide complementary mechanisms for spatial consistency. Cube-map methods coordinate generation across faces (Huang et al., 2025b; Kalischek et al., 2025), whereas spherical and correspondence-aware methods model panoramic geometry directly (Ye et al., 2024; Sun et al., 2025; Park et al., 2025; Zhang et al., 2024). Other work develops panoramic synthesis, completion, and scene construction over ERP images (Chen et al., 2022; Shum et al., 2023; Feng et al., 2023; Ai et al., 2024; Wang et al., 2024; Yang et al., 2024a; Ni et al., 2025; Wang et al., 2025; Xie, 2025; HunyuanWorld Team et al., 2025; Yang et al., 2024b). DiT360 combines hybrid training with cube-space supervision to improve panoramic detail (Feng et al., 2025), although conversion between cube-map and ERP representations requires resampling.

Recent in-context panoramic methods adapt image editing to the spherical domain with cube-map processing (Yang et al., 2025) or 3D spherical positional encodings (Zhong et al., 2025). They use task-specific pairs to supervise changes in scene content. *Canvas360* complements this paired supervision with RGB–depth pretraining that supplies an explicit scene-structure signal, then transfers the resulting prior to four in-context tasks through a shared RGB-only formulation.

3 METHOD

Canvas360 first learns a panoramic prior with auxiliary depth generation and boundary-aware velocity prediction. It then transfers this prior to four in-context tasks through a shared RGB-only representation (Fig. 2).

3.1 PRELIMINARIES

Flow Matching. We adopt the flow-matching (FM) formulation (Lipman et al., 2022; Liu et al., 2022) used by the FLUX backbone (Black Forest Labs, 2024). Let $x_0 \sim \pi_0$ denote a data latent and $x_1 \sim \pi_1$ a Gaussian noise sample. Following rectified flow (Liu et al., 2022), we use the linear interpolation

$$x_t = (1-t)x_0 + tx_1, \quad t \in [0,1]. \tag{1}$$

The derivative of Eq. (1) is $x_1 - x_0$. We train $v_\theta(x_t, t)$ to predict this velocity by minimizing

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{t,\, x_0\sim\pi_0,\, x_1\sim\pi_1}\left[\left\|v_\theta(x_t,t) - (x_1 - x_0)\right\|^2\right]. \tag{2}$$

Generation integrates the learned field from $t = 1$ to $t = 0$. Text conditioning is omitted from Eq. (2) for clarity.

Diffusion Transformer. Our backbone follows the diffusion transformer (DiT) architecture (Peebles & Xie, 2023; Feng et al., 2025). A variational autoencoder (VAE) (Kingma & Welling, 2013) encodes each image into latent features that form a sequence $\mathbf{x} \in \mathbb{R}^{N\times d}$ for transformer processing (Dosovitskiy et al., 2021). Here, N is the number of tokens and d is the feature dimension. Rotary positional embeddings (RoPE) encode token positions through coordinate-dependent rotations (Su et al., 2024). For multiple images, 3D RoPE uses coordinates (T, h, w), where T identifies the input and (h, w) specifies its spatial position (Black Forest Labs et al., 2025). These coordinates index tokens rather than points in 3D scene space.

3.2 GEOMETRY-AWARE TEXT-TO-PANORAMA PRETRAINING

Pretraining combines joint RGB–depth generation, boundary-aware supervision, and modality separation to model scene structure and horizontal periodicity.

Parallel Depth Generation. Depth offers a compact representation of scene structure for panoramic generation (Wu et al., 2024; Lin et al., 2025), and depth-based objectives can strengthen spatial perception (Tan et al., 2026). Following joint image and dense-prediction modeling (Qi et al., 2024; Zhang et al., 2023a), we generate RGB and depth within one transformer.

We estimate panoramic pseudo-depth with DAP (Lin et al., 2025) and encode the processed depth maps with the VAE. Let $\mathbf{x}_{\text{rgb}}, \mathbf{x}_{\text{depth}} \in \mathbb{R}^{N\times d}$ denote the RGB and depth token sequences. Following a multi-image token formulation (Black Forest Labs et al., 2025), we concatenate them along the token dimension as

$$\mathbf{x} = \left[\mathbf{x}_{\text{rgb}}\,;\,\mathbf{x}_{\text{depth}}\right]. \tag{3}$$

Each modality has a separate flow-matching target. Both modalities are generated from noise, so clean depth is never supplied as a condition. The concatenation in Eq. (3) leaves the backbone channel dimension unchanged.

We distinguish RGB and depth tokens by offsetting the first coordinate of their 3D RoPE indices. For a spatial position (h, w), we define

$$\mathbf{u}_{\text{rgb}} = (0, h, w), \qquad \mathbf{u}_{\text{depth}} = (T_d, h, w), \qquad T_d \in \mathbb{N}_{>0}. \tag{4}$$

The positive offset $T_d = 50$ in Eq. (4) separates the modalities while preserving their spatial coordinates. Attention lets corresponding RGB and depth tokens exchange information, while the positional indices keep each modality identifiable.

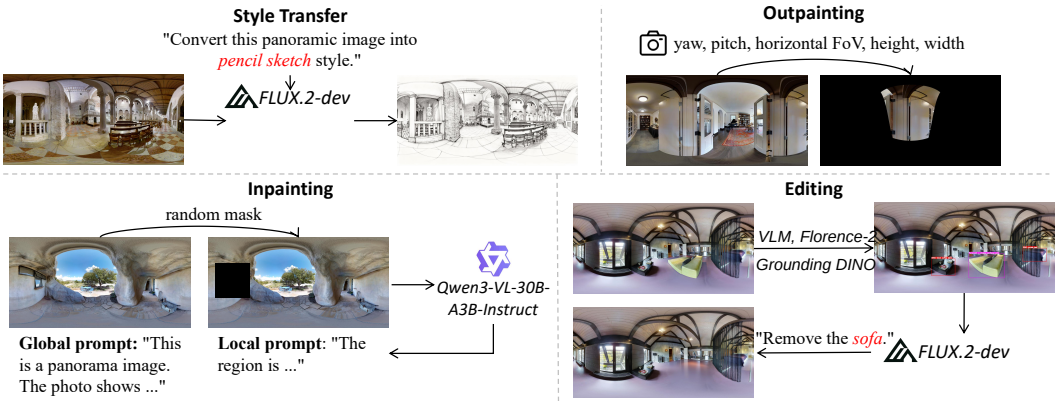

Figure 3: Task-specific procedures create the 900K in-context training pairs while retaining panoramic structure. (a) Style transfer. FLUX.2-dev (Black Forest Labs, 2025) generates stylized panoramas from 12 style prompts. (b) Outpainting. We sample camera parameters and use inverse projection to obtain visible-region masks in ERP. (c) Inpainting. We mask rectangular regions and provide a prompt describing either the scene or the missing content. (d) Editing. Vision-language and grounding models localize objects for removal with FLUX.2-dev, followed by NanoBanana refinement for difficult cases. Reversing the pairs provides object-addition examples.

Velocity Circular Padding. Prior work applies circular padding to panorama latents to promote boundary continuity (Feng et al., 2025; Zhong et al., 2025). We extend this idea by combining circular padding with consecutive longitude indices and direct supervision of the padded velocity field. This construction represents adjacency across the ERP boundary explicitly. For each modality, after reshaping the interpolated latent $x_t \in \mathbb{R}^{N\times d}$ onto an $H \times W$ token grid with $N = HW$, we index the original longitude columns as $1, \ldots, W$ and append two ghost columns with longitude indices 0 and $W{+}1$. Before every transformer block, we softly synchronize the ghost-column hidden states with their circular counterparts. For a padded hidden-state sequence h, the update is

$$\tilde{h}^j = \begin{cases} \alpha_{\text{blend}} h^0 + (1-\alpha_{\text{blend}}) h^W, & j = 0, \\ h^j, & 1 \le j \le W, \\ \alpha_{\text{blend}} h^{W+1} + (1-\alpha_{\text{blend}}) h^1, & j = W+1, \end{cases} \qquad \alpha_{\text{blend}} = 0.3. \tag{5}$$

Here, $\alpha_{\text{blend}} = 0$ gives a hard copy and $\alpha_{\text{blend}} = 1$ leaves the ghost columns unchanged. We apply hard circular padding to the target velocity $v = x_1 - x_0$,

$$\tilde{v}^0 = v^W, \qquad \tilde{v}^j = v^j,\ j = 1, \ldots, W, \qquad \tilde{v}^{W+1} = v^1. \tag{6}$$

The padded sequence in Eq. (5) links hidden states across the $0°/360°$ boundary while retaining consecutive longitude indices $0, 1, \ldots, W, W{+}1$. The model observes local transitions $(0, 1)$ and $(W, W{+}1)$. Feature synchronization enforces the wrap-around correspondences $0 \equiv W$ and $W{+}1 \equiv 1$. The flow-matching loss supervises all columns of the padded target in Eq. (6). The soft synchronization is applied independently to RGB and depth sections before every transformer block.

Similarity Regularization. Because the two modalities share one transformer, their velocity predictions can collapse toward each other despite having distinct targets. We therefore regularize their normalized alignment.

Let $\mathbf{v}_{\text{rgb}}$ and $\mathbf{v}_{\text{depth}}$ denote the predicted RGB and depth velocity fields. We flatten each field and standardize it per sample as $\hat{\mathbf{v}}_m = (\mathbf{v}_m - \mu(\mathbf{v}_m))/(\sigma(\mathbf{v}_m) + \varepsilon)$ for $m \in \{\text{rgb}, \text{depth}\}$ and $\varepsilon = 10^{-6}$, then penalize the squared magnitude of their normalized correlation,

$$\mathcal{L}_{\text{sim}} = \mathbb{E}\left[\left(\frac{1}{D}\sum_{i=1}^{D} \hat{v}_{\text{rgb},i}\hat{v}_{\text{depth},i}\right)^2\right], \tag{7}$$

where μ and σ are the per-sample mean and standard deviation over the flattened fields, and D is their number of elements. This penalty acts on the predictions, while each modality's flow-matching

target preserves its correspondence to the paired scene. The pretraining objective is

$$\mathcal{L} = \mathcal{L}_{\text{FM}} + \lambda \mathcal{L}_{\text{sim}}. \tag{8}$$

Here, $\mathcal{L}_{\text{FM}} = \mathcal{L}_{\text{rgb}} + \alpha_d \mathcal{L}_{\text{depth}}$ with $\alpha_d = 0.5$, and $\lambda = 1.0$ weights the regularizer. Thus, the objective in Eq. (8) matches the reported stage-one configuration.

3.3 Unified In-context Panoramic Fine-tuning

We omit the depth tokens and fine-tune the pretrained model on style transfer, inpainting, outpainting, and editing. All four tasks share an interface of RGB context and text instructions, represented as tuples of the form

$$(\mathbf{x}_{\text{cxt}}, c, \mathbf{x}_{\text{tgt}}), \tag{9}$$

where $\mathbf{x}_{\text{cxt}}$ and $\mathbf{x}_{\text{tgt}}$ are the VAE-encoded context and target panoramas and c is the text prompt. In Eq. (9), the context is a full panorama for style transfer and editing or a masked panorama for inpainting and outpainting.

We retain the token-concatenation design from pretraining. At time t, the input combines the noisy target $\mathbf{x}_{\text{tgt},t}$ with the clean context as

$$\mathbf{x} = [\mathbf{x}_{\text{tgt},t}\,;\,\mathbf{x}_{\text{cxt}}]\,. \tag{10}$$

Their positional coordinates are

$$\mathbf{u}_{\text{tgt}} = (0, h, w), \qquad \mathbf{u}_{\text{cxt}} = (T_c, h, w), \qquad T_c \in \mathbb{N}_{>0}. \tag{11}$$

The offset $T_c = 1$ in Eq. (11) separates target and context tokens while preserving their spatial correspondence. Flow matching is applied only to target tokens. The context remains clean and fixed throughout denoising, and depth generation with $\mathcal{L}_{\text{sim}}$ is omitted in this stage (Eq. (10)).

Inpainting and outpainting supply visible regions that constrain missing content. Style transfer conditions on the complete panorama and changes its appearance. Editing uses the source panorama to preserve content outside the requested modification. All four tasks use the same context-and-prompt interface.

3.4 Data Pipeline

Canvas360Dataset provides 100K pretraining panoramas and 900K downstream pairs derived from this source pool. The pretraining set combines Matterport3D images, web panoramas, and generated panoramas. Each panorama is annotated with a caption and pseudo-depth. The pair-construction procedures in Fig. 3 provide task-specific supervision. Quality control and pseudo-depth processing are described in Sec. B.

Style Transfer Data. We apply FLUX.2-dev (Black Forest Labs, 2025) with 12 style prompts to transform panorama appearance. We compare line drawings from source and stylized images to reject large structural changes, retaining 200K style-transfer pairs.

Outpainting Data. We construct outpainting examples by sampling perspective views and deriving their visible-region masks in ERP. For each sample, we draw the yaw, pitch, field of view, and output dimensions from predefined distributions. After applying the sampled yaw shift, we extract a centered perspective view and project it back to ERP to obtain the visible-region mask. Each sample contains the yaw-shifted panorama, its visibility mask, and a scene caption. The masked panorama provides the context, and the complete panorama provides the target. We construct 250K outpainting pairs.

Inpainting Data. Following prior work (Rombach et al., 2022; Suvorov et al., 2022), we construct inpainting pairs with global or local prompts. Global prompts describe the complete panorama, whereas local prompts describe the masked content. For both settings, we sample rectangular masks with random area ratios, aspect ratios, and locations. Global-prompt examples use larger masks and require more extensive scene completion. We retain scene captions for global prompts and use Qwen3-VL-30B-A3B-Instruct (Bai et al., 2025a) to caption masked regions for local prompts. This procedure produces 250K inpainting pairs.



Figure 4: *Canvas360* preserves local detail and panoramic structure across the displayed scenes. Red boxes highlight artifacts and geometric inconsistencies in the generated scenes. Full comparisons appear in Sec. D.

Table 1: Text-to-panorama generation results. CS denotes CLIP score and QA denotes Q-Align. FID_{pole} and FID_{equ} evaluate polar and equatorial regions. Red shading with bold text marks the best value, and orange shading with underlined text marks the second-best value in each column; ties receive the same marking. Arrows indicate the preferred metric direction.

Methods	FID↓	FID $_{\text{pole}}$↓	FID $_{\text{equ}}$↓	FAED↓	IS↑	CS↑	QA $_{\text{quality}}$↑	QA $_{\text{aesthetic}}$↑	BRISQUE↓	NIQE↓
PanFusion	124.87	182.09	108.12	11.06	1.30	28.35	3.83	3.56	27.38	4.31
SMGD	46.72	65.69	34.84	3.29	1.40	31.14	4.05	3.77	30.35	4.75
PAR	47.72	76.93	27.39	2.97	1.34	33.85	3.91	3.54	32.26	4.38
WorldGen	67.11	79.32	33.45	3.29	1.40	34.61	4.30	3.59	32.31	4.82
LayerPano3D	62.82	80.37	38.67	2.98	1.50	34.40	4.73	3.93	33.91	3.79
HunyuanWorld	76.75	106.58	41.75	2.91	1.53	34.73	4.67	3.85	39.12	5.18
DiT360	**42.88**	**50.88**	**24.77**	2.91	1.60	34.68	4.69	4.19	**10.25**	3.72
Ours	47.22	54.64	29.52	**1.47**	**2.44**	**35.03**	**4.75**	**4.22**	15.04	**3.65**

Editing Data. Following SE360 (Zhong et al., 2025), we combine vision-language captioning with Florence-2 (Xiao et al., 2024) and Grounding DINO (Liu et al., 2024) to obtain object bounding boxes and segmentation masks. These annotations guide object removal with FLUX.2-dev (Black Forest Labs, 2025). For small objects, fine structures, or complex backgrounds, we use NanoBanana (Google, 2025) for refinement. Reversing each original–edited pair provides an object-addition example whose target retains the original panoramic geometry. This process produces 200K object-removal and object-addition pairs.

4 Experiments

We evaluate the learned panoramic prior through text-to-panorama generation and its transfer to four RGB-only in-context tasks, then use targeted ablations to examine panoramic fidelity, boundary continuity, and task-level reconstruction quality.

4.1 Setup

We build *Canvas360* on FLUX.1-dev (Black Forest Labs, 2024) and optimize LoRA adapters (Hu et al., 2021). We fine-tune on the task pairs in *Canvas360Dataset*. For text-to-panorama evaluation, we follow the Matterport3D training and validation setting of DiT360 (Feng et al., 2025; Chang et al., 2017). The main experiments use 1024×2048 panoramas, and ablations use 512×1024 panoramas. We assess distributional fidelity, scene diversity, text alignment, perceptual quality, and horizontal boundary continuity. Implementation details and metric descriptions are provided in Sec. C.

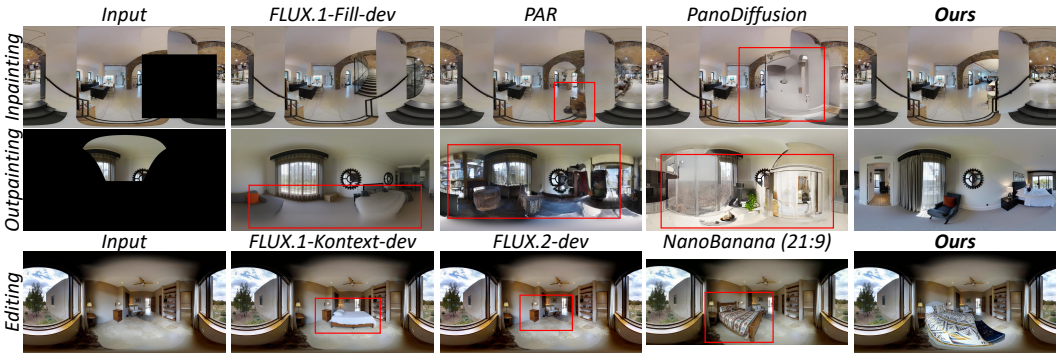

Figure 5: *Canvas360* preserves surrounding panoramic structure in the displayed inpainting, outpainting, and editing examples. Red boxes mark representative artifacts. Additional comparisons appear in Sec. F.

4.2 Main Results

Text-to-Panorama Generation. *Canvas360* combines detailed appearance with coherent panoramic structure (Fig. 4 and Tab. 1). In the visual comparisons, SMGD (Sun et al., 2025) and PAR (Wang et al., 2025) lose fine detail, while WorldGen (Xie, 2025), HunyuanWorld (HunyuanWorld Team et al., 2025), and DiT360 (Feng et al., 2025) exhibit local artifacts or geometric inconsistencies in the marked regions. *Canvas360* produces sharper details and object shapes that conform more closely to ERP. Its FAED of 1.47 is 49.4% below the next-best value of 2.91, indicating a substantial gain in panorama-specific distributional fidelity.

The gains extend beyond panorama-specific fidelity. *Canvas360* achieves the highest CLIP score of 35.03 and scene-classification IS of 2.44, supporting text alignment and confident, diverse scene predictions. It also leads in Q-Align quality and aesthetics, with scores of 4.75 and 4.22, and attains the lowest NIQE of 3.65. The agreement across these measures connects the gain in panoramic fidelity with improved appearance and semantics. DiT360 retains lower FID and BRISQUE, so the advantage is strongest in panorama-specific fidelity and the complementary quality measures rather than uniform across every metric.

Horizontal Boundary Continuity. Global distributional metrics do not directly measure the ERP seam. With RGB left-to-right consistency error (LRCE-RGB), *Canvas360* attains the lowest error of 0.0061, compared with 0.0094 for HunyuanWorld and 0.0101 for DiT360 (Tab. 6), a 35.1% reduction relative to the next-best method. This complements the FAED result with evidence of local boundary continuity.

In-context Panoramic Generation. One RGB-only model handles all four downstream tasks, with 500 generated results per task. Against PAR (Wang et al., 2025), PanoDiffusion (Wu et al., 2024), and FLUX.1-Fill-dev (Black Forest Labs, 2026), *Canvas360* produces sharper completions that preserve surrounding structure (Fig. 5). It reduces LPIPS by 34.7% for inpainting and 18.3% for outpainting relative to the next-best results, while lower FAED and higher PSNR support gains in perceptual similarity and panoramic reconstruction.

Editing and style transfer test controlled changes to a complete panorama. Against FLUX.1-Kontext-dev (Black Forest Labs et al., 2025), FLUX.2-dev (Black Forest Labs, 2025), NanoBanana (Google, 2025), SE360 (Zhong et al., 2025), and Omni2 (Yang et al., 2025), *Canvas360* preserves local ERP geometry and surrounding layout, achieving LPIPS 0.084, FAED 0.358, and PSNR 26.40. The visual comparisons show why this matters for editing. Objects must fit the local projection while content outside the requested change remains consistent. Style-transfer examples retain scene structure while changing appearance (Sec. F), showing that one RGB-only formulation accommodates both localized edits and full-panorama changes.

User Study. Human preferences provide a complementary assessment of panoramic coherence. The study involved 71 participants and 10 test cases, comparing *Canvas360* with Matrix-3D, Hun-

Table 2: Text-to-panorama user-study preferences.

Methods	TA ↑	BC ↑	PA ↑	OQ ↑
Matrix-3D	23.0%	20.0%	16.8%	15.3%
HunyuanWorld	26.9%	20.9%	20.5%	18.1%
DiT360	23.9%	28.0%	28.6%	30.7%
Ours	26.2%	31.1%	34.1%	35.9%

Table 3: Cumulative RGB–depth ablations (PO: positional offset). Red shading and bold text mark the strongest value in each column.

Methods	FID ↓	FAED ↓	BRISQUE ↓	NIQE ↓
Baseline	**51.16**	5.37	**14.62**	3.86
+ Depth	51.57	**4.74**	17.84	3.85
+ PO	57.04	4.81	23.64	4.12
+ $\mathcal{L}_{\text{sim}}$	51.48	4.93	16.23	**3.84**

yuanWorld, and DiT360 (Lu et al., 2025; HunyuanWorld Team et al., 2025; Feng et al., 2025). Participants selected among four anonymized outputs for text alignment (TA), boundary continuity (BC), panorama awareness (PA), and overall quality (OQ). *Canvas360* received the largest preference shares for BC, PA, and OQ (Tab. 2). Its PA and OQ shares exceeded DiT360's by 5.5 and 5.2 percentage points, supporting the visual improvements in spherical structure and overall quality observed in these cases. The protocol appears in Sec. H.

4.3 Ablations

The ablations examine auxiliary depth, horizontal boundary treatment, and transfer to RGB-only completion. Full results appear in Sec. E.

Parallel Depth Generation. The cumulative ablation in Tab. 3 shows how the RGB–depth design balances panoramic fidelity with appearance. Adding depth lowers FAED from 5.37 to 4.74, supporting geometric supervision as a useful training signal, but also increases BRISQUE and produces dark artifacts (Fig. 11). Positional separation raises QA aesthetics from 3.41 to 4.13 while increasing FID and BRISQUE. Similarity regularization subsequently lowers FID from 57.04 to 51.48 and BRISQUE from 23.64 to 16.23. The complete variant attains the highest QA quality and aesthetic scores, 4.75 and 4.22, while retaining lower FAED than the RGB-only baseline (Tab. 7). Depth alone gives the lowest FAED, whereas the combined design improves the perceptual quality of joint generation. This balance motivates combining geometric supervision with modality separation and regularization.

Velocity Circular Padding. VCP targets the periodic constraint at the ERP boundary. Moving the seam to the center with a 180° yaw rotation makes this effect visible. Compared with standard circular padding, VCP improves edge alignment and reduces seam discontinuities (Fig. 10). In the separate baseline ablation, adding VCP lowers BRISQUE from 14.62 to 13.85 and NIQE from 3.86 to 3.79 (Tab. 7). FID and FAED do not improve. The visual comparison thus supports better boundary continuity, complemented by gains in no-reference image quality.

Transfer of the Pretrained Prior. Under identical completion fine-tuning, the pretrained *Canvas360* model produces clearer inpainting and outpainting results than FLUX.1-dev (Fig. 9). Its completions retain more coherent panoramic structure with fewer perspective-like patterns. This comparison shows that the full pretrained prior transfers to RGB-only completion. A separate qualitative comparison of generated RGB panoramas with and without depth supervision finds more coherent structures in the depth-supervised examples (Sec. A).

5 Conclusion

We presented *Canvas360*, which learns a geometry-aware panoramic prior before adapting it to four in-context tasks. Parallel RGB–depth pretraining supplies scene-structure supervision, while modality separation and velocity circular padding address cross-modal learning and horizontal continuity. Supported by *Canvas360Dataset*, the pretrained model transfers to style transfer, inpainting, outpainting, and editing through one RGB-only formulation. Experiments show improved panorama-specific fidelity and boundary consistency, alongside gains in completion and editing quality. The framework demonstrates that depth can guide panoramic prior learning while downstream generation uses only RGB context and text instructions, connecting geometric supervision with flexible panoramic image manipulation.

AI USE STATEMENT

Generative AI models are used in the research pipeline, and language models assisted with manuscript preparation.

Use in the research pipeline. FLUX.1-dev is the generative backbone of *Canvas360* (Black Forest Labs, 2024). FLUX.2-dev (Black Forest Labs, 2025) and NanoBanana (Google, 2025) support synthesis of the style-transfer and editing splits of *Canvas360Dataset*. Qwen3-VL provides panorama and masked-region captions (Bai et al., 2025a). DAP estimates pseudo-depth (Lin et al., 2025), while Florence-2 (Xiao et al., 2024) and Grounding DINO (Liu et al., 2024) provide object grounding. Their roles are described in Sec. 3. Model-generated training data underwent the task-specific quality-control procedures in Sec. B. We also manually inspected a 50K subset to assess pair quality.

Use in manuscript preparation. We used large language models to edit author-written text for grammar, wording, clarity, and consistency and to organize and check interpretations of author-provided experimental results. AI assistance also supported citation and submission-format checks. We did not use generative AI tools to generate research ideas, design the method, derive technical claims, or conduct the experiments reported in this paper. Experimental measurements, tables, and figures were produced by author-written code and author-run experiments. The authors reviewed all AI-assisted text and verified that it reflects the underlying work. We take full responsibility for the final content of this paper, including all text, claims, and artifacts.

ETHICS STATEMENT

Potential misuse and broader impact. Improved geometric consistency and visual fidelity in panoramic generation support VR/AR authoring, simulation, digital-twin prototyping, immersive scene creation, and 360-degree content design. The system may also be misused for deceptive scene manipulation, synthetic visual misinformation, or unauthorized content creation. Our method focuses on scene-level panoramic content, which limits exposure to identity-related imagery but does not eliminate these risks. We discuss safeguards for responsible release in Sec. I.

Data provenance and licensing. The dataset combines existing panoramas (Chang et al., 2017; Feng et al., 2025), web images, and model-synthesized samples. Large-scale curation can raise licensing and attribution concerns. We filter unsafe, sensitive, and low-quality samples during construction. Dataset releases will document sources, preparation procedures, filtering, and intended-use restrictions (Secs. B and I).

Human subjects. The human preference study reported in Tab. 2 involved volunteer participants who compared anonymized generated panoramic images. Participation was voluntary and did not affect employment, compensation, or performance evaluation. No personally identifiable information, sensitive attributes, or free-form personal responses were collected, and all responses were aggregated before analysis. The study was reviewed through the authors' organizational review process. Full details, including the evaluation criteria and interface, are provided in Sec. H. Institution-identifying details are omitted here to preserve anonymity during review.

REPRODUCIBILITY STATEMENT

The two-stage training formulation, including parallel RGB–depth generation, velocity circular padding, and similarity regularization, is specified in Sec. 3. Sec. C reports the backbone, LoRA configuration, optimizer settings, per-stage batch sizes, gradient accumulation, learning rates, training schedules, resolutions, and inference parameters. Metric descriptions and sources appear there and in the corresponding evaluation sections of the appendix. The construction of *Canvas360Dataset* is described in Sec. 3.4 and Sec. B, including the per-task synthesis procedures, quality-control checks, the manual-inspection sample size, and pseudo-depth processing. The full ablation results appear in Sec. E. We will release the code, the trained model weights, and the dataset together with documentation of data sources and intended-use restrictions. Release links are omitted to preserve anonymity during double-blind review.

REFERENCES

Hao Ai, Zidong Cao, Haonan Lu, Chen Chen, Jian Ma, Pengyuan Zhou, Tae-Kyun Kim, Pan Hui, and Lin Wang. Dream360: Diverse and immersive outdoor virtual scene creation via transformer-based 360 image outpainting. *IEEE Transactions on Visualization and Computer Graphics*, 2024. doi: 10.1109/TVCG.2024.3372085.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Yunpeng Bai, Shaoheng Fang, Chaohui Yu, Fan Wang, and Qixing Huang. Geovideo: Introducing geometric regularization into video generation model. *arXiv preprint arXiv:2512.03453*, 2025b.

Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. MultiDiffusion: Fusing diffusion paths for controlled image generation. In *arXiv*, 2023.

Shariq Farooq Bhat, Niloy Mitra, and Peter Wonka. Loosecontrol: Lifting controlnet for generalized depth conditioning. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–11, 2024.

Black Forest Labs. Flux. `https://github.com/black-forest-labs/flux`, 2024. Accessed: 2024-09-23.

Black Forest Labs. FLUX.2. `https://github.com/black-forest-labs/flux2`, 2025. Accessed: 2026-01-21.

Black Forest Labs. Flux.1-fill. `https://huggingface.co/black-forest-labs/FLUX.1-Fill-dev`, 2026. Accessed: 2026-01-21.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. In *arXiv*, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18392–18402, 2023.

ByteDance Seed. Seededit. `https://seed.bytedance.com/en/tech/seededit`, 2026. Accessed: 2026-01-21.

Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *3DV*, 2017.

Hangwei Chen, Feng Shao, Xiongli Chai, Yuese Gu, Qiuping Jiang, Xiangchao Meng, and Yo-Sung Ho. Quality evaluation of arbitrary style transfer: Subjective study and objective metric. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3055–3070, 2023. doi: 10.1109/TCSVT.2022.3231041.

Hao Chen, Yuqi Hou, Chenyuan Qu, Irene Testini, Xiaohan Hong, and Jianbo Jiao. 360+ x: A panoptic multi-modal scene understanding dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19373–19382, 2024.

Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2light: Zero-shot text-driven hdr panorama generation. *ACM Transactions on Graphics*, 2022. doi: 10.1145/3550454.3555447.

Yen-Chi Cheng, Chieh Hubert Lin, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, and Ming-Hsuan Yang. Inout: Diverse image outpainting via gan inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11431–11440, 2022.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *arXiv*, 2024.

Chuan Fang, Yuan Dong, Kunming Luo, Xiaotao Hu, Rakesh Shrestha, and Ping Tan. Ctrl-room: Controllable text-to-3d room meshes generation with layout constraints. In *arXiv*, 2023.

Haoran Feng, Dizhe Zhang, Xiangtai Li, Bo Du, and Lu Qi. Dit360: High-fidelity panoramic image generation via hybrid training. *arXiv preprint arXiv:2510.11712*, 2025.

Mengyang Feng, Jinlin Liu, Miaomiao Cui, and Xuansong Xie. Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models. In *arXiv*, 2023.

Google. Nano Banana. `https://blog.google/products/gemini/updated-image-editing-model/`, 2025. Accessed: 2026-01-21.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *arXiv*, 2015.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NeurIPS*, 2017.

Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In *ICCV*, 2023.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *arXiv*, 2021.

Jiehui Huang, Yuechen Zhang, Xu He, Yuan Gao, Zhi Cen, Bin Xia, Yan Zhou, Xin Tao, Pengfei Wan, and Jiaya Jia. Unityvideo: Unified multi-modal multi-task learning for enhancing world-aware video generation. *arXiv preprint arXiv:2512.07831*, 2025a.

Yukun Huang, Yanning Zhou, Jianan Wang, Kaiyi Huang, and Xihui Liu. DreamCube: RGB-D panorama generation via multi-plane synchronization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b. doi: 10.1109/ICCV51701.2025.02311.

HunyuanWorld Team, Zhenwei Wang, Yuhao Liu, Junta Wu, Zixiao Gu, Haoyuan Wang, Xuhui Zuo, Tianyu Huang, Wenhuan Li, Sheng Zhang, Yihang Lian, Yulin Tsai, Lifu Wang, Sicong Liu, Puhua Jiang, Xianghui Yang, Dongyuan Guo, Yixuan Tang, Xinyue Mao, Jiaao Yu, Junlin Yu, Jihong Zhang, Meng Chen, Liang Dong, Yiwen Jia, Chao Zhang, Yonghao Tan, Hao Zhang, Zheng Ye, Peng He, Runzhou Wu, Minghui Chen, Zhan Li, Wangchen Qin, Lei Wang, Yifu Sun, Lin Niu, Xiang Yuan, Xiaofeng Yang, Yingping He, Jie Xiao, Yangyu Tao, Jianchen Zhu, Jinbao Xue, Kai Liu, Chongqing Zhao, Xinming Wu, Tian Liu, Peng Chen, Di Wang, Yuhong Liu, Linus, Jie Jiang, Tengfei Wang, and Chunchao Guo. Hunyuanworld 1.0: Generating immersive, explorable, and interactive 3d worlds from words or pixels. In *arXiv*, 2025.

Nikolai Kalischek, Michael Oechsle, Fabian Manhardt, Philipp Henzler, Konrad Schindler, and Federico Tombari. CubeDiff: Repurposing diffusion-based image models for panorama generation. In *The Thirteenth ICLR*, 2025. URL https://arxiv.org/abs/2501.17162.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *arXiv*, 2013.

Yuseung Lee, Kunho Kim, Hyunjin Kim, and Minhyuk Sung. Syncdiffusion: Coherent montage via synchronized joint diffusions. In *NeurIPS*, 2023.

Jialu Li and Mohit Bansal. Panogen: Text-conditioned panoramic environment generation for vision-and-language navigation. In *NeurIPS*, 2023.

Kang Liao, Size Wu, Zhonghua Wu, Linyi Jin, Chao Wang, Yikai Wang, Fei Wang, Wei Li, and Chen Change Loy. Thinking with camera: A unified multimodal model for camera-centric understanding and generation. *arXiv preprint arXiv:2510.08673*, 2025.

Xin Lin, Meixi Song, Dizhe Zhang, Wenxuan Lu, Haodong Li, Bo Du, Ming-Hsuan Yang, Truong Nguyen, and Lu Qi. Depth any panoramas: A foundation model for panoramic depth estimation. *arXiv preprint arXiv:2512.16913*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pp. 38–55. Springer, 2024.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, et al. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *arXiv*, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Yuanxun Lu, Jingyang Zhang, Tian Fang, Jean-Daniel Nahmias, Yanghai Tsin, Long Quan, Xun Cao, Yao Yao, and Shiwei Li. Matrix3d: Large photogrammetry model all-in-one. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025. doi: 10.1109/CVPR52734.2025.01051.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *arXiv*, 2024.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 2012. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 2013. doi: 10.1109/LSP.2012.2227726.

Jinhong Ni, Chang-Bin Zhang, Qiang Zhang, and Jing Zhang. What makes for text to 360-degree panorama generation with stable diffusion? In *arXiv*, 2025.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *arXiv*, 2022.

Changgyoon Oh, Wonjune Cho, Daehee Park, Yujeong Chae, Lin Wang, and Kuk-Jin Yoon. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *European Conference on Computer Vision*, 2022. doi: 10.1007/978-3-031-19787-1_20.

Minho Park, Taewoong Kang, Jooyeol Yun, Sungwon Hwang, and Jaegul Choo. Spherediff: Tuning-free omnidirectional panoramic image and video generation via spherical latent representation. In *arXiv*, 2025.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *arXiv*, 2023.

Lu Qi, Lehan Yang, Weidong Guo, Yu Xu, Bo Du, Varun Jampani, and Ming-Hsuan Yang. Unigs: Unified representation for image generation and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *arXiv*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. In *arXiv*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.01042.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *arXiv*, 2016.

Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: panorama transformer for indoor 360 depth estimation. In *European Conference on Computer Vision*, pp. 195–211. Springer, 2022.

Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. In *arXiv*, 2023.

Ka Chun Shum, Hong-Wing Pang, Binh-Son Hua, Duc Thanh Nguyen, and Sai-Kit Yeung. Conditional 360-degree image synthesis for immersive indoor scene decoration. In *ICCV*, 2023.

Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Feng Tang, Junsong Yuan, and Yang Zhao. Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture. In *arXiv*, 2023.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. doi: 10.1016/j.neucom.2023.127063.

Xiancheng Sun, Mai Xu, Shengxi Li, Senmao Ma, Xin Deng, Lai Jiang, and Gang Shen. Spherical manifold guided diffusion model for panoramic image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025.

Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2149–2159, 2022.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *arXiv*, 2015.

Bin Tan, Changjiang Sun, Xiage Qin, Hanat Adai, Zelin Fu, Tianxiang Zhou, Han Zhang, Yinghao Xu, Xing Zhu, Yujun Shen, and Nan Xue. Masked depth modeling for spatial perception. *arXiv preprint arXiv:2601.17895*, 2026.

Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. In *arXiv*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

Chaoyang Wang, Xiangtai Li, Lu Qi, Xiaofan Lin, Jinbin Bai, Qianyu Zhou, and Yunhai Tong. Conditional panoramic image generation via masked autoregressive modeling. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-0932.

Hai Wang, Xiaoyu Xiang, Yuchen Fan, and Jing-Hao Xue. Customizing 360-degree panoramas through text-to-image diffusion models. In *WACV*, 2024.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025a.

Chenyuan Wu, Pengfei Zheng, Ruiran Yan, Shitao Xiao, Xin Luo, Yueze Wang, Wanli Li, Xiyan Jiang, Yexin Liu, Junjie Zhou, et al. OmniGen2: Towards instruction-aligned multimodal generation. *arXiv preprint arXiv:2506.18871*, 2025b.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *arXiv*, 2023.

Tianhao Wu, Chuanxia Zheng, and Tat-Jen Cham. PanoDiffusion: 360-degree panorama outpainting via diffusion. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.03177.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4818–4829, 2024.

Ziyang Xie. Worldgen: Generate any 3d scene in seconds. https://github.com/ZiYang-xie/WorldGen, 2025.

Bangbang Yang, Wenqi Dong, Lin Ma, Wenbo Hu, Xiao Liu, Zhaopeng Cui, and Yuewen Ma. Dreamspace: Dreaming your room space with text-driven panoramic texture propagation. In *IEEE VR*, 2024a.

Liu Yang, Huiyu Duan, Yucheng Zhu, Xiaohong Liu, Lu Liu, Zitong Xu, Guangji Ma, Xiongkuo Min, Guangtao Zhai, and Patrick Le Callet. Omni2: Unifying omnidirectional image generation and editing in an omni model. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10103–10112, 2025.

Shuai Yang, Jing Tan, Mengchen Zhang, Tong Wu, Yixuan Li, Gordon Wetzstein, Ziwei Liu, and Dahua Lin. Layerpano3d: Layered 3d panorama for hyper-immersive scene generation. In *arXiv*, 2024b. URL https://arxiv.org/abs/2408.13252.

Weicai Ye, Chenhao Ji, Zheng Chen, Junyao Gao, Xiaoshui Huang, Song-Hai Zhang, Wanli Ouyang, Tong He, Cairong Zhao, and Guofeng Zhang. Diffpano: Scalable and consistent text to panorama generation with spherical epipolar-aware diffusion. In *NeurIPS*, 2024.

Heyang Yu, Yinan Han, Xiangyu Zhang, Baiqiao Yin, Bowen Chang, Xiangyu Han, Xinhao Liu, Jing Zhang, Marco Pavone, Chen Feng, et al. Thinking in 360°: Humanoid visual search in the wild. *arXiv preprint arXiv:2511.20351*, 2025a. URL https://arxiv.org/abs/2511.20351.

Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 5916–5926, 2025b.

Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *ICCV*, 2023.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*, 2025c.

Cheng Zhang, Qianyi Wu, Camilo Cruz Gambardella, Xiaoshui Huang, Dinh Phung, Wanli Ouyang, and Jianfei Cai. Taming stable diffusion for text to 360 panorama image generation. In *CVPR*, 2024.

Jingyang Zhang, Shiwei Li, Yuanxun Lu, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, and Yao Yao. Jointnet: Extending text-to-image diffusion for dense distribution modeling. *arXiv preprint arXiv:2310.06347*, 2023a.

Qinsheng Zhang, Jiaming Song, Xun Huang, Yongxin Chen, and Ming-Yu Liu. Diffcollage: Parallel generation of large content with diffusion models. In *CVPR*, 2023b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10146–10156, 2023c.

Haoyi Zhong, Fang-Lue Zhang, Andrew Chalmers, and Taehyun Rhee. Se360: Semantic edit in 360 panoramas via hierarchical data construction. *arXiv preprint arXiv:2512.19943*, 2025.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. doi: 10.1109/TPAMI.2017.2723009.

A Analysis of Geometry-aware Pretraining

We examine scene structure in two complementary ways. Jointly generated RGB–depth pairs illustrate cross-modal correspondence. Depth estimates derived from generated RGB panoramas provide a separate qualitative check of visible scene structure.

Visualization of Predicted Depth. In Fig. 6, generated depth maps follow the main scene layout and object boundaries of the corresponding RGB panoramas. The alignment is qualitative evidence that the shared model captures correspondence between appearance and depth.

Geometry in Generated RGB Panoramas. We compare models trained with and without auxiliary depth supervision to examine its effect on RGB geometry. For each generated panorama, we retain the four side faces of its cube map, estimate their depth, and stitch the estimates back into ERP. The resulting projections appear in Fig. 7. The depth-supervised model produces more coherent structures near stitching boundaries in the displayed views. These estimates are separate from the jointly generated depth maps and remain qualitative because they are affected by depth-estimation and reprojection errors.

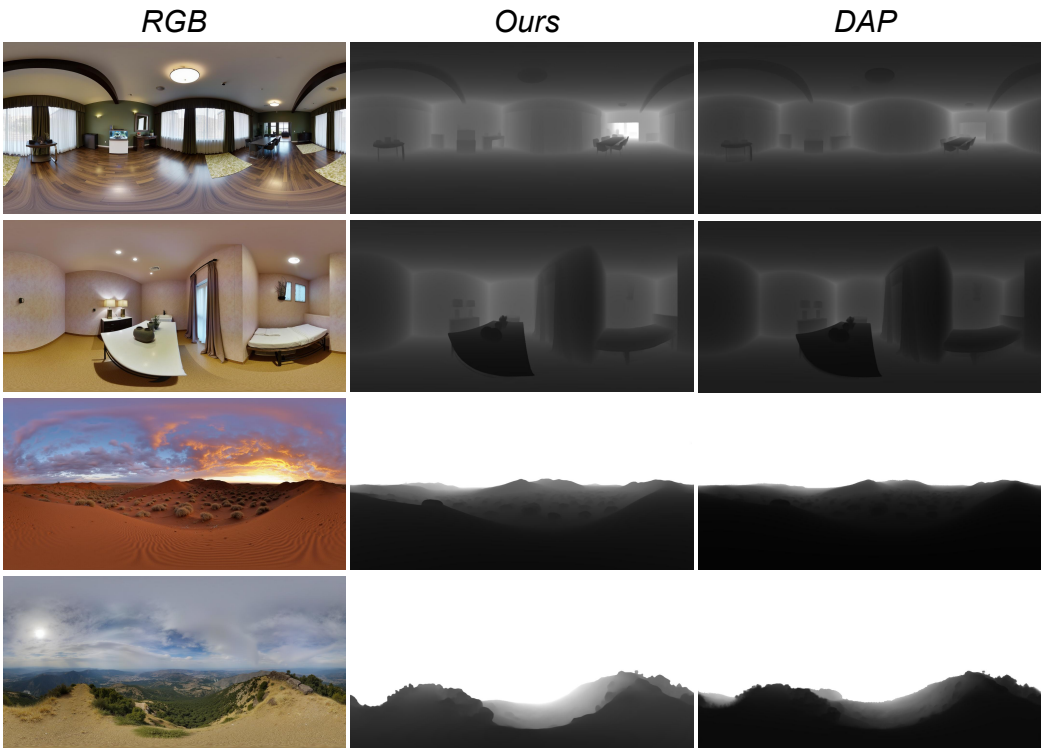

Figure 6: Joint RGB and depth generation by the pretrained model. The depth maps follow the scene layout and object boundaries of the corresponding RGB panoramas. Depth generation supplies auxiliary supervision during pretraining and is omitted during downstream in-context generation.

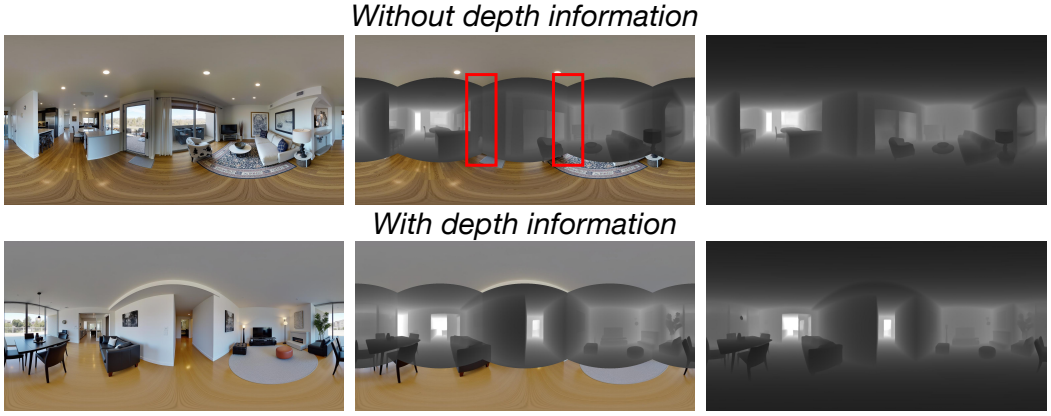

Figure 7: Comparison of RGB geometry with and without depth-supervised pretraining. Each generated panorama is projected onto cube-map faces. We estimate depth for the four side faces and stitch the estimates back into ERP. The depth-supervised model produces more coherent structures across the displayed views, particularly near stitching boundaries.

B Dataset Construction Details

Comparison with Existing Data Pipelines. Existing scene datasets support perception, embodied reasoning, multimodal understanding, and camera-controlled generation. Their annotations serve different objectives from the paired supervision required for task-conditioned panoramic synthesis. H*Bench (Yu et al., 2025a) supports embodied visual search with human-annotated reasoning, without

providing paired supervision for panoramic generation. Puffin-4M (Liao et al., 2025) supports camera-controlled generation with explicit camera modeling. Its pairs do not target panoramic inpainting, outpainting, or editing. The 360+x Dataset (Chen et al., 2024) emphasizes multimodal collection and alignment, without task-specific generative pairs or explicit geometric supervision for generation. *Canvas360Dataset* combines 100K RGB–depth pretraining examples with 900K context–target pairs for style transfer, inpainting, outpainting, and editing. The corresponding construction properties are summarized in Tab. 4.

Table 4: Comparison of dataset construction pipelines. *Canvas360Dataset* provides geometric supervision for panoramic pretraining and paired examples for four in-context generation tasks.

Data pipelines	Generation-oriented	Panoramic Data	Explicit Camera Modeling	Multi-task Design	Human-in-the-loop	Language Annotation	In-context Design
H*Bench	✗	✓	✓	✗	✓	✓	✗
Puffin-4M	✓	✗	✓	✓	✗	✓	✓
360+x Dataset	✗	✓	✗	✓	✓	✗	✗
Ours	✓	✓	✓	✓	✓	✓	✓

Pair Construction and Quality Control. Image generators and editors trained on perspective images can introduce geometric artifacts when applied to ERP panoramas. Our task-specific construction procedures reduce these artifacts in the training pairs. For inpainting and outpainting, we mask a source panorama to obtain the context and retain the complete panorama as the target. This construction preserves source geometry while defining missing content from the original panorama. For editing, models trained on perspective images can introduce geometric inconsistencies when adding objects to panoramas. We use grounding annotations to guide object removal, then reverse the resulting pairs to obtain addition examples whose targets are the original panoramas. For style transfer, we compare extracted line drawings between the original and stylized images and remove samples with large structural inconsistencies.

Manual inspection identified 48K valid examples in a 50K-sample subset, corresponding to a 96% acceptance rate within that subset. Separately, to assess the sensitivity of editing fine-tuning to pair cleaning, we compared 20K manually cleaned editing samples with 20K randomly selected editing samples. Cleaning reduced LPIPS from 0.096 to 0.093 and FAED from 0.396 to 0.391, while increasing PSNR from 25.34 to 25.70 (Tab. 5). The performance change was modest in this editing comparison.

Table 5: Data cleaning produces modest improvements in the 20K-sample editing comparison. Arrows indicate the preferred direction of each metric.

Method	LPIPS↓	FAED↓	PSNR↑
Randomly selected samples	0.096	0.396	25.34
Cleaned samples	**0.093**	**0.391**	**25.70**

Pseudo-depth Processing. We estimate panoramic metric depth with DAP (Lin et al., 2025). Very large depth estimates in distant regions can dominate the target distribution and destabilize optimization. We therefore clip depth values at 100 for outdoor scenes and at 10 for indoor scenes before normalization. Clipping limits the influence of extreme estimates on the auxiliary targets. We normalize the clipped maps to a fixed range before VAE encoding and use the normalized maps as targets for joint RGB–depth training.

C Experiment Settings

Model and Optimization. *Canvas360* uses FLUX.1-dev (Black Forest Labs, 2024) as its backbone. We insert LoRA adapters (Hu et al., 2021) into the attention blocks and input/output embedding layers, with rank $r = 64$, LoRA alpha $\alpha = 128$, and dropout 0.10. We train both stages with bfloat16 mixed precision on 8 NVIDIA H20 GPUs. Only the LoRA parameters are optimized. We use AdamW (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-6}$, and no weight decay.

Training Schedule and Inference. Stage one uses a learning rate of 2×10^{-5} for 30 epochs, with a per-GPU batch size of 2 and gradient accumulation over 2 steps. Stage two initializes from the stage-one adapter and uses a learning rate of 5×10^{-5} for 30 epochs, with a per-GPU batch size of 1 and gradient accumulation over 4 steps. Both stages therefore use an effective batch size of 32 across 8 GPUs. We apply a 200-step linear learning-rate warmup followed by a constant schedule. The guidance value supplied to the model during training is 1.0. The main experiments use 1024×2048 panoramas, and ablations use 512×1024 panoramas. Inference uses 28 steps with guidance scale 3.0.

Evaluation Metrics. We use complementary metrics to assess distributional fidelity, scene diversity, text alignment, and perceptual quality. Fréchet Inception Distance (FID) (Heusel et al., 2017) measures distributional fidelity. Following SMGD (Sun et al., 2025), we also report FID_{pole} and FID_{equ} to assess the polar and equatorial regions separately. Because the Inception features used by FID are learned from perspective images, we additionally report Fréchet Auto-Encoder Distance (FAED), which uses a representation tailored to panoramas (Oh et al., 2022). We compute Inception Score (IS) (Salimans et al., 2016) with a ResNet pretrained on Places365 (He et al., 2015; Zhou et al., 2018), replacing the conventional Inception-v3 classifier (Szegedy et al., 2015). This scene-classification variant jointly measures confidence for individual images and diversity across predicted scene categories. Its values are not directly comparable with conventional Inception-v3 IS. CLIP Score (CS) measures text–image alignment (Radford et al., 2021). Following HunyuanWorld (HunyuanWorld Team et al., 2025), we evaluate perceptual quality with Q-Align (QA) (Wu et al., 2023), BRISQUE (Mittal et al., 2012), and NIQE (Mittal et al., 2013). QA provides separate quality and aesthetic scores. The perceptual and reconstruction metrics used for downstream tasks, together with the style-transfer criteria, are described in Sec. F.

D Extended Text-to-Panorama Comparisons



Figure 8: Extended text-to-panorama comparisons. Red boxes mark local distortions and artifacts discussed in the text.

Table 6: Canvas360 obtains the lowest left-to-right boundary discrepancy among the evaluated methods. Following PanoFormer (Shen et al., 2022), we extend LRCE to RGB panoramas by measuring the discrepancy between the left and right boundary regions. Lower LRCE-RGB indicates better boundary consistency.

Metric	PanFusion	SMGD	PAR	WorldGen	LayerPano3D	HunyuanWorld	DiT360	Ours
LRCE-RGB↓	0.0154	0.0146	0.0171	0.0152	0.0188	0.0094	0.0101	**0.0061**

Fig. 8 extends the qualitative comparisons in the main paper. The marked regions show noisy textures, local geometric distortions, and boundary discontinuities. Across the displayed prompts, *Canvas360* generates sharper details and fewer visible seam discontinuities. Its FAED of 1.47, IS of 2.44, and CLIP score of 35.03 correspond to the quantitative comparison in Tab. 1. The FID results show that this advantage does not extend to every feature space or image region.

To quantify horizontal boundary consistency, we adapt the left-to-right consistency error (LRCE) from PanoFormer (Shen et al., 2022) to RGB panoramas. LRCE-RGB measures the discrepancy between the left and right boundary regions. *Canvas360* obtains the lowest error of 0.0061 in Tab. 6, compared with 0.0094 for the next-best method. The metric provides a quantitative counterpart to the visual evidence of reduced seam artifacts.

E Additional Ablations



Figure 9: Transfer of the pretrained panoramic prior. Under the same completion fine-tuning setting, *Canvas360* produces clearer and more geometrically coherent completions than FLUX.1-dev.



Figure 10: Velocity circular padding improves seam continuity relative to standard circular padding. A 180° yaw rotation places the original horizontal boundary at the center for inspection.



Figure 11: Cumulative RGB–depth ablations of scene geometry and dark outputs. The final column completes this sequence. VCP is evaluated separately in Fig. 10.

The visual comparisons examine transfer of the pretrained backbone, treatment of the horizontal boundary, and the joint RGB–depth design. Tab. 7 reports extended quantitative results for the component ablations.

Table 7: Extended quantitative ablations at 512×1024 resolution. B, VCP, D, and PO denote the baseline, velocity circular padding, parallel depth generation, and positional offset. VCP is evaluated independently. The RGB–depth variants are cumulative, adding PO to D and then $\mathcal{L}_{\text{sim}}$ to both. QA_{qua} and QA_{ae} denote QA quality and aesthetic scores. Shaded cells identify the two strongest values in each column. Arrows indicate the preferred metric direction.

Methods	FID↓	FID_{pole} ↓	FID_{equ} ↓	FAED↓	IS↑	CS↑	QA_{qua} ↑	QA_{ae} ↑	BRISQUE↓	NIQE↓
B	51.16	53.33	28.10	5.37	1.83	34.62	4.40	3.97	14.62	3.86
B + VCP	53.53	54.13	28.49	5.43	1.77	34.58	4.52	3.92	13.85	3.79
B + D	51.57	55.47	29.02	4.74	1.98	34.81	4.23	3.41	17.84	3.85
B + D + PO	57.04	62.82	29.44	4.81	1.85	34.72	4.44	4.13	23.64	4.12
B + D + PO + $\mathcal{L}_{\text{sim}}$	51.48	55.63	29.74	4.93	1.88	34.73	4.75	4.22	16.23	3.84

Velocity Circular Padding. Fig. 10 compares velocity circular padding with standard circular padding. Synchronizing boundary features while retaining consecutive longitude indices represents the wrap-around boundary as a local coordinate transition. The displayed results show improved edge alignment and fewer visible seam artifacts.

The B versus B + VCP comparison in Tab. 7 evaluates adding VCP to the baseline. BRISQUE decreases from 14.62 to 13.85 and NIQE from 3.86 to 3.79, while QA quality increases from 4.40 to 4.52. FID, its regional variants, and FAED do not improve. These no-reference scores assess image quality rather than seam continuity directly. The quantitative ablation therefore complements the visual comparison with standard circular padding in Fig. 10.

Parallel RGB–Depth Generation. Fig. 11 and Tab. 7 examine the cumulative RGB–depth design. Adding depth reduces FAED from 5.37 to 4.74 and improves IS, CS, and NIQE, which is consistent with its value as auxiliary supervision for panoramic generation. Depth alone nevertheless reduces QA quality and aesthetic scores and increases BRISQUE. Some displayed outputs contain dark regions or other artifacts.

Noisy depth targets and ERP distortion may make joint optimization sensitive to cross-modal coupling. This possibility motivates positional separation and regularization, but the ablation does not distinguish it from other potential causes of the visual failures.

Positional offsets distinguish the RGB and depth token groups. Adding them increases the aesthetic score from 3.41 to 4.13 but worsens FID, regional FID, BRISQUE, and NIQE. Similarity regularization then reduces FID from 57.04 to 51.48, BRISQUE from 23.64 to 16.23, and NIQE from 4.12 to 3.84. The combined RGB–depth variant achieves the highest QA quality and aesthetic scores, at 4.75 and 4.22. Its FAED of 4.93 remains below the RGB-only baseline's 5.37, while depth alone attains the lowest FAED. Together with fewer excessively dark outputs in the visual comparison, these results support combining geometric supervision with modality separation and regularization.

Transfer of the Pretrained Prior. Under matched completion fine-tuning, *Canvas360* produces clearer, more coherent panoramas than FLUX.1-dev (Black Forest Labs, 2024), whose outputs show blur and perspective-like patterns (Fig. 9). This comparison evaluates the pretrained prior as a whole. A matched RGB-only panoramic pretraining control would be needed to isolate the downstream contribution of depth supervision.

F Extended In-context Panoramic Comparisons

We evaluate style transfer, inpainting, outpainting, and editing with task-specific baselines and metrics. Each task uses 500 generated results.

F.1 Style Transfer

We compare with FLUX.1-Kontext-dev (Black Forest Labs et al., 2025), FLUX.2-dev (Black Forest Labs, 2025), and Qwen-Image-Edit (Wu et al., 2025a). Following SRQE (Chen et al., 2023), we report content preservation (CP), style resemblance (SR), and overall vision (OV). CP assesses

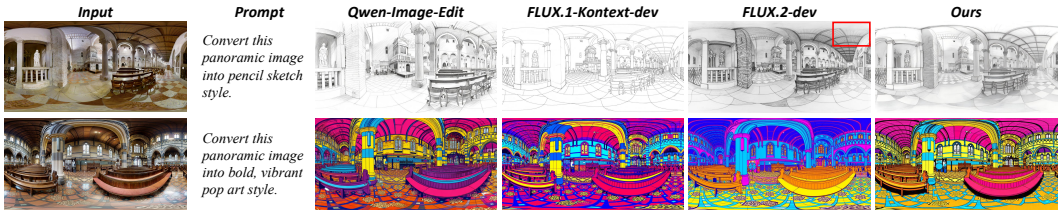

Figure 12: Style-transfer comparisons. *Canvas360* changes scene appearance while retaining the major geometric structure and content layout of the input panoramas.

Table 8: Style-transfer results. CP, SR, and OV denote content preservation, style resemblance, and overall vision. Higher values indicate better performance for all three measures.

Method	CP↑	SR↑	OV↑
FLUX.1-Kontext-dev	0.457	0.482	0.467
FLUX.2-dev	0.490	**0.503**	0.495
Qwen-Image-Edit	0.464	0.483	0.471
Ours	**0.502**	0.491	**0.497**

retention of the source content and layout, SR measures resemblance to the target style, and OV evaluates overall visual quality. Visual and quantitative comparisons appear in Fig. 12 and Tab. 8.

The displayed examples and scores indicate a trade-off between content preservation and style resemblance. *Canvas360* records CP and OV scores of 0.502 and 0.497, whereas FLUX.2-dev records the highest SR score of 0.503, compared with 0.491 for *Canvas360*. Because the OV difference is small, these point estimates do not establish a statistically reliable preference.

F.2 Inpainting and Outpainting

Table 9: Inpainting and outpainting results. LPIPS and FAED measure perceptual and panoramic feature-space distance, respectively, and PSNR measures pixel-level reconstruction quality.

Task	Method	LPIPS↓	FAED↓	PSNR↑
Inpainting	FLUX.1-Fill-dev	0.171	0.461	24.23
	PAR	0.158	0.455	23.76
	PanoDiffusion	0.147	0.523	23.93
	Ours	**0.096**	**0.371**	**25.87**
Outpainting	FLUX.1-Fill-dev	0.509	1.916	16.32
	PAR	0.553	1.849	16.71
	PanoDiffusion	0.674	1.989	15.21
	Ours	**0.416**	**1.791**	**17.16**

We compare with FLUX.1-Fill-dev, PAR, and PanoDiffusion, as in the main paper. LPIPS measures perceptual distance to the reference (Zhang et al., 2018), FAED assesses panoramic distributional fidelity (Oh et al., 2022), and PSNR measures pixel-level reconstruction quality.

The table reports lower distances and higher PSNR for the proposed method across both tasks. Inpainting LPIPS is 0.096 and PSNR is 25.87, compared with 0.147 and 24.23 for the strongest corresponding baseline. For outpainting, LPIPS is 0.416 and PSNR is 17.16, compared with 0.509 and 16.71. The corresponding FAED values also improve, supporting stronger panoramic fidelity for local completion and scene extension.

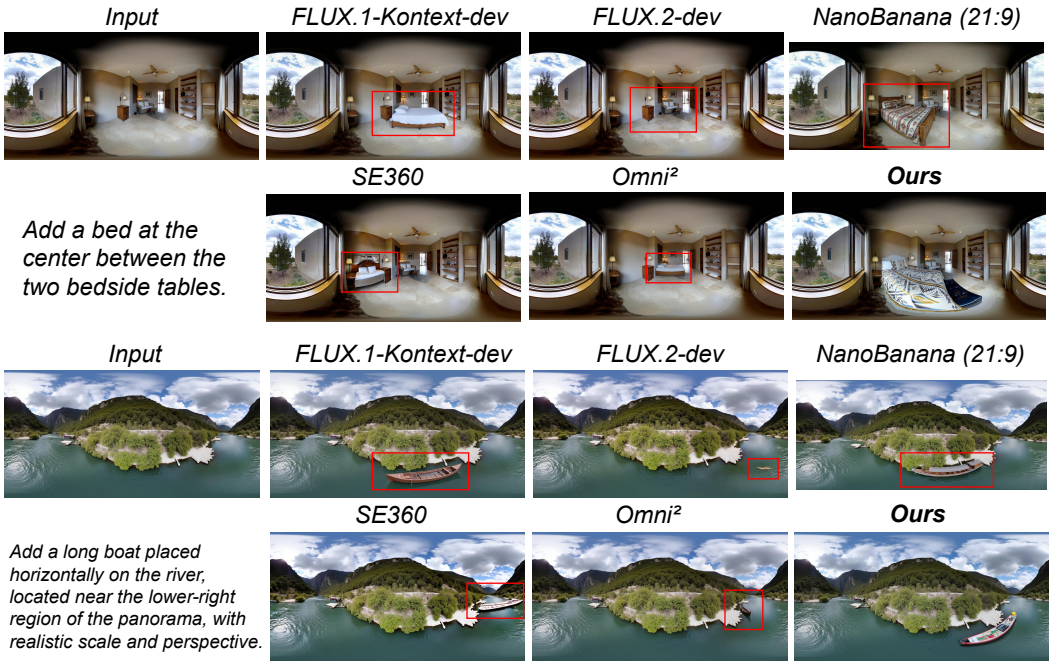

Figure 13: Panoramic editing comparisons. The examples examine whether the requested changes preserve local ERP geometry and the surrounding scene content.

Table 10: Panoramic editing results. LPIPS, FAED, and PSNR measure perceptual distance, panoramic feature-space distance, and pixel-level reconstruction quality, respectively.

Method	LPIPS↓	FAED↓	PSNR↑
FLUX.1-Kontext-dev	0.102	0.458	25.77
FLUX.2-dev	0.099	0.410	26.17
NanoBanana	0.094	0.395	25.91
SE360	0.138	0.386	25.16
Omni2	0.105	0.392	25.03
Ours	**0.084**	**0.358**	**26.40**

F.3 Editing

We compare with FLUX.1-Kontext-dev (Black Forest Labs et al., 2025), FLUX.2-dev (Black Forest Labs, 2025), and NanoBanana (Google, 2025), as well as the panorama editors SE360 (Zhong et al., 2025) and Omni2 (Yang et al., 2025). We use LPIPS, FAED, and PSNR, as in the completion evaluations. Visual and quantitative comparisons appear in Fig. 13 and Tab. 10.

Canvas360 performs the requested edits while retaining panoramic geometry and surrounding content in the displayed examples. Its LPIPS, FAED, and PSNR are 0.084, 0.358, and 26.40, respectively. The corresponding strongest baseline values are 0.094, 0.386, and 26.17. The gains span perceptual similarity, panoramic distributional fidelity, and pixel-level reconstruction accuracy.

G Additional Generation Results

Figs. 14 and 15 show additional text-to-panorama results and examples of the four in-context tasks. Across scene types and transformations, the examples retain scene detail, ERP-consistent object shapes, and boundary continuity.

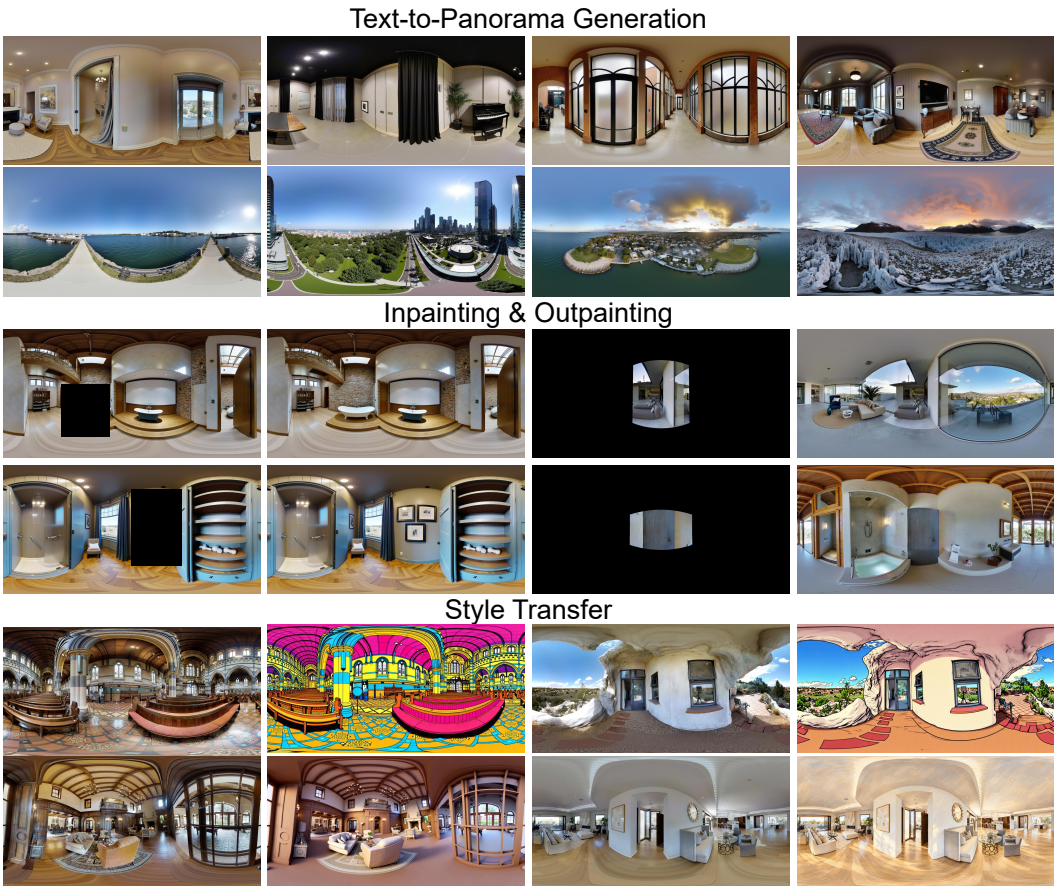

Figure 14: Additional text-to-panorama and in-context results from *Canvas360*, illustrating scene detail and structure across the supported tasks.

H Human Preference Study Details

The study in Tab. 2 compared outputs from *Canvas360*, Matrix-3D, HunyuanWorld, and DiT360. Participants viewed four anonymized panoramas for each question and selected the result that best satisfied the displayed criterion.

Evaluation Criteria. Participants evaluated text alignment (TA), boundary continuity (BC), panorama awareness (PA), and overall quality (OQ). Text alignment measures whether the generated panorama is consistent with the input prompt. Boundary continuity measures whether the panorama is seamless near the horizontal wrap-around boundary. Panorama awareness measures whether image structure is consistent with panoramic geometry, including projection distortion and the spatial layout across the field of view. Overall quality measures the general perceptual fidelity, realism, and visual appeal of the generated panorama.

Study Interface. Fig. 16 shows the evaluation interface. Each page presented one criterion and four candidate images. Participants selected one preferred result. Method names were hidden to reduce identification bias.

Participants and Procedure. The study involved 71 volunteers and 10 test cases. Participants were informed of the study procedure before starting the evaluation. Participation was voluntary and had no effect on employment, compensation, or performance evaluation. Each participant compared generated panoramic images through four-choice questions. We used the responses only for aggregate statistical analysis.

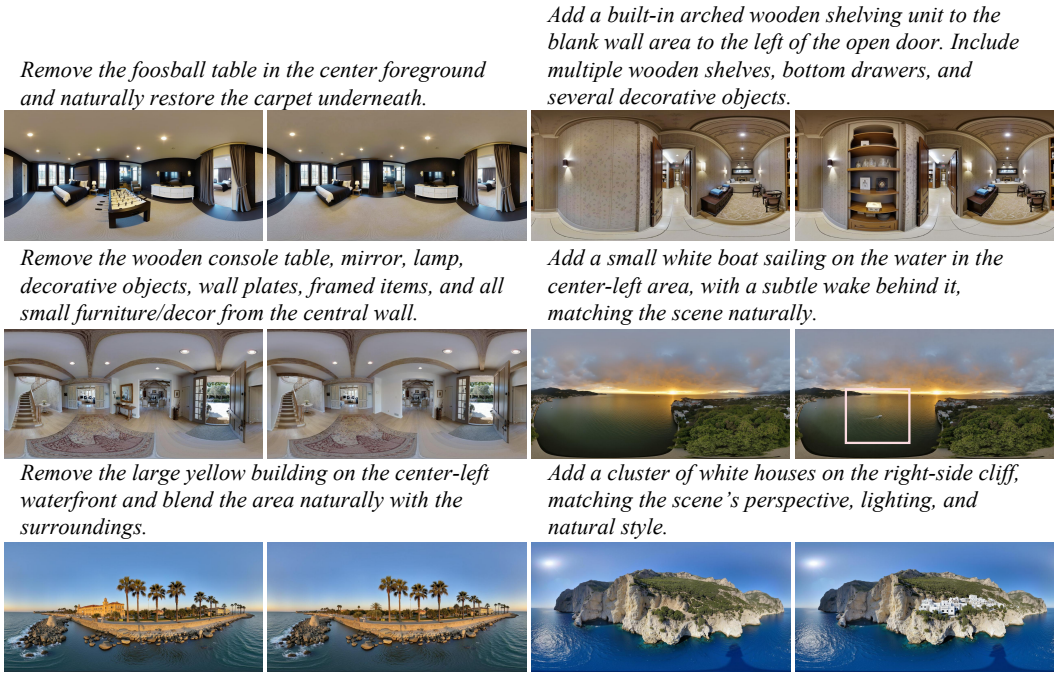

Figure 15: Additional panoramic edits from *Canvas360*. The examples illustrate object manipulation within the surrounding ERP scene.

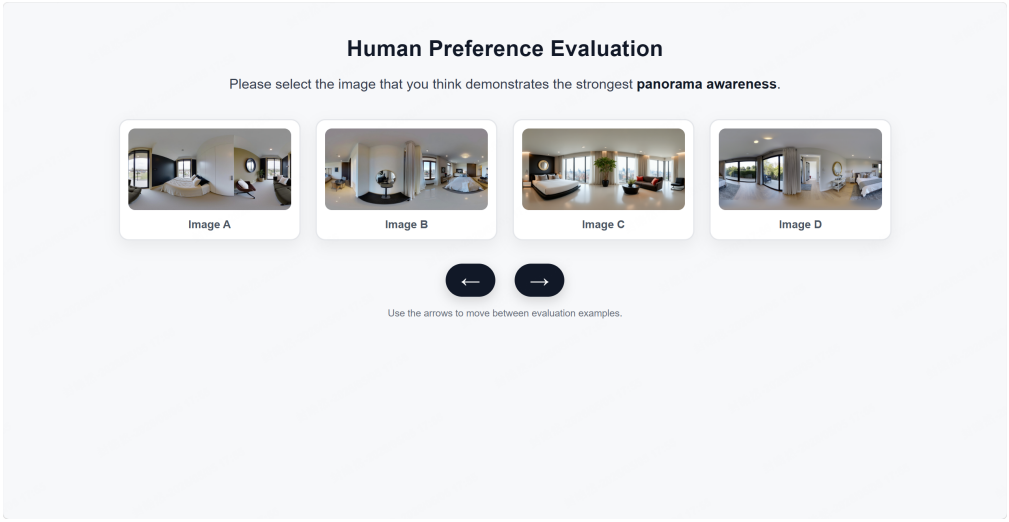

Figure 16: Screenshot of the human preference study interface. For each question, participants were shown four anonymized candidate panoramic images and selected the one that best satisfied the displayed evaluation criterion.

Risk, Privacy, and Review. The study involved minimal risk. Participants compared generated panoramas without exposure to sensitive content and provided only image-preference choices. We did not collect personally identifiable information, sensitive attributes, private data, or free-form personal responses. We aggregated all responses across participants before analysis and reporting. The study was submitted to and reviewed through the authors' organizational review process. To preserve anonymity in the initial submission, institution-identifying details are omitted.

I SAFEGUARDS FOR RESPONSIBLE RELEASE

Panoramic generation can be misused for deceptive scene manipulation or synthetic visual misinformation. During dataset construction, we filter unsafe, sensitive, and low-quality samples and focus on scene-level content. This scope reduces exposure to identity-related imagery, although it does not eliminate privacy or misuse risks.

We will accompany released resources with usage guidelines addressing deceptive manipulation, visual misinformation, and unauthorized content generation. Recommended safeguards include provenance tracking, watermarking, transparent attribution, and human review for sensitive applications. Dataset documentation will describe the sources, preparation procedures, filtering steps, and intended-use restrictions.

J LIMITATIONS AND FUTURE WORK

The training corpus is imbalanced across scene categories, and high-quality panoramas are scarce for some content. *Canvas360* can struggle with underrepresented details, including high-resolution faces and text-rich signage, especially under strong ERP distortion. Expanding and rebalancing the corpus may improve these cases. The completion comparison evaluates the pretrained model as a whole, so it does not isolate the downstream contribution of depth supervision. Geometry checks remain qualitative, and the human study covers 10 cases. Matched RGB-only pretraining controls, direct geometric measurements, and broader preference testing would strengthen the evaluation. Distributional metrics also show trade-offs. DiT360 retains lower overall and regional FID despite *Canvas360*'s lower FAED. Repeated evaluations are needed to characterize uncertainty in close metric comparisons.