SE360: Semantic Edit in 360° Panoramas via Hierarchical Data Construction

Haoyi Zhong[1], **Fang-Lue Zhang**[1*], **Andrew Chalmers**[1], **Taehyun Rhee**[2]

[1]Victoria University of Wellington, New Zealand
[2]The University of Melbourne, Australia
{haoyi.zhong, fanglue.zhang, andrew.chalmers}@vuw.ac.nz, taehyun.rhee@unimelb.edu.au

Abstract

While instruction-based image editing is emerging, extending it to 360° panoramas introduces additional challenges. Existing methods often produce implausible results in both equirectangular projections (ERP) and perspective views. To address these limitations, we propose SE360, a novel framework for multi-condition guided object editing in 360° panoramas. At its core is a novel coarse-to-fine autonomous data generation pipeline without manual intervention. This pipeline leverages a Vision-Language Model (VLM) and adaptive projection adjustment for hierarchical analysis, ensuring the holistic segmentation of objects and their physical context. The resulting data pairs are both semantically meaningful and geometrically consistent, even when sourced from unlabeled panoramas. Furthermore, we introduce a cost-effective, two-stage data refinement strategy to improve data realism and mitigate model overfitting to erase artifacts. Based on the constructed dataset, we train a Transformer-based diffusion model to allow flexible object editing guided by text, mask, or reference image in 360° panoramas. Our experiments demonstrate that our method outperforms existing methods in both visual quality and semantic accuracy.

Introduction

While instruction-based image editing has made substantial progress in conventional image settings, the growing demand for immersive applications—such as virtual reality (VR)—has shifted attention toward the 360° panoramic images. Enabling diverse, high-fidelity content manipulation in omnidirectional environments presents additional challenges that remain underexplored.

State-of-the-art instruction-based image editing methods often produce artifacts when applied to 360° imagery in equirectangular projection (ERP). As shown in Figure 1, these models struggle with the domain gap between perspective and spherical views—edited objects appear plausible in ERP but become geometrically distorted when viewed through VR or panoramic viewers. Lacking awareness of spherical continuity, they also treat ERP boundaries as disconnected, causing visual breaks. Cubemap-based approaches like Omni^2(Yang et al. 2025b) offer partial relief by reducing distortion but yield moderate gains in realism.

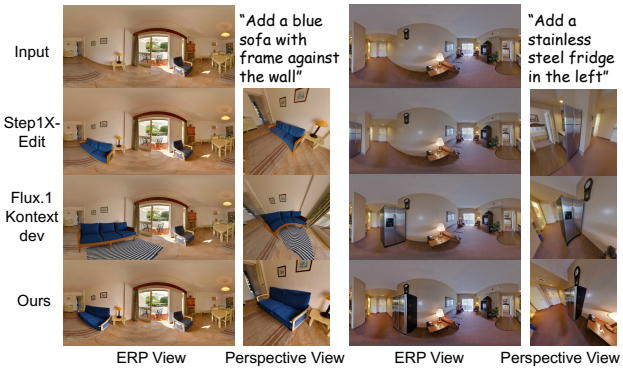

Figure 1: Comparison with state-of-the-art models.

However, Cubemap-based approaches introduce their own limitations, particularly in handling large objects that span multiple cubemap faces. More critically, automated data annotation methods based on Vision-Language Models (VLMs), such as Erasedraw (Canberk et al. 2024), Omni^2 and InsightEdit (Xu et al. 2025) often rely on bounding box (BBox)-guided segmentation for mask generation. These methods tend to overlook object hierarchies and physical containment relationships. For example, a sofa might be segmented while ignoring a cushion resting on it, resulting in semantically inconsistent training data. While such issues can be mitigated with manual filtering in perspective image datasets, this is impractical for the sparsely annotated 360° domain.

To address these challenges, we propose **SE360**, a scalable framework for instruction-guided object editing in 360° panoramas. At its core is a coarse-to-fine autonomous data generation pipeline that creates high-quality training pairs from unlabeled ERP images. The pipeline proceeds as follows: (1) *Object identification*: a VLM, aided by referential detection models, extracts candidate objects and generates descriptions in the ERP domain. (2) *Containment-aware projection*: candidate objects are projected into perspective views, where hierarchical relationships (e.g., sofa and cushion) are identified via VLM-driven analysis. (3) *View adjustment and segmentation*: the system adjusts viewpoint and field of view (FOV) to fully capture the object and its con-

*Corresponding author

text, followed by segmentation to obtain object masks. (4) *Erasing and prompt synthesis*: the object is removed using the mask and an erasing model, while the VLM regenerates descriptive prompts based on the original attributes and spatial context. This process requires no manual labeling and enables large-scale, semantically aligned 360° editing data generation.

Beyond the base pipeline, we address limitations in erasing artifacts and model overfitting through refinement and training design. Using these two datasets, we train a multi-conditional Transformer diffusion model in a two-stage process to perform object editing. To enable 360° spatial understanding, we replace conventional positional embeddings with 3D spherical positional embedding that projects each pixel to unique coordinates on a unit sphere. This provides explicit geometric priors about distortion patterns and boundary continuity in panoramic imagery.

Our main contributions are summarized as follows:

- We propose a novel, coarse-to-fine automated 360° data generation pipeline. By leveraging adaptive projection adjustment and VLM-driven analysis, it achieves a holistic understanding of object hierarchies and addresses the challenge of editing large or cross-face objects.
- We introduce a cost-effective data refinement strategy that combines traditional erasing with advanced instruction-based models, significantly improving training data quality.
- We build and train a multi-conditional diffusion-based 360° panorama editing model. The model incorporates spherical positional priors and topological continuity handling, enabling instruction-guided editing across arbitrary locations.

Our experiments demonstrate that SE360 outperforms existing methods across multiple metrics, excelling in visual quality, semantic accuracy, and edit plausibility.

Related Works

Instruction-Guided Image Editing. Instruction-guided image editing aims to modify images via natural language instructions. The foundational paradigm for this task was established by the pioneering work InstructPix2Pix (Brooks, Holynski, and Efros 2023), which auto-generates large-scale (instruction, image) training data, surpassing earlier explorations like SDEdit (Meng et al. 2022; Hertz et al. 2023). To enhance the quality of this synthetic data, subsequent works have pursued multiple avenues, including leveraging high-quality human annotations (Zhang et al. 2023) and proposing innovative automatic data generation strategies (Zhao et al. 2024; Hui et al. 2024; Canberk et al. 2024; Wasserman et al. 2025; Xu et al. 2025). More recently, the advent of Diffusion Transformers (Peebles and Xie 2023) has propelled new paradigms like "in-context editing" and spurred research into building unified models for both generation and editing (Xiao et al. 2025; Chen et al. 2025; Zhang et al. 2025; Labs et al. 2025; Liu et al. 2025). However, all these methods face fundamental challenges when applied to 360° panoramas. As they are not designed to account for the unique properties of spherical geometry, their edits result in severe perspective distortions and visible seam artifacts. Our work is proposed to address this specific challenge.

360° Panorama Generation and Editing. 360° panorama editing has unique challenges related to continuity and geometric consistency for generative models. To ensure seamless connectivity between the left and right boundaries, existing methods either employ an autoregressive outpainting strategy to progressively expand from a narrow FoV image (Lu et al. 2024; Wu, Zheng, and Cham 2024), or leverage techniques such as circular padding and blending during the holistic generation process to enforce topological continuity on the model (Feng et al. 2023; Zhang et al. 2024). Some methods circumvent the distortion issues inherent in the ERP format by decomposing the panorama into multiple perspective views with lower distortion, such as cubemaps, thereby better leveraging powerful diffusion models pre-trained on standard perspective images (Tang et al. 2023; Kalischek et al. 2025). However, the task of editing panoramas, particularly at the semantic level, remains underexplored. Early panoramic editing techniques (Zhang et al. 2021, 2022) allow simple color or style adjustments, lacking semantic content understanding. Approaches like (Yang et al. 2025b) have begun to unify generation and editing tasks; however, a dedicated framework for high-fidelity, instruction-guided, object-level editing—particularly one capable of handling objects that span across multiple views—remains an open challenge. Our work is proposed to bridge this gap to enable consistent semantic editing of arbitrary objects within 360° panoramas.

Method

SE360 employs a transformer-based learning framework capable of handling diverse input conditions, thereby enabling flexible and high-fidelity panoramic image editing. To train the model, we adopt a two-phase, automated data generation pipeline to construct 360° panoramic datasets. The initial phase, *SE360-Base*, is designed to extract fine-grained, instance-level annotations from large-scale 360° panorama datasets. The subsequence phase, *SE360-HF*, builds upon the output of the SE360-Base, refines the data via enhancing the images' visual quality. Finally, we obtain paired image data for the most common editing tasks: object removal and addition.

SE360-Base: Large-Scale Dataset Generation

Through several dedicated data processing stages, SE360-Base enables robust object grounding, captures complete composite objects, and ensures full visibility in 360° scenes. This results in high-quality, context-rich annotations that support accurate and flexible image editing.

Stage 1: Object Extraction. As in Figure 2, we first guide the VLM model (Qwen2.5-VL-32B) (Bai et al. 2025) to identify and describe the primary foreground objects within each scene, yielding a structured list of objects, each annotated with a detailed description and a category. Subsequently, we employ a multi-model fusion strategy for robust phrase grounding, which integrates the strengths of three

distinct models: the initial VLM, Grounding DINO (Liu et al. 2024), and Florence-2 (Xiao et al. 2024). We adopt a consensus-based filtering strategy to enhance grounding precision. An object is accepted only if Grounding DINO or Florence-2 produces a bounding box with high Intersection over Union (IoU) overlap with the VLM's initial detection, and the largest valid box is selected for coverage. For objects missed by the VLM but detected by both others, we retain only high-confidence overlapping results.

Stage 2: Physical Affiliation Analysis. Recent methods (Canberk et al. 2024; Xu et al. 2025) using VLMs and Grounded-SAM (Ren et al. 2024) often produce incomplete masks, especially for composite objects with varied textures—e.g., segmenting a side table but missing items placed on it (Figure 3). Even replacing Segment Anything (SAM) (Kirillov et al. 2023) with advanced models like Semantic-SAM (Li et al. 2023) struggles to unify such objects into a single entity. To overcome this, we propose Physical Affiliation Analysis (PAA), which uses VLM-guided prompts to identify and list items physically associated with a target object. It enables SAM to segment each part individually, preserving object completeness for fine-grained detection. We automate this process using a VLM guided by a rule-based prompt. The VLM is prompted to list items that are on, inside, or attached to the target object. The results are saved in a structured JSON file for use in the next fine-grained detection stage.

Stage 3: Adaptive Projection Adjustment. We then perform fine-grained detection within an adaptively adjusted perspective viewport. Rather than relying on a static projection, our method iteratively adjusts the Field of View (FOV) and projection center to ensure all target objects are fully visible, as illustrated in the "Adaptive Adjustment" panel of Figure 2. Grounding DINO is applied to detect primary and affiliated items in the initial projection. If any bounding box touches the viewport edge—indicating partial visibility—the FOV is expanded and the center refined. This loop continues until all objects are fully enclosed or a maximum number of iterations is reached. This adaptive mechanism ensures complete object views for reliable segmentation in complex 360° scenes.

Stage 4: Mask-Guided Erasing. To support mask-guided erasing, we employ a progressive segmentation strategy with SAM2 (Ravi et al. 2024). An initial coarse mask, generated by SAM2 using the object's bounding box, is refined by re-prompting SAM2 with positive points sampled from its interior. We then remove noises and dilate the refined mask to create the final erasing mask. Instead of erasing on the perspective image, we reproject the final mask to the full ERP image to guide LaMa (Suvorov et al. 2022), leveraging global context for more coherent results.

Stage 5: Instruction Recaption. We finally introduce an Instruction Recaption module to enhance the descriptive richness of our dataset. Leveraging the original description and the adaptively-adjusted perspective view, we utilize a VLM to generate two new, hierarchical descriptions:

- Standard Refined Description, which integrates verifiable visual details from the high-quality perspective view with a pre-established spatial context.
- Brief Description, which offers a succinct summary of the object's core attributes and location.

This dual-level description provides each object with labels that are both detailed and context-aware, as well as concise and easily parsable—allowing models to handle instructions of varying complexity. The resulting JSON file is then processed by Qwen3-8B (Yang et al. 2025a), which classifies the localization method as either absolute (relative to the image frame) or relative (with respect to other objects). This classification guides our image rotation strategy during model training. From the initial data generation phase, we obtained 195,504 editing triplets ($37{,}304 \times 2 + 60{,}448 \times 2$) for addition and removal tasks. This total includes both the detailed and brief variants of each instruction.

SE360-HF: High-Fidelity Refinement

Object erasing in complex scenes often results in visual artifacts, such as residual shadows, which can cause the model to overfit to these imperfections when generating novel content, as shown in our ablation studies in later sections. Although downsampling helps mitigate such artifacts during training, it proves unreliable for large-scale objects. To address this, we employ a high-fidelity data generation phase, SE360-HF. We replace LaMa with a more robust, instruction-driven editor—Flux.1 Kontext max—to produce cleaner image-edit pairs. As shown in Figure 4, we first generate a template instruction based on the object's category (e.g., "remove the table with chairs in the center of the image"), which, along with the perspective view from the SE360-Base pipeline, guides the editing process. The resulting images undergo a two-stage filtering process:

1. SSIM Filtering: We first compute the Structural Similarity Index (SSIM) (Hore and Ziou 2010) between the erased and original images, discarding those with scores above a predefined threshold.
2. Region-Constrained Feature Similarity Check: We perform a localized similarity check using DINOv2 (Oquab et al. 2023) features within the masked bounding box and at the image borders. This region-constrained strategy, unlike a global comparison, allows for consistent filtering across objects of varying sizes with a uniform threshold. Images with high similarity scores in these areas are excluded.

We sampled 10,000 images with a distribution of 70% large, 20% medium, and 10% small. After applying the automated filtering pipeline, 8,420 high-quality images were retained, resulting in 16,840 edit pairs using the same instructions as SE360-Base. Training on this refined dataset reduces overfitting to artifacts and improves the realism of generated objects.

Model Design

Here, we introduce SE360, our learning framework that adapts powerful pre-trained diffusion models for multi-

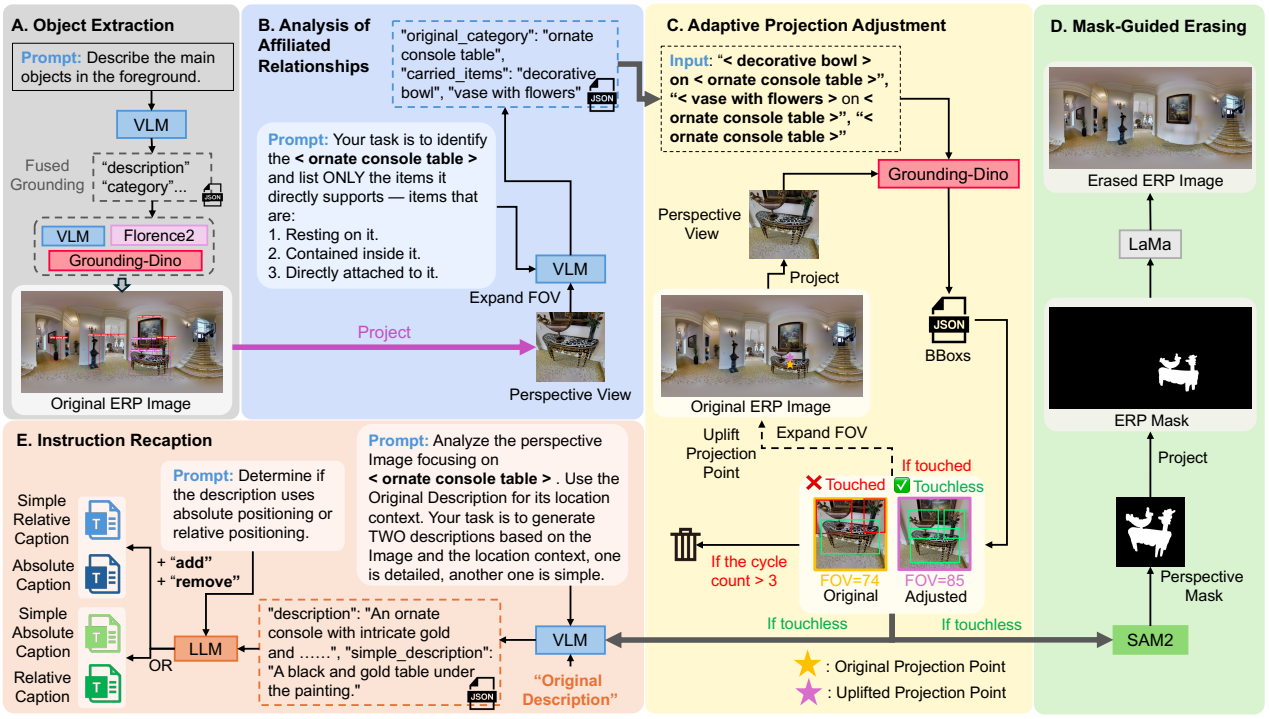

Figure 2: The overview diagram of SE360-Base.

Figure 3: Comparison of different segmentation strategies.

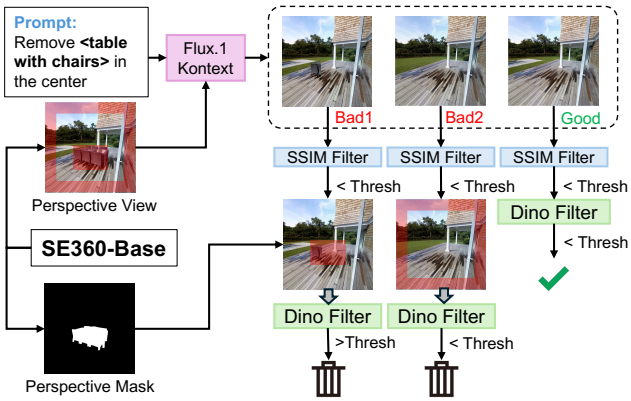

Figure 4: The overview diagram of SE360-HF.

conditional, multi-modal editing of 360° panoramic images through a series of designs tailored to the spherical domain.

Diffusion Transformer. As illustrated in Figure 5, the core of our model is a Diffusion Transformer that drives the generative process. To incorporate strong editing priors, we initialize it with weights from the pre-trained OmniGen (Xiao et al. 2025). The Transformer operates on a concatenated 1D sequence that fuses multimodal information, including token embeddings of text prompts and latent visual features extracted from the source panorama and reference images—produced by a frozen, location-aware Variational Autoencoder (VAE) (Kingma and Welling 2014) of SDXL (Podell et al. 2024) . This sequence is further enhanced with 3D spherical positional embeddings for spatial awareness, along with timestep embeddings and an editing mask map to guide the diffusion.

For parameter-efficient adaptation to panoramic editing, we adopt Low-Rank Adaptation (LoRA) (Hu et al. 2021), fine-tuning only the query (Q), key (K), and value (V) projection matrices within attention layers, as well as the feed-forward networks. This strategy preserves most of the pre-trained knowledge while efficiently adapting the model to the unique properties of the 360° spherical domain.

Location-Aware Encoding and Decoding. To eliminate visible artifacts at the longitudinal seams of panoramic images caused by standard VAE convolutional kernels failing to account for spherical topology (Wu, Zheng, and Cham 2023), we introduce a symmetric "Pad-and-Unpad" strategy. During encoding, we apply circular padding by prepending and appending content from the opposite horizontal edges, providing the encoder with adjacent 360° context. This results in a semantically seamless latent representation, which

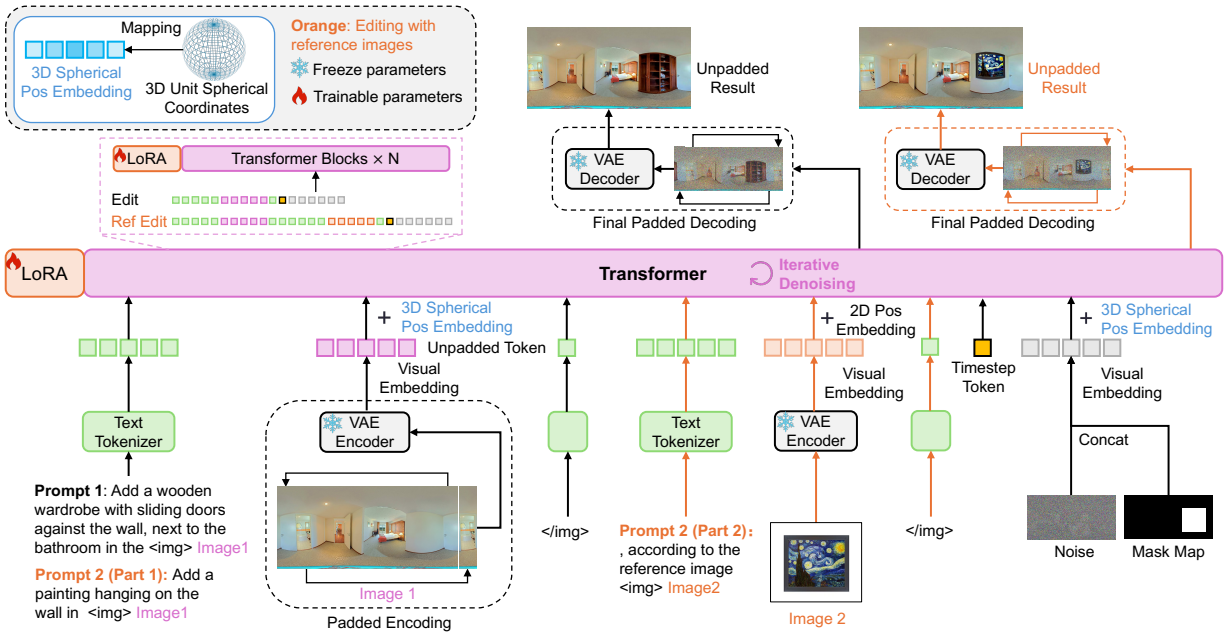

Figure 5: The model architecture of SE360.

is then cropped back to the original dimensions. In decoding, we apply the same circular padding to the latent features, allowing the decoder to reconstruct smooth boundary transitions at the pixel level. Finally, we crop the central portion of the slightly wider output to produce a seamless and visually coherent panoramic image, effectively removing artifacts introduced by standard convolution.

3D Spherical Positional Embedding. Conventional 2D sinusoidal positional embeddings struggle with panoramas, as they fail to capture the true spatial relationships inherent in spherical geometry. To overcome this, we introduce a 3D Spherical Positional Embedding (SPE) strategy. Specifically, we map ERP pixel coordinates to 3D Cartesian coordinates on a unit sphere. Given pixel coordinates (i, j) in an ERP image of size $H \times W$, where $i \in [0, H-1]$ and $j \in [0, W-1]$, we first convert them to spherical coordinates—longitude (λ) and latitude (ϕ)—and then transform them into 3D Cartesian coordinates $\mathbf{p}$ on the unit sphere as:

$$\mathbf{p} = (x, y, z) = (\cos\phi \cos\lambda, \cos\phi \sin\lambda, \sin\phi). \quad (1)$$

We extend the standard frequency-based positional encoding to 3D by dividing the embedding dimension d equally among the x, y, and z axes, applying sinusoidal functions independently to each. This preserves spatial proximity on the spherical manifold while remaining compatible with Transformer architectures, offering a more faithful spatial prior for panoramic understanding.

Training Methodology. SE360 contains 3.9 billion parameters and is initialized from OmniGen. We adopt a two-stage training strategy: first, foundational training on SE360-Base for 10K iterations; second, fine-tuning on SE360-HF for 1,000 iterations, using weights from stage one. Training is based on the Flow Matching (Lipman et al. 2023) framework and incorporates LoRA with rank and alpha set to 16. To enable precise local editing, a mask map is applied with 20% probability. For reference-based addition, we use object masks from perspective views in SE360-Base to remove backgrounds and construct inputs. We also conduct orientation-decoupled data augmentation by latitudinal rotation to preserve directional consistency with the text prompt containing absolute instructions. Further details are provided in the supplementary.

Experiments

We conduct a series of experiments to comprehensively evaluate SE360. Our evaluation focuses on three aspects: comparison with state-of-the-art image editing models for panoramic object removal and addition, demonstration of multi-conditional input editing, and ablation studies analyzing each data generation pipeline's contribution.

Dataset

SE360 builds a large-scale editing dataset of over 30,000 images from Matterport3D (Chang et al. 2017) and Structured3D (Zheng et al. 2020). For evaluation, we curated a high-quality test set from the PanoContext (Zhang et al. 2014) dataset, named by *PanoEval*. We randomly selected 250 panoramas and applied the same data creation pipeline as SE360-HF to generate editing pairs. These pairs were then manually inspected to ensure correctness and quality, resulting in a test set of 448 high-fidelity editing pairs.

Comparison with State-of-the-art

To evaluate the model's handling of panoramic boundaries, we created a challenging subset, PanoEval-Boundary, containing 200 editing pairs from our test set that were rotated

Figure 6: The comparison with SOTA models.

horizontally so the object of interest lies at or crosses the left–right seam of the ERP image. The main text reports only average results across the two test sets; detailed results are provided in the supplementary material.

To evaluate the performance of SE360 in 360° panoramic image editing, we benchmark it against several state-of-the-art open-source perspective image editing models, including InstructPix2Pix (Brooks, Holynski, and Efros 2023), Erasedraw (Canberk et al. 2024), ICEdit (Zhang et al. 2025), MagicBrush (Zhang et al. 2023), PaintByInpaint (Wasserman et al. 2025), Omnigen (Xiao et al. 2025), Flux.1 Kontext dev (Labs et al. 2025) and Step1X-Edit (Liu et al. 2025), on two primary tasks: object removal and object addition.

Evaluation Metrics. We use PSNR (Hore and Ziou 2010) and LPIPS (Zhang et al. 2018) to measure reconstruction quality and perceptual similarity in panoramas, respectively. $\text{LPIPS}_{\text{pers}}$ are used to measure perceptual similarity in perspective view. FAED (Oh et al. 2022) assesses the realism of the generated panoramas. CS_{pers} measures the semantic consistency between the edited result and the text description by CLIP (Radford et al. 2021) in perspective views. For the removal task, we utilize $\text{CS-No}_{\text{pers}}$, which measures the similarity of the edited image to a scene description that excludes the target object in perspective view.

Model	$\text{CS}_{\text{pers}}\uparrow$	LPIPS↓	$\text{LPIPS}_{\text{pers}}\downarrow$
InstructPix2Pix	23.691	0.783	0.798
Erasedraw	25.262	0.733	0.737
ICEdit	27.196	0.286	0.356
MagicBrush	27.618	0.109	0.219
PaintByInpaint	28.124	0.094	0.198
OmniGen	28.370	0.091	0.209
Flux.1 Kontext dev	29.295	0.141	0.265
Step1X-Edit	28.919	0.091	0.211
SE360 (Ours)	**29.354**	**0.076**	**0.177**

Table 1: Quantitative comparison for the object addition task on our test set. Best results are in **bold**.

Results. SE360 outperforms existing methods in object addition and removal (Tables 1 and 3), achieving superior perceptual quality and semantic accuracy. These advantages are shown in Figure 6. Other models often produce geometrically inconsistent objects or leave visual artifacts when handling large boundary-crossing edits. In contrast, SE360 leverages its 3D spatial understanding to generate perspectively accurate objects, seamlessly inpaint occluded regions, and produce visually coherent results. For object removal across stitching seams, methods like Step1X-Edit often fail

	Addition			Removal				
Trainnig Type	CS_{pers}↑	LPIPS↓	$LPIPS_{pers}$↓	$CS\text{-}No_{pers}$↑	FAED↓	LPIPS↓	$LPIPS_{pers}$↓	PSNR↑
SE360-Base w/artifacts	**29.377**	**0.053**	**0.125**	-	-	-	-	-
SE360-Base	27.965	**0.074**	0.184	73.055	0.498	**0.057**	**0.128**	25.854
SE360-HF	28.402	0.171	0.285	**73.820**	2.570	0.170	0.268	19.419
Full	**29.354**	0.076	**0.177**	73.143	**0.314**	0.079	0.179	**26.394**

Table 2: Ablation studies on the object addition and removal tasks.

Model	$CS\text{-}No_{pers}$↑	FAED↓	LPIPS↓	$LPIPS_{pers}$↓	PSNR↑
MagicBrush	71.319	2.798	0.119	0.223	20.268
ICEdit	72.232	1.545	0.268	0.311	23.058
Omnigen	72.256	0.857	0.066	0.142	25.200
Flux.1 Kontext dev	71.975	1.580	0.115	0.191	19.890
Step1X-Edit	72.606	0.748	**0.057**	**0.095**	**28.471**
SE360 (Ours)	**73.143**	**0.314**	0.079	0.179	26.394

Table 3: Quantitative comparison for the object removal task on our test set. Best results are in **bold**.

Figure 7: Addition with reference image and mask map.

to synthesize continuous textures, whereas SE360 removes objects cleanly and generates plausible background content. For object addition, especially with perspective distortion (e.g., inserting a painting high on a wall), existing methods tend to produce flat, unrealistic results, while SE360 generates properly warped, well-lit additions.

Multi-Condition Editing

Beyond single-object manipulation, SE360 proves capable of processing complex directives with multiple concurrent conditions. As showcased in Figure 7, the model performs object insertion guided by both a mask map and a reference image, generating panoramas with geometric and photometric consistency.

Ablation Studies

To validate the effectiveness of our training method, we conducted a series of ablation studies on two test datasets. The average results in Table 2 reveal a clear trade-off in SE360's performance: the model performs best on data containing LaMa's erasing artifacts but worst when those artifacts are absent. This indicates that training exclusively on SE360-Base leads to overfitting to residual artifacts, even though the model learns object-level texture and geometry. Conversely, training solely on SE360-HF mitigates this overfitting but degrades generation quality due to insufficient training data. Fine-tuning a model pre-trained on SE360-Base with SE360-HF alleviates artifact overfitting while preserving the model's core generative abilities.

Model	Consistency ↑	Quality ↑	Plausibility ↑
PaintByInpaint	2.04 (14.5%)	2.34 (15.6%)	2.44 (16.1%)
Omnigen	2.11 (15.0%)	2.19 (14.6%)	2.34 (15.4%)
Flux.1 Kontext dev	2.91 (20.7%)	3.44 (22.9%)	3.11 (20.5%)
Step1X-Edit	2.50 (17.8%)	2.94 (19.6%)	2.90 (19.1%)
SE360 (Ours)	**4.47 (31.9%)**	**4.09 (27.3%)**	**4.35 (28.7%)**

Table 4: User Study results for object addition.

User Study

We conducted a comprehensive user study to evaluate our method on the task of instruction-based object addition. In the study, 22 participants compared results from SE360 against four top-performing models across 30 distinct samples. Evaluations were performed on a 5-point Likert scale, assessing three key metrics: Consistency (alignment between the generated object and the instruction), Quality (visual fidelity of the added object), and Plausibility (spatial and perspective coherence). As in Table 4, SE360 outperforms all competing methods. These results confirm that users perceive our generated objects as more faithful to the given instructions and visually superior.

Conclusion

We presented SE360, a semantic editing framework for 360° panoramas that resolves plausible and discontinuity issues from existing methods. Its core is a coarse-to-fine data pipeline ensuring object integrity and realism via hierarchical analysis and two-stage refinement, which also mitigates artifact overfitting. The resulting high-quality dataset enables training of a geometry-aware Transformer diffusion model. Experimental results and user studies demonstrate that SE360 outperforms prior approaches in visual quality, semantic accuracy, and geometric plausibility.

Limitations. Despite impressive performance, our method still has limitations. First, in reference-based tasks, the output of SE360 may bring background elements from the reference image to the result. Second, while our data generation pipeline incorporates multi-stage filtering, it may still introduce occasional errors.

Acknowledgments

This work was supported by the Marsden Fund Council managed by the Royal Society of New Zealand (No. MFP-20-VUW-180).

References

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 18392–18402.

Canberk, A.; Bondarenko, M.; Ozguroglu, E.; Liu, R.; and Vondrick, C. 2024. Erasedraw: Learning to insert objects by erasing them from images. In *European Conference on Computer Vision*, 144–160. Springer.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Niessner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*.

Chen, X.; Zhang, Z.; Zhang, H.; Zhou, Y.; Kim, S. Y.; Liu, Q.; Li, Y.; Zhang, J.; Zhao, N.; Wang, Y.; et al. 2025. Unireal: Universal image generation and editing via learning real-world dynamics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 12501–12511.

Feng, M.; Liu, J.; Cui, M.; and Xie, X. 2023. Diffusion360: Seamless 360 Degree Panoramic Image Generation based on Diffusion Models. *arXiv preprint arXiv:2311.13141*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Hore, A.; and Ziou, D. 2010. Image quality metrics: PSNR vs. SSIM. In *2010 20th international conference on pattern recognition*, 2366–2369.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. *arXiv preprint arXiv:2106.09685*.

Hui, M.; Yang, S.; Zhao, B.; Shi, Y.; Wang, H.; Wang, P.; Zhou, Y.; and Xie, C. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990*.

Kalischek, N.; Oechsle, M.; Manhardt, F.; Henzler, P.; Schindler, K.; and Tombari, F. 2025. Cubediff: Repurposing diffusion-based image models for panorama generation. In *The Thirteenth International Conference on Learning Representations*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4015–4026.

Labs, B. F.; Batifol, S.; Blattmann, A.; Boesel, F.; Consul, S.; Diagne, C.; Dockhorn, T.; English, J.; English, Z.; Esser, P.; et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742*.

Li, F.; Zhang, H.; Sun, P.; Zou, X.; Liu, S.; Yang, J.; Li, C.; Zhang, L.; and Gao, J. 2023. Semantic-SAM: Segment and Recognize Anything at Any Granularity. *arXiv preprint arXiv:2307.04767*.

Lipman, Y.; Chen, R. T. Q.; Ben-Hamu, H.; Nickel, M.; and Le, M. 2023. Flow Matching for Generative Modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Liu, S.; Han, Y.; Xing, P.; Yin, F.; Wang, R.; Cheng, W.; Liao, J.; Wang, Y.; Fu, H.; Han, C.; et al. 2025. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2024. Grounding DINO: Marrying DINO with Grounded Pre-training for Open-Set Object Detection. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLVII*, volume 15105 of *Lecture Notes in Computer Science*, 38–55.

Lu, Z.; Hu, K.; Wang, C.; Bai, L.; and Wang, Z. 2024. Autoregressive omni-aware outpainting for open-vocabulary 360-degree image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 14211–14219.

Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.; and Ermon, S. 2022. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*.

Oh, C.; Cho, W.; Chae, Y.; Park, D.; Wang, L.; and Yoon, K.-J. 2022. Bips: Bi-modal indoor panorama synthesis via residual depth-aided adversarial learning. In *European Conference on Computer Vision*, 352–371.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

Peebles, W.; and Xie, S. 2023. Scalable Diffusion Models with Transformers. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 4172–4182.

Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2024. SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis. In *The Twelfth International Conference on*

Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; et al. 2024. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2149–2159.

Tang, S.; Zhang, F.; Chen, J.; Wang, P.; and Furukawa, Y. 2023. MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Wasserman, N.; Rotstein, N.; Ganz, R.; and Kimmel, R. 2025. Paint by Inpaint: Learning to Add Image Objects by Removing Them First. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, 18313–18324.

Wu, T.; Zheng, C.; and Cham, T. 2023. IPO-LDM: Depth-aided 360-degree Indoor RGB Panorama Outpainting via Latent Diffusion Model. *arXiv preprint arXiv:2307.03177*.

Wu, T.; Zheng, C.; and Cham, T. 2024. PanoDiffusion: 360-degree Panorama Outpainting via Diffusion. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Xiao, B.; Wu, H.; Xu, W.; Dai, X.; Hu, H.; Lu, Y.; Zeng, M.; Liu, C.; and Yuan, L. 2024. Florence-2: Advancing a Unified Representation for a Variety of Vision Tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 4818–4829.

Xiao, S.; Wang, Y.; Zhou, J.; Yuan, H.; Xing, X.; Yan, R.; Li, C.; Wang, S.; Huang, T.; and Liu, Z. 2025. Omnigen: Unified image generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 13294–13304.

Xu, Y.; Kong, J.; Wang, J.; Pan, X.; Lin, B.; and Liu, Q. 2025. Insightedit: Towards better instruction following for image editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2694–2703.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yang, L.; Duan, H.; Zhu, Y.; Liu, X.; Liu, L.; Xu, Z.; Ma, G.; Min, X.; Zhai, G.; and Callet, P. L. 2025b. Omni^2: Unifying Omnidirectional Image Generation and Editing in an Omni Model. *arXiv preprint arXiv:2504.11379*.

Zhang, C.; Wu, Q.; Gambardella, C. C.; Huang, X.; Phung, D.; Ouyang, W.; and Cai, J. 2024. Taming stable diffusion for text to 360 {\deg} panorama image generation. *arXiv preprint arXiv:2404.07949*.

Zhang, K.; Mo, L.; Chen, W.; Sun, H.; and Su, Y. 2023. MagicBrush: A Manually Annotated Dataset for Instruction-Guided Image Editing. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Y.; Song, S.; Tan, P.; and Xiao, J. 2014. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European conference on computer vision*, 668–686.

Zhang, Y.; Zhang, F.; Lai, Y.; and Zhu, Z. 2021. Efficient propagation of sparse edits on 360○ panoramas. *Comput. Graph.*, 96: 61–70.

Zhang, Y.; Zhang, F.; Zhu, Z.; Wang, L.; and Jin, Y. 2022. Fast Edit Propagation for 360 Degree Panoramas Using Function Interpolation. *IEEE Access*, 10: 43882–43894.

Zhang, Z.; Xie, J.; Lu, Y.; Yang, Z.; and Yang, Y. 2025. In-Context Edit: Enabling Instructional Image Editing with In-Context Generation in Large Scale Diffusion Transformer. *arXiv preprint arXiv:2504.20690*.

Zhao, H.; Ma, X. S.; Chen, L.; Si, S.; Wu, R.; An, K.; Yu, P.; Zhang, M.; Li, Q.; and Chang, B. 2024. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems*, 37: 3058–3093.

Zheng, J.; Zhang, J.; Li, J.; Tang, R.; Gao, S.; and Zhou, Z. 2020. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, 519–535. Springer.