EditBridge: Towards Faithful and Efficient Ultra-High-Resolution Image Editing

Jiayi Song[1,2], Shijie Huang[2], Fangtai Wu[2], Yubo Huang[2], Zhenxiong Tan[3], Songhua Liu[1]*, Jiaming Liu[2]†, Ruihua Huang[2]
[1]School of Artificial Intelligence, Shanghai Jiao Tong University
[2]Qwen Business Unit of Alibaba
[3]National University of Singapore
liusonghua@sjtu.edu.cn

Abstract

High-resolution image editing is increasingly demanded in professional workflows, yet existing diffusion-based models remain constrained to resolutions below 1K due to quadratic attention complexity and prohibitive memory requirements. A prevalent workaround employs a two-stage pipeline: editing at low resolution followed by independent super-resolution. However, this approach suffers from two critical issues: *information divergence*, where hallucinated details contradict the original high-resolution (HR) source, and *texture degradation*, manifesting as over-smoothed or over-sharpened artifacts. We propose **EditBridge**, a diffusion bridge framework for efficient ultra high-resolution editing. Unlike conventional diffusion that regenerates from noise, we formulate refinement as structured data-to-data translation from the low-resolution (LR) edited result to its HR counterpart, explicitly conditioned on the original HR source to preserve authentic details. To efficiently incorporate HR source guidance, we introduce a prior-guided block-wise sparse attention mechanism that exploits semantic correspondence from first-stage editing to constrain cross-image interactions to spatially aligned regions, significantly reducing computational overhead. Extensive experiments demonstrate that EditBridge achieves high-fidelity editing with superior perceptual quality at resolutions up to 4K, delivering 3.6–8.4$\times$ speedup at 2K and enabling practical 4K editing in 61 seconds. Project page: https://editbridge.github.io/.

1 Introduction

High-resolution visual content generation has become a central topic in modern generative modeling, spanning both image [1–3] and video [4–6] synthesis. However, due to the quadratic computational complexity of attention mechanisms and substantial memory requirements [7–9], most existing image editing models [10–14] remain constrained to resolutions no higher than 1K (1024×1024) during inference. To obtain higher-resolution outputs, a common paradigm relies on a two-stage pipeline: performing the edit at a lower resolution and subsequently employing an independent super-resolution(SR) model for upscaling [15].

While super-resolution increases the output resolution, it tends to hallucinate high-frequency textures due to the absence of high-resolution(HR) source guidance. This introduces two primary challenges: (1) *Information Divergence*: the hallucinated details inevitably diverge from the original HR source content, contradicting the fundamental goal of faithful editing—as shown in Fig. 1(c) left, where the facial details produced by SR are entirely inconsistent with the HR source; and (2) *Texture Degradation*: the generation of over-smoothed or over-sharpened artifacts that degrade visual fidelity, as illustrated in Fig. 1(c) right.

*Corresponding Author.
†Project Lead.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

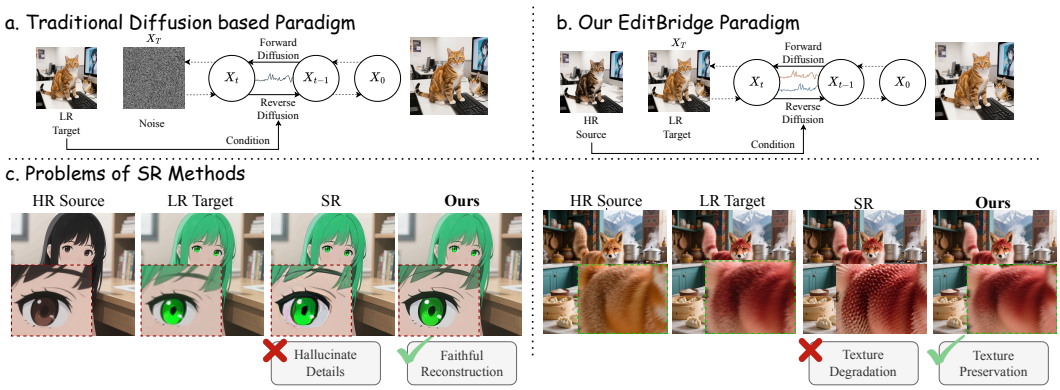

Figure 1: Motivation of the proposed method. (a) and (b) compare the standard high-resolution image editing paradigm with our approach. (c) analyzes the inherent limitations of existing methods.

A straightforward alternative would be to train an image editing model natively at high resolutions. However, such an approach demands exorbitant computational resources and large-scale, high-resolution editing datasets, making it prohibitively expensive in practice. Moreover, even with a trained model, high-resolution inference remains computationally intensive due to the quadratic complexity of attention mechanisms, resulting in severe efficiency bottlenecks. This naturally raises a key question: *Can we achieve high-resolution image editing in a more faithful and efficient manner?*

Motivated by this question, we propose **EditBridge**, a prior-guided diffusion bridge framework for ultra-high-resolution image editing. Since the core limitation of SR methods lies in the absence of HR source guidance, our key idea is to explicitly leverage the original high-resolution source image as conditional guidance during refinement. EditBridge addresses this through two core designs: (1) a diffusion bridge formulation that models the structured refinement from LR edited images to HR outputs, and (2) a prior-guided sparse attention mechanism that efficiently incorporates HR source information without prohibitive computational overhead.

Specifically, we formulate high-resolution editing as a refinement process that transforms the low-resolution (LR) edited image into its high-resolution (HR) counterpart. Unlike generic upsampling that operates solely on the LR input, our goal is to synthesize fine-grained textures guided by the original HR source while preserving consistency with the edited semantic content.

A key observation is that this task differs fundamentally from standard conditional generation: we have a concrete LR edited image as the starting point, not random noise. This motivates us to adopt diffusion bridges [16, 17], which model stochastic transitions between two data distributions rather than from noise to data. Compared to conventional diffusion models that must "re-generate" the entire image from scratch, diffusion bridges preserve the structural information from the LR input throughout the refinement trajectory, enabling more faithful and efficient high-resolution synthesis.

While the diffusion bridge provides a principled formulation, a key challenge remains: how to efficiently incorporate the HR source image as guidance. A naive approach would concatenate contextual tokens directly within the Diffusion Transformer (DiT) [18], but this incurs prohibitive computational overhead due to token expansion and the quadratic complexity of global attention. We observe that high-resolution editing does not necessitate exhaustive all-to-all attention between target queries and source keys. Instead, interactions should be governed by semantic relevance: for unedited regions, the model primarily requires high-fidelity preservation, which can be achieved by attending strictly to semantically corresponding anchors in the HR source image; for edited regions, the model must synthesize new high-frequency details by assimilating local contextual cues.

Building upon this insight, we propose a prior-guided block-wise sparse attention mechanism that explicitly routes information based on spatial semantic alignment. During the LR source-to-target editing stage, our approach extracts semantic correspondence priors from the attention maps. These priors serve as a spatial roadmap, identifying the precise regions relevant to the modification and locating their corresponding anchors in the HR source. Guided by these cues, the sparse attention mechanism constrains the HR receptive field to aligned areas, effectively pruning redundant interactions with irrelevant tokens. Consequently, EditBridge achieves a superior balance between computational efficiency and high-fidelity texture synthesis.

In summary, the main contributions of this paper are:

- We present a DiT-based framework for high-resolution image editing that achieves both computational efficiency and faithful detail preservation.
- We propose a diffusion-bridge-based refinement stage equipped with a prior-guided block-wise sparse attention mechanism, facilitating semantically selective cross-image interaction.
- Our method achieves state-of-the-art high-resolution editing performance while drastically reducing computational costs, yielding up to a $2.2\times$ speedup compared to conventional diffusion-based approaches.

2 Related Work

Instruction Based Image Editing. The rapid advancement of diffusion models has significantly advanced instruction-based image editing. Early works such as InstructPix2Pix [19] and SuTI [20] introduced the paradigm of editing images directly from instructions rather than descriptive captions. MagicBrush [21] further improved editing capabilities by introducing high-quality, diverse training datasets. Beyond dataset-driven improvements, subsequent approaches [22–25] integrated Multimodal Large Language Models (MLLMs) [26–28] with diffusion models [11, 12] to enhance instruction understanding and controllability. With the introduction of Diffusion Transformers (DiT) [18], research efforts have shifted toward developing unified DiT-based foundation models for conditional image generation [29, 30] and editing [10, 29, 31], injecting conditional signals directly into attention layers for fine-grained guidance. Despite this flexibility, these approaches incur substantial computational overhead due to dense global attention operations, a bottleneck that becomes prohibitive at high resolutions.

High-Resolution Visual Generation. Recent works have explored high-resolution visual generation [2, 32, 33]. Early approaches [5, 34] primarily rely on U-Net [35] architectures. With the rise of foundation diffusion models, DiT [18] has become the dominant backbone, as its attention mechanism excels at modeling complex token-wise dependencies. To improve inference efficiency at high resolutions, several acceleration methods have emerged. For instance, I-Max [36] projects high-resolution flows into a lower-resolution latent space to reduce computational costs. HiFlow [1] introduces a model-agnostic acceleration framework for scalable generation, while CLEAR [37] proposes a linearization strategy tailored to mitigate the quadratic complexity of global attention. URAE [38] enables resource-efficient adaptation for ultra-high-resolution image generation.

However, these methods are primarily designed for pure image generation tasks. In contrast, image editing strictly requires fine-grained consistency between the source and target images, rendering generation-oriented acceleration strategies sub-optimal for editing scenarios. To bridge this gap, works like MobilePicasso [39] introduce hybrid pipelines with hallucination-aware training and adaptive tiling. Similarly, ScaleEdit [40] proposes a training-free transfer function for high-resolution editing. However, these methods are built upon U-Net architectures and cannot be directly applied to the increasingly prevalent DiT-based editing frameworks. In this work, we address this gap by proposing EditBridge, a diffusion bridge framework specifically designed for efficient and faithful high-resolution editing within the DiT paradigm.

3 Preliminary: Diffusion Bridge

Diffusion Bridge. Standard diffusion models learn to transport samples from an uninformative prior (e.g., Gaussian noise) to the data distribution. In contrast, *diffusion bridges* (also known as Brownian bridges) model stochastic paths between two structured data domains, making them naturally suited for data-to-data translation tasks. Formally, given paired endpoints (x_0, x_1) from source and target domains, the Brownian bridge defines a stochastic interpolation path. The intermediate state X_t at time $t \in [0, 1]$ follows:

$$X_t \mid (x_0, x_1) \sim \mathcal{N}\left((1-t)x_0 + tx_1,\ t(1-t)I\right). \tag{1}$$

Intuitively, X_t is a noisy interpolation between the two endpoints, with the noise level peaking at $t = 0.5$ and vanishing at the boundaries. The corresponding instantaneous velocity that drives this transition is:

$$u_t(X_t \mid x_0, x_1) = \frac{x_1 - X_t}{1-t}. \tag{2}$$

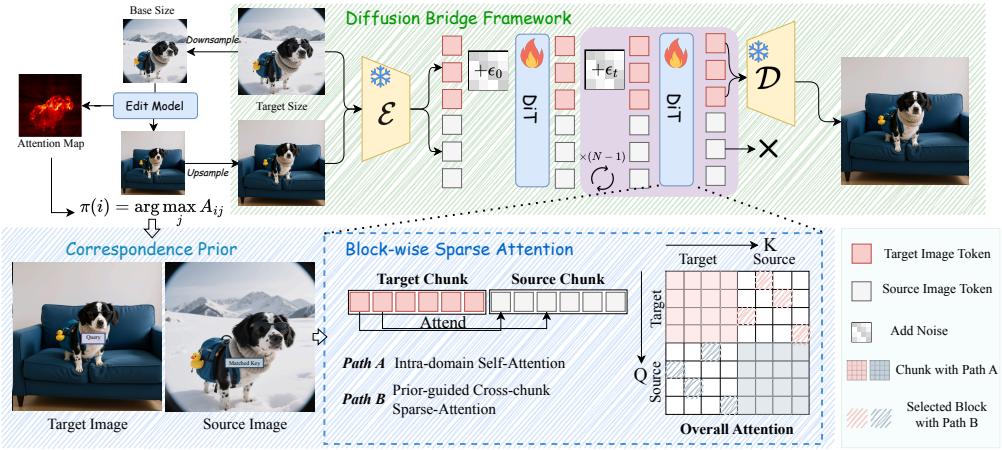

Figure 2: **Overview of our proposed EditBridge.** The upper section illustrates the inference process of the diffusion bridge, which transports the low-resolution edit to its high-resolution counterpart. The lower section details the construction of the correspondence prior and the proposed Prior-Guided Sparse Attention (PG-BSA) mechanism.

In practice, we parameterize a neural network v_θ to approximate this velocity field via the matching objective:

$$\mathcal{L}(\theta) = \mathbb{E}_{x_0, x_1, t, X_t} \left\| v_\theta(X_t, t) - u_t(X_t \mid x_0, x_1) \right\|^2. \tag{3}$$

Compared to standard diffusion that reconstructs signals from noise, the bridge formulation preserves structural information from x_0 throughout the trajectory, enabling efficient and faithful transformations.

Problem Formulation. High-resolution (HR) image editing aims to map an HR source x_s^{HR} and instruction c to an edited HR output x_t^{HR}. To bypass the computational burden of direct HR optimization, we adopt a coarse-to-fine paradigm:

1. **LR Editing:** $x_t^{LR} = \mathcal{G}(x_s^{LR}, c)$, where $\mathcal{G}$ is a pre-trained low-resolution (LR) model and x_s^{LR} is the downsampled source.
2. **HR Refinement:** $x_t^{HR} = \mathcal{H}(\tilde{x}_t^{HR}, x_s^{HR})$, where $\tilde{x}_t^{HR}$ is the upsampled x_t^{LR}.

Our objective is to learn the refinement mapping $\mathcal{H}$. Notably, this HR refinement task naturally fits the diffusion bridge formulation: the upsampled LR edit $\tilde{x}_t^{HR}$ serves as the source endpoint x_0, and the target HR output x_t^{HR} serves as the target endpoint x_1.

4 Method

Based on the formulation, we present two key components of our framework: Bridge-based HR Refinement (Sec. 4.1) and Prior-Guided Sparse Attention (Sec. 4.2). An overview is shown in Fig. 2.

4.1 Bridge-based High-Resolution Refinement

As illustrated in Fig. 1b, based on the diffusion bridge formulation introduced in Sec. 3, we instantiate a conditional bridge model tailored for high-resolution refinement. In our framework, the source endpoint x_0 is defined as upsampled low-resolution edited result $\tilde{x}_t^{HR}$. The target endpoint x_1 corresponds to the desired high-resolution edited image x_t^{HR}. The objective of the second stage is to model a probability path that characterizes the transition from $\tilde{x}_t^{HR}$ to x_t^{HR}.

Unlike unconditional generative modeling, our refinement process is conditioned on the high-resolution source image x_s^{HR}, which serves as a high-fidelity reference to restore details lost during the initial LR editing stage. Specifically, the bridge model learns a conditional velocity field:

$$v_\theta(X_t, t \mid x_s^{HR}) \approx u_t(X_t \mid \tilde{x}_t^{HR}, x_t^{HR}), \tag{4}$$

where the network progressively transforms the noisy state X_t into a high-fidelity result by integrating information from x_s^{HR}.

We implement the bridge model using a Diffusion Transformer (DiT) [18] backbone. In Fig. 2, at each timestep t, the model receives the current state X_t along with contextual features extracted from x_s^{HR}. The network predicts the velocity term according to the matching objective in Eq. (3). The final refined image is obtained by integrating the learned probability path from $t = 0$ to $t = 1$.

Compared to generating high-resolution images from pure noise (Fig. 1a), the bridge formulation shortens the generative trajectory. Furthermore, explicitly conditioning on the uncorrupted x_s^{HR} ensures the faithful recovery of fine-grained details and high-frequency textures.

4.2 Prior-Guided Block-wise Sparse Attention

In the context of high-resolution image editing, where the source image is typically integrated as a conditional context via long-sequence concatenation, standard attention mechanisms exhibit a prohibitive quadratic computational complexity $\mathcal{O}(N^2)$ relative to the sequence length N. As the token count N scales quadratically with the image resolution (e.g., 2K or 4K), this leads to unsustainable memory footprints and severe latency bottlenecks. To mitigate these challenges, we introduce **Prior-Guided Block-wise Sparse Attention (PG-BSA)**, a sparsity strategy specifically tailored for high-resolution refinement. Unlike generic sparse attention, PG-BSA leverages the intrinsic semantic alignment between the source and target domains to prune redundant interactions, ensuring both computational efficiency and reconstruction fidelity.

4.2.1 Attention-Induced Correspondence Prior

The cornerstone of our method is the utilization of semantic knowledge from the first-stage low-resolution editing. During the initial stage, the model establishes a coarse-grained mapping between the high-resolution source and the edited target. We capture this relationship through the cross-domain attention affinity maps.

Let $Q_{LR} \in \mathbb{R}^{n_t \times d}$ and $K_{LR} \in \mathbb{R}^{n_s \times d}$ denote the queries from the target domain and keys from the source domain in the first stage, respectively. The attention matrix A is computed as:

$$A = \text{Softmax}\left(\frac{Q_{LR}K_{LR}^{\top}}{\sqrt{d}}\right) \in \mathbb{R}^{n_t \times n_s}, \tag{5}$$

where n_t and n_s represent the number of tokens at the native low resolution. To derive a robust spatial prior, we identify the source anchor that yields the maximum response for each target query, effectively establishing a hard correspondence roadmap:

$$\pi(i) = \arg\max_j A_{ij}. \tag{6}$$

This mapping $\pi(\cdot)$ serves as a dense guidance field. Since the second-stage refinement operates at a significantly higher resolution, we apply a coordinate-based upscaling to $\hat{\pi}(\cdot)$. Specifically, with an upscaling factor of $S = 2$, each low-resolution token j at spatial coordinate (x, y) is mapped to a 2×2 cluster of high-resolution tokens. Formally, for any high-resolution token k located at position (x', y'), its prior is assigned as $\hat{\pi}(k) = \pi(\lfloor x'/S \rfloor, \lfloor y'/S \rfloor)$. This nearest-neighbor expansion strategy ensures that the semantic guidance remains spatially aligned across scales, providing a robust and piece-wise constant search space for the subsequent prior-guided sparse attention.

4.2.2 Prior-Guided Sparsification Mechanism

In the high-resolution refinement stage, the input sequence is structured as a concatenated multi-modal embedding: [Text | Target Image | Source Image]. To implement the prior-guided sparsification while preserving global semantic consistency, we decouple the attention mechanism into two distinct computational paths:

- **Intra-domain Self-attention:** To maintain internal structural coherence, we apply localized self-attention independently within the Text ($\mathcal{A}_{txt}$), Target ($\mathcal{A}_{tgt}$), and Source ($\mathcal{A}_{src}$) segments. This domain-isolated processing ensures that each modality focuses on its intrinsic dependencies.

Table 1: Quantitative comparison results. ↑/↓ indicate whether higher/lower values are better. Bold numbers represent the best results.

Method	Haar ↑	mPSNR ↑	mSSIM ↑	mMSE ↓	mLPIPS ↓	Time (s) ↓
1K Resolution						
Direct Inference	0.3973	16.8078	0.7539	0.0549	0.4116	20.00
DiT4SR [41]	0.4127	14.5031	0.7092	0.0554	0.4818	28.47
DiT-SR [15]	0.4523	16.7152	0.7955	0.0430	0.4403	2.44
PiSA-SR [42]	0.4247	15.2115	0.7481	0.0486	0.4492	**0.30**
TSD-SR [43]	0.4127	15.3864	0.7531	0.0516	0.4628	0.42
HiFlow [1]	0.3012	11.6964	0.6489	0.1323	0.6012	5.12
Ours	**0.4992**	**18.6536**	**0.8686**	**0.0369**	**0.3829**	0.84
2K Resolution						
Direct Inference	0.2254	13.5262	0.7491	0.0880	0.6700	82.14
DiT4SR [41]	0.3806	15.1298	0.7468	0.0516	0.5348	113.32
DiT-SR [15]	0.4210	17.2430	0.8265	0.0423	0.4738	14.82
PiSA-SR [42]	0.4046	16.2433	0.8061	0.0446	0.4702	20.22
TSD-SR [43]	0.4046	16.2021	0.7952	0.0467	0.4910	34.20
HiFlow [1]	0.3144	13.9314	0.7377	0.0759	0.5113	33.11
Ours	**0.4673**	**18.6366**	**0.8712**	**0.0377**	**0.3903**	**4.08**
4K Resolution						
Direct Inference	0.2134	10.7086	0.5411	0.1281	0.7774	695.23
DiT4SR [41]	0.4002	20.2771	0.7535	0.0133	0.3510	533.83
DiT-SR [15]	0.4267	16.8615	0.6818	0.0415	0.5436	119.68
PiSA-SR [42]	0.4267	21.6200	0.7974	0.0108	0.2687	111.42
TSD-SR [43]	0.4124	22.2766	0.8035	0.0074	0.2399	108.72
HiFlow [1]	0.3307	14.4400	0.6719	0.0664	0.6038	129.30
Ours	**0.4279**	**24.1380**	**0.8347**	**0.0049**	**0.2112**	**61.10**

- **Prior-guided Cross-domain Attention:** To facilitate information flow, interactions between Target and Source tokens are strictly constrained by the upscaled prior $\hat{\pi}(\cdot)$. For a target query token q_j at position j, the attention is computed over a sparse set of source keys:

$$\text{Attn}(q_j, K, V) = \sum_{k \in \mathcal{S}_j} \text{Softmax}\left(\frac{q_j k_k^\top}{\sqrt{d}}\right) v_k, \tag{7}$$

  where the receptive field $\mathcal{S}_j$ is defined as a $k \times k$ local window centered at the semantically aligned anchor $\hat{\pi}(j)$. This window size compensates for potential spatial misalignments in the upscaled prior, with larger windows at higher resolutions accounting for increased discretization errors, while maintaining $k^2 \ll N$ to ensure computational efficiency.

By restricting the interaction to a local neighborhood $\mathcal{S}_j$, we reduce the complexity of cross-domain attention from $\mathcal{O}(N^2)$ to $\mathcal{O}(N \cdot k^2)$. This formulation ensures that each target patch synthesizes high-frequency details by consulting only its most relevant source counterpart, significantly reducing the memory footprint during inference.

5 Experiments

5.1 Implementation Details

Training Settings. We build our proposed method upon Qwen-Image-Edit [12], a state-of-the-art image editing model. To realize the sparse attention mechanism, we adopt the VMoBA framework [44] and leverage FlashAttention [45] to optimize the underlying kernel operations. During training, we employ the Prodigy optimizer [46] with a learning rate of 1.0, a weight decay of 0.01, enabling a



Figure 3: Qualitative comparison at 1K resolution. Our method better preserves fidelity while enhancing visual clarity.

safeguard warmup. Additionally, we apply Low-Rank Adaptation (LoRA) [47] to all linear layers, setting the rank parameter $r = 128$ and $\alpha = 128$. All experiments are conducted on 8 NVIDIA H800 GPUs (80GB) with a per-GPU batch size of 1.

Dataset. To construct our training corpus, we curate high-resolution source images from publicly available datasets, including Aesthetic-4k [48] and Aesthetic-Train-V2 [49]. For images satisfying the resolution criteria, we employ a center-cropping strategy to crop them to the target resolutions. Subsequently, we utilize Gemini 3 [50] to automatically generate detailed editing instructions. The corresponding target images are synthesized via Nano Banana Pro [51]. The final dataset consists of $5,000$ image pairs at resolutions of 1K and 2K, supplemented by $1,500$ image pairs at 4K resolution.

Evaluation Metrics. To comprehensively evaluate the performance of our method, we follow the evaluation protocol established by ScaleEdit [40]. Specifically, we employ HaarPSI [52] to directly assess the perceptual similarity between the source and the synthesized images. Furthermore, we incorporate several region-aware metrics to provide a granular analysis of the editing quality. Utilizing a binary mask to bifurcate the image, we compute **M-PSNR**, **M-SSIM**, and **M-MSE** over the *unedited regions* to quantify the fidelity. Conversely, **M-LPIPS** is restricted to the *edited regions* to evaluate the generative fidelity and perceptual quality of the specific modifications. This dual-region evaluation ensures that our model not only achieves high-quality refinement but also maintains strict consistency in non-target areas.

5.2 Main Comparison

We benchmark our method against several state-of-the-art diffusion-based super-resolution (SR) models, including **DiT-SR** [15], **DiT4SR** [41], **PiSA-SR** [42], and **TSD-SR** [43]. Additionally, we compare our approach with high-resolution generative models **HiFlow** [1] to underscore the difference between generic upscaling and fidelity-preserving editing. A Direct Inference baseline(using the base model without modifications) is also included.

Tab. 1 presents the quantitative evaluation of our method at 1k, 2k, and 4k resolutions, respectively. As shown, our approach consistently outperforms the baseline models across all resolution settings. Furthermore, we evaluated the inference time of our method, which explicitly accounts for the computational overhead of the prior estimation process. Our method achieves 3.6–8.4$\times$ speedup at 2K compared to diffusion-based SR methods, requiring only 4.08 seconds. At 4K, while the

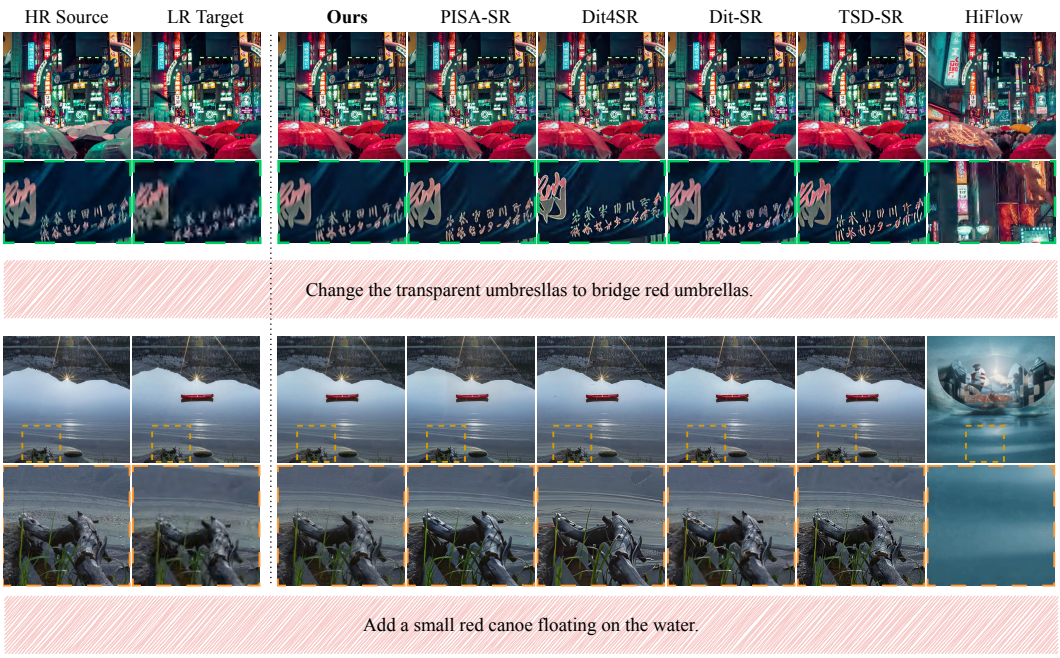

Figure 4: Qualitative comparison at 2K resolution. Our method better preserves fidelity while enhancing visual clarity.

speedup over SR methods is 1.8–2.1×, our method achieves 11× acceleration compared to direct high-resolution inference, completing 4K editing in 61 seconds.

We further provide a qualitative evaluation of our results. As illustrated in Fig. 3, our method achieves exceptional visual clarity with richer fine-grained details compared to the low-resolution targets, while strictly maintaining the highest structural and semantic consistency with the high-resolution source. Taking the first example in Fig. 3, conventional super-resolution methods suffer from excessive sharpening artifacts, leading to unnatural texture degradation and severely compromised visual fidelity. In the second example, the baseline models frequently hallucinate arbitrary details, causing the generated door to diverge significantly from the original HR source. Furthermore, direct high-resolution visual generation methods tend to alter the overall global semantics of the image. As illustrated in Fig. 4, similar advantages and consistent visual improvements are also observed under the 2k resolution settings.

5.3 Ablation Study

Sparse Attention vs. Full Attention. We further investigate the impact of different attention mechanisms at 1K and 2K resolutions to validate our design. As shown in Tab. 2, Full Attention achieves slightly higher quantitative scores in pixel-level reconstruction metrics (e.g., PSNR, HaarPSI). However, this quantitative superiority can be deceptive in the context of local editing. Since full attention computes global interactions without spatial constraints, the model tends to indiscriminately over-attend to semantically irrelevant tokens from the high-resolution source image. This contextual bleeding introduces noticeable ghosting and structural artifacts in the newly synthesized areas, as illustrated in Fig. 5.

Table 2: Performance comparison between our PG-BSA and full attention across 1K and 2K settings. ↑/↓ indicate whether higher/lower values are better. Bold numbers represent the best results within each resolution block.

Setting	Method	HaarPSI↑	mPSNR↑	mSSIM↑	mMSE↓	mLPIPS↓	Time (s)↓
1K	Ours	0.499	18.653	**0.868**	0.036	**0.382**	0.84
	FullAttention	**0.539**	**19.080**	0.857	**0.030**	0.599	**0.73**
2K	Ours	0.467	**18.636**	**0.871**	**0.037**	**0.390**	**4.08**
	FullAttention	**0.505**	17.918	0.705	0.041	0.425	4.71



Figure 5: **Ablation of Attention Sparsity.** Full attention leads to noticeable source-induced artifacts, whereas our sparse attention maintains superior visual fidelity and alignment.

Moreover, the quadratic computational complexity of global attention imposes severe memory footprints and latency bottlenecks, rendering it computationally prohibitive for high-resolution inference. In contrast, our proposed sparse attention explicitly constrains cross-image interactions using semantic registration priors. By routing queries strictly to their semantically aligned anchors, it effectively prunes noisy global contexts. This design not only curtails the computational complexity to a linear scale—yielding the substantial speedups discussed earlier—but also strictly preserves local detail fidelity without unwanted interference from unedited regions.

Impact of Inference Steps. We evaluate the model performance across varying numbers of inference steps (N), as illustrated in Tab. 3. Interestingly, empirical results indicate that single-step inference ($N = 1$) achieves the optimal trade-off between computational efficiency and perceptual quality. Unlike multi-step sampling, which may accumulate quantization errors at extreme resolutions, single-step inference yields the highest fidelity to the source's high-frequency structures while maintaining minimal computational cost (e.g., a mere 0.84s for 1K images). Furthermore, as illustrated in Fig. 6, the setting $N = 1$ enables the model to exhibit the richest fine-grained details.

Table 3: Quantitative evaluation of varying inference steps. ↑/↓ indicate whether higher/lower values are better. Bold numbers represent the best results.

Method	HaarPSI↑	mPSNR↑	mSSIM↑	mMSE↓	mLPIPS↓	Time(s)↓
5 steps	0.4131	15.7269	0.7763	0.0474	0.4739	4.03
10 steps	0.4020	15.4190	0.7636	0.0494	0.4662	8.10
Ours (1 step)	**0.4991**	**18.6535**	**0.8685**	**0.0368**	**0.3829**	**0.84**

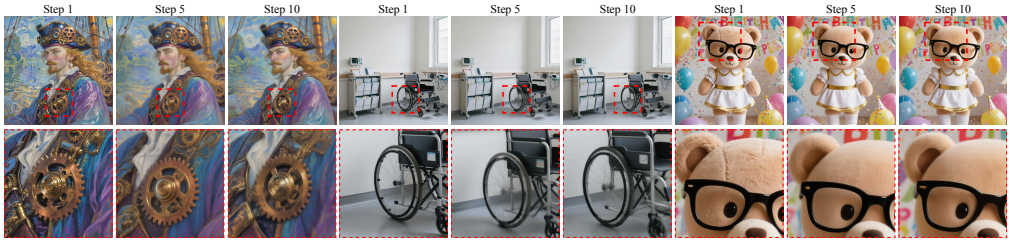

Figure 6: **Impact of sampling steps.** Perceptual quality and high-frequency details are maximized at $T = 1$. Increasing the number of inference steps does not yield significant visual gains, further validating the efficiency of our refinement framework.

6 Conclusion

In this paper, we introduced **EditBridge**, a novel prior-guided diffusion bridge framework tailored for ultra-high-resolution image editing. Our approach formulates high-resolution refinement as a continuous, data-to-data translation process. By leveraging the original high-resolution source image as conditional guidance, EditBridge overcomes the critical limitations of conventional two-stage super-resolution pipelines, namely information divergence and texture hallucination, thereby ensuring faithful detail preservation. Furthermore, to address the prohibitive computational costs associated with high-resolution processing, we proposed a task-specific, prior-guided sparse attention mechanism. By dynamically routing information based on spatial semantic alignment extracted from the low-resolution editing stage, our model selectively focuses only on relevant high-resolution anchors, effectively pruning redundant global interactions. Overall, our framework establishes a highly efficient and scalable paradigm for high-fidelity image editing.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grant No. 62606320).

References

[1] Jiazi Bu, Pengyang Ling, Yujie Zhou, Pan Zhang, Tong Wu, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Hiflow: Training-free high-resolution image generation with flow-aligned guidance. *arXiv preprint arXiv:2504.06232*, 2025.

[2] Tian Ye, Song Fei, and Lei Zhu. Ultraflux: Data-model co-design for high-quality native 4k text-to-image generation across diverse aspect ratios. *arXiv preprint arXiv:2511.18050*, 2025.

[3] Haonan Qiu, Shiwei Zhang, Yujie Wei, Ruihang Chu, Hangjie Yuan, Xiang Wang, Yingya Zhang, and Ziwei Liu. Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16893–16903, 2025.

[4] Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, and Sergey Tulyakov. Hierarchical patch diffusion models for high-resolution video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7569–7579, 2024.

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22563–22575, 2023.

[6] Yunfeng Wu, Jiayi Song, Zhenxiong Tan, Zihao He, and Songhua Liu. Freeswim: revisiting sliding-window attention mechanisms for training-free ultra-high-resolution video generation. In *European Conference on Computer Vision*, pages 76–93. Springer, 2026.

[7] Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster, 2023. URL `https://arxiv.org/abs/2210.09461`.

[8] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL `https://arxiv.org/abs/1706.03762`.

[9] Qingyan Wei, Guangzhao Li, Xiaobing Tu, Yinggui Wang, Xiantao Zhang, Jinkui Ren, Xiaohong Liu, and Linfeng Zhang. Step-opd: Rethinking output targets and internal dynamics in on-policy distillation for diffusion models. *arXiv preprint arXiv:2608.04887*, 2026.

[10] 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, 2025. URL `https://arxiv.org/abs/2506.15742`.

[11] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining, 2025. URL `https://arxiv.org/abs/2505.14683`.

[12] 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, 2025. URL `https://arxiv.org/abs/2508.02324`.

[13] Super Intelligence Team, Changhao Qiao, Chao Hui, Chen Li, Cunzheng Wang, Dejia Song, Jiale Zhang, Jing Li, Qiang Xiang, Runqi Wang, Shuang Sun, Wei Zhu, Xu Tang, Yao Hu, Yibo Chen, Yuhao Huang, Yuxuan Duan, Zhiyi Chen, and Ziyuan Guo. Firered-image-edit-1.0 technical report, 2026. URL `https://arxiv.org/abs/2602.13344`.

[14] Guangzhao Li, Qingyan Wei, Huayu Zheng, Yige Zheng, Chaoyang Zhang, Jie Yang, Yunan Ding, Yan Tai, Siqi Luo, and Xiaohong Liu. Insertfuse: A unified framework for multi-category reference-guided image insertion. *arXiv preprint arXiv:2608.06490*, 2026.

[15] Kun Cheng, Lei Yu, Zhijun Tu, Xiao He, Liyu Chen, Yong Guo, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Effective diffusion transformer architecture for image super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pages 2455–2463, 2025.

[16] Bo Li, Kaitao Xue, Bin Liu, and Yu-Kun Lai. Bbdm: Image-to-image translation with brownian bridge diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern Recognition*, pages 1952–1961, 2023.

[17] Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *The Twelfth International Conference on Learning Representations*.

[18] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[19] 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*, pages 18392–18402, 2023.

[20] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *Advances in Neural Information Processing Systems*, 36:30286–30305, 2023.

[21] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems*, 36:31428–31449, 2023.

[22] Yuzhou Huang, Liangbin Xie, Xintao Wang, Ziyang Yuan, Xiaodong Cun, Yixiao Ge, Jiantao Zhou, Chao Dong, Rui Huang, Ruimao Zhang, et al. Smartedit: Exploring complex instruction-based image editing with multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8362–8371, 2024.

[23] Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Yang, and Zhe Gan. Guiding instruction-based image editing via multimodal large language models. *arXiv preprint arXiv:2309.17102*, 2023.

[24] Shufan Li, Harkanwar Singh, and Aditya Grover. Instructany2pix: Flexible visual editing via multimodal instruction following. *arXiv preprint arXiv:2312.06738*, 2023.

[25] Yujia Hu, Songhua Liu, Zhenxiong Tan, Xingyi Yang, and Xinchao Wang. Image editing as programs with diffusion models. *Advances in Neural Information Processing Systems*, 38:18405–18436, 2026.

[26] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), November 2024. ISSN 2053-714X. doi: 10.1093/nsr/nwae403. URL `http://dx.doi.org/10.1093/nsr/nwae403`.

[27] 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, 2025. URL `https://arxiv.org/abs/2511.21631`.

[28] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL `https://arxiv.org/abs/2502.13923`.

[29] Yuxuan Zhang, Yirui Yuan, Yiren Song, Haofan Wang, and Jiaming Liu. Easycontrol: Adding efficient and flexible control for diffusion transformer, 2025. URL `https://arxiv.org/abs/2503.07027`.

[30] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer, 2025. URL `https://arxiv.org/abs/2411.15098`.

[31] Fangtai Wu, Hailong Guo, Shijie Huang, Jiayi Song, Yubo Huang, Mushui Liu, Zhao Wang, Yunlong Yu, Jiaming Liu, and Ruihua Huang. Collectionlora: Collecting 50 effects in 1 lora via multi-teacher on-policy distillation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2026.

[32] Ruonan Yu, Zigeng Chen, Zhenxiong Tan, Songhua Liu, and Xinchao Wang. Apertureattn: Native 4k video generation with image-only supervision. In *The Fortieth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=9pq44O4Vzq.

[33] Yunfeng Wu, Hongying Cheng, Zihao He, and Songhua Liu. Vibe: ultra-high-resolution video synthesis born from pure images. In *European Conference on Computer Vision*, pages 337–353. Springer, 2026.

[34] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[36] Ruoyi Du, Dongyang Liu, Le Zhuo, Qin Qi, Hongsheng Li, Zhanyu Ma, and Peng Gao. I-max: Maximize the resolution potential of pre-trained rectified flow transformers with projected flow. 2024.

[37] Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Clear: Conv-like linearization revs pre-trained diffusion transformers up. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

[38] Ruonan Yu, Songhua Liu, Zhenxiong Tan, and Xinchao Wang. Ultra-resolution adaptation with ease. *International Conference on Machine Learning*, 2025.

[39] Young D. Kwon, Abhinav Mehrotra, Malcolm Chadwick, Alberto Gil Ramos, and Sourav Bhattacharya. Efficient high-resolution image editing with hallucination-aware loss and adaptive tiling, 2025. URL https://arxiv.org/abs/2510.06295.

[40] Junsung Lee, Hyunsoo Lee, Yong Jae Lee, and Bohyung Han. Low-resolution editing is all you need for high-resolution editing, 2025. URL https://arxiv.org/abs/2511.19945.

[41] Zheng-Peng Duan, Jiawei Zhang, Xin Jin, Ziheng Zhang, Zheng Xiong, Dongqing Zou, Jimmy S Ren, Chunle Guo, and Chongyi Li. Dit4sr: Taming diffusion transformer for real-world image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18948–18958, 2025.

[42] Lingchen Sun, Rongyuan Wu, Zhiyuan Ma, Shuaizheng Liu, Qiaosi Yi, and Lei Zhang. Pixel-level and semantic-level adjustable super-resolution: A dual-lora approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2333–2343, June 2025.

[43] Linwei Dong, Qingnan Fan, Yihong Guo, Zhonghao Wang, Qi Zhang, Jinwei Chen, Yawei Luo, and Changqing Zou. Tsd-sr: One-step diffusion with target score distillation for real-world image super-resolution. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23174–23184, 2025.

[44] Jianzong Wu, Liang Hou, Haotian Yang, Xin Tao, Ye Tian, Pengfei Wan, Di Zhang, and Yunhai Tong. Vmoba: Mixture-of-block attention for video diffusion models, 2025. URL https://arxiv.org/abs/2506.23858.

[45] Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35: 16344–16359, 2022.

[46] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner, 2024. URL https://arxiv.org/abs/2306.06101.

[47] 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 *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[48] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23464–23473, 2025.

[49] Jinjin Zhang, Qiuyu Huang, Junjie Liu, Xiefan Guo, and Di Huang. Ultra-high-resolution image synthesis: Data, method and evaluation. *arXiv preprint arXiv:2506.01331*, 2025.

[50] Google. Gemini 3. https://aistudio.google.com/models/gemini-3, 2026. Accessed: March 3, 2026.

[51] Google DeepMind. Gemini image pro. https://deepmind.google/models/gemini-image/pro/, 2026. Accessed: March 3, 2026.

[52] Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61: 33–43, February 2018. ISSN 0923-5965. doi: 10.1016/j.image.2017.11.001. URL http://dx.doi.org/10.1016/j.image.2017.11.001.

[53] Yuan Wang, Yuhao Wan, Siming Zheng, Bo Li, Qibin Hou, and Peng-Tao Jiang. Trust but verify: Adaptive conditioning for reference-based diffusion super-resolution via implicit reference correlation modeling. In *The Fourteenth International Conference on Learning Representations*, 2026.

Appendices

A Supplementary Experiments **15**

A.1 Comparison with ScaleEdit 15

A.2 Comparison with Reference-Based Super-Resolution 15

A.3 Single-Step Refinement and Bridge Training 15

B User study **16**

C Training Configuration **16**

D Pseudo Code **17**

E Prompts **19**

F Limitation and Future Work **19**

G More Cases **20**

A Supplementary Experiments

We provide additional comparisons with ScaleEdit and reference-based super-resolution, followed by ablations on the low-resolution editing resolution and the refinement formulation. We use the evaluation metrics defined in Sec. 5; bold values indicate the best result within each resolution or ablation setting.

A.1 Comparison with ScaleEdit

We evaluate ScaleEdit [40] using its official implementation under the same evaluation protocol. Table 4 reports results at 1K, 2K, and 4K. EditBridge achieves higher HaarPSI, M-PSNR, and M-SSIM and lower M-MSE at all three resolutions. It also achieves lower M-LPIPS at 2K and 4K, while ScaleEdit obtains a slightly lower M-LPIPS at 1K. The reported inference times are substantially lower for EditBridge across all resolutions.

Table 4: Comparison with the official implementation of ScaleEdit.

Resolution	Method	HaarPSI ↑	M-PSNR ↑	M-SSIM ↑	M-MSE ↓	M-LPIPS ↓	Time (s) ↓
1K	ScaleEdit [40]	0.4824	16.8461	0.8206	0.0408	**0.3784**	181.2
	Ours	**0.4992**	**18.6536**	**0.8686**	**0.0369**	0.3829	**0.84**
2K	ScaleEdit [40]	0.4583	16.5510	0.6818	0.0485	0.5197	662.5
	Ours	**0.4673**	**18.6366**	**0.8712**	**0.0377**	**0.3903**	**4.08**
4K	ScaleEdit [40]	0.4116	18.5714	0.7638	0.0390	0.6015	3,601.2
	Ours	**0.4279**	**24.1380**	**0.8347**	**0.0049**	**0.2112**	**61.1**

A.2 Comparison with Reference-Based Super-Resolution

Our refinement stage is related to reference-based super-resolution: it upsamples an LR edited image with guidance from an HR source image. In image editing, however, the HR source can conflict with the desired output in edited regions. Refinement must therefore preserve useful source details while allowing the edited content to differ from the reference. We evaluate AdaRefSR [53] in the same pipeline, using the LR edited result as its input and the HR source as its reference. As shown in Table 5, EditBridge achieves better results on all five metrics at each resolution, supporting the effectiveness of its editing-specific refinement design.

Table 5: Comparison with reference-based super-resolution using AdaRefSR.

Resolution	Method	HaarPSI ↑	M-PSNR ↑	M-SSIM ↑	M-MSE ↓	M-LPIPS ↓
1K	AdaRefSR [53]	0.4626	16.8071	0.8125	0.0413	0.3836
	Ours	**0.4992**	**18.6536**	**0.8686**	**0.0369**	**0.3829**
2K	AdaRefSR [53]	0.4557	16.0025	0.6502	0.0544	0.4596
	Ours	**0.4673**	**18.6366**	**0.8712**	**0.0377**	**0.3903**
4K	AdaRefSR [53]	0.3923	16.9001	0.6763	0.0541	0.5164
	Ours	**0.4279**	**24.1380**	**0.8347**	**0.0049**	**0.2112**

A.3 Single-Step Refinement and Bridge Training

To examine whether our single-step performance can be reproduced by deterministic refinement, we compare two alternatives: a feed-forward refiner that directly regresses the HR target from the upsampled LR edited result, and a deterministic bridge trained without the stochastic Brownian perturbation. All three variants use one inference update.

Table 6 shows that EditBridge achieves the best M-PSNR, M-SSIM, M-MSE, and M-LPIPS at both 1K and 2K, while the deterministic bridge achieves the highest HaarPSI. The results support the benefit of stochastic bridge training for the evaluated region-aware quality metrics, even when

inference uses only one update. Stochastic perturbations during training should be distinguished from inference-time noise injection: these results do not imply that the deployed single-step sampler injects stochastic noise. For the degradation observed with additional sampling steps in Table 3, a possible explanation is the accumulation of prediction errors across repeated model evaluations; we treat this as an interpretation rather than an established cause.

Table 6: Comparison of refinement formulations with one inference update. The deterministic bridge removes the Brownian perturbation during training.

Resolution	Method	HaarPSI ↑	M-PSNR ↑	M-SSIM ↑	M-MSE ↓	M-LPIPS ↓
1K	Feed-forward refiner	0.4860	18.5485	0.8622	0.0373	0.3928
	Deterministic bridge	**0.5002**	16.2357	0.8439	0.0383	0.3923
	Ours	0.4992	**18.6536**	**0.8686**	**0.0369**	**0.3829**
2K	Feed-forward refiner	0.3868	11.1048	0.5208	0.1091	0.6674
	Deterministic bridge	**0.4788**	16.9659	0.6778	0.0506	0.4563
	Ours	0.4673	**18.6366**	**0.8712**	**0.0377**	**0.3903**

B User study

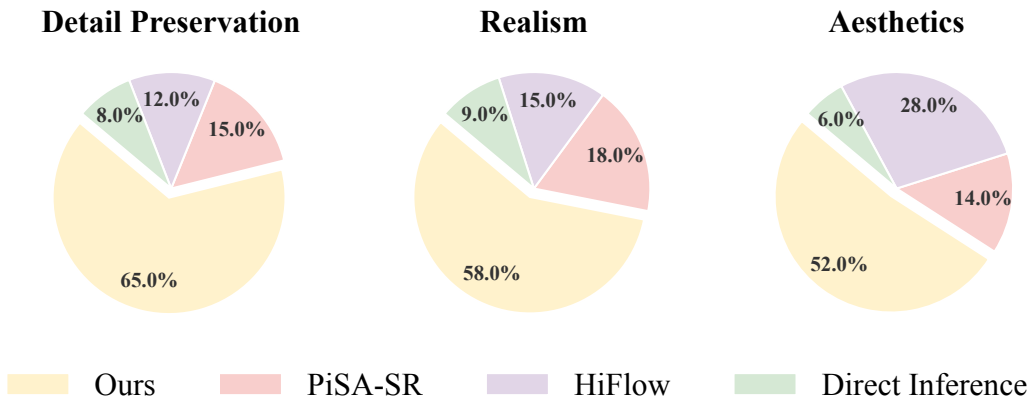

Figure 7: Results of the user study across three dimensions: *Detail Preservation*, *Realism*, and *Aesthetics*. Our method consistently outperforms all baselines in human preference.

To evaluate subjective visual quality beyond quantitative metrics, we conducted a blind Four-Alternative Forced Choice user study. Thirty participants evaluated 50 randomly sampled image pairs. We compared EditBridge against direct inference and PiSA-SR, HiFlow representing SR method, high-resolution visual generation method, respectively. Participants were provided with both full images and 100% cropped patches and voted based on three criteria: *Detail Preservation* (fidelity to unedited high-frequency source textures), *Realism* (naturalness of micro-structures), and *Aesthetics* (overall composition and structural integrity).

As shown in Fig. 7, EditBridge consistently achieves the highest preference rates across all evaluated dimensions. Specifically, our method significantly outperforms PiSA-SR and HiFlow in Detail Preservation and Realism. This confirms that our prior-guided sparse attention effectively prevents the texture hallucination typical of SR pipelines and the information divergence common in pure generative models. Furthermore, while HiFlow exhibits strong performance in the *Aesthetics* dimension—owing to the remarkable aesthetic priors inherent in large generative models—EditBridge still secures the highest overall aesthetic preference. This demonstrates that our approach successfully marries superior visual aesthetics with strict structural fidelity.

C Training Configuration

We provide the detailed training configuration below.

Training Configuration

```yaml
lora_config:
  r: 128
  lora_alpha: 128
  init_lora_weights: "gaussian"
  target_modules: "all-linear"

optimizer:
  type: "Prodigy"
  params:
    lr: 1
    use_bias_correction: true
    safeguard_warmup: true
    weight_decay: 0.01
```

D Pseudo Code

We present our training procedure in Algorithm 1 and the inference sampling process in Algorithm 2.

For training, we learn a velocity field that bridges the ground truth distribution and the conditional distribution. Specifically, we define the ground truth latent z_{gt} as the flow source z_0 and the low-resolution target z_{lr_tgt} as the flow target z_1. We sample a timestep t and construct a fused latent z_t through the Bridge forward process to train the model to predict the corresponding vector field. To preserve high-frequency details, we concatenate the high-resolution source z_{hr} along the sequence dimension. We employ a sparse attention mechanism where the attention mask Ω is pre-computed based on the correspondence between the low-resolution versions of the source and target, which regularizes the model during optimization of the weighted MSE loss.

For inference, we perform progressive denoising. Starting from the initial high-resolution source latent z_1, we iteratively solve the stochastic differential equation (SDE) toward the target state z_0. In each denoising step, we downsample the current latent z_t to re-estimate the sparse attention mask Ω, ensuring that the Transformer operates under accurate structural constraints relative to the fixed high-resolution reference. We incorporate a Brownian Bridge update rule, where a predicted velocity field and a scaled noise term are combined to update the latent state. This sampling strategy, stabilized by a noise rescaling factor, enables high-fidelity reconstruction of the edited image while adhering to the input text and reference source.

Algorithm 1: Training Procedure of Resolution-Bridge Model with Sparse Attention

Input: High-res Source I_{hr}, Low-res Target I_{lr_tgt}, Ground Truth I_{gt}, Text Prompt y
Output: Optimized Loss $\mathcal{L}$

$I_{lr_src} \leftarrow \text{Downsample}(I_{hr})$

$z_{hr}, z_{gt}, z_{lr_tgt}, z_{lr_src} \leftarrow \text{VAE}(I_{hr}, I_{gt}, I_{lr_tgt}, I_{lr_src})$;
$c \leftarrow \text{TextEncoder}(y, z_{lr_tgt})$;

$z_0 \leftarrow z_{gt}$;
$z_1 \leftarrow z_{lr_tgt}$;

Sample time $t \sim \mathcal{U}(0, 1)$ and noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$;
$z_t \leftarrow (1 - t)z_0 + tz_1 + \sigma_t\epsilon$

$x_{in} \leftarrow \text{Concat}([z_t, z_{hr}], \text{dim} = \text{sequence})$;
$\Omega \leftarrow \text{ExtractSparseIndices}(z_{lr_tgt}, z_{lr_src})$;

$v_\theta \leftarrow \text{Transformer}_\theta(x_{in}, t, c\,; \Omega)$;
$\mathcal{L}_{mse} \leftarrow \left\| v_\theta - \frac{z_t - z_0}{t} \right\|^2$;

if *stabilized* **then**
  $w_t \leftarrow \frac{t\|z_1 - z_0\|^2}{t\|z_1 - z_0\|^2 + (1-t)}$;
  $\mathcal{L} \leftarrow w_t \cdot \mathcal{L}_{mse}$;
else
  $\mathcal{L} \leftarrow \mathcal{L}_{mse}$;
return $\mathcal{L}$

Algorithm 2: Inference Procedure of Model with Sparse Attention

Input: Initial latent z_1, High-res reference z_{hr}, Low-res source z_{lr_src}, Text embedding c, Timesteps $\mathcal{T} = \{t_1, \ldots, t_N\}$, Noise scale σ, Rescale flag $rescale_noise$
Output: Denoised final latent z_0

$z_t \leftarrow z_1$;

for $i = 1$ **to** N **do**
  $t \leftarrow \mathcal{T}[i]$;
  $t_{next} \leftarrow \mathcal{T}[i + 1]$ (or 0 if $i = N$);
  $\Delta t \leftarrow t_{next} - t$;

  $z_{lr_tgt} \leftarrow \text{Downsample}(z_t)$;
  $\Omega \leftarrow \text{ExtractSparseIndices}(z_{lr_tgt}, z_{lr_src})$;

  $x_{in} \leftarrow \text{Concat}([z_t, z_{hr}], \text{dim} = \text{sequence})$;
  $v_\theta \leftarrow \text{Transformer}_\theta(x_{in}, t, c\,; \Omega)$;

  $\eta \leftarrow \sqrt{-\Delta t \cdot \frac{t_{next}}{t}}$;
  $\epsilon \sim \mathcal{N}(0, \mathbf{I})$;
  if *rescale_noise* **then**
    $\gamma \leftarrow \text{Clip}(|v_\theta|, 0, 1)$;
  else
    $\gamma \leftarrow 1$;

  $z_t \leftarrow z_t + v_\theta \cdot \Delta t + \sigma \cdot \eta \cdot \gamma \cdot \epsilon$;

return z_t

E Prompts

We present detailed prompts used in our experiments below. The prompts correspond to the cases shown in the main paper, ordered by their appearance.

Prompts for Editing

System Prompt Input to Gemini

You are an expert image editing instruction writer.

IMPORTANT

- You must output STRICT JSON.
- Do NOT output any explanation outside JSON.
- Do NOT use markdown.
- Do NOT include text before or after JSON.
- If you violate this format, the output is invalid.

You are given one image.
Your task: generate **ONE high-quality image editing prompt** based on this image. The prompt should describe **plausible, clear, and specific edits**. Keep it in **one sentence**. The instruction should be in **English and concise**.
Example: "Change the background to a sunny day with blue sky and white clouds."

Final output format

```
{   "edit_prompt":  "your single-sentence editing prompt" }
```

Editing Instructions

1. Transform the scene into an elementary school hallway with all doors wide open. Show the subject in a confused state, and add question marks above their head.
2. Transform the fox's monochrome fur into a red gradient, transitioning from deep amber to light coral. Position its front paws to be pinching a Nepalese dumpling, featuring a stylized fox-face motif rendered in white ink on the dumpling's skin.
3. Repaint the starry sky scene in a thick impasto oil painting style. Retain Van Gogh's *Starry Night* brushwork while blending the sky with amber dusk hues, forming a warm sunset gradient within the swirling nebulae.
4. Soften the giant surveillance eye motif into a starlit moon shape. Add colorful mosaic decorations with traditional White Wolf tribe patterns to the building base, and lower the overall color saturation to create a melancholic Eastern classical atmosphere.
5. Change the transparent umbrellas to bright red umbrellas.
6. Add a small red canoe floating on the water.
7. Remove the square card the character is holding.
8. Transform the main character from a Chibi (Q-version) style to a classical Eastern aesthetic. Add vine-like gold ornaments and red gemstones to the hair and accessories, replace the clothing with a cyan Hanfu featuring intricate embroidery, and adjust the environment to a natural green-toned setting with shallow depth of field.
9. Create an impressionistic portrait of a Steampunk Pirate Captain combining the blue-violet palette of Monet's *Water Lilies* with Van Gogh's swirling brushstrokes. Metallic gears should shimmer with cadmium yellow impasto while steampunk structures with starry gradients appear within the background foliage.
10. Seamlessly integrate a second image of a bedside monitor into the ward scene with automatically matched lighting and perspective.
11. Dress the bear in a classic maid-inspired outfit featuring a white apron with gold piping and white stockings. Decorate the environment with balloons and birthday banners to create a whimsical fairytale party atmosphere.

F Limitation and Future Work

Limitations Despite the superior editing performance demonstrated by our framework, several constraints remain to be addressed. First, our method currently relies on a pre-defined indices prior,

which must be extracted or synthesized before the translation process. This dependency introduces a bottleneck in fully automated pipelines, as the quality of the final output is inherently bounded by the precision of these initial guidance signals. Second, the iterative nature of the Diffusion Bridge sampling process, while robust, still incurs a non-negligible computational overhead, particularly when dealing with high-resolution imagery or complex multi-step transitions.

Future Work: Automated Prior Estimation To enhance the autonomy of our approach, a promising future direction is to integrate a learnable prior estimation module. By leveraging self-supervised representation learning, we aim to enable the model to infer the necessary structural or semantic indices directly from the source image, thereby eliminating the need for manual or external prior acquisition. This would significantly broaden the applicability of Diffusion Bridge models in "in-the-wild" image editing scenarios.

Future Work: Post-hoc Refinement and Localized Repair Beyond global image-to-image translation, we envision the Diffusion Bridge as a powerful tool for post-hoc refinement and interactive content repair. Specifically, if certain regions of a generated image suffer from "mode collapse" or textural artifacts, the Diffusion Bridge can be re-initialized within a masked area to perform localized resampling. By bridging the intermediate latent states of a suboptimal generation back to the clean data manifold, the model can “correct" specific flaws while preserving the global coherence of the scene.

Generalizing the Bridge Mechanism Furthermore, we plan to investigate the potential of the Diffusion Bridge in broader domains, such as video-to-video editing and 3D asset stylization. Extending the bridge trajectory to maintain temporal and spatial consistency across frames or viewpoints remains a challenging yet rewarding frontier. We believe that the mathematical elegance of the Diffusion Bridge offers a versatile foundation for a wide range of controllable generative tasks.

G More Cases

We provide an extensive collection of additional samples to further validate the efficacy of our proposed method. As illustrated in Fig. 8 and Fig. 9, our results achieve fidelity to the source images while maintaining the utmost perceptual naturalness in the edited regions.

Our Diffusion Bridge framework demonstrates a superior ability to preserve the original structural integrity and identity of the subjects. Meanwhile, the synthesized content integrates seamlessly with the existing background without introducing common generative artifacts such as blurriness or texture repetition. These extensive cases substantiate that our approach strikes an optimal balance between strictly following complex textual instructions and faithfully retaining the essential characteristics of the source content.



Figure 8: Visual results at 1K resolution.

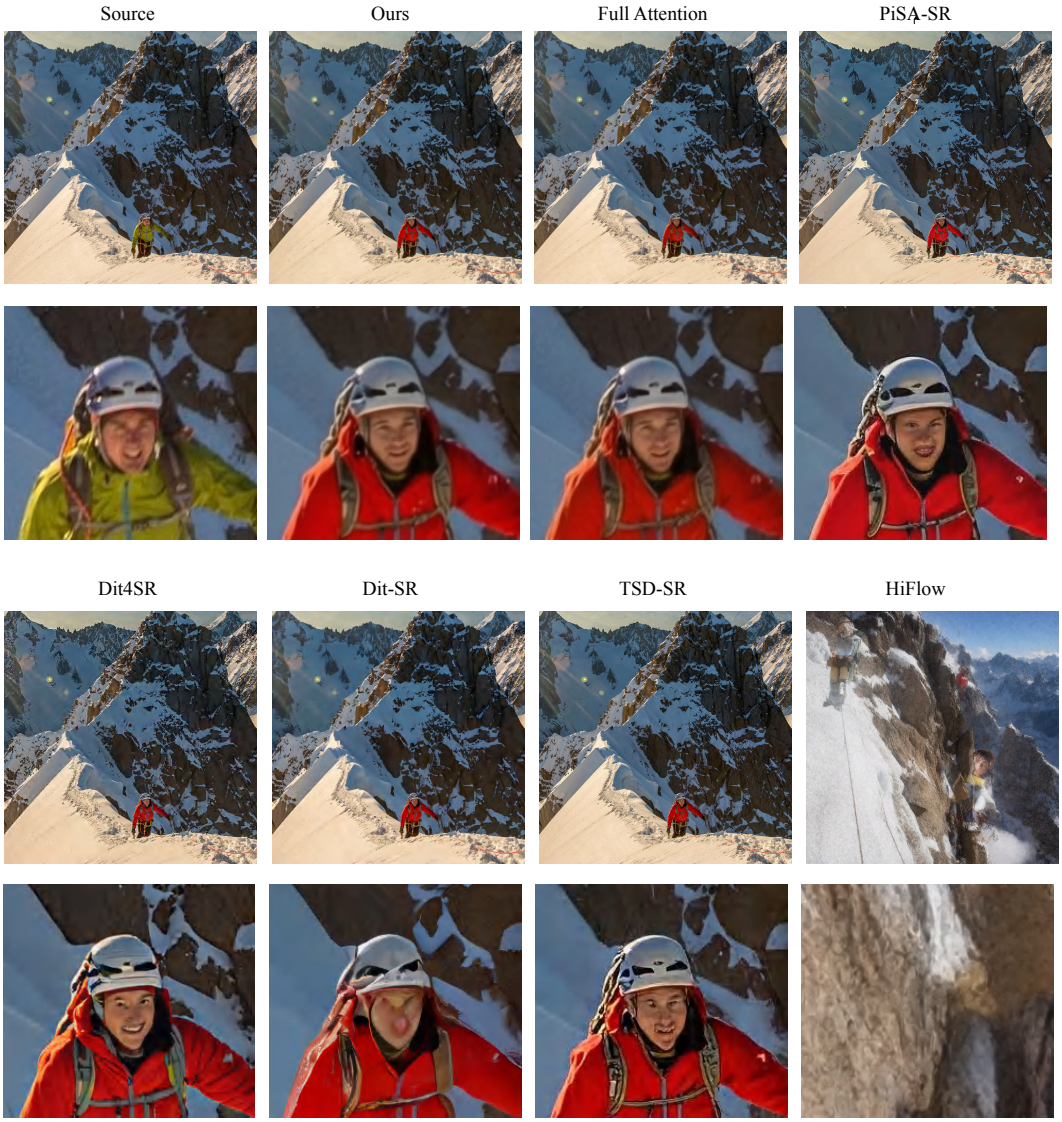

Figure 9: Visual results at 2K resolution.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our motivation, contribution, and scope are clearly described in both the abstract and the introduction.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our work in Appendix. Specifically, we analyze the performance dependency on the quality of spatial priors and acknowledge the constant computational overhead for prior estimation, which makes our method less optimal for very low-resolution tasks. We also discuss how significant spatial discrepancies in LR edited images might affect the high-resolution synthesis.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: Our paper is primarily an empirical study focused on architectural design and optimization. We do not introduce new theoretical theorems that require formal proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide comprehensive implementation details in Section 4 and the Appendix. This includes the precise architecture of our Prior-Guided Sparse Attention, hyperparameters for the Diffusion Bridge training, and the pre-processing protocols for our 1K/2K/4K high-resolution benchmarks.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The source code, including training and inference scripts and environment setup instructions, is publicly available at `https://github.com/songyangyifei/EditBridge`. However, the complete set of data and artifacts needed to faithfully reproduce all main experimental results is not yet publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Justification: We present our experimental setting and details in Appendix. C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: In our user study, we report the preference distributions collected from 30 participants across 50 test cases. For quantitative metrics in Tab. 1, we report the mean values calculated over the entire evaluation dataset to ensure the stability of the observed performance gains.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We show the experiments compute resources in Sec. 5

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics, and strictly followed it in this work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Yes, we we talk about boarder impacts in Appendix. F

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Our research complies with the safety guidelines of the underlying pre-trained models. Upon release, we will implement a restrictive user license to prevent the generation of malicious content.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All third-party assets, including pre-trained diffusion backbones (e.g., Qwen-Image-Edit), evaluation suites, and public datasets, are cited in the references. We have verified their respective licenses (e.g., MIT, Apache 2.0, or Open RAIL-M) and ensure that our use and intended redistribution strictly adhere to their terms.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All new assets introduced in this work, including the Prior-Guided Sparse Attention implementation, the specialized high-resolution evaluation benchmarks (1K/2K/4K), and the corresponding rewritten caption sets, are thoroughly documented in the Appendix. The documentation includes data organization schemas, pre-processing logic for spatial priors, and detailed environment setup instructions.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Details of the user study are included in the Appendix. B. We also specify the number of participants and the compensation provided for their time.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The user study conducted in this work poses minimal risk to participants as it only involves visual assessment of synthetic images. All participants were informed of the study's scope and purpose. The study was conducted in accordance with the ethical guidelines of our home institution, and no sensitive personal data was collected.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We utilize Large Language Models (LLMs) to perform semantic caption rewriting for our high-resolution training and evaluation datasets. This process is essential for generating the fine-grained textual descriptions required for high-fidelity editing.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.