ELASTIC BRIDGE MODEL FOR IMAGE EDITING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image editing has two competing goals: it should preserve unedited regions with high fidelity while generating realistic and coherent content in edited regions. Existing data-to-data bridge paradigms struggle to satisfy both. Flow matching often under-edits due to overly deterministic trajectories, whereas Brownian bridges introduce stochasticity via global noise, which harms regions that should remain intact. We interpret this stochasticity-determinism tension through a spatial signal-to-noise ratio (SNR) lens. High SNR yields more deterministic trajectories and better fidelity in unedited regions, while lower SNR increases stochasticity and enables substantial edits. As a result, edited and unedited regions require different SNR levels. To resolve this, we propose **Elastic Bridge**, which replaces the global noise scale with a spatially adaptive elasticity map to balance SNR across regions. To make this design practical, we introduce Source-Grounded Elasticity, a simple approximation that mitigates the resulting train-inference mismatch. Under fair comparison, Elastic Bridge achieves the best accuracy-preservation trade-off among all compared paradigms and scales effectively to state-of-the-art large-scale models.

1 INTRODUCTION

Recent work has shown that data-to-data bridge models (Liu et al., 2022; Li et al., 2023; Tan et al., 2025b; Zhou et al., 2023) are a promising approach to complex vision tasks such as image editing (Tan et al., 2025b), inpainting (Liu et al., 2023), and cross-domain translation (Wang et al., 2024b), offering advantages in training efficiency, inference speed, and structural preservation over noise-to-data generation methods (Tan et al., 2025b). Unlike noise-to-data regeneration methods (Brooks et al., 2023; Rombach et al., 2022; Tan et al., 2025a) that synthesize the target from structureless Gaussian noise, data-to-data bridge models build a direct transport trajectory from the source image to the target. Intuitively, this paradigm is also the more principled choice for editing, as it simply asks the model to transport what should change rather than reconstruct what should be preserved.

Instruction-guided image editing (Ye et al., 2025; Li et al., 2025; Yu et al., 2025), however, exposes a tension within this paradigm. On one hand, unedited regions must be preserved with near-perfect fidelity: any unwanted perturbation constitutes a failure. On the other hand, edited regions must be generated with high quality and sufficient diversity to represent all plausible outcomes. These two demands are inherently contradictory, since fidelity requires determinism while diversity requires stochasticity.

Existing data-to-data bridge models (Zheng et al., 2024; Wang et al., 2024a; Vasilev et al., 2025; Ji et al., 2024; Serrano-Lozano et al., 2026; Chadebec et al., 2025) address this tension through two strategies, each of which resolves one problem while introducing another (see Figure 1). Deterministic bridges model the trajectory as a straight line from source to target, following an Ordinary Differential Equation (ODE) (Lipman et al., 2022; Albergo et al., 2023). This approach keeps trajectories stable but averages over all possible edits, causing the output to collapse and fail to correspond to any realistic edit. Stochastic bridges like Brownian Bridge introduce random noise into the trajectory to restore diversity (Li et al., 2023; Shi et al., 2023; Liu et al., 2023). However, this noise is applied uniformly to every pixel. For instance, when adding a scarf to a dog, only the neck region should change, while the face, ears, and background must remain exactly the same. Uniform noise disturbs all regions equally, so the preserved regions are corrupted to the same degree as the edited ones.

We trace this failure to the uniform application of a single scalar σ, even though different regions require different amounts of noise. Our instantaneous signal-to-noise ratio (SNR) (Girsanov, 1960;

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

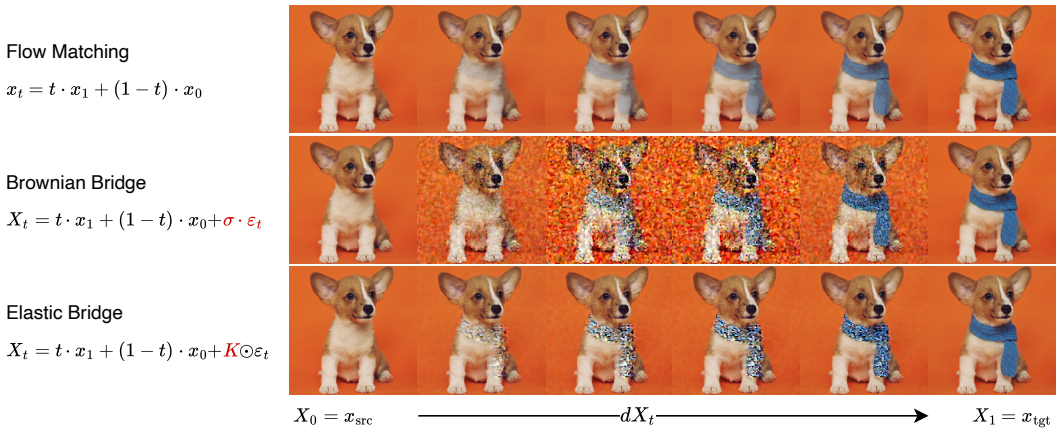

Figure 1: Comparison of transport paradigms on an editing task (adding a scarf to a dog). Flow Matching (ODE) follows a straight, deterministic path. Brownian Bridge (SDE) introduces global noise to restore diversity, but corrupts the entire image. Elastic Bridge (Ours) employs a spatially adaptive noise map K that concentrates stochasticity on the neck region while keeping the remainder regions stable.

Hang et al., 2023) analysis reveals the underlying mechanism. In regions with little change, the diffusion noise overwhelms the transport drift and erases the source structure, whereas in regions with large change, the transport drift overwhelms the diffusion noise and suppresses output diversity. This suggests that a well-designed bridge should maintain a balanced SNR in every region, neither too low nor too high (Li et al., 2023; Tan et al., 2025b).

Motivated by this insight, we propose **Elastic Bridge**. Rather than relying on a single scalar σ, we introduce a spatial elasticity map K that assigns each region its own noise level, chosen so that every region operates at a well-balanced SNR. Regions that should be preserved receive near-zero noise to maintain fidelity, while regions that should be edited receive sufficient noise to produce diverse and realistic outputs. While Elastic Bridge is elegant in theory, deploying it in practice introduces a significant challenge. Computing K requires the target image, which is available during training but unknown at inference, creating a train-inference mismatch that substantially degrades real-world performance. To close this gap, we introduce **Source-Grounded Elasticity**, which employs an approximation to resolve the mismatch without any auxiliary networks.

Under identical training conditions, we compare Elastic Bridge against existing transport paradigms and find that it consistently achieves the best trade-off between editing accuracy and structural preservation. We further adapt it to a state-of-the-art large-scale model (Wu et al., 2025), demonstrating that the benefits of Elastic Bridge scale effectively to larger architectures.

Our main contributions are summarized as follows. **SNR analysis.** We show that a fixed noise level in existing bridge models produces a spatially imbalanced SNR, explaining the preservation-diversity tension. **Elastic Bridge.** We replace the scalar σ with $K \propto |\delta|$, keeping the SNR uniform across all locations and resolving both failure modes in a single step. **Source-Grounded Elasticity.** We introduce a parameter-free mechanism that grounds the noise map in the source image alone, closing the train-inference mismatch without any auxiliary network.

2 Related Works

Generative models. Early generative models, such as Variational Autoencoders (VAEs) (Pu et al., 2016) and Generative Adversarial Networks (GANs) (Goodfellow et al., 2014; Karras et al., 2019), enabled high-quality synthesis through latent modeling and adversarial training. Diffusion models (Ho et al., 2020; Rombach et al., 2022) later introduced iterative denoising processes, significantly advancing image and video generation. Flow-matching models (Esser et al., 2024; Lipman et al., 2022) further reframed generation as learning deterministic or stochastic paths between distributions. More recently, transformer-based architectures trained at scale have further enhanced the fidelity and diversity of generative models (Peebles & Xie, 2023; Wu et al., 2025; Wan et al., 2025).

Instruction-based Image Editing Instruction-based image editing (Brooks et al., 2023; Zhang et al., 2023; Hu et al., 2025) fine-tunes diffusion models on paired datasets so that users can modify images via natural language commands. Recent methods further integrate MLLMs for stronger instruction comprehension (Fu et al., 2024; Huang et al., 2024), while ImgEdit (Ye et al., 2025), UniWorld (Li et al., 2025), AnyEdit (Yu et al., 2025) and others (Liu et al., 2025; Labs et al., 2025; Kara et al., 2024; Bai et al., 2025a) continue to push editing quality. However, all these approaches remain within the noise-to-data paradigm, where uniform noise addition inevitably corrupts unedited regions, leaving the accuracy-preservation trade-off fundamentally unresolved.

Bridge models. Bridge Models (Liu et al., 2023; Li et al., 2023; Zhou et al., 2023) construct stochastic processes directly connecting source and target distributions, providing an alternative to noise-driven generation. Early approaches employed Schrödinger bridges (De Bortoli et al., 2021), stochastic interpolants (Albergo et al., 2023), and entropic optimal transport (Cuturi, 2013). Recent diffusion-based variants, such as Denoising Diffusion Bridge Models (DDBM) (Zhou et al., 2023) and Brownian Bridge methods (Li et al., 2023), demonstrated promising results for conditional generation and image translation tasks. Several recent works have highlighted the potential of Bridge Models in vision tasks, demonstrating improved structural and stylistic consistency in exemplar-guided image translation (Lee et al., 2024), enhanced temporal coherence in video synthesis (Vasilev et al., 2025), and increased efficiency during training and inference for image translation tasks (Ji et al., 2024; Chadebec et al., 2025; Tan et al., 2025b; Li et al., 2023).

3 Elastic Bridge

Given a source image I_{src} and a condition c (e.g., a text prompt or mask), *conditional image-to-image translation* generates a target I_{tgt} satisfying c while preserving unrelated regions.

3.1 Preliminaries

Data-to-data bridges (Tan et al., 2025b; Albergo et al., 2023; De Bortoli et al., 2021; Li et al., 2023; Shi et al., 2023; Liu et al., 2023) define a continuous latent trajectory X_t ($t \in [0, 1]$) from $X_0 = x_0 = \mathcal{E}(I_{\text{src}})$ to $x_1 = \mathcal{E}(I_{\text{tgt}})$, where $\mathcal{E}$ is the encoder of a pretrained autoencoder.

Deterministic Transport. Following the flow matching framework (Liu et al., 2022; Lipman et al., 2022), the most direct path is a straight-line interpolation from source to target, modeled as an Ordinary Differential Equation (ODE), given by

$$\mathrm{d}x_t = (x_1 - x_0)\,\mathrm{d}t. \tag{1}$$

This equation describes a particle moving at constant velocity towards x_1. While conceptually simple, this deterministic path is limited. Because one source image can map to many valid edits, the optimal vector field under an L_2 objective collapses to the conditional expectation of the target distribution (Liu et al., 2022; Esser et al., 2024). In practice, averaging over all plausible modes causes the output to collapse into a degenerate mean, failing to represent valid edits (Albergo et al., 2023; Tan et al., 2025b).

Stochastic Transport. To mitigate this issue, one common approach is to inject random noise into the trajectory, converting the deterministic ODE into a Stochastic Differential Equation (SDE). The Brownian Bridge process (Li et al., 2023; Maruyama, 1955; Zhou et al., 2023) provides a canonical construction for data-to-data transport, formulated as

$$\mathrm{d}X_t = \underbrace{\frac{x_1 - X_t}{1 - t}}_{\text{Drift}}\,\mathrm{d}t + \underbrace{\sigma}_{\text{Diffusion}}\,\mathrm{d}W_t, \tag{2}$$

where W_t is standard Brownian motion and $\sigma > 0$ is the diffusion scale, typically set to 1. While stochasticity is essential for generative diversity, injecting noise uniformly across all spatial locations introduces two problems. First, at the *task level*, different translation tasks empirically demand different noise levels; no single global configuration works universally (Tan et al., 2025b; Li et al., 2023). Second, at the *spatial level*, uniform noise inevitably corrupts regions that should remain untouched and degrades the fine-grained details of the source content.

These limitations share a common root: a fixed noise level σ is *rigid*, incapable of distinguishing regions that must stay unchanged from regions that must change. **Can we build a bridge that is *elastic*, adapting its noise to each region's demand for change?**

3.2 SNR Analysis of Transport Dynamics

To find such a bridge, we start from an analysis of the *Signal-to-Noise Ratio (SNR)* of the transport process. Based on Girsanov's theorem (Girsanov, 1960), the SNR of a stochastic transport $\mathrm{d}X_t = v\,\mathrm{d}t + \sigma\,\mathrm{d}W_t$, conditioned on the boundary pair (x_0, x_1), is defined as the ratio of expected drift energy to diffusion variance:

$$\mathrm{SNR} \triangleq \frac{\|\mathbb{E}[v \mid x_0, x_1]\|^2}{\sigma^2}. \tag{3}$$

Since each spatial location evolves independently, we conduct the analysis element-wise. Let $\delta = x_1 - x_0$ denote the displacement from source to target. For the Brownian Bridge SDE (Equation (2)), the drift is $v = (x_1 - X_t)/(1-t)$. Its conditional mean, given (x_0, x_1), can be derived from the closed-form solution $\mathbb{E}[X_t \mid x_0, x_1] = (1-t)x_0 + tx_1$, giving

$$\mathbb{E}\left[\frac{x_1 - X_t}{1-t} \,\middle|\, x_0, x_1\right] = \frac{x_1 - \mathbb{E}[X_t \mid x_0, x_1]}{1-t} = \frac{(1-t)\delta}{1-t} = \delta. \tag{4}$$

This is independent of t. Substituting into Equation (3) yields a time-invariant SNR:

$$\mathrm{SNR} = \frac{\delta^2}{\sigma^2}. \tag{5}$$

When a single global σ is used, the SNR at each spatial location i, $\mathrm{SNR}^{(i)} = (\delta^{(i)})^2/\sigma^2$, varies entirely with the local displacement $|\delta^{(i)}|$. This leads to two failure modes. In regions with little or no change ($|\delta^{(i)}| \approx 0$), $\mathrm{SNR}^{(i)}$ is near-zero and noise dominates, causing a random walk that corrupts the source content. In regions with large displacements, $\mathrm{SNR}^{(i)}$ becomes high and drift dominates, producing outputs that lack diversity. This explains also why low-displacement tasks favor small σ while high-displacement tasks require large σ (Albergo et al., 2023; Tan et al., 2025b).

3.3 The Elastic Bridge SDE

The analysis above suggests that transport quality depends on a well-balanced SNR: too low leads to random walks, too high leads to oversmoothing. A single scalar σ cannot achieve this balance at every location, because different locations have different displacements $|\delta^{(i)}|$.

We solve this by replacing σ with a spatial map $K \in \mathbb{R}^{H\times W\times C}$ that scales the noise independently at each location. The resulting *Elastic Bridge* SDE is:

$$\mathrm{d}X_t = \frac{x_1 - X_t}{1-t}\,\mathrm{d}t + \underbrace{K}_{\text{Elasticity}} \odot \mathrm{d}W_t, \quad t \in [0, 1), \tag{6}$$

where $\odot$ denotes element-wise multiplication. The SNR at each location i is now $(\delta^{(i)})^2/(K^{(i)})^2$. To make it uniform, we set a target value $\gamma > 0$, requiring that

$$(\delta^{(i)})^2/(K^{(i)})^2 = \gamma \quad \Longrightarrow \quad K = |\delta|/\sqrt{\gamma}. \tag{7}$$

A larger γ makes the SDE more deterministic; a smaller γ allows more noise.

Because K scales with the displacement, the Elastic Bridge adapts automatically. Where nothing changes ($|\delta| \approx 0$), the noise vanishes and the source is preserved exactly. Where editing occurs ($|\delta| > 0$), the noise grows proportionally, enabling diverse outputs. The name *Elastic Bridge* reflects this adaptive nature. The transport is *stiff* where the image stays the same and *flexible* where it needs to change.

4 SOURCE-GROUNDED ELASTICITY

The Elastic Bridge SDE (Eq. (6)) requires the spatial elasticity map $K = |\delta|/\sqrt{\gamma}$, which depends on the oracle displacement $\delta = x_1 - x_0$. During training, the target x_1 is available, so δ is known exactly. During inference, however, x_1 is the quantity we aim to generate and is strictly unavailable. This section identifies the resulting discrepancy and proposes a principled resolution.

4.1 THE ELASTICITY GAP

With access to the oracle pair (x_0, x_1), training states X_t can be sampled directly from the analytical marginal distribution of the Elastic Bridge, given by

$$X_t = (1-t)x_0 + tx_1 + \sqrt{t(1-t)} \cdot K \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{8}$$

where $K = |\delta|/\sqrt{\gamma}$ is the oracle elasticity map derived in Section 3.3. The velocity network v_θ is trained to predict the analytical drift via the Flow Matching objective (Lipman et al., 2022), formulated as

$$\mathcal{L}(\theta) = \mathbb{E}_{t,x_0,x_1,\epsilon}\left[\left\| v_\theta(X_t, t, c) - \frac{x_1 - X_t}{1-t} \right\|_2^2\right]. \tag{9}$$

At inference time, a natural strategy is to estimate the target on-the-fly from the current state and the predicted velocity using

$$\hat{x}_1(X_t) = X_t + (1-t)\, v_\theta(X_t, t, c), \tag{10}$$

which yields a running displacement estimate $\hat{\delta}_t = \hat{x}_1(X_t) - x_0$. However, $\hat{x}_1(X_t)$ inevitably contains prediction error, so $\hat{\delta}_t$ deviates from the oracle δ. Because the network was optimized exclusively on trajectories generated with the oracle K (Eq. (8)), scaling the noise by an inaccurate $\hat{\delta}_t$ pushes the trajectory off the training distribution. We call this train-inference mismatch the *Elasticity Gap*.

4.2 GROUNDING ELASTICITY IN THE SOURCE

To close the Elasticity Gap, we propose *Source-Grounded Elasticity*. Instead of relying on the oracle δ during training or on the estimate $\hat{\delta}_t$ during inference, we anchor the elasticity map to the source x_0 and the condition c alone. Concretely, we predict the displacement once at $t = 0$ using the trained velocity network: $\hat{\delta}_0 = v_\theta(x_0, 0, c)$. This prediction defines the *Source-Grounded Elasticity Map*:

$$\hat{K} = \frac{|\hat{\delta}_0|}{\sqrt{\gamma}} = \frac{|v_\theta(x_0, 0, c)|}{\sqrt{\gamma}}, \tag{11}$$

which replaces the oracle map K in both training and inference. Substituting $\hat{K}$ into Eq. (6) yields the Source-Grounded Elastic Bridge SDE, formalized as

$$\mathrm{d}X_t = \frac{x_1 - X_t}{1-t}\,\mathrm{d}t + \hat{K} \odot \mathrm{d}W_t, \quad t \in [0, 1). \tag{12}$$

Why the velocity network at $t = 0$? By definition, the target velocity is $v_t = (x_1 - X_t)/(1-t)$. At the boundary $t = 0$ where $X_0 = x_0$, this reduces to $v_0 = \frac{x_1 - x_0}{1-0} = x_1 - x_0 = \delta$. Since v_θ is trained to approximate this vector field (Eq. (9)), evaluating it at $t = 0$ yields the conditional expectation of the displacement, such that $\hat{\delta}_0 \approx \mathbb{E}_{p(x_1|x_0,c)}[x_1 - x_0]$. This provides a principled spatial prior without any auxiliary network.

Why does SGE close the gap? The key property is that $\hat{\delta}_0$ depends only on x_0 and c, both of which are available at inference. By using the same $\hat{\delta}_0$ to construct the elasticity map $\hat{K}$ during both training and inference, the noise distribution governing the trajectory is identical in both phases. The inference trajectory therefore lies on the same distribution that the network was trained on, and the Elasticity Gap vanishes.

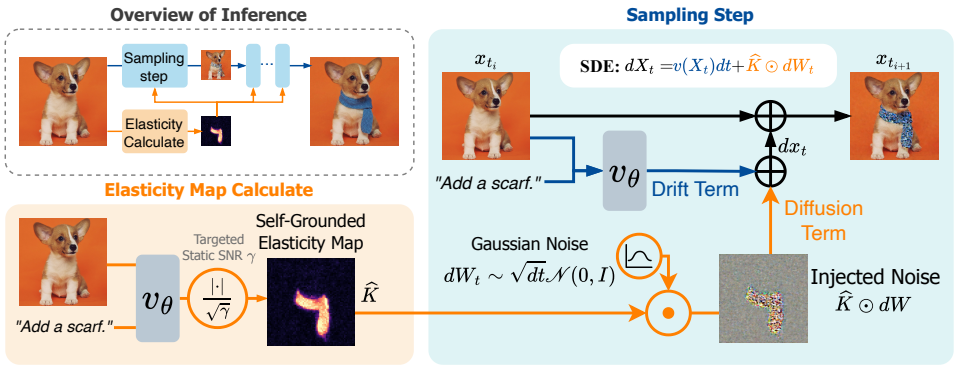

Figure 2: Overview of the Elastic Bridge inference pipeline. *Left*: the Source-Grounded Elasticity Map $\hat{K}$ is computed once at $t{=}0$ by evaluating $v_\theta(x_0, 0, c)$ and normalizing by $\sqrt{\gamma}$. *Right*: at each sampling step, the SDE integrates a drift term $v_\theta(X_t, t, c)\,\mathrm{d}t$ with a spatially modulated diffusion term $\hat{K} \odot \mathrm{d}W_t$, concentrating noise on regions that require editing while preserving the rest.

Training and inference. At each training iteration, we compute $\hat{K}$ from the current network under stop-gradient (Eq. (11)), draw $t \sim \mathcal{U}(0, 1)$, and sample X_t from Eq. (8) with K replaced by $\hat{K}$. We then optimize the flow matching loss in Eq. (9), updating the elasticity estimate online as the network learns. At inference, a single forward pass at $t = 0$ computes $\hat{K}$ from x_0 and c; this map remains fixed throughout SDE integration (Figure 2). See Algorithms 1 and 2 in Appendix A.4.

5 Experimental Analysis

We validate Elastic Bridge through a controlled comparison of transport paradigms (Section 5.2), ablation studies on Source-Grounded Elasticity and the SNR parameter γ (Section 5.3), and extended experiments on external public benchmarks and adaptation of a pretrained editing model (Section 5.4).

5.1 Experimental Setup

Training Details. For the main experiments, we train all transport variants (Flow Matching, Brownian Bridge, and Elastic Bridge) initialized from the same text-to-image model `Stable-Diffusion-3.5-Large` (Esser et al., 2024) using LoRA adapters with lora rank of 128 to enable strict comparison. This setting learns image-editing capability from a pretrained text-to-image backbone. By default, all models are trained on a 8-GPU cluster for 20k steps with the AdamW optimizer (`lr=1e-4`) and EMA decay of `0.999`, the batch size is set to 8.

Benchmark Construction. For matched training and test distributions, we use 10k OpenImages (Kuznetsova et al., 2020) images, Qwen-VL (Bai et al., 2025b) instructions across 8 editing categories, and Qwen-Image-Edit (Wu et al., 2025) targets. Filtering for generation quality and instruction alignment yields approximately 50k image pairs at 1024×1024, with 400 for testing and the remainder for training (construction details in the Appendix).

Evaluation Metrics. We evaluate models along three dimensions: Perceptual Quality, Editing Accuracy, and Structural Preservation. Editing accuracy is measured by CLIP Similarity (Radford et al., 2021) to ground-truth edited images and instruction-following Accuracy, the percentage of edits accepted by a vision-language model that judges whether the editing instruction is correctly followed. Structural preservation is measured using SSIM (Wang et al., 2004), which captures pixel-level luminance, contrast, and structural fidelity, on both the full image and the unedited regions. To obtain the unedited region, we compute the patch-level cosine similarity between `DINOv2-Base` (Oquab et al., 2023) features of the source and edited images, which is sensitive to semantic content changes but robust to minor pixel variations. Thresholding this similarity map at $\tau{=}0.5$ followed by 4 iterations of morphological dilation produces a binary unedited-region mask used to compute the unedited-region SSIM. Perceptual quality is evaluated using BRISQUE (Mittal et al., 2012) and PSNR (Hore & Ziou, 2010); inference time is reported to compare efficiency.

Table 1: Quantitative comparison across editing paradigms. CLIP Similarity measures similarity to ground-truth edited images; Accuracy is assessed by a vision-language model that judges whether the editing instruction is correctly followed. SSIM measures structural preservation against the source image on the full image and unedited regions. BRISQUE and PSNR report perceptual quality, and inference time reports efficiency.

	Editing Accuracy		Preservation		Perceptual Quality		Efficiency
Paradigm	CLIP Sim ↑	Accuracy ↑	$\text{SSIM}_{\text{full}}$ ↑	$\text{SSIM}_{\text{unedited}}$ ↑	BRISQUE ↓	PSNR ↑	Inference Time ↓
Conditional DiT	0.8849	73.75%	0.6051	0.6858	39.4887	15.3062	14.5
Flow Matching	0.9211	45.00%	**0.8224**	**0.8683**	37.8715	19.0590	**6.2**
Brownian Bridge	0.8667	72.25%	0.7094	0.7445	52.3533	18.0906	**6.2**
Elastic Bridge (Ours)	**0.9331**	**75.50%**	0.7549	0.8126	**37.4636**	**19.8780**	**6.2**

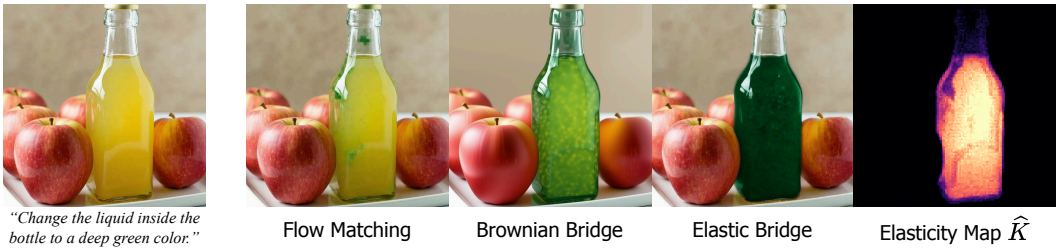

Figure 3: Qualitative comparison across editing paradigms. Flow Matching leaves the scene nearly unchanged (high scene fidelity, failed editing), Brownian Bridge introduces a noticeable texture shift (on apples and background) and fails to complete the edit, while Elastic Bridge correctly performs the edit with high scene fidelity.

5.2 Main Comparison

We first compare our Elastic Bridge against three existing paradigms, which consist of two data-to-data transport paradigms (Flow Matching (Lipman et al., 2022) and Brownian Bridge (Li et al., 2023; Tan et al., 2025b; Zhou et al., 2023)) and one noise-to-data generation paradigm (Conditional DiT based on Flow Matching). Specifically, the implementation of Flow Matching and Brownian Bridge is based on the design of ViBT (Tan et al., 2025b). And the implementation of Conditional DiT is based on the design of OminiControl (Tan et al., 2025a) which concatenates the conditional image tokens into the input tokens.

Editing Performance. Elastic Bridge achieves the best perceptual quality and editing accuracy across all metrics, and the best structural preservation among SDE-based paradigms (Table 1). It also matches the lowest reported inference time of Flow Matching and Brownian Bridge. Flow Matching records the highest SSIM scores, but its Accuracy of 45.0% (far below our 75.5%) reveals that this preservation is achieved by leaving the source largely unchanged rather than by performing the edit. The qualitative comparison in Figure 3 illustrates these differences: Flow Matching fails to perform the edit because its noise-free ODE trajectory collapses to the conditional mean, while Brownian Bridge corrupts unedited regions because a uniform σ drives the local SNR toward zero wherever $|\delta| \approx 0$, causing noise to dominate the drift. By setting $K \propto |\delta|$, Elastic Bridge avoids both failure modes, correctly performing the edit while preserving scene fidelity. The extended experiments in Section 5.4 provide complementary evidence on external public benchmarks and when adapting a pretrained editing model: Elastic Bridge improves reference-image fidelity on MagicBrush and the reported ImgEdit-Bench and Kontext Bench scores over Brownian Bridge, although its MagicBrush CLIP-T is lower.

Qualitative Results of Elastic Bridge. Figure 4 shows additional editing results from Elastic Bridge across diverse tasks. In each case, the edit is correctly applied to the target region while the rest of the scene remains intact. The elasticity map $\hat{K}$ confirms this by concentrating on the to-be-edited region and remaining near-zero elsewhere, reflecting the spatial structure of $\hat{\delta}_0 = v_\theta(x_0, 0, c)$.

Table 2: Training with Oracle vs. Source-Grounded Elasticity.

Variant	BRISQUE ↓	PSNR ↑	$\text{SSIM}_{\text{unedited}}$ ↑	Accuracy ↑
Oracle	36.40	19.24	**0.90**	36.75%
Source-Grounded	**35.91**	**20.10**	0.84	**70.75%**

Table 3: Sensitivity to γ on ImgEdit-Bench. Higher ImgEdit Score is better.

γ	100	4	1	0.25
ImgEdit Score ↑	1.95	2.98	**3.58**	3.52



Figure 4: Qualitative results of Elastic Bridge (SD3.5 based).

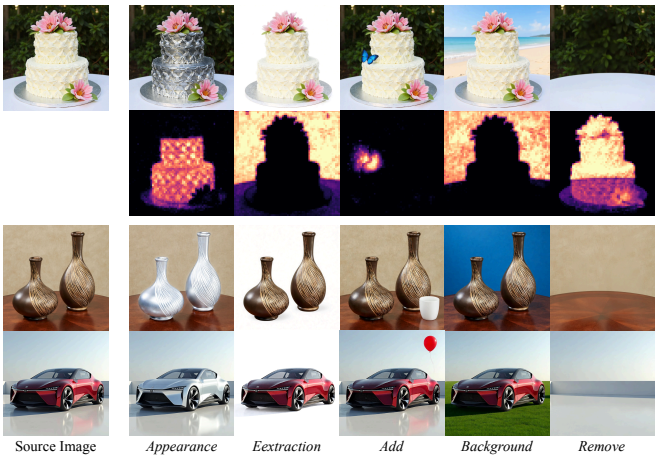

Figure 5: Qualitative results of Elastic Bridge (Qwen-Image-Edit based, ViBT) across diverse editing tasks.

Accuracy-Preservation Trade-off. Figure 6 plots all paradigms in the edit-success vs. unedited-region-preservation space. Flow Matching achieves high preservation but low edit success, consistent with its ODE trajectory collapsing to the conditional mean. Conditional DiT achieves high edit success but the lowest preservation, as it generates from noise with no structural anchor. Brownian Bridge's edit success initially increases with noise, peaking at $\sigma = 1$ before declining at $\sigma = 2$, while its unedited-region SSIM decreases throughout the sweep. For Elastic Bridge, reducing γ from 100 to 0.25 increases edit success with a more gradual decline in unedited-region SSIM. This is because a uniform σ drives the local SNR toward zero in low-displacement regions ($|\delta| \approx 0$), corrupting unedited content, whereas $K \propto |\delta|$ allocates noise according to the local editing displacement and keeps the SNR constant at γ across all locations. At their respective highest-accuracy settings, Elastic Bridge achieves both higher edit success and higher unedited-region SSIM than Brownian Bridge, demonstrating a more favorable balance between editing accuracy and preservation.

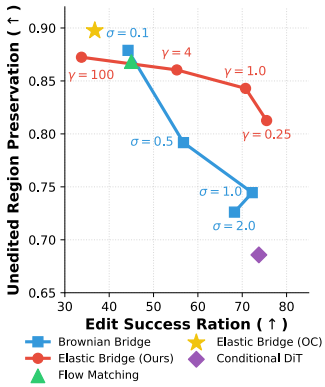

Figure 6: Edit success vs. unedited-region preservation across tested noise settings.

5.3 Ablation Studies

Oracle vs. Source-Grounded. We compare two training–inference designs on the same benchmark. *Oracle* trains with ground-truth elasticity K but uses online estimates $\hat{K}_t$ at inference (Equation (10)), whereas *Source-Grounded* uses the source-conditioned $\hat{K}$ in both phases (Section 4.2). Neither variant accesses the target image at inference; Oracle denotes oracle training, not a target-aware inference upper bound. The full parameter sweep is in Table 7. Table 2 shows that Source-Grounded improves editing accuracy while maintaining competitive perceptual quality, supporting a shared elasticity construction across training and inference.

Sensitivity to γ. We evaluate Elastic Bridge on ImgEdit-Bench with $\gamma \in \{100, 4, 1, 0.25\}$. As shown in Table 3, the ImgEdit Score increases from 1.95 at γ=100 to 3.58 at γ=1, then decreases slightly to 3.52 at γ=0.25. Thus, γ=1 achieves the highest ImgEdit Score among the tested settings on this benchmark.

Table 4: Evaluation on public benchmarks under two training settings. The upper group learns image editing from the SD3 text-to-image backbone; the lower group adapts the pretrained Qwen-Image-Edit model into Brownian and Elastic bridges. Results are compared within each group.

Method	MagicBrush					ImgEdit-Bench	Kontext Bench
	L1 ↓	L2 ↓	CLIP-I ↑	CLIP-T ↑	DINO ↑	OVERALL ↑	AVG (%) ↑
Learning image editing (SD3 text-to-image backbone)							
Conditional DiT	0.0977	0.0331	0.8425	**0.2150**	0.6778	3.04	43.8
Brownian Bridge	0.0650	0.0184	0.8783	0.2035	0.7522	3.47	49.2
Elastic Bridge (Ours)	**0.0484**	**0.0155**	**0.9176**	0.1970	**0.8332**	**3.58**	**52.2**
Adapting a pretrained editing model (Qwen-Image-Edit 20B)							
Qwen-Image-Edit (Base)	0.0790	0.0273	0.9020	**0.2024**	0.8087	**3.90**	57.50
ViBT (Brownian)	0.0836	0.0220	0.8682	0.2007	0.7555	3.76	56.63
ViBT (Elastic, Ours)	**0.0555**	**0.0175**	**0.9252**	0.1995	**0.8443**	3.84	**58.28**

5.4 Extended Experiments

Evaluation on Public Benchmarks. We evaluate on MagicBrush (Zhang et al., 2023), ImgEdit-Bench (Ye et al., 2025), and Kontext Bench (Labs et al., 2025), which are out-of-distribution relative to our editing training data. MagicBrush measures fidelity to reference edited images and text alignment (CLIP-T), while ImgEdit-Bench combines instruction adherence, editing quality, and detail preservation; these complement the main comparison's source-relative SSIM and instruction-following Accuracy. In the SD3 group of Table 4, Elastic Bridge leads in reference-image fidelity and both editing benchmark scores, showing that its benefits extend to external benchmarks. CLIP-T is lower than both baselines.

Paradigm Adaptation. We next start from `Qwen-Image-Edit` (Wu et al., 2025), a pretrained Conditional DiT that already has image-editing capability. Following the adaptation protocol of ViBT (Tan et al., 2025b), we convert it into Brownian Bridge and Elastic Bridge models through lightweight post-training, obtaining ViBT (Brownian) and ViBT (Elastic). This setting tests how well each bridge retains existing editing capability while changing the transport paradigm from noise-to-data to data-to-data. The lower group of Table 4 compares both adaptations with the base model on the same three public benchmarks.

Both adaptations eliminate the need for conditional image tokens, achieving a 2.2× speedup in training and inference. On MagicBrush, ViBT (Elastic) achieves the best L1, L2, CLIP-I, and DINO within the Qwen group, while CLIP-T is slightly lower than both alternatives. On Kontext Bench, it scores 58.28%, exceeding ViBT (Brownian) at 56.63% and the base model at 57.50%. On ImgEdit-Bench, ViBT (Elastic) scores 3.84, with an adaptation cost of only ↓0.06 from the base model's 3.90, compared with ↓0.14 for ViBT (Brownian) at 3.76. Thus, Elastic Bridge retains more of the base model's editing performance under the same adaptation protocol. Table 5 compares the overall ImgEdit Score with additional public baselines; the category-level breakdown is provided in Table 11 in the appendix. Figure 5 shows qualitative results of the adapted Elastic Bridge across diverse editing tasks.

Table 5: Overall comparison on *ImgEdit-Bench*. ImgEdit Score reports the overall score.

Model	ImgEdit Score ↑
MagicBrush	1.83
Ins.Pix2Pix	1.88
AnyEdit	2.45
Step1X-Edit	3.06
UniWorld-V1	3.26
FLUX Kontext	3.71
ViBT (Brownian)	3.76
ViBT (Elastic)	3.84
Qwen-Image-Edit (Base)	**3.90**

6 Conclusion and Limitations

Elastic Bridge resolves the preservation-diversity tension caused by a global σ through an adaptive map $K \propto |\delta|$ that equalizes SNR. Source-Grounded Elasticity closes the train-inference gap with one forward pass at $t{=}0$, yielding a parameter-free drop-in replacement for standard bridge SDEs. Experiments show the best accuracy-preservation trade-off among compared paradigms, supported by external public benchmarks and effective adaptation of a pretrained editing model to bridges.

Our fixed elasticity estimate and shared γ leave room for refinement on complex edits. Future work could explore adaptive elasticity estimation and category-aware SNR control.

AI USE STATEMENT

We used generative AI tools to assist with manuscript editing and LaTeX preparation. For benchmark construction, Qwen-VL and Qwen-Image-Edit generated editing instructions and target images, respectively; a vision-language model also judged instruction following during evaluation. The models, data filtering, and judging protocol are described in Section 5.1 and Appendix A.2. The authors take responsibility for the final text, claims, and results of this work.

ETHICS STATEMENT

This work studies instruction-guided image editing using public image datasets and model-generated image pairs. Like other image editing systems, it may be misused for deceptive manipulation or non-consensual edits, and may inherit biases from its data and pretrained models. Responsible use should respect image rights and privacy, disclose AI-edited content where appropriate, and include safeguards against harmful editing requests.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 define Elastic Bridge and Source-Grounded Elasticity. Section 5.1 and Appendix A.2 document training and inference settings, benchmark construction, evaluation metrics, the accuracy judge prompt, and the preservation-mask procedure. Appendix A.3 provides the SNR derivation, and Appendix A.4 gives the complete training and inference algorithms.

REFERENCES

Michael S Albergo, Nicholas M Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *arXiv preprint arXiv:2303.08797*, 2023.

Qingyan Bai, Qiuyu Wang, Hao Ouyang, Yue Yu, Hanlin Wang, Wen Wang, Ka Leong Cheng, Shuailei Ma, Yanhong Zeng, Zichen Liu, Yinghao Xu, Yujun Shen, and Qifeng Chen. Scaling instruction-based video editing with a high-quality synthetic dataset. *arXiv preprint arXiv:2510.15742*, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

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

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. Lbm: Latent bridge matching for fast image-to-image translation. *arXiv preprint arXiv:2503.07535*, 2025.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

Valentin De Bortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in neural information processing systems*, 34:17695–17709, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

Tsu-Jui Fu, Wenze Hu, Xianzhi Du, William Yang Wang, Yinfei Wang, and Yinfei Yang. Guiding instruction-based image editing via multimodal large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

Igor Vladimirovich Girsanov. On transforming a certain class of stochastic processes by absolutely continuous substitution of measures. *Theory of Probability & Its Applications*, 5(3):285–301, 1960.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Tiankai Hang, Shuyang Gu, Chen Li, Jianmin Bao, Dong Chen, Han Hu, Xin Geng, and Baining Guo. Efficient diffusion training via min-snr weighting strategy. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7441–7451, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pp. 2366–2369. IEEE, 2010.

Yujia Hu, Songhua Liu, Zhenxiong Tan, Xingyi Yang, and Xinchao Wang. Image editing as programs with diffusion models. *arXiv preprint arXiv:2506.04158*, 2025.

Yuzhou Huang, Liangbin Wang, Sheng Chen, Hao Yang, Ming Yan, Jianwei Wu, Hao Zou, Ji Zhou, Baigui Zhao, Jiesi Lin, 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*, pp. 8358–8368, 2024.

Haorui Ji, Tao Jun Lin, and Hongdong Li. Dpbridge: Latent diffusion bridge for dense prediction. *ArXiv*, abs/2412.20506, 2024. URL https://api.semanticscholar.org/CorpusID:275134211.

Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M. Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4401–4410, 2019.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International journal of computer vision*, 128(7):1956–1981, 2020.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Eungbean Lee, Somi Jeong, and Kwanghoon Sohn. Ebdm: exemplar-guided image translation with brownian-bridge diffusion models. In *European Conference on Computer Vision*, pp. 306–323. Springer, 2024.

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*, 2023.

Zongjian Li, Zheyuan Liu, Qihui Zhang, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Yang Ye, Wangbo Yu, Yuwei Niu, and Li Yuan. Uniworld-v2: Reinforce image editing with diffusion negative-aware finetuning and mllm implicit feedback. *arXiv preprint arXiv:2510.16888*, 2025.

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

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I ^{2} sb: Image-to-image schr\" odinger bridge. *arXiv preprint arXiv:2302.05872*, 2023.

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

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022.

Gisiro Maruyama. Continuous markov processes and stochastic equations. *Rendiconti del Circolo Matematico di Palermo*, 4(1):48–90, 1955.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing*, 21(12):4695–4708, 2012.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. *Advances in neural information processing systems*, 29, 2016.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

David Serrano-Lozano, Anand Bhattad, Luis Herranz, Jean-François Lalonde, and Javier Vazquez-Corral. Synclight: Controllable and consistent multi-view relighting. *arXiv preprint arXiv:2601.16981*, 2026.

Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36:62183–62223, 2023.

Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14940–14950, 2025a.

Zhenxiong Tan, Zeqing Wang, Xingyi Yang, Songhua Liu, and Xinchao Wang. Vision bridge transformer at scale. *arXiv preprint arXiv:2511.23199*, 2025b.

Viacheslav Vasilev, Arseny Ivanov, Nikita Gushchin, Maria Kovaleva, and Alexander Korotin. Time-correlated video bridge matching. *ArXiv*, abs/2510.12453, 2025. URL https://api.semanticscholar.org/CorpusID:282064795.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, 2023.

Peiyong Wang, Bohan Xiao, Qisheng He, Carri Glide-Hurst, and Ming Dong. Score-based image-to-image brownian bridge. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10765–10773, 2024a.

Yuji Wang, Zehua Chen, Xiaoyu Chen, Yixiang Wei, Jun Zhu, and Jianfei Chen. Framebridge: Improving image-to-video generation with bridge models. *arXiv preprint arXiv:2410.15371*, 2024b.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. *arXiv preprint arXiv:2505.20275*, 2025.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26125–26135, 2025.

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.

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

Kaiwen Zheng, Guande He, Jianfei Chen, Fan Bao, and Jun Zhu. Diffusion bridge implicit models. *arXiv preprint arXiv:2405.15885*, 2024.

Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. *ArXiv*, abs/2309.16948, 2023. URL https://api.semanticscholar.org/CorpusID:263310775.