Stabilizing Holistic Semantics in Diffusion Bridge for Image Inpainting

Jinjia Peng[1] , **Mengkai Li**[1*] , **Huibing Wang**[2*]
[1]School of Cyber Security and Computer, Hebei University, China
[2]College of Information Science and Technology, Dalian Maritime University, China
pengjinjia@hbu.edu.cn, limengkai@stumail.hbu.edu.cn, huibing.wang@dlmu.edu.cn

Abstract

Image inpainting aims to restore the original image from a damaged version. Recently, a special type of diffusion bridge model has achieved promising performance by directly mapping the degradation process and restoring corrupted images through the corresponding reverse process. However, due to the lack of explicit semantic priors during the denoising process, the inpainted results typically exhibit inferior context-stability and semantic consistency. To this end, this paper proposes a novel Global Structure-Guided Diffusion Bridge framework (GSGDiff), which incorporates an additional structure restorer to stabilize the generation of holistic semantics. Specifically, to acquire richer semantic structure priors, this paper proposes a posterior sampling approach that captures semantically global and consistent structures at each timestep, efficiently integrating them into the texture generation through the corresponding guidance module. Additionally, considering the characteristics of diffusion models with low denoising levels at larger timesteps, this paper proposes a semantic fusion schedule to avoid noise interference by reducing the weight of ineffective guided semantics in the early stages. By applying the proposed posterior sampling to the texture denoising process, GSGDiff can achieve more stable and superior inpainting results over competitive baselines. Experiments on Places2, Paris Street View and CelebA-HQ datasets validate the efficacy of the proposed method.

1 Introduction

Image inpainting refers to reconstructing a high-quality image from an incomplete one, and it has a wide range of applications in many fields such as image editing [Li *et al.*, 2021], artifact restoration [Quan *et al.*, 2024], and object removal. It is an inverse problem with an ill-posed nature. To resolve this challenging issue, conventional algorithms [Criminisi *et al.*, 2004; Komodakis and Tziritas, 2007; Barnes *et al.*, 2009] mainly utilize low-level visual assumptions to heuristically reconstruct the damaged regions or search and copy similar image patches from of the undamaged source image to fill the target region. However, due to limited feature representation, these methods typically struggle to generate accurate semantics within the hole. To this end, later works [Krizhevsky *et al.*, 2012; Liu *et al.*, 2021; Liu *et al.*, 2024b; Suvorov *et al.*, 2022; Zeng *et al.*, 2022; Yu *et al.*, 2018; Ko and Kim, 2023; Li *et al.*, 2022; Yao *et al.*, 2024; Peng *et al.*, 2023] have attempted to design advanced components to enhance feature representations or introduce self-attention mechanisms into GAN-based [Goodfellow *et al.*, 2014] conditional generative models, achieving better performance. Nonetheless, such strategies often lead to weak semantics correlation among various patches within the masked regions. To address this, some approaches [Nazeri *et al.*, 2019; Dong *et al.*, 2022; Liu *et al.*, 2022; Wang *et al.*, 2023] explore the utilization of additional sparse structure priors as a means to strengthen the correlations between the inpainted and masked regions. However, due to their limited semantic generation capabilities and heavy reliance on the semantic consistency between structure priors and texture, these methods inevitably suffer from meaningless artifacts.

Recently, diffusion models [Ho *et al.*, 2020; Song *et al.*, 2021] have shown state-of-the-art performance in generative tasks, exhibiting excellent semantic generation capability and pattern convergence, thus effectively addressing poor semantic generation in image inpainting [Lugmayr *et al.*, 2022; Luo *et al.*, 2023; Liu *et al.*, 2024a; Yue *et al.*, 2024; Wang *et al.*, 2022]. Within these models, a notable development is the diffusion bridges [Liu *et al.*, 2023; Shi *et al.*, 2024; Zhou *et al.*, 2024; Han *et al.*, 2025], which skillfully integrate the end-to-end training paradigm of CNN-based models with the denoising concept of diffusion models, creating a point-to-point diffusion process between original and damaged images. Among these developments, GOUB [Yue *et al.*, 2024] proposes applying Doob's h-transform to the Generalized Ornstein-Uhlenbeck (GOU) process, achieving superiority over other diffusion bridge models. However, due to the absence of richer contextual semantics in the denoising process, all of the above diffusion-based methods tend to either exhibit poor semantics consistency or fill the masked area with inferior context-stability (see Figures 1 (a) and (b)).

[*]Corresponding author.

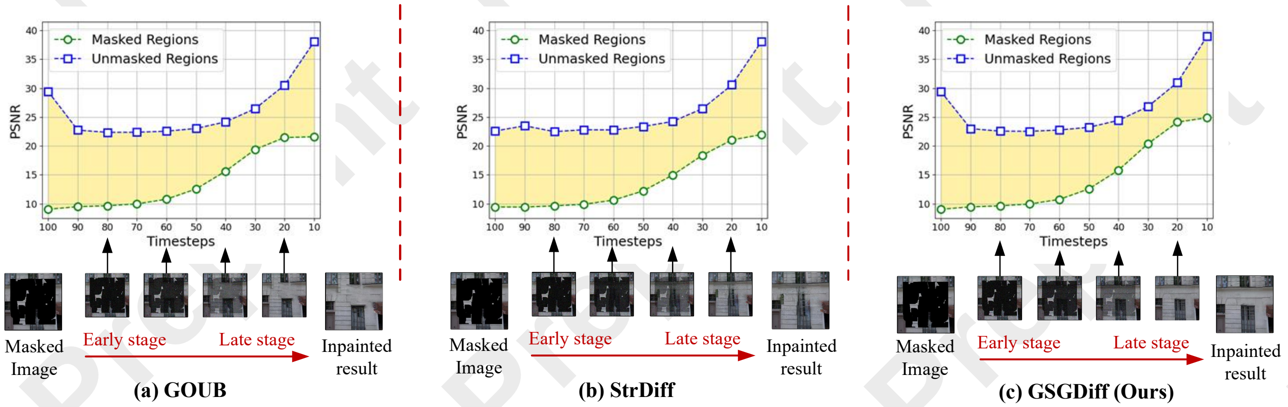


Figure 1: Illustration of the denoised results for existing arts on PSV [Doersch *et al.*, 2012], *e.g.,* GOUB [Yue *et al.*, 2024] (a) fails to generate consistent semantics between the masked and unmasked regions due to missing structure guidance; equipped with the guidance of sparse consistent structure, StrDiff [Liu *et al.*, 2024a] (b) tackles the semantic discrepancy with consistent structure guidance but introduces blurring and artifacts. Our GSGDiff (c) achieves coherent denoised results via auxiliary global and consistent structure guidance.

This, in practice, limits their performance capabilities, especially in complex degradation scenarios.

In this paper, we are committed to stabilizing the holistic semantics by exploiting additional semantic priors in texture denoising and propose an intuitive and effective inpainting architecture named the Global Structure-Guided Diffusion Bridge framework (GSGDiff). Specifically, we employ a guidance scheme that diffuses from grayscale images to masked edge maps within the structure denoising network to produce auxiliary structural information. To obtain global semantic prior over time, we propose a posterior sampling approach for the generalized ornstein-uhlenbeck bridge model. This approach enables the structure denoising network to capture global and consistent semantics at each timestep. Moreover, given the unique characteristics of the diffusion models, where higher values of timesteps correspond to lower levels of denoising, the guidance semantics often contain more irrelevant noise. To mitigate this issue, we propose a semantic fusion schedule that reduces the weight of ineffective semantics in the early stages to improve the effectiveness of guided information. With the assistance of structures in the texture denoise process, GSGDiff can generate meaningful results.

The main contributions of the paper are as follows:

- This paper proposes a novel Global Structure-Guided Diffusion Bridge framework that leverages a pre-trained auxiliary network to acquire structure semantics and efficiently injects them into texture generation, achieving superior restoration in texture and structure.
- To obtain time-dependent global semantics guidance, this paper proposes a posterior sampling approach tailored to the generalized ornstein-uhlenbeck bridge model, allowing the structure denoising network to capture global and consistent semantics at each timestep.
- Considering the lower denoising capability in the initial stages of denoising, this paper proposes a semantic fusion schedule to improve guided semantics more effectively while reducing the impact of irrelevant noise on texture generation.

2 Methodology

2.1 Preliminaries: Generalize Ornstein-Uhlenbeck Bridge for Image inapinting

Given a ground-truth image $\mathbf{I_{gt}} \in \mathbb{R}^{3\times H\times W}$ and a binary mask $\mathbf{M} \in \mathbb{R}^{1\times H\times W}$ that indicates the region to be inpainted, the goal of image inpainting is to reconstruct the masked image $\mathbf{I_m} = \mathbf{I_{gt}} \odot \mathbf{M}$ into a fully inpainted image. Typical diffusion bridge models for inpainting consist of two main processes: the forward texture diffusion process and the reverse texture denoising process.

For the forward diffusion process, given an initial texture state $\mathbf{x}_0$ representing the ground truth image $\mathbf{I_{gt}}$ and a final state $\mathbf{x}_T$ representing the corresponding masked image $\mathbf{I_m}$, with both ends of the bridge (*i.e.,* $\mathbf{x}_0$ *and* $\mathbf{x}_T$) known, the diffusion process $\{\mathbf{x}_t\}_{t=0}^{T}$ for any time $t \in [0, T]$ can be expressed as:

$$\begin{aligned} \mathrm{d}\mathbf{x}_t &= \left(\theta_t + g_t^2 \frac{e^{-2\bar{\theta}_{t:T}}}{\bar{\sigma}_{t:T}^2}\right)(\mathbf{x}_T - \mathbf{x}_t)\mathrm{d}t + g_t \mathrm{d}\mathbf{w}_t, \\ \bar{\theta}_{t:T} &= \int_t^T \theta_z dz, \quad \bar{\sigma}_{t:T}^2 = \frac{g_T^2}{2\theta_T}\left(1 - e^{-2\bar{\theta}_{t:T}}\right), \end{aligned} \tag{1}$$

where θ_t and g_t refer to the drift and diffusion coefficients, respectively. They are positive time-dependent parameters that satisfy $2\lambda^2 = g_t^2/\theta_t$. λ^2 is a given constant scalar and $\mathbf{w}_t$ is a standard Brownian motion that introduces randomness to the differential equation. The stochastic differential equation (SDE) defined above will necessarily pass through the final state $\mathbf{x}_T$ when $t = T$, creating a bridge connecting the points $\mathbf{x}_0$ and $\mathbf{x}_T$, hence this type of model is referred to as a diffusion bridge model. The forward process at any given moment t can be defined as follows:

$$\begin{aligned} p(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_T) &= N(\bar{\mathbf{m}}'_t, \bar{\sigma}_t'^2 \mathbf{I}), \\ \bar{\mathbf{m}}'_t = e^{-\bar{\theta}_t} \frac{\bar{\sigma}_{t:T}^2}{\bar{\sigma}_T^2}\mathbf{x}_0 &+ \left[\left(1 - e^{-\bar{\theta}_t}\right)\frac{\bar{\sigma}_{t:T}^2}{\bar{\sigma}_T^2} + e^{-2\bar{\theta}_{t:T}}\frac{\bar{\sigma}_t^2}{\bar{\sigma}_T^2}\right]\mathbf{x}_T \\ \bar{\sigma}_t'^2 &= \frac{\bar{\sigma}_t^2 \bar{\sigma}_{t:T}^2}{\bar{\sigma}_T^2} \end{aligned} \tag{2}$$

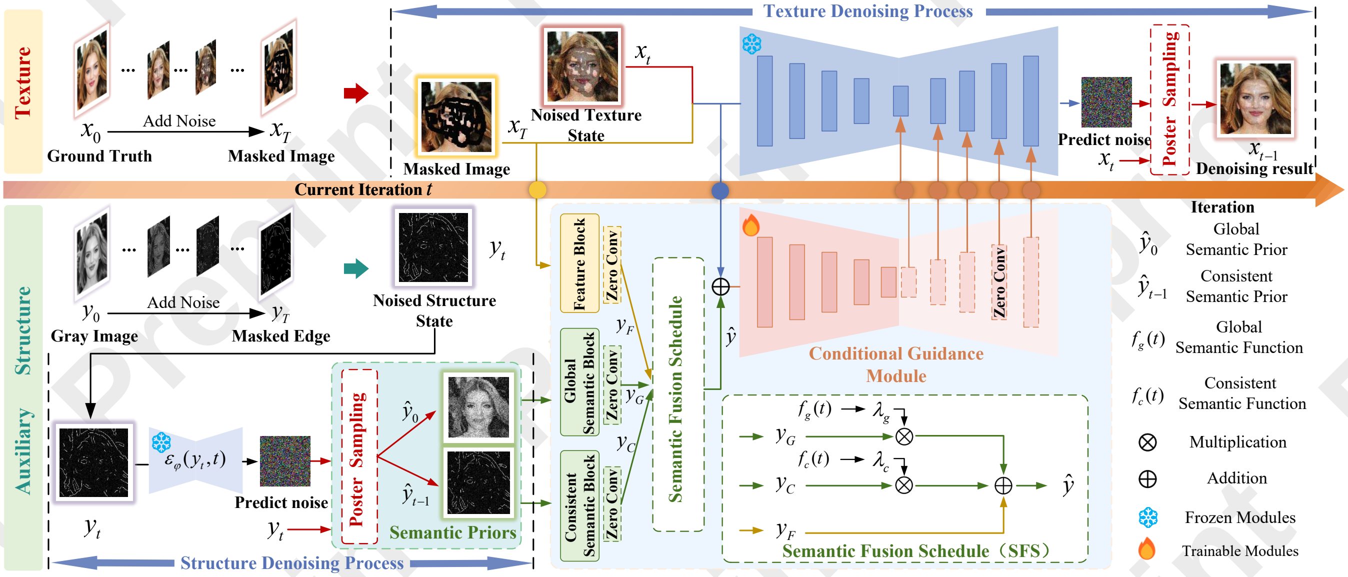


Figure 2: Overview pipeline of proposed GSGDiff, which utilizes a posterior sampling approach in the structure denoising network to generate semantically global and consistent structure priors. These semantic priors extract feature representations through their respective Semantic Blocks and then perform feature fusion. In this process, a Semantic Fusion Schedule is used to mitigate the impact of invalid noise within guidance semantics in the early denoising stages. Subsequently, the guidance information is injected into the texture denoising network through a Conditional Guidance Module, ultimately yielding consistent and meaningful denoising results.

For the reverse texture denoising process, we can reverse the diffusion SDE from Eq. (1) to denoise starting from the final state $\mathbf{x}_T$ and obtain the restored image $\mathbf{x}_0$, this process can be defined as:

$$\mathrm{d}\mathbf{x}_t = \left[\left(\theta_t + g_t^2 \frac{e^{-2\bar{\theta}_{t:T}}}{\bar{\sigma}_{t:T}^2} \right) (\mathbf{x}_T - \mathbf{x}_t) \right. \\ \left. - g_t^2 \nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{x}_T) \right] \mathrm{d}t + g_t \mathrm{d}\mathbf{w}_t. \tag{3}$$

Since the conditional score function is unknown, we can train a conditional time-dependent neural network by parameterizing the noise as $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{x}_T, t)$ to estimate the score. Thus, the ultimate training objective can be expressed as:

$$\mathcal{L} = \mathbb{E}_{t,\mathbf{x}_0,\mathbf{x}_t,\mathbf{x}_T} \left[\frac{1}{2g_t^2} \left\| \frac{1}{\bar{\sigma}_t'^2} \left[\bar{\sigma}_{t-1}'^2 (\mathbf{x}_t - b\mathbf{x}_T) a \right. \right. \right. \\ \left. + (\bar{\sigma}_t'^2 - \bar{\sigma}_{t-1}'^2 a^2) \bar{\mathbf{m}}_t' \right] - \mathbf{x}_t \\ \left. \left. + \left(\theta_t + g_t^2 \frac{e^{-2\bar{\theta}_{t:T}}}{\bar{\sigma}_{t:T}^2} \right) (\mathbf{x}_T - \mathbf{x}_t) \quad + \frac{g_t^2}{\bar{\sigma}_t'} \boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{x}_T, t) \right\| \right], \tag{4}$$

where a and b is positive weight and the conditional score $\nabla_{\mathbf{x}_t} \log p(\mathbf{x}_t \mid \mathbf{x}_T) \approx \nabla_{\mathbf{x}_t} \log p_\theta(\mathbf{x}_t \mid \mathbf{x}_T) = -\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{x}_T, t)/\bar{\sigma}_t'$. Therefore, starting from masked image $\mathbf{x}_T$, we can recover $\mathbf{x}_0$ by utilizing Eq. (3) to perform reverse iteration.

2.2 How to obtain Time-dependent Global Semantics for Texture Denoising?

Recent work [Liu *et al.*, 2024a] indicates that progressively incorporating the semantically sparse structure into texture generation over time can encourage consistent semantics in the inpainted regions. However, in more complex degradation scenarios, merely relying on consistent semantics established at an early stage often fails to align the generated content with the global semantics well.

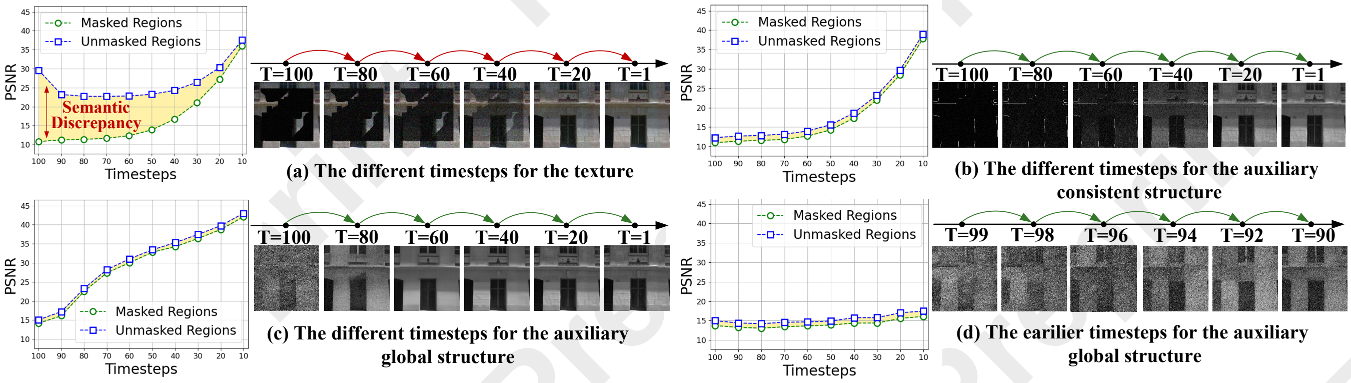
As shown in Figure 1, both GOUB [Yue *et al.*, 2024] (a), which lacks additional semantic prior, and StrDiff [Liu *et al.*, 2024a] (b), which relies on consistency-based semantics structure guidance, fail to achieve optimal inpainted result. To this end, an intuitive idea is to introduce global semantics, building upon the consistent sparse structure guidance provided in the early stage.

To achieve this, this paper proposes a posterior sampling approach for GOUB to generate richer global semantic prior from the posterior distribution $p(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{x}_T)$ conditioned on $\mathbf{x}_0$. Specifically, given an initial state $\mathbf{x}_0$ and a finite random diffusion state $\mathbf{x}_t$ at time $t \in [0, T]$, we can prove that the posterior of GOUB is tractable, and this posterior distribution is given by:

$$p(\mathbf{x}_{t-1} \mid \mathbf{x}_0, \mathbf{x}_t, \mathbf{x}_T) = \mathcal{N}(\mathbf{x}_{t-1} \mid \tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0, \mathbf{x}_T), \tilde{\beta}_t I). \tag{5}$$

From Bayes' formula, we can infer that:

$$p(\mathbf{x}_{t-1} \mid \mathbf{x}_0, \mathbf{x}_t, \mathbf{x}_T) = \frac{p(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_{t-1}, \mathbf{x}_T) p(\mathbf{x}_{t-1} \mid \mathbf{x}_0, \mathbf{x}_T)}{p(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_T)} \\ = \frac{p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_T) p(\mathbf{x}_{t-1} \mid \mathbf{x}_0, \mathbf{x}_T)}{p(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_T)}. \tag{6}$$



(a) The different timesteps for the texture

(b) The different timesteps for the auxiliary consistent structure

(c) The different timesteps for the auxiliary global structure

(d) The earilier timesteps for the auxiliary global structure

Figure 3: Illustration of the dense texture and auxiliary structures at various timesteps during training. It can be seen that in the early stages (large timesteps) the global structure (c) contains richer semantics compared to the sparse consistent structure (b). (d) shows a view of the earlier global structure when t decreases from 100 to 90.

Since $p(\mathbf{x}_{t-1} \mid \mathbf{x}_0, \mathbf{x}_T)$ and $p(\mathbf{x}_t \mid \mathbf{x}_0, \mathbf{x}_T)$ are Gaussian distributions (2), by employing the reparameterization technique, we can obtain $p(\mathbf{x}_t \mid \mathbf{x}_{t-1}, \mathbf{x}_T) = N(a\mathbf{x}_{t-1} + b\mathbf{x}_T, (\bar{\sigma}_t'^2 - a^2\bar{\sigma}_{t-1}'^2)\,\boldsymbol{I})$. Thus, this posterior distribution with mean and variance is given by:

$$\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0, \mathbf{x}_T) = \frac{1}{\bar{\sigma}_t'^2}\Big[\bar{\sigma}_{t-1}'^2(\mathbf{x}_t - b\mathbf{x}_T)a + (\bar{\sigma}_t'^2 - \bar{\sigma}_{t-1}'^2 a^2)\bar{\mathbf{m}}_t'\Big] \tag{7}$$

$$\text{and} \quad \tilde{\beta}_t = \frac{\bar{\sigma}_{t-1}'^2\left(\bar{\sigma}_t'^2 - a^2\bar{\sigma}_{t-1}'^2\right)}{\bar{\sigma}_t'^2}, \tag{8}$$

where,

$$a = \frac{e^{-\bar{\theta}_{t-1:t}}\bar{\sigma}_{t:T}^2}{\bar{\sigma}_{t-1:T}^2},$$

$$b = \frac{1}{\bar{\sigma}_T^2}\Big\{(1 - e^{-\bar{\theta}_t})\bar{\sigma}_{t:T}^2 + e^{-2\bar{\theta}_{t:T}}\bar{\sigma}_t^2 - \Big[(1 - e^{-\bar{\theta}_{t-1}})\bar{\sigma}_{t-1:T}^2 + e^{-2\bar{\theta}_{t-1:T}}\bar{\sigma}_{t-1}^2\Big]a\Big\}.$$

Moreover, thanks to the reparameterization technique ($\mathbf{x}_t = \bar{\mathbf{m}}_t' + \bar{\sigma}_t'\,\epsilon_t$), we can combine it with the noise prediction network $\tilde{\epsilon}_\phi(\mathbf{x}_t, \mathbf{x}_T, t)$ to estimate the variable $\mathbf{x}_0$ at time t:

$$\hat{\mathbf{x}}_\mathbf{0} = \Bigg[x_t - \left((1 - e^{-\bar{\theta}_t})\frac{\bar{\sigma}_{t:T}^2}{\bar{\sigma}_T^2} + e^{-2\bar{\theta}_{t:T}}\frac{\bar{\sigma}_t^2}{\bar{\sigma}_T^2}\right)\mathbf{x}_T - \bar{\sigma}_t'\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\mathbf{x}_t, \mathbf{x}_T, t)\Bigg]e^{\bar{\theta}_t\frac{\bar{\sigma}_T^2}{\bar{\sigma}_{t:T}^2}}, \tag{9}$$

where $\bar{\mathbf{m}}_t'$ and $\bar{\sigma}_t'$ are the forward mean and variance in Eq. (2). Then we can iteratively sample reverse states by combining Eq. (9) with Eq. (5) to construct the sampling process.

By applying this sampling strategy to the structure denoising U-Net $\boldsymbol{\epsilon}_{\boldsymbol{\varphi}}(\mathbf{y}_t, t)$, as depicted in Figure 2, we can obtain the global semantics $\hat{\mathbf{y}}_\mathbf{0}$ and the consistent semantics $\hat{\mathbf{y}}_{\mathbf{t-1}}$ at any time t to assist the texture generation.

2.3 Structure-Guided Denoising Process and Semantic Fusion Schedule

Considering that the semantic discrepancy between the masked and unmasked regions gradually increases as the timestep increases (see Figure 3 (a)), our aim is to alleviate this discrepancy in the early denoising through the aid of structures. This led to the introduction of the following:

How does the Structure Guide the Texture Denoising Process?

For the choice of guidance approaches, previous works [Liu *et al.*, 2024a; Dong *et al.*, 2022] used designed feature fusion strategies or modules to incorporate guidance information. However, these approaches typically require retraining the entire model, resulting in both limited flexibility and high computational costs. To address this, we introduce some new guidance modules inspired by ControlNet [Zhang *et al.*, 2023], which guide the model's generation without modifying the original pre-trained diffusion model, thus fully leveraging its capabilities.

Specifically, as illustrated in Figure 2, the global and consistent semantics $\hat{\mathbf{y}}_\mathbf{0}$ and $\hat{\mathbf{y}}_{\mathbf{t-1}}$ are first encoded by their corresponding semantic blocks. These extracted features, combined with features of available regions, serve as inputs to the Conditional Guidance Module (CGM). The CGM is specifically designed to process and integrate the guidance information. Its left side aligns with and is initialized from the downsampling module of the pre-trained texture denoising U-Net, while its right side comprises zero-initialized convolutional layers. Due to the zero initialization, the initial influence of the conditional guidance on the denoising process is zero, which allows the model to maintain stability during the early stages of training. As the model learns, it gradually incorporates the guidance information, enabling it to approach optimal parameters more stably.

How to avoid noise interference within the obtained structure priors?

Due to the limited denoising capabilities in the early stages, the obtained structure semantics usually lack effective information. As depicted in Figure 3 (b), as t decreases from

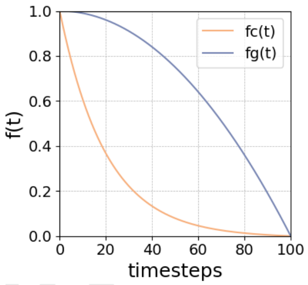


Figure 4: Illustration of the weight function for consistent and global features guidance.

100 to 60, the masked regions within the auxiliary consistent structure predominantly contain invalid noise, thus it is difficult to perform any guidance function. To minimize the influence of invalid information during training, we propose a specific Semantic Fusion Schedule for the extracted global and consistent structure features, aiming to assign appropriate weights to incorporate them into the fusion process. Specifically, for the weight function of the consistent features, we desire it to smoothly approximate zero in the early stages and eventually converge to one. Given the natural properties of the exponential function, it can be set as follows:

$$f_c(t) = e^{-at} - e^{-a}t, \tag{10}$$

where the parameter a is set to 5 for all experiments. Compared to the sparse consistent structure, the global structure contains richer semantics (see Figure 3 (b) and (c)). Noisy and ineffective information primarily appears in the earlier stages (e.g., when t decreases from 100 to 90), as illustrated in Figure 3 (d). Therefore, our idea is to initially assign higher weights to the global features, ensuring robust guidance from global semantics throughout training. The weight function for global features is given by:

$$f_g(t) = -t^2 + 1. \tag{11}$$

The function curves of the two weight functions are shown in Figure 4. Thus, the final fusion strategy is defined as follows:

$$\hat{y} = y_F + f_c(t) * y_C + f_g(t) * y_G. \tag{12}$$

As a result, each semantic block is updated only when the guidance features become more distinctive, effectively avoiding the impact of irrelevant information.

2.4 Sampling Strategy for Texture Denoising Process

Generally, we can utilize the trained network to generate high-quality images by sampling a state $\mathbf{x}_T$ and then iteratively solving the Eq. (3) with a numerical scheme. However, recent studies [Zhang *et al.*, 2024; Luo *et al.*, 2024] have shown that iterative sampling using the posterior sampling method during the reverse process can accelerate the convergence of the generative process and improve sample efficiency. Therefore, this paper applies the proposed posterior sampling strategy to texture denoising networks as well, aiming to achieve more excellent inpainting results. However, as the denoising proceeds, the correlation between the semantics generated by the auxiliary structure branch ($\hat{\mathbf{y}}_\mathbf{0}$ and $\hat{\mathbf{y}}_\mathbf{t-1}$) and the texture ($\hat{\mathbf{x}}_\mathbf{t-1}$) gradually weakens. Continually injecting the guidance semantics throughout the entire texture denoising process may result in color distortion and blurring in the masked regions [Liu *et al.*, 2024a].

To address this, we adopt a simple yet effective stage-wise semantic injection strategy, utilizing the stage point α to control the guided timesteps. Specifically, during the early denoising stages (i.e., when $t \in [T, \alpha]$), we leverage sparse consistent and global structures to generate coherent contents align well with the overall semantics. In the later stage (i.e., when $t \in (\alpha, 0]$), we rely solely on dense textures to generate meaningful semantic details. By doing so, we avoid semantic discrepancy between the generated textures and structures in the later stage, ultimately eliminating color distortion and achieving excellent results.

2.5 Overall Architecture

The pipeline of proposed GSGDiff is illustrated in Figure 2. Specifically, we first employ a pre-trained structure denoising U-Net $\boldsymbol{\epsilon}_{\boldsymbol{\varphi}}(\mathbf{y}_t, t)$, which models the transformation from gray image to masked edge image, to acquire structural semantic priors at each timestep. Then, these structural priors, combined with the masked image, are encoded by their respective feature extract blocks. Each block consists of four convolutional layers followed by a zero-initialized convolution layer and utilizes the SiLU [Elfwing *et al.*, 2018] activation function. Afterward, the guidance module takes these encoded features to guide the texture generation. We initialize the texture denoising network with a pre-trained model and freeze its parameters, preserving its base denoising capability while enabling guidance learning and reducing training overhead.

3 Experiments

3.1 Experimental Settings

We validate our method and various baselines on three typical datasets, including Paris Street View (PSV) [Doersch *et al.*, 2012], which consists of street photos taken in Paris and contains 14,900 training images and 100 validation images; CelebA-HQ [Karras *et al.*, 2018] is a dataset containing 30,000 high-quality human face images, divided into 28k training images and 2k validation images; Places2 [Zhou *et al.*, 2017] is a collection of more than 1.8 million natural images in multiple scenes. Following previous research, we use PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) [Wang *et al.*, 2004] to measure pixel and structural similarity. FID (Fréchet Inception Distance) [Heusel *et al.*, 2017] and LPIPS (Learned Perceptual Image Patch Similarity) [Zhang *et al.*, 2018] are used to measure perceptual disparity and visual effect.

3.2 Comparison with the State of the Arts

To validate GSGDiff's effectiveness, various typical inpainting models are compared, including LaMa [Suvorov *et al.*, 2022] which employs Fourier convolution to enlarge the receptive field; MAT [Li *et al.*, 2022], CMT [Ko and Kim, 2023] which uses a self-attention mechanism to model the long-distance dependencies between masked and unmasked

Places2		0.01%-20%				20%-40%				40%-60%			
Method	Venue	PSNR↑	SSIM↑	FID↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	LPIPS↓
LaMa [Suvorov *et al.*, 2022]	WACV'22	31.7873	0.9576	9.5640	0.0399	25.9291	0.8740	26.0854	0.1060	22.1447	0.7706	55.5161	0.1849
MAT [Li *et al.*, 2022]	ECCV'22	30.1129	0.9243	19.5104	0.0790	25.7473	0.8575	32.3014	0.1278	21.9842	0.7571	59.8215	0.2022
CMT [Ko and Kim, 2023]	ICCV'23	32.2212	0.9568	9.7308	0.0388	25.7010	0.8706	29.3242	0.1099	22.2333	0.7634	58.9052	0.1959
CTSDG [Guo *et al.*, 2021]	ICCV'21	30.8392	0.9520	16.0018	0.0523	24.7299	0.8544	53.6743	0.1492	21.2436	0.7362	102.6408	0.2500
DGTS [Liu *et al.*, 2022]	MM'22	32.1521	0.9577	7.6521	0.0338	25.6256	0.8732	26.7281	0.1007	21.2612	0.7552	69.9825	0.2042
ZITS [Dong *et al.*, 2022]	CVPR'22	32.0579	0.9575	8.6984	0.0384	26.2415	0.8758	23.5181	0.1026	22.1872	0.7700	50.2054	0.1825
Repaint [Lugmayr *et al.*, 2022]	CVPR'22	32.8635	0.9611	7.2329	0.0347	25.2324	0.8781	25.4316	0.1001	20.6880	0.7550	64.4779	0.2053
IR-SDE [Luo *et al.*, 2023]	ICML'23	33.3146	0.9632	6.6301	0.0320	25.6727	0.8830	24.0067	0.0915	20.9667	0.7656	59.0665	0.1821
StrDiff [Liu *et al.*, 2024a]	CVPR'24	33.3335	0.9625	8.2728	0.0330	26.3314	0.8770	29.0277	0.1050	21.6504	0.7581	64.4334	0.2028
GOUB [Yue *et al.*, 2024]	ICML'24	33.5159	0.9637	6.5709	0.0305	25.9980	0.8832	22.9658	0.0906	21.2422	0.7670	56.6660	0.1802
GSGDiff (Ours)	-	**34.0562**	**0.9688**	**5.7582**	**0.0237**	**26.6631**	**0.8946**	**21.6388**	**0.0814**	**22.2646**	**0.7882**	**49.5415**	**0.1669**
Paris Street View (PSV)		0.01%-20%				20%-40%				40%-60%			
Method	Venue	PSNR↑	SSIM↑	FID↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	LPIPS↓
CTSDG [Guo *et al.*, 2021]	ICCV'21	32.6122	0.9605	14.4908	0.0496	26.4840	0.8831	39.0581	0.1369	22.3543	0.7812	74.5036	0.2323
DGTS [Liu *et al.*, 2022]	MM'22	27.3439	0.8972	37.2752	0.1315	23.5567	0.8295	47.7383	0.1873	20.4743	0.7299	71.8430	0.2666
IR-SDE [Luo *et al.*, 2023]	ICML'23	33.2511	0.9586	13.2687	0.0466	26.8808	0.8836	35.3289	0.1216	22.9834	0.7811	64.3514	0.2135
StrDiff [Liu *et al.*, 2024a]	CVPR'24	32.9251	0.9552	15.5061	0.0503	26.7581	0.8704	39.6069	0.1389	23.2642	0.7630	76.8962	0.2387
GOUB [Yue *et al.*, 2024]	ICML'24	32.8552	0.9582	13.6456	0.0461	26.4787	0.8814	33.4046	0.1231	23.1824	0.7820	66.3349	0.2125
GSGDiff (Ours)	-	**33.2808**	**0.9622**	**12.2068**	**0.0421**	**27.0950**	**0.8913**	**32.4010**	**0.1186**	**23.7311**	**0.7981**	**63.8277**	**0.2082**
CelebA-HQ		0.01%-20%				20%-40%				40%-60%			
Method	Venue	PSNR↑	SSIM↑	FID↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	LPIPS↓	PSNR↑	SSIM↑	FID↓	LPIPS↓
LaMa [Suvorov *et al.*, 2022]	WACV'22	36.7654	0.9728	6.5111	0.0322	29.2867	0.9088	15.2416	0.0821	**24.6776**	0.8340	22.3708	0.1378
DGTS [Liu *et al.*, 2022]	MM'22	35.2689	0.9694	4.9180	0.0271	27.8547	0.9032	14.4271	0.0793	22.8301	0.8085	31.7918	0.1651
CMT [Ko and Kim, 2023]	ICCV'23	37.2173	0.9753	5.1643	0.0271	29.2500	0.9113	14.3871	0.0779	24.4221	0.8309	22.8132	0.1391
Repaint [Lugmayr *et al.*, 2022]	CVPR'22	36.2290	0.9730	5.2734	0.0291	27.8263	0.9035	13.9523	0.0832	22.7000	0.8111	23.7413	0.1524
IR-SDE [Luo *et al.*, 2023]	ICML'23	37.4388	0.9753	4.8706	0.0303	28.7787	0.9082	13.1708	0.0804	23.6836	0.8146	22.3751	0.1475
StrDiff [Liu *et al.*, 2024a]	CVPR'24	38.2595	0.9737	4.7226	0.0237	**29.4721**	0.9087	14.5869	0.0789	24.3714	0.8149	25.5039	0.1549
GOUB [Yue *et al.*, 2024]	ICML'24	37.4265	0.9755	4.5224	0.0284	28.5941	0.9076	13.7277	0.0796	24.0265	0.8158	21.2560	0.1461
GSGDiff (Ours)	-	**38.6694**	**0.9807**	**3.8291**	**0.0188**	29.1570	**0.9181**	**12.1898**	**0.0677**	24.3745	**0.8344**	**20.0816**	**0.1305**

Table 1: Comparison results on Places2, PSV and CelebA-HQ. The **bold** and underline indicate the best and the second best respectively.

regions; CTSDG [Guo *et al.*, 2021], DGTS [Liu *et al.*, 2022], ZITS [Dong *et al.*, 2022] which focuses on using the structure to assist texture generation; Repaint [Lugmayr *et al.*, 2022] and IR-SDE [Luo *et al.*, 2023], which benefit from DDPM but overlook the semantic consistency; StrDiff [Liu *et al.*, 2024a], which recently exploits consistent structure to assist texture generation, and GOUB [Yue *et al.*, 2024], which performs well as a diffusion bridge model in inpainting.

Quantitative Results Table 1 shows that GSGDiff achieves near-optimal performance across all three benchmark datasets as masking ratios increase. On the Places2 dataset, we achieve PSNR improvements of 1.6%, 2.6%, and 4.8% for three different masking ratios compared to GOUB. Additionally, for street-view and natural image scenarios, our method outperforms StrDiff in all metrics. This superiority is attributed to StrDiff's lack of holistic semantics in its denoising process, suggesting that our method is capable of generating more stable and semantically consistent contexts. For texture-rich face inpainting, GSGDiff demonstrates superiority over other methods in terms of perceptual performance, as evidenced by lower FID and LPIPS indices.

Qualitative Results In Figures 5 and 6, some visualization results are presented for our method and some representative methods. For complex structural degradation, our method avoids structural inconsistencies and blurring, yielding sharper results. Examples include the wall texture in the first row of Figure 5 and the second row of Figure 6. In addition, our method excels in restoring natural scenery, such as the wheat field and the chair depicted in Figure 5, demonstrating that our method can obtain more contextually semantic restoration results. For face restoration, GSGDiff also shows excellent texture generation ability, and compared with other methods, our method generates more natural face features with more realistic expressions.

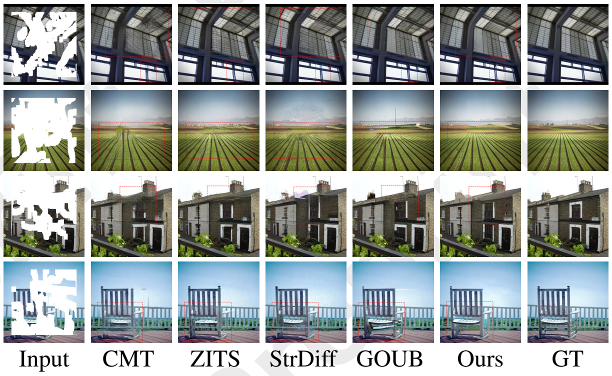


Figure 5: Qualitative comparison on the Places2 dataset among CMT [Ko and Kim, 2023], ZITS [Dong *et al.*, 2022], StrDiff [Liu *et al.*, 2024a], GOUB [Yue *et al.*, 2024], and our model.

3.3 Ablation and analysis

In this section, several ablated methods are conducted to assess the efficacy of the proposed algorithm and its components. We also analyze the selection of stage points α in texture denoising. Experiments use the PSV [Doersch *et al.*, 2012] dataset with mix mask ratios of 20%-60%, and the baseline model is obtained by removing the proposed prior guidance and the semantic fusion schedule from our model.

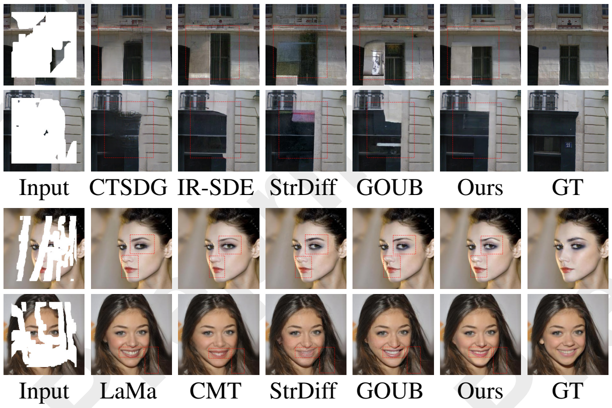


Figure 6: Qualitative comparison on the PSV and CelebA-HQ datasets among CTSDG [Guo *et al.*, 2021], IR-SDE [Luo *et al.*, 2023], LaMa [Suvorov *et al.*, 2022], CMT [Ko and Kim, 2023], StrDiff [Liu *et al.*, 2024a], GOUB [Yue *et al.*, 2024], and our model.

Setup	Semantic Prior		Fusion Schedule		Mix mask (20%-60%)			
	$\hat{y}_{t-1}$	$\hat{y}_0$	$f_c(t)$	$f_g(t)$	PSNR↑	SSIM↑	FID↓	LPIPS↓
Baseline					23.8064	0.8057	54.0858	0.1896
A	✓				24.4789	0.8081	53.2989	0.1865
B		✓			24.5202	0.8091	53.1829	0.1864
C	✓	✓			24.5131	0.8098	52.5058	0.1870
D	✓	✓	✓		24.6488	0.8106	51.6009	0.1857
E	✓	✓		✓	24.6257	0.8104	52.6605	0.1861
F	✓	✓	fusion w concat		24.6143	0.8091	51.8449	0.1874
Ours	✓	✓	✓	✓	**24.7712**	**0.8121**	**51.1518**	**0.1853**

Table 2: Ablation studies. Setup A only uses consistent semantics $\hat{y}_{t-1}$, B only uses global semantics $\hat{y}_0$, C uses both $\hat{y}_{t-1}$ and $\hat{y}_0$, D and E use respective semantic weighting functions, and F refers to replace the semantic fusion schedule with concatenating these semantic priors for fusion.

Semantic prior guidance Table 2 shows that both semantic priors yield improvements over the baseline model. The global prior notably enhances the performance, with model B achieving a gain of 0.7 dB in PSNR over the baseline. Notably, by combining consistent and global semantics, model C exhibits better FID and SSIM, but performs slightly worse on the other metrics. The reason for this is that the consistent semantics are too sparse in the early denoising stage and provide limited effective guidance to the texture denoising network. Hence, the performance improvement of the model C is not as significant as that of the model B.

Semantic fusion schedule Table 2 shows that both model D and model E show better performance than model C when the respective weighting functions are applied to the semantic priors. At the beginning of the denoising process, the consistent semantics are sparse compared to the global semantics. Thus, after adjusting the weights to reduce the interference of invalid noise, the model D exhibits higher performance metrics than the model E. Ultimately, our GSGDiff exhibits optimal performance when equipped with a corresponding fusion schedule for each semantic prior. In contrast, model F, which utilizes a direct concatenation of semantic priors, is shown to be inferior to our proposed semantic fusion schedule.

The analysis of sampling strategies in texture denoising To explore the practical performance of the proposed posterior sampling in texture denoising, we compare sampling strategies across models. The results, shown in Table 3, demonstrate that using our posterior sampling during inference significantly improves the model's performance, yielding better PSNR, SSIM, and LPIPS metrics.

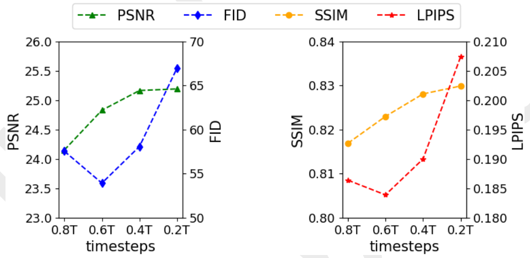


Figure 7: Comparison with different stage point α in texture denoising.

Methods	Mix Mask (20%-60%)			
	PSNR↑	SSIM↑	FID↓	LPIPS↓
GSGDiff	24.7712	0.8121	**51.1518**	0.1853
GSGDiff-posterior	**24.8301**	**0.8230**	53.9333	**0.1839**

Table 3: Comparison with sampling strategies in texture denoising. 'posterior' means we use the proposed posterior sampling approach for texture denoising.

The choices for stage point α In Figure 7, we set α to take values at [0.2T, 0.4T, 0.6T, 0.8T] to evaluate its effect on the final inpainted results. From the figure, it can be seen that the values of distortion metrics (PSNR and SSIM) tend to increase as the value of α decreases. Notably, the FID and LPIPS metrics show a significant advantage at $\alpha = 0.6T$; however, as the guided timestep is prolonged, the perceptual performance (FID and LPIPS) deteriorates. This is due to the correlation between the semantics generated by the auxiliary structure branch and the texture gradually weakens. Continually injecting the guidance semantics throughout the entire texture denoising process may result in color distortion and blurring. Considering the limitations of the traditional metrics (i.e., PSNR, SSIM) which tend to assign higher values to smoothed results and do not reflect human perception well [Zhang *et al.*, 2018], we finally choose $\alpha = 0.6T$, as the stage point for phased injection of guidance information.

4 Conclusion

We propose a novel diffusion bridge for inpainting, which aims to stabilize the inpainted result by integrating time-dependent holistic semantics in texture denoising. A posterior sampling approach is tailored to the Generalized Ornstein-Uhlenbeck bridge, which acquires semantically global and consistent structure priors. Considering the limited guidance caused by noise in the early denoising, a semantic fusion schedule is designed to reduce the weight of ineffective guided semantics. Experimental results demonstrate that our method achieves superior performance compared to previous methods.

Acknowledgments

This work was jointly supported by the National Key Research and Development Program of China Grant 2024YFB4710800; Liaoning Provincial Natural Science Foundation Grant 2024-MS-012; Dalian Science and Technology Talent Innovation Support Plan, Grant 2024RY010; Basic Research Project of Shijiazhuang Municipal Universities in Hebei Province (241791387A); Interdisciplinary Research Program of Hebei University (DXK202404).

References

[Barnes *et al.*, 2009] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009.

[Criminisi *et al.*, 2004] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004.

[Doersch *et al.*, 2012] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A Efros. What makes paris look like paris? *ACM Transactions on graphics*, 31(4), 2012.

[Dong *et al.*, 2022] Qiaole Dong, Chenjie Cao, and Yanwei Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11368, 2022.

[Elfwing *et al.*, 2018] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural networks*, 107:3–11, 2018.

[Goodfellow *et al.*, 2014] Ian 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.

[Guo *et al.*, 2021] Xiefan Guo, Hongyu Yang, and Di Huang. Image inpainting via conditional texture and structure dual generation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 14134–14143, 2021.

[Han *et al.*, 2025] Zihao Han, Baoquan Zhang, Lisai Zhang, Shanshan Feng, Kenghong Lin, Guotao Liang, Yunming Ye, Xiaochen Qi, and Guangming Ye. Asyncdsb: Schedule-asynchronous diffusion schrödinger bridge for image inpainting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.

[Heusel *et al.*, 2017] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

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

[Karras *et al.*, 2018] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[Ko and Kim, 2023] Keunsoo Ko and Chang-Su Kim. Continuously masked transformer for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13169–13178, 2023.

[Komodakis and Tziritas, 2007] Nikos Komodakis and Georgios Tziritas. Image completion using efficient belief propagation via priority scheduling and dynamic pruning. *IEEE Transactions on Image Processing*, 16(11):2649–2661, 2007.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[Li *et al.*, 2021] Honglei Li, Wenmin Wang, Cheng Yu, and Shixiong Zhang. Swapinpaint: Identity-specific face inpainting with identity swapping. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(7):4271–4281, 2021.

[Li *et al.*, 2022] Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. Mat: Mask-aware transformer for large hole image inpainting. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10758–10768, 2022.

[Liu *et al.*, 2021] Yi Liu, Dingwen Zhang, Qiang Zhang, and Jungong Han. Part-object relational visual saliency. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3688–3704, 2021.

[Liu *et al.*, 2022] Haipeng Liu, Yang Wang, Meng Wang, and Yong Rui. Delving globally into texture and structure for image inpainting. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 1270–1278, 2022.

[Liu *et al.*, 2023] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I2sb: image-to-image schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning*, pages 22042–22062, 2023.

[Liu *et al.*, 2024a] Haipeng Liu, Yang Wang, Biao Qian, Meng Wang, and Yong Rui. Structure matters: Tackling the semantic discrepancy in diffusion models for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8038–8047, 2024.

[Liu *et al.*, 2024b] Yi Liu, De Cheng, Dingwen Zhang, Shoukun Xu, and Jungong Han. Capsule networks with residual pose routing. *IEEE Transactions on Neural Networks and Learning Systems*, 2024.

[Lugmayr *et al.*, 2022] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF* *conference on computer vision and pattern recognition*, pages 11461–11471, 2022.

[Luo *et al.*, 2023] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Image restoration with mean-reverting stochastic differential equations. In *Proceedings of the 40st International Conference on Machine Learning*, pages 23045–23066, 2023.

[Luo *et al.*, 2024] Ziwei Luo, Fredrik K Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B Schön. Photo-realistic image restoration in the wild with controlled vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6641–6651, 2024.

[Nazeri *et al.*, 2019] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019.

[Peng *et al.*, 2023] Jinjia Peng, Guangqi Jiang, and Huibing Wang. Adaptive memorization with group labels for unsupervised person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10):5802–5813, 2023.

[Quan *et al.*, 2024] Weize Quan, Jiaxi Chen, Yanli Liu, Dong-Ming Yan, and Peter Wonka. Deep learning-based image and video inpainting: A survey. *International Journal of Computer Vision*, pages 1–34, 2024.

[Shi *et al.*, 2024] Yuyang Shi, Valentin De Bortoli, Andrew Campbell, and Arnaud Doucet. Diffusion schrödinger bridge matching. *Advances in Neural Information Processing Systems*, 36, 2024.

[Song *et al.*, 2021] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

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

[Wang *et al.*, 2004] 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.

[Wang *et al.*, 2022] Huibing Wang, Guangqi Jiang, Jinjia Peng, Ruoxi Deng, and Xianping Fu. Towards adaptive consensus graph: multi-view clustering via graph collaboration. *IEEE Transactions on Multimedia*, 25:6629–6641, 2022.

[Wang *et al.*, 2023] Huibing Wang, Mingze Yao, Guangqi Jiang, Zetian Mi, and Xianping Fu. Graph-collaborated auto-encoder hashing for multiview binary clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Yao *et al.*, 2024] Mingze Yao, Huibing Wang, Yawei Chen, and Xianping Fu. Between/within view information completing for tensorial incomplete multi-view clustering. *IEEE Transactions on Multimedia*, 2024.

[Yu *et al.*, 2018] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[Yue *et al.*, 2024] Conghan Yue, Zhengwei Peng, Junlong Ma, Shiyan Du, Pengxu Wei, and Dongyu Zhang. Image restoration through generalized ornstein-uhlenbeck bridge. In *Proceedings of the 41st International Conference on Machine Learning*, 2024.

[Zeng *et al.*, 2022] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Aggregated contextual transformations for high-resolution image inpainting. *IEEE Transactions on Visualization and Computer Graphics*, 2022.

[Zhang *et al.*, 2018] 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*, pages 586–595, 2018.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[Zhang *et al.*, 2024] Ruoqi Zhang, Ziwei Luo, Jens Sjölund, Thomas B Schön, and Per Mattsson. Entropy-regularized diffusion policy with q-ensembles for offline reinforcement learning. *Advances in neural information processing systems*, 2024.

[Zhou *et al.*, 2017] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1452–1464, 2017.

[Zhou *et al.*, 2024] Linqi Zhou, Aaron Lou, Samar Khanna, and Stefano Ermon. Denoising diffusion bridge models. In *The Twelfth International Conference on Learning Representations*, 2024.