Improving Diffusion-Based Image Synthesis with Context Prediction

Ling Yang[1]* **Jingwei Liu**[1]† **Shenda Hong**[1] **Zhilong Zhang**[1] **Zhilin Huang**[2]
Zheming Cai[1] **Wentao Zhang**[1] **Bin Cui**[1] ‡
[1]Peking University   [2] Tsinghua University
yangling0818@163.com, jingweiliu1996@163.com, zhilong.zhang@bjmu.edu.cn
{hongshenda, wentao.zhang, bin.cui}@pku.edu.cn

Abstract

Diffusion models are a new class of generative models, and have dramatically promoted image generation with unprecedented quality and diversity. Existing diffusion models mainly try to reconstruct input image from a corrupted one with a pixel-wise or feature-wise constraint along spatial axes. However, such point-based reconstruction may fail to make each predicted pixel/feature fully preserve its neighborhood context, impairing diffusion-based image synthesis. As a powerful source of automatic supervisory signal, *context* has been well studied for learning representations. Inspired by this, we for the first time propose CONPREDIFF to improve diffusion-based image synthesis with context prediction. We explicitly reinforce each point to predict its neighborhood context (*i.e.*, multi-stride features/tokens/pixels) with a context decoder at the end of diffusion denoising blocks in training stage, and remove the decoder for inference. In this way, each point can better reconstruct itself by preserving its semantic connections with neighborhood context. This new paradigm of CONPREDIFF can generalize to arbitrary discrete and continuous diffusion backbones without introducing extra parameters in sampling procedure. Extensive experiments are conducted on unconditional image generation, text-to-image generation and image inpainting tasks. Our CON-PREDIFF consistently outperforms previous methods and achieves a new SOTA text-to-image generation results on MS-COCO, with a zero-shot FID score of 6.21.

1 Introduction

Recent diffusion models [99, 5, 63, 4, 47, 10, 22] have made remarkable progress in image generation. They are first introduced by Sohl-Dickstein et al. [76] and then improved by Song & Ermon [78] and Ho et al. [28], and can now generate image samples with unprecedented quality and diversity [24, 68, 67]. Numerous methods have been proposed to develop diffusion models by improving their empirical generation results [53, 77, 79] or extending the capacity of diffusion models from a theoretical perspective [80, 81, 47, 46, 108]. We revisit existing diffusion models for image generation and break them into two categories, pixel- and latent-based diffusion models, according to their diffusing spaces. Pixel-based diffusion models directly conduct continuous diffusion process in the pixel space, they incorporate various conditions (*e.g.*, class, text, image, and semantic map) [29, 70, 51, 2, 66] or auxiliary classifiers [80, 14, 27, 54, 40] for conditional image generation.

On the other hand, latent-based diffusion models [65] conduct continuous or discrete diffusion process [87, 30, 1] on the semantic latent space. Such diffusion paradigm not only significantly reduces the

---

*Contact: Ling Yang, yangling0818@163.com.

†Contributed equally.

‡Corresponding Authors: Wentao Zhang, Bin Cui.

computational complexity for both training and inference, but also facilitates the conditional image generation in complex semantic space [62, 38, 58, 19, 98]. Some of them choose to pre-train an autoencoder [41, 64] to map the input from image space to the continuous latent space for continuous diffusion, while others utilize a vector quantized variational autoencoder [88, 17] to induce the token-based latent space for discrete diffusion [24, 75, 114, 85].

Despite all these progress of pixel- and latent-based diffusion models in image generation, both of them mainly focus on utilizing a point-based reconstruction objective over the spatial axes to recover the entire image in diffusion training process. This point-wise reconstruction neglects to fully preserve local context and semantic distribution of each predicted pixel/feature, which may impair the fidelity of generated images. Traditional non-diffusion studies [15, 45, 32, 50, 110, 8] have designed different *context*-preserving terms for advancing image representation learning, but few researches have been done to constrain on context for diffusion-based image synthesis.

In this paper, we propose CONPREDIFF to explicitly force each pixel/feature/token to predict its local neighborhood context (*i.e.*, multi-stride features/tokens/pixels) in image diffusion generation with an extra context decoder near the end of diffusion denoising blocks. This explicit *context prediction* can be extended to existing discrete and continuous diffusion backbones without introducing additional parameters in inference stage. We further characterize the neighborhood context as a probability distribution defined over multi-stride neighbors for efficiently decoding large context, and adopt an optimal-transport loss based on Wasserstein distance [21] to impose structural constraint between the decoded distribution and the ground truth. We evaluate the proposed CONPREDIFF with the extensive experiments on three major visual tasks, unconditional image generation, text-to-image generation, and image inpainting. Notably, our CONPREDIFF consistently outperforms previous diffusion models by a large margin regarding generation quality and diversity.

Our main contributions are summarized as follows: **(i):** To the best of our knowledge, we for the first time propose CONPREDIFF to improve diffusion-based image generation with context prediction; **(ii):** We further propose an efficient approach to decode large context with an optimal-transport loss based on Wasserstein distance; **(iii):** CONPREDIFF substantially outperforms existing diffusion models and achieves new SOTA image generation results, and we can generalize our model to existing discrete and continuous diffusion backbones, consistently improving their performance.

2   Related Work

Diffusion Models for Image Generation   Diffusion models [99, 76, 78, 28] are a new class of probabilistic generative models that progressively destruct data by injecting noise, then learn to reverse this process for sample generation. They can generate image samples with unprecedented quality and diversity [24, 68, 67], and have been applied in various applications [99, 9, 6]. Existing pixel- and latent-based diffusion models mainly utilize the discrete diffusion [30, 1, 24] or continuous diffusion [87, 65] for unconditional or conditional image generation [80, 14, 27, 54, 40, 68]. Discrete diffusion models were also first described in [76], and then applied to text generation in Argmax Flow [30]. D3PMs [1] applies discrete diffusion to image generation. VQ-Diffusion [24] moves discrete diffusion from image pixel space to latent space with the discrete image tokens acquired from VQ-VAE [88]. Latent Diffusion Models (LDMs) [87, 65] reduce the training cost for high resolution images by conducting continuous diffusion process in a low-dimensional latent space. They also incorporate conditional information into the sampling process via cross attention [89]. Similar techniques are employed in DALLE-2 [62] for image generation from text, where the continuous diffusion model is conditioned on text embeddings obtained from CLIP latent codes [59]. Imagen [68] implements text-to-image generation by conditioning on text embeddings acquired from large language models (*e.g.*, T5 [60]). Despite all this progress, existing diffusion models neglect to exploit rich neighborhood context in the generation process, which is critical in many vision tasks for maintaining the local semantic continuity in image representations [111, 45, 32, 50]. In this paper, we firstly propose to explicitly preserve local neighborhood context for diffusion-based image generation.

Context-Enriched Representation Learning   *Context* has been well studied in learning representations, and is widely proved to be a powerful automatic supervisory signal in many tasks. For example, language models [52, 13] learn word embeddings by predicting their context, *i.e.*, a few words before and/or after. More utilization of contextual information happens in visual tasks, where spatial context
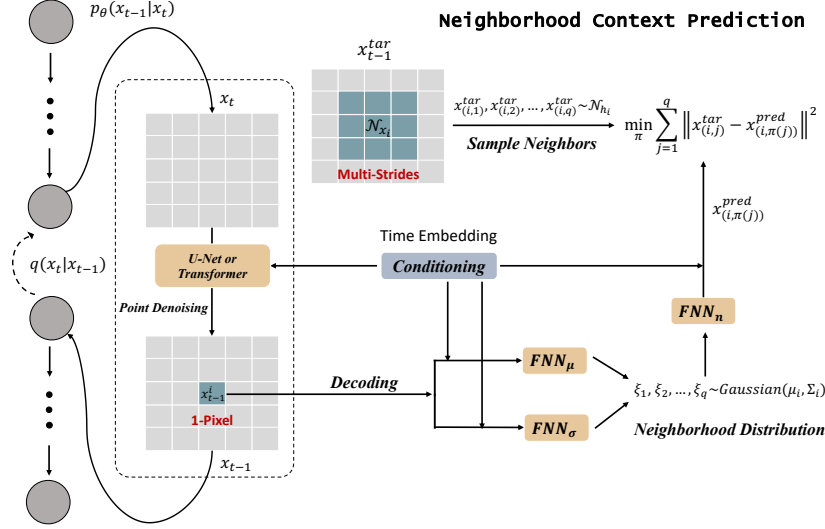
Figure 1: In training stage, CONPREDIFF first performs self-denoising as standard diffusion models, then it conducts neighborhood context prediction based on denoised point $\boldsymbol{x}_{t-1}^i$. In inference stage, CONPREDIFF only uses its self-denoising network for sampling.

is vital for image domain. Many studies [15, 111, 45, 32, 50, 110, 8, 106, 95, 94, 44] propose to leverage context for enriching learned image representations. Doersch et al. [15] and Zhang et al. [110] make predictions from visible patches to masked patches to enhance the self-supervised image representation learning. Hu et al. [32] designs local relation layer to model the context of local pixel pairs for image classification, while Liu et al. [45] preserves contextual structure to guarantee the local feature/pixel continuity for image inpainting. Inspired by these studies, in this work, we propose to incorporate neighborhood context prediction for improving diffusion-based generative modeling.

3 Preliminary

Discrete Diffusion We briefly review a classical discrete diffusion model, namely Vector Quantized Diffusion (VQ-Diffusion) [24]. VQ-Diffusion utilizes a VQ-VAE to convert images x to discrete tokens $x_0 \in \{1, 2, ..., K, K+1\}$, K is the size of codebook, and $K + 1$ denotes the [MASK] token. Then the forward process of VQ-Diffusion is given by:

$$q(x_t|x_{t-1}) = \boldsymbol{v}^\top(x_t)\boldsymbol{Q}_t\boldsymbol{v}(x_{t-1}) \tag{1}$$

where $\boldsymbol{v}(x)$ is a one-hot column vector with entry 1 at index x. And $\boldsymbol{Q}_t$ is the probability transition matrix from x_{t-1} to x_t with the mask-and-replace VQ-Diffusion strategy. In the reverse process, VQ-Diffusion trains a denoising network $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ that predicts noiseless token distribution $p_\theta(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t)$ at each step:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) = \sum_{\tilde{\boldsymbol{x}}_0=1}^{K} q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \tilde{\boldsymbol{x}}_0)p_\theta(\tilde{\boldsymbol{x}}_0|\boldsymbol{x}_t), \tag{2}$$

which is optimized by minimizing the following variational lower bound (VLB) [76]:

$$\mathcal{L}_{t-1}^{dis} = D_{KL}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \,||\, p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)). \tag{3}$$

Continuous Diffusion A continuous diffusion model progressively perturbs input image or feature map $\boldsymbol{x}_0$ by injecting noise, then learn to reverse this process starting from $\boldsymbol{x}_T$ for image generation. The forward process can be formulated as a Gaussian process with Markovian structure:

$$\begin{aligned} q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) &:= \mathcal{N}(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t\mathbf{I}), \\ q(\boldsymbol{x}_t|\boldsymbol{x}_0) &:= \mathcal{N}(\boldsymbol{x}_t; \sqrt{\overline{\alpha}_t}\boldsymbol{x}_0, (1-\overline{\alpha}_t)\mathbf{I}), \end{aligned} \tag{4}$$

where $\beta_1, \ldots, \beta_T$ denotes fixed variance schedule with $\alpha_t := 1 - \beta_t$ and $\overline{\alpha}_t := \prod_{s=1}^{t} \alpha_s$. This forward process progressively injects noise to data until all structures are lost, which is well approximated by $\mathcal{N}(0, \mathbf{I})$. The reverse diffusion process learns a model $p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t)$ that approximates the

3

true posterior:

$$p_\theta(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t) := \mathcal{N}(\boldsymbol{x}_{t-1}; \mu_\theta(\boldsymbol{x}_t), \Sigma_\theta(\boldsymbol{x}_t)), \tag{5}$$

Fixing Σ_θ to be untrained time dependent constants $\sigma_t^2 \boldsymbol{I}$, Ho *et al.* [28] improve the diffusion training process by optimizing following objective:

$$\mathcal{L}_{t-1}^{con} = \mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})} \left[\frac{1}{2\sigma_t^2} ||\mu_\theta(\boldsymbol{x}_t, t) - \hat{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0)||^2 \right] + C, \tag{6}$$

where C is a constant that does not depend on θ. $\hat{\mu}(\boldsymbol{x}_t, \boldsymbol{x}_0)$ is the mean of the posterior $q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0, \boldsymbol{x}_t)$, and $\mu_\theta(\boldsymbol{x}_t, t)$ is the predicted mean of $p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ computed by neural networks.

4 The Proposed CONPREDIFF

In this section, we elucidate the proposed CONPREDIFF as in Figure 1. In Sec. 4.1, we introduce our proposed context prediction term for explicitly preserving local neighborhood context in diffusion-based image generation. To efficiently decode large context in training process, we characterize the neighborhood information as the probability distribution defined over multi-stride neighbors in Sec. 4.2, and theoretically derive an optimal-transport loss function based on Wasserstein distance to optimize the decoding procedure. In Sec. 4.3, we generalize our CONPREDIFF to both existing discrete and continuous diffusion models, and provide optimization objectives.

4.1 Neighborhood Context Prediction in Diffusion Generation

We use unconditional image generation to illustrate our method for simplicity. Let $\boldsymbol{x}_{t-1}^i \in \mathbb{R}^d$ to denote i-th pixel of the predicted image, i-th feature point of the predicted feature map, or i-th image token of the predicted token map in spatial axes. Let $\mathcal{N}_i^s$ denote the s-stride neighborhoods of $\boldsymbol{x}_{t-1}^i$, and K denotes the total number of $\mathcal{N}_i^s$. For example, the number of 1-stride neighborhoods is $K = 8$, and the number of 2-stride ones is $K = 24$.

S-**Stride Neighborhood Reconstruction** Previous diffusion models make point-wise reconstruction, *i.e.*, reconstructing each pixel, thus their reverse learning processes can be formulated by $p_\theta(\boldsymbol{x}_{t-1}^i|\boldsymbol{x}_t)$. In contrast, our context prediction aims to reconstruct $\boldsymbol{x}_{t-1}^i$ and further predict its s-stride neighborhood contextual representations $\boldsymbol{H}_{\mathcal{N}_i^s}$ based on $\boldsymbol{x}_{t-1}^i$: $p_\theta(\boldsymbol{x}_{t-1}^i, \boldsymbol{H}_{\mathcal{N}_i^s}|\boldsymbol{x}_t)$, where p_θ is parameterized by two reconstruction networks (ψ_p, ψ_n). ψ_p is designed for the point-wise denoising of $\boldsymbol{x}_{t-1}^i$ in $\boldsymbol{x}_t$, and ψ_n is designed for decoding $\boldsymbol{H}_{\mathcal{N}_i^s}$ from $\boldsymbol{x}_{t-1}^i$. For denoising i-th point in $\boldsymbol{x}_t$, we have:

$$\boldsymbol{x}_{t-1}^i = \psi_p(\boldsymbol{x}_t, t), \tag{7}$$

where t is the time embedding and ψ_p is parameterized by a U-Net or transformer with an encoder-decoder architecture. For reconstructing the entire neighborhood information $\boldsymbol{H}_{\mathcal{N}_i^s}$ around each point $\boldsymbol{x}_{t-1}^i$, we have:

$$\boldsymbol{H}_{\mathcal{N}_i^s} = \psi_n(\boldsymbol{x}_{t-1}^i, t) = \psi_n(\psi_p(\boldsymbol{x}_t, t)), \tag{8}$$

where $\psi_n \in \mathbb{R}^{Kd}$ is the neighborhood decoder. Based on Equation (7) and Equation (8), we unify the point- and neighborhood-based reconstruction to form the overall training objective:

$$\mathcal{L}_{\text{CONPREDIFF}} = \sum_{i=1}^{x \times y} \left[\underbrace{\mathcal{M}_p(\boldsymbol{x}_{t-1}^i, \hat{\boldsymbol{x}}^i)}_{point\ denoising} + \underbrace{\mathcal{M}_n(\boldsymbol{H}_{\mathcal{N}_i^s}, \hat{\boldsymbol{H}}_{\mathcal{N}_i^s})}_{context\ prediction} \right], \tag{9}$$

where x, y are the width and height on spatial axes. $\hat{\boldsymbol{x}}^i$ ($\hat{\boldsymbol{x}}_0^i$) and $\hat{\boldsymbol{H}}_{\mathcal{N}_i^s}$ are ground truths. $\mathcal{M}_p$ and $\mathcal{M}_n$ can be Euclidean distance. In this way, CONPREDIFF is able to maximally preserve local context for better reconstructing each pixel/feature/token.

Interpreting Context Prediction in Maximizing ELBO We let $\mathcal{M}_p, \mathcal{M}_n$ be square loss, $\mathcal{M}_n(\boldsymbol{H}_{\mathcal{N}_i^s}, \hat{\boldsymbol{H}}_{\mathcal{N}_i^s}) = \sum_{j \in \mathcal{N}_i} (\boldsymbol{x}_0^{i,j} - \hat{\boldsymbol{x}}_0^{i,j})^2$, where $\hat{\boldsymbol{x}}_0^{i,j}$ is the j-th neighbor in the context of $\hat{\boldsymbol{x}}_0^i$ and $\boldsymbol{x}_0^{i,j}$ is the prediction of $\boldsymbol{x}_0^{i,j}$ from a denoising neural network. Thus we have:

$$\boldsymbol{x}_0^{i,j} = \psi_n(\psi_p(\boldsymbol{x}_t, t)(i))(j). \tag{10}$$

4

Compactly, we can write the denoising network as:

$$\Psi(x_t, t)(i, j) = \begin{cases} \psi_n(\psi_p(\boldsymbol{x}_t, t)(i))(j), & j \in \mathcal{N}_i, \\ \psi_p(\boldsymbol{x}_t, t)(i), & j = i. \end{cases} \quad (11)$$

We will show that the DDPM loss is upper bounded by ConPreDiff loss, by reparameterizing $\boldsymbol{x}_0(\boldsymbol{x}_t, t)$. Specifically, for each unit i in the feature map, we use the mean of predicted value in its neighborhood as the final prediction:

$$\boldsymbol{x}_0(\boldsymbol{x}_t, t)(i) = 1/(|\mathcal{N}_i| + 1) * \sum_{j \in \mathcal{N}_i \cup \{i\}} \Psi(\boldsymbol{x}_t, t)(i, j). \quad (12)$$

Now we can show the connection between the DDPM loss and ConPreDiff loss:

$$
\begin{aligned}
||\hat{\boldsymbol{x}}_0 - \boldsymbol{x}_0(\boldsymbol{x}_t, t)||_2^2 &= \sum_i (\hat{\boldsymbol{x}}_0^i - \boldsymbol{x}_0(\boldsymbol{x}_t, t)(i))^2, \\
&= \sum_i (\hat{\boldsymbol{x}}_0^i - \sum_{j \in \mathcal{N}_i \cup \{i\}} \Psi(\boldsymbol{x}_t, t)(i, j)/(|\mathcal{N}_i| + 1))^2, \\
&= \sum_i (\sum_{j \in \mathcal{N}_i \cup \{i\}} (\Psi(\boldsymbol{x}_t, t)(i, j) - \hat{\boldsymbol{x}}_0^i))^2/(|\mathcal{N}_i| + 1)^2, \\
\text{(Cauchy Inequality)} &\leq \sum_i \sum_{j \in \mathcal{N}_i \cup \{i\}} (\Psi(\boldsymbol{x}_t, t)(i, j) - \hat{\boldsymbol{x}}_0^i)^2/(|\mathcal{N}_i| + 1), \\
&= 1/(|\mathcal{N}_i| + 1) \sum_i [(\hat{\boldsymbol{x}}_0^i - \psi_p(\boldsymbol{x}_t, t)(i))^2 + \sum_{j \in \mathcal{N}_i} (\hat{\boldsymbol{x}}_0^{i,j} - \boldsymbol{x}_0^{i,j})^2]
\end{aligned}
\quad (13)
$$

In the last equality, we assume that the feature is padded so that each unit i has the same number of neighbors $|\mathcal{N}|$. As a result, the ConPreDiff loss is an upper bound of the negative log likelihood.

Complexity Problem We note that directly optimizing the Equation (9) has a complexity problem and it will substantially lower the efficiency of CONPREDIFF in training stage. Because the network $\psi_n : \mathbb{R}^d \to \mathbb{R}^{Kd}$ in Equation (8) needs to expand the channel dimension by K times for large-context neighborhood reconstruction, it significantly increases the parameter complexity of the model. Hence, we seek for another way that is efficient for reconstructing neighborhood information.

We solve the challenging problem by changing the direct prediction of entire neighborhoods to the prediction of neighborhood distribution. Specifically, for each $\boldsymbol{x}_{t-1}^i$, the neighborhood information is represented as an empirical realization of i.i.d. sampling Q elements from $\mathcal{P}_{\mathcal{N}_i^s}$, where $\mathcal{P}_{\mathcal{N}_i^s} \triangleq \frac{1}{K} \sum_{u \in \mathcal{N}_i^s} \delta_{h_u}$. Based on this view, we are able to transform the neighborhood prediction $\mathcal{M}_n$ into the neighborhood distribution prediction. **However, such sampling-based measurement loses original spatial orders of neighborhoods, and thus we use a permutation invariant loss (Wasserstein distance) for optimization**. Wasserstein distance [23, 21] is an effective metric for measuring structural similarity between distributions, which is especially suitable for our neighborhood distribution prediction. And we rewrite the Equation (9) as:

$$\mathcal{L}_{\text{CONPREDIFF}} = \sum_{i=1}^{x \times y} \left[\underbrace{\mathcal{M}_p(\boldsymbol{x}_{t-1}^i, \hat{\boldsymbol{x}}^i)}_{point\ denoising} + \underbrace{\mathcal{W}_2^2(\psi_n(\boldsymbol{x}_{t-1}^i, t), \mathcal{P}_{\mathcal{N}_i^s})}_{neighborhood\ distribution\ prediction} \right], \quad (14)$$

where $\psi_n(\boldsymbol{x}_{t-1}^i, t)$ is designed to decode neighborhood distribution parameterized by feedforward neural networks (FNNs), and $\mathcal{W}_2(\cdot, \cdot)$ is the 2-Wasserstein distance as defined below. We provide a more explicit formulation of $\mathcal{W}_2^2(\psi_n(\boldsymbol{x}_{t-1}^i, t), \mathcal{P}_{\mathcal{N}_i^s})$ in Sec. 4.2.

Definition 4.1. Let $\mathcal{P}, \mathcal{Q}$ denote two probability distributions with finite second moment defined on $\mathcal{Z} \subseteq \mathbb{R}^m$. The 2-Wasserstein distance between $\mathcal{P}$ and $\mathcal{Q}$ defined on $\mathcal{Z}, \mathcal{Z}' \subseteq \mathbb{R}^m$ is the solution to the optimal mass transportation problem with ℓ_2 transport cost [90]:

$$\mathcal{W}_2(\mathcal{P}, \mathcal{Q}) = \left(\inf_{\gamma \in \Gamma(\mathcal{P}, \mathcal{Q})} \int_{\mathcal{Z} \times \mathcal{Z}'} ||Z - Z'||_2^2 d\gamma(Z, Z') \right)^{1/2} \quad (15)$$

where $\Gamma(\mathcal{P}, \mathcal{Q})$ contains all joint distributions of (Z, Z') with marginals $\mathcal{P}$ and $\mathcal{Q}$ respectively.

4.2 Efficient Large Context Decoding

Our CONPREDIFF essentially represents the node neighborhood $\hat{\boldsymbol{H}}_{\mathcal{N}_i^s}$ as a distribution of neighbors' representations $\mathcal{P}_{\mathcal{N}_i^s}$ (Equation (14)). We adopt Wasserstein distance to characterize the distribution reconstruction loss because $\mathcal{P}_{\mathcal{N}_i^s}$ has atomic non-zero measure supports in a continuous space, where the family of f-divergences such as KL-divergence cannot be applied. Maximum mean discrepancy may be applied but it needs to choose a specific kernel function.

We define the decoded distribution $\psi_n(\boldsymbol{x}_{t-1}^i, t)$ as an FNN-based transformation of a Gaussian distribution parameterized by $\boldsymbol{x}_{t-1}^i$ and t. The reason for choosing this setting stems from the fact that the universal approximation capability of FNNs allows to (approximately) reconstruct any distributions in 1-Wasserstein distance, as formally stated in Theorem 4.2, proved in Lu & Lu [48]. To enhance the empirical performance, our case adopts the 2-Wasserstein distance and an FNN with d-dim output instead of the gradient of an FNN with 1-dim outout. Here, the reparameterization trick [42] needs to be used:

$$
\begin{aligned}
\psi_n(\boldsymbol{x}_{t-1}^i, t) &= \text{FNN}_n(\xi), \ \xi \sim \mathcal{N}(\mu_i, \Sigma_i), \\
\mu_i &= \text{FNN}_\mu(\boldsymbol{x}_{t-1}^i), \Sigma_i = \text{diag}(\exp(\text{FNN}_\sigma(\boldsymbol{x}_{t-1}^i))).
\end{aligned}
\tag{16}
$$

Theorem 4.2. *For any $\epsilon > 0$, if the support of the distribution $\mathcal{P}_v^{(i)}$ is confined to a bounded space of $\mathbb{R}^d$, there exists a FNN $u(\cdot) : \mathbb{R}^d \to \mathbb{R}$ (and thus its gradient $\nabla u(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$) with sufficiently large width and depth (depending on ϵ) such that $\mathcal{W}_2^2(\mathcal{P}_v^{(i)}, \nabla u(\mathcal{G})) < \epsilon$ where $\nabla u(\mathcal{G})$ is the distribution generated through the mapping $\nabla u(\xi)$, $\xi \sim$ a d-dim non-degenerate Gaussian distribution.*

Another challenge is that the Wasserstein distance between $\psi_n(\boldsymbol{x}_{t-1}^i, t)$ and $\mathcal{P}_{\mathcal{N}_i^s}$ does not have a closed form. Thus, we utilize the empirical Wasserstein distance that can provably approximate the population one as in Peyré et al. [57]. For each forward pass, our CONPREDIFF will get q sampled target pixel/feature points $\{\boldsymbol{x}_{(i,j)}^{tar} | 1 \le j \le q\}$ from $\mathcal{P}_{\mathcal{N}_i^s}$; Next, get q samples from $\mathcal{N}(\mu_i, \Sigma_i)$, denoted by $\xi_1, \xi_2, ..., \xi_q$, and thus $\{\boldsymbol{x}_{(i,j)}^{pred} = \text{FNN}_n(\xi_j) | 1 \le j \le q\}$ are q samples from the prediction $\psi_n(\boldsymbol{x}_{t-1}^i, t)$; Adopt the following empirical surrogated loss of $\mathcal{W}_2^2(\psi_n(\boldsymbol{x}_{t-1}^i, t), \mathcal{P}_{\mathcal{N}_i^s})$ in Equation (14):

$$
\min_\pi \sum_{j=1}^q \|\boldsymbol{x}_{(i,j)}^{tar} - \boldsymbol{x}_{(i,\pi(j))}^{pred}\|^2, \quad \text{s.t. } \pi \text{ is a bijective mapping:} [q] \to [q].
\tag{17}
$$

The loss function is based on solving a matching problem and needs the Hungarian algorithm with $O(q^3)$ complexity [33]. A more efficient surrogate loss may be needed, such as Chamfer loss based on greedy approximation [18] or Sinkhorn loss based on continuous relaxation [11], whose complexities are $O(q^2)$. In our study, as q is set to a small constant, we use Equation (17) based on a Hungarian matching and do not introduce much computational overhead. The computational efficiency of design is empirically demonstrated in Sec. 5.3.

4.3 Discrete and Continuous CONPREDIFF

In training process, given previously-estimated $\boldsymbol{x}_t$, our CONPREDIFF simultaneously predict both $\boldsymbol{x}_{t-1}$ and the neighborhood distribution $\mathcal{P}_{\mathcal{N}_i^s}$ around each pixel/feature. Because $\boldsymbol{x}_{t-1}^i$ can be pixel, feature or discrete token of input image, we can generalize the CONPREDIFF to existing discrete and continuous backbones to form discrete and continuous CONPREDIFF. More concretely, we can substitute the point denoising part in Equation (14) alternatively with the discrete diffusion term $\mathcal{L}_{t-1}^{dis}$ (Equation (3)) or the continuous (Equation (6)) diffusion term $\mathcal{L}_{t-1}^{con}$ for generalization:

$$
\begin{aligned}
\mathcal{L}_{\text{CONPREDIFF}}^{dis} &= \mathcal{L}_{t-1}^{dis} + \lambda_t \cdot \sum_{i=1}^{x \times y} \mathcal{W}_2^2(\psi_n(\boldsymbol{x}_{t-1}^i, t), \mathcal{P}_{\mathcal{N}_i^s}), \\
\mathcal{L}_{\text{CONPREDIFF}}^{con} &= \mathcal{L}_{t-1}^{con} + \lambda_t \cdot \sum_{i=1}^{x \times y} \mathcal{W}_2^2(\psi_n(\boldsymbol{x}_{t-1}^i, t), \mathcal{P}_{\mathcal{N}_i^s}),
\end{aligned}
\tag{18}
$$

where $\lambda_t \in [0, 1]$ is a time-dependent weight parameter. Note that our CONPREDIFF only performs context prediction in training for optimizing the point denoising network ψ_p, and thus does not
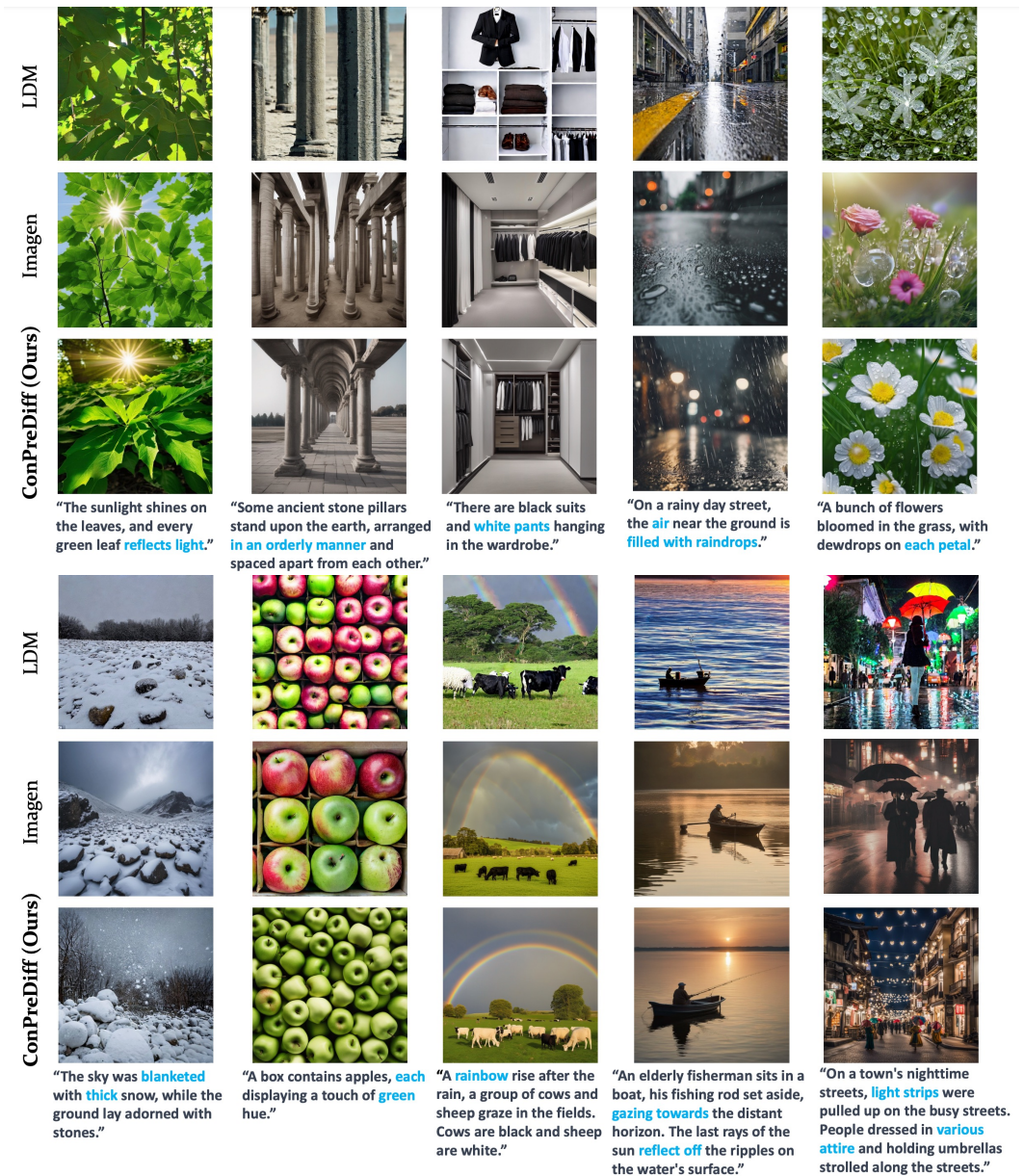
Figure 2: Synthesis examples demonstrating text-to-image capabilities of for various text prompts with LDM, Imagen, and ConPreDiff (Ours). Our model can better express local contexts and semantics of the texts marked in blue.

introduce extra parameters to the inference stage, which is computationally efficient. Equipped with our proposed context prediction term, existing diffusion models consistently gain performance promotion. Next, we use extensive experimental results to prove the effectiveness.

5   Experiments

5.1   Experimental Setup

Datasets and Metrics   Regarding unconditional image generation, we choose four popular datasets for evaluation: CelebA-HQ [34], FFHQ [35], LSUN-Church-outdoor [102], and LSUN-bedrooms [102]. We evaluate the sample quality and their coverage of the data manifold using FID [26] and Precision-and-Recall [43]. For text-to-image generation, we train the model with LAION [73, 74]

7

Table 1: Quantitative evaluation of FID on MS-COCO for 256 × 256 image resolution.

Approach	Model Type	FID-30K	Zero-shot FID-30K
AttnGAN [96]	GAN	35.49	-
DM-GAN [113]	GAN	32.64	-
DF-GAN [86]	GAN	21.42	-
DM-GAN + CL [100]	GAN	20.79	-
XMC-GAN [107]	GAN	9.33	-
LAFITE [112]	GAN	8.12	-
Make-A-Scene [22]	Autoregressive	7.55	-
DALL-E [61]	Autoregressive	-	17.89
LAFITE [112]	GAN	-	26.94
LDM [65]	Continuous Diffusion	-	12.63
GLIDE [54]	Continuous Diffusion	-	12.24
DALL-E 2 [62]	Continuous Diffusion	-	10.39
Improved VQ-Diffusion [85]	Discrete Diffusion	-	8.44
Simple Diffusion [31]	Continuous Diffusion	-	8.32
Imagen [69]	Continuous Diffusion	-	7.27
Parti [104]	Autoregressive	-	7.23
Muse [7]	Non-Autoregressive	-	7.88
eDiff-I [3]	Continuous Diffusion	-	6.95
CONPREDIFF$_{dis}$	Discrete Diffusion	-	**6.67**
CONPREDIFF$_{con}$	Continuous Diffusion	-	**6.21**

and some internal datasets, and conduct evaluations on MS-COCO dataset with zero-shot FID and CLIP score [25, 59], which aim to assess the generation quality and resulting image-text alignment. For image inpainting, we choose CelebA-HQ [34] and ImageNet [12] for evaluations, and evaluate all 100 test images of the test datasets for the following masks: Wide, Narrow, Every Second Line, Half Image, Expand, and Super-Resolve. We report the commonly reported perceptual metric LPIPS [109], which is a learned distance metric based on the deep feature space.

Baselines To demonstrate the effectiveness of CONPREDIFF, we compare with the latest diffusion and non-diffusion models. Specifically, for unconditional image generation, we choose ImageBART[16], U-Net GAN (+aug) [72], UDM [39], StyleGAN [36], ProjectedGAN [71], DDPM [28] and ADM [14] for comparisons. As for text-to-image generation, we choose DM-GAN [113], DF-GAN [86], DM-GAN + CL [100], XMC-GAN [107] LAFITE [112], Make-A-Scene [22], DALL-E [61], LDM [65], GLIDE [54], DALL-E 2 [62], Improved VQ-Diffusion [85], Imagen-3.4B [69], Parti [104], Muse [7], and eDiff-I [3] for comparisons. For image inpainting, we choose autoregressive methods(DSI [56] and ICT [91]), the GAN methods (DeepFillv2 [103], AOT [105], and LaMa [84]) and diffusion based model (RePaint [49]). All the reported results are collected from their published papers or reproduced by open source codes.

Implementation Details For text-to-image generation, similar to Imagen [68], our continuous diffusion model CONPREDIFF$_{con}$ consists of a base text-to-image diffusion model (64×64) [53], two super-resolution diffusion models [29] to upsample the image, first 64×64 → 256×256, and then 256×256 → 1024×1024. The model is conditioned on both T5 [60] and CLIP [59] text embeddings. The T5 encoder is pre-trained on a C4 text-only corpus and the CLIP text encoder is trained on an image-text corpus with an image-text contrastive objective. We use the standard Adam optimizer with a learning rate of 0.0001, weight decay of 0.01, and a batch size of 1024 to optimize the base model and two super-resolution models on NVIDIA A100 GPUs, respectively, equipped with multi-scale training technique (6 image scales). We generalize our context prediction to discrete diffusion models [24, 85] to form our CONPREDIFF$_{dis}$. For image inpainting, we adopt a same pipeline as RePaint [49], and retrain its diffusion backbone with our context prediction loss. We use T = 250 time steps, and applied r = 10 times resampling with jumpy size j = 10. For unconditional generation tasks, we use the same denoising architecture like LDM [65] for fair comparison. The max channels are 224, and we use T=2000 time steps, linear noise schedule and an initial learning rate of 0.000096.

Table 2: Quantitative evaluation of image inpainting on CelebA-HQ and ImageNet.

CelebA-HQ Method	Wide LPIPS ↓	Narrow LPIPS ↓	Super-Resolve 2× LPIPS ↓	Altern. Lines LPIPS ↓	Half LPIPS ↓	Expand LPIPS ↓
AOT [105]	0.104	0.047	0.714	0.667	0.287	0.604
DSI [56]	0.067	0.038	0.128	0.049	0.211	0.487
ICT [91]	0.063	0.036	0.483	0.353	0.166	0.432
DeepFillv2 [103]	0.066	0.049	0.119	0.049	0.209	0.467
LaMa [84]	0.045	0.028	0.177	0.083	**0.138**	0.342
RePaint [49]	0.059	0.028	0.029	**0.009**	0.165	0.435
CONPREDIFF	**0.042**	**0.022**	**0.023**	0.022	**0.139**	**0.297**

ImageNet Method	Wide LPIPS ↓	Narrow LPIPS ↓	Super-Resolve 2× LPIPS ↓	Altern. Lines LPIPS ↓	Half LPIPS ↓	Expand LPIPS ↓
DSI [56]	0.117	0.072	0.153	0.069	0.283	0.583
ICT [91]	0.107	0.073	0.708	0.620	0.255	0.544
LaMa [84]	0.105	0.061	0.272	0.121	**0.254**	0.534
RePaint [49]	0.134	0.064	0.183	**0.089**	0.304	0.629
CONPREDIFF	**0.098**	**0.057**	**0.129**	0.107	0.285	**0.506**

Our context prediction head contains two non-linear blocks (*e.g.*, Conv-BN-ReLU, resnet block or transformer block), and its choice can be flexible according to specific task. The prediction head does not incur significant training costs, and can be removed in inference stage without introducing extra testing costs. We set the neighborhood stride to 3 for all experiments, and carefully choose the specific layer for adding context prediction head near the end of denoising networks.

5.2   Main Results

Text-to-Image Synthesis   We conduct text-to-image generation on MS-COCO dataset, and quantitative comparison results are listed in Tab. 1. We observe that both discrete and continuous CONPREDIFF substantially surpasses previous diffusion and non-diffusion models in terms of FID score, demonstrating the new state-of-the-art performance. Notably, our discrete and continuous CONPREDIFF achieves **an FID score of 6.67 and 6.21 which are better than the score of 8.44 and 7.27 achieved by previous SOTA discrete and continuous diffusion models**. We visualize text-to-image generation results in Figure 2, and find that our CONPREDIFF can synthesize images that are semantically better consistent with text prompts. It demonstrates our CONPREDIFF can make promising cross-modal semantic understanding through preserving visual context information generating process, and effectively associating with contextual text information. Moreover, we observe that CONPREDIFF can synthesize complex objects and scenes consistent with text prompts as demonstrated by Figure 7 in Appendix A.4, proving the effectiveness of our designed neighborhood context prediction. Human evaluations are provided in Appendix A.5.

Image Inpainting   Our CONPREDIFF naturally fits image inpainting task because we directly predict the neighborhood context of each pixel/feature in diffusion generation. We compare our CONPREDIFF against state-of-the-art on standard mask distributions, commonly employed for benchmarking. As in Tab. 2, our CONPREDIFF outperforms previous SOTA method for most kinds of masks. We also put some qualitative results in Figure 3, and observe that CONPREDIFF produces a semantically meaningful filling, demonstrating the effectiveness of our context prediction.

Unconditional Image Synthesis   We list the quantitative results about unconditional image generation in Tab. 3 of Appendix A.3. We observe that our CONPREDIFF significantly improves upon the state-of-the-art in FID and Precision-and-Recall scores on FFHQ and LSUN-Bedrooms datasets. The CONPREDIFF obtains high perceptual quality superior to prior GANs and diffusion models, while maintaining a higher coverage of the data distribution as measured by recall.

5.3   The Impact and Efficiency of Context Prediction

In Sec. 4.2, we tackle the complexity problem by transforming the decoding target from entire neighborhood features to neighborhood distribution. Here we investigate both impact and efficiency of the proposed neighborhood context prediction. For fast experiment, we conduct ablation study with the diffusion backbone of LDM [65]. As illustrated in Figure 4, the FID score of CONPREDIFF

Figure 3: Inpainting examples generated by our CONPREDIFF.

is better with the neighbors of more strides and 1-stride neighbors contribute the most performance gain, revealing that preserving local context benefits the generation quality. Besides, we observe that increasing neighbor strides significantly increases the training cost when using feature decoding, while it has little impact on distribution decoding with comparable FID score. To demonstrate the generalization ability, we equip previous diffusion models with our context prediction head. From the results in Figure 5, we find that our context prediction can consistently and significantly improve the FID scores of these diffusion models, sufficiently demonstrating the effectiveness and extensibility of our method. We further conduct ablation study on the trade-off between CLIP and FID scores across a range of guidance weights in Appendix A.2, the results exhibit our superior generation ability.
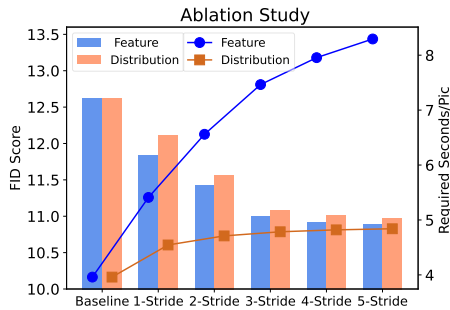

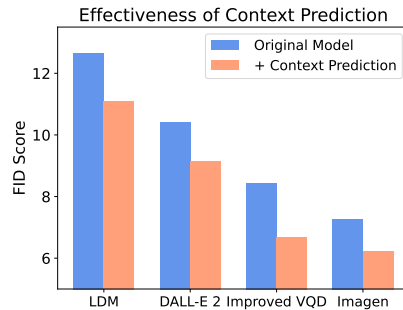Figure 4: Bar denotes FID and line denotes time cost.

Figure 5: Equip diffusion models with our context prediction.

6 Conclusion

In this paper, we for the first time propose CONPREDIFF to improve diffusion-based image synthesis with context prediction. We explicitly force each point to predict its neighborhood context with an efficient context decoder near the end of diffusion denoising blocks, and remove the decoder for inference. CONPREDIFF can generalize to arbitrary discrete and continuous diffusion backbones and consistently improve them without extra parameters. We achieve new SOTA results on unconditional image generation, text-to-image generation and image inpainting tasks.

Acknowledgement

References

[1] Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, 2021. 1, 2

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022. 1

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022. 8

[4] Fan Bao, Chongxuan Li, Jun Zhu, and Bo Zhang. Analytic-dpm: an analytic estimate of the optimal reverse variance in diffusion probabilistic models. In *International Conference on Learning Representations*, 2021. 1

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2018. 1

[6] Hanqun Cao, Cheng Tan, Zhangyang Gao, Guangyong Chen, Pheng-Ann Heng, and Stan Z Li. A survey on generative diffusion model. *arXiv preprint arXiv:2209.02646*, 2022. 2

[7] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023. 8

[8] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022. 2, 3

[9] Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *arXiv preprint arXiv:2209.04747*, 2022. 2

[10] Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. Vqgan-clip: Open domain image generation and editing with natural language guidance. *arXiv preprint arXiv:2204.08583*, 2022. 1

[11] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 2013. 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. 8

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. 2

[14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021. 1, 2, 8, 19

[15] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pp. 1422–1430, 2015. 2, 3

[16] Patrick Esser, Robin Rombach, Andreas Blattmann, and Bjorn Ommer. Imagebart: Bidirectional context with multinomial diffusion for autoregressive image synthesis. *Advances in Neural Information Processing Systems*, 34:3518–3532, 2021. 8, 19

[17] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 12873–12883, 2021. 2, 19

[18] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 6

[19] Wan-Cyuan Fan, Yen-Chun Chen, DongDong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *arXiv preprint arXiv:2208.13753*, 2022. 2

[20] Mary Anne Franks and Ari Ezra Waldman. Sex, lies, and videotape: Deep fakes and free speech delusions. *Md. L. Rev.*, 78:892, 2018. 18

[21] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso A Poggio. Learning with a wasserstein loss. In *NeurIPS*, 2015. 2, 5

[22] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022. 1, 8

[23] Clark R Givens and Rae Michael Shortt. A class of wasserstein metrics for probability distributions. *Michigan Mathematical Journal*, 31(2):231–240, 1984. 5

[24] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10696–10706, 2022. 1, 2, 3, 8

[25] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, 2021. 8

[26] 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. 7, 18

[27] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 2

[28] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020. 1, 2, 4, 8, 19

[29] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23:47–1, 2022. 1, 8

[30] Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12454–12465, 2021. 1, 2

[31] Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*, 2023. 8

[32] Han Hu, Zheng Zhang, Zhenda Xie, and Stephen Lin. Local relation networks for image recognition. In *ICCV*, pp. 3464–3473, 2019. 2, 3

[33] Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987. 6

[34] 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. 7, 8

[35] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4401–4410, 2019. 7

[36] 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. 8, 19

[37] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8110–8119, 2020. 19

[38] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022. 2

[39] Dongjun Kim, Seungjae Shin, Kyungwoo Song, Wanmo Kang, and Il-Chul Moon. Score matching model for unbounded data score. *arXiv preprint arXiv:2106.05527*, 2021. 8, 19

[40] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2426–2435, 2022. 1, 2

[41] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[42] Diederik P Kingma and Max Welling. Auto-encoding variational bayess. In *ICLR*, 2014. 6

[43] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, 32, 2019. 7, 19

[44] Zhi Lei, Guixian Zhang, Lijuan Wu, Kui Zhang, and Rongjiao Liang. A multi-level mesh mutual attention model for visual question answering. *Data Science and Engineering*, 7(4): 339–353, 2022. 3

[45] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *ICCV*, pp. 4170–4179, 2019. 2, 3

[46] Cheng Lu, Kaiwen Zheng, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Maximum likelihood training for score-based diffusion odes by high order denoising score matching. In *International Conference on Machine Learning*, pp. 14429–14460, 2022. 1

[47] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 1

[48] Yulong Lu and Jianfeng Lu. A universal approximation theorem of deep neural networks for expressing probability distributions. *NeurIPS*, 2020. 6

[49] 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 (CVPR)*, pp. 11461–11471, June 2022. 8, 9

[50] Jiayi Ma, Ji Zhao, Junjun Jiang, Huabing Zhou, and Xiaojie Guo. Locality preserving matching. *International Journal of Computer Vision*, 127(5):512–531, 2019. 2, 3

[51] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021. 1

[52] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 2

[53] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pp. 8162–8171, 2021. 1, 8

[54] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pp. 16784–16804, 2022. 1, 2, 8

[55] Gaurav Parmar, Dacheng Li, Kwonjoon Lee, and Zhuowen Tu. Dual contradistinctive generative autoencoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 823–832, 2021. 19

[56] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10775–10784, 2021. 8, 9

[57] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019. 6

[58] Konpat Preechakul, Nattanat Chatthee, Suttisak Wizadwongsa, and Supasorn Suwajanakorn. Diffusion autoencoders: Toward a meaningful and decodable representation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10619–10629, 2022. 2

[59] 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, 2021. 2, 8

[60] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020. 2, 8

[61] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021. 8

[62] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 8, 21

[63] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[64] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pp. 1278–1286, 2014. 2

[65] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022. 1, 2, 8, 9, 19

[66] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 1

[67] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Proceedings*, pp. 1–10, 2022. 1, 2

[68] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 2, 8, 21

[69] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Seyedeh Sara Mahdavi, Raphael Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. *ArXiv*, abs/2205.11487, 2022. 8

[70] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 1

[71] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. *Advances in Neural Information Processing Systems*, 34:17480–17492, 2021. 8, 19

[72] Edgar Schönfeld, Bernt Schiele, and Anna Khoreva. A u-net based discriminator for generative adversarial networks. In *CVPR*, pp. 8204–8213, 2020. 8, 19

[73] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021. 7

[74] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 7

[75] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022. 2

[76] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pp. 2256–2265, 2015. 1, 2, 3

[77] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020. 1

[78] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019. 1, 2

[79] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12438–12448, 2020. 1

[80] 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*, 2020. 1, 2

[81] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1415–1428, 2021. 1

[82] Ramya Srinivasan and Kanji Uchino. Biases in generative art: A causal look from the lens of art history. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 41–51, 2021. 18

[83] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 701–713, 2021. 18

[84] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2149–2159, 2022. 8, 9

[85] Zhicong Tang, Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Improved vector quantized diffusion models. *arXiv preprint arXiv:2205.16007*, 2022. 2, 8

[86] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16515–16525, 2022. 8

[87] Arash Vahdat, Karsten Kreis, and Jan Kautz. Score-based generative modeling in latent space. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11287–11302, 2021. 1, 2, 19

[88] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 2

[89] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[90] Cédric Villani. *Optimal transport: old and new*, volume 338. 2008. 5

[91] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4692–4701, 2021. 8, 9

[92] Haixin Wang, Jianlong Chang, Xiao Luo, Jinan Sun, Zhouchen Lin, and Qi Tian. Lion: Implicit vision prompt tuning. *arXiv preprint arXiv:2303.09992*, 2023. 18

[93] Haixin Wang, Xinlong Yang, Jianlong Chang, Dian Jin, Jinan Sun, Shikun Zhang, Xiao Luo, and Qi Tian. Mode approximation makes good vision-language prompts. *arXiv preprint arXiv:2305.08381*, 2023. 18

[94] Jing Wang, Yehao Li, Yingwei Pan, Ting Yao, Jinhui Tang, and Tao Mei. Contextual and selective attention networks for image captioning. *Science China Information Sciences*, 65 (12):222103, 2022. 3

[95] Meng Wang, Yinghui Shi, Han Yang, Ziheng Zhang, Zhenxi Lin, and Yefeng Zheng. Probing the impacts of visual context in multimodal entity alignment. *Data Science and Engineering*, 8(2):124–134, 2023. 3

[96] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018. 8

[97] Ling Yang, Liangliang Li, Zilun Zhang, Xinyu Zhou, Erjin Zhou, and Yu Liu. Dpgn: Distribution propagation graph network for few-shot learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 13390–13399, 2020. 18

[98] Ling Yang, Zhilin Huang, Yang Song, Shenda Hong, Guohao Li, Wentao Zhang, Bin Cui, Bernard Ghanem, and Ming-Hsuan Yang. Diffusion-based scene graph to image generation with masked contrastive pre-training. *arXiv preprint arXiv:2211.11138*, 2022. 2

[99] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*, 2022. 1, 2

[100] Hui Ye, Xiulong Yang, Martin Takac, Rajshekhar Sunderraman, and Shihao Ji. Improving text-to-image synthesis using contrastive learning. *arXiv preprint arXiv:2107.02423*, 2021. 8

[101] Bruce XB Yu, Jianlong Chang, Haixin Wang, Lingbo Liu, Shijie Wang, Zhiyu Wang, Junfan Lin, Lingxi Xie, Haojie Li, Zhouchen Lin, et al. Visual tuning. *arXiv preprint arXiv:2305.06061*, 2023. 18

[102] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 7

[103] 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*, pp. 5505–5514, 2018. 8, 9

[104] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022. 8, 21

[105] 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. 8, 9

[106] Dong Zhang, Liyan Zhang, and Jinhui Tang. Augmented fcn: rethinking context modeling for semantic segmentation. *Science China Information Sciences*, 66(4):142105, 2023. 3

[107] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 833–842, 2021. 8

[108] Qinsheng Zhang and Yongxin Chen. Fast sampling of diffusion models with exponential integrator. *arXiv preprint arXiv:2204.13902*, 2022. 1

[109] 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. 8

[110] Xinyu Zhang, Jiahui Chen, Junkun Yuan, Qiang Chen, Jian Wang, Xiaodi Wang, Shumin Han, Xiaokang Chen, Jimin Pi, Kun Yao, et al. Cae v2: Context autoencoder with clip target. *arXiv preprint arXiv:2211.09799*, 2022. 2, 3

[111] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *CVPR*, pp. 5565–5573, 2019. 2, 3

[112] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17907–17917, 2022. 8

[113] Minfeng Zhu, Pingbo Pan, Wei Chen, and Yi Yang. Dm-gan: Dynamic memory generative adversarial networks for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5802–5810, 2019. 8

[114] Ye Zhu, Yu Wu, Kyle Olszewski, Jian Ren, Sergey Tulyakov, and Yan Yan. Discrete contrastive diffusion for cross-modal and conditional generation. *arXiv preprint arXiv:2206.07771*, 2022. 2

A   Appendix

A.1   Limitations and Broader Impact

Limitations   While our ConPreDiff boosts performance of both discrete and continuous diffusion models without introducing additional parameters in model inference, our models still have more trainable parameters than other types of generative models, e.g GANs. Furthermore, we note the long sampling times of both and compared to single step generative approaches like GANs or VAEs. However, this drawback is inherited from the underlying model class and is not a property of our context prediction approach. Neighborhood context decoding is fast and incurs negligible computational overhead in training stage. For future work, we will try to find more intrinsic information to preserve for improving existing point-wise denoising diffusion models, and extend to more challenging tasks like text-to-3D and text-to-video generation.

Broader Impact   Recent generative image models enable creative applications and autonomous media creation, but can also be viewed as a dual-use technology with negative implications. In this paper, we use human face datasets only for evaluating the image inpainting performance of our method, and our method is not intended to create content that is used to mislead or deceive. However, like other related image generation methods, it could still potentially be misused in the realm of human impersonation. A notorious example are so-called "deep fakes" that have been used, for example, to create pornographic "undressing" applications. We strongly disapprove of any actions aimed at producing deceptive or harmful content featuring real individuals. Besides, generative methods have the capacity to be harnessed for other malicious intentions, including harassment and misinformation spread [20], and give rise to significant concerns pertaining to societal and cultural exclusion as well as biases [83, 82]. These considerations guide our decision not to release the source code or a public demo at this point in time.

Furthermore, the immediate availability of mass-produced high-quality images can be used to spread misinformation and spam, which in turn can be used for targeted manipulation in social media. Datasets are crucial for deep learning as they are the main input of information [101, 92, 93, 97]. Large-scale data requirements of text-to-image models have led researchers to rely heavily on large, mostly uncurated, web-scraped datasets. While this approach has enabled rapid algorithmic advances recently, datasets of this nature have been critiqued and contested along various ethical dimensions. One should consider the ability to curate the database to exclude (or explicitly contain) potential harmful source images. Creating a public API could offer a cheaper way to offer a safe model than retraining a model on a filtered subset of the training data or doing difficult prompt engineering. Conversely, including only harmful content is an easy way to build a toxic model.

A.2   Guidance Scale *vs*. FID

To further demonstrate the effectiveness of our proposed context prediction, we quantitatively conduct evaluations about the trade-off between MS-COCO zero-shot FID [26] and CLIP scores. The results in Figure 6 indicate that the guidance hurts the diversity of GLIDE much more than DALL-E 2 and CONPREDIFF. The phenomenon reveals that the proposed CONPREDIFF can overall improve the generation quality of diffusion models.
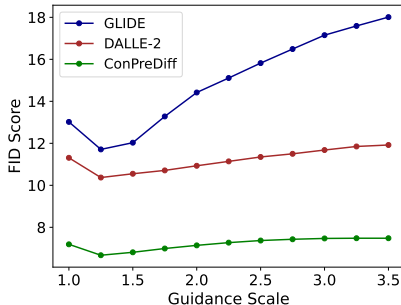

Figure 6: Trade-off between guidance scale and FID.

A.3 More Quantitative Results

We list the unconditional generation results on FFHQ, CelebA-HQ, LSUN-Churches, and LSUN-Bedrooms in Tab. 3. We find CONPREDIFF consistently outperforms previous methods, demonstrating the effectiveness of the CONPREDIFF.

Table 3: Evaluation results for unconditional image synthesis.

FFHQ 256×256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART[16]	9.57	-	-
U-Net GAN (+aug) [72]	7.6	-	-
UDM [39]	5.54	-	-
StyleGAN [36]	4.16	0.71	0.46
ProjectedGAN [71]	3.08	0.65	0.46
LDM [65]	4.98	0.73	0.50
CONPREDIFF	**2.24**	**0.81**	**0.61**
LSUN-Bedrooms 256×256			
Method	FID ↓	Prec. ↑	Recall ↑
ImageBART [16]	5.51	-	-
DDPM [28]	4.9	-	-
UDM [39]	4.57	-	-
StyleGAN [36]	2.35	0.59	0.48
ADM [14]	1.90	0.66	0.51
ProjectedGAN [71]	1.52	0.61	0.34
LDM-4 [65]	2.95	0.66	0.48
CONPREDIFF	**1.12**	**0.73**	**0.59**
CelebA-HQ 256×256			
Method	FID ↓	Prec. ↑	Recall ↑
DC-VAE [55]	15.8	-	-
VQGAN+T. [17] (k=400)	10.2	-	-
PGGAN [43]	8.0	-	-
LSGM [87]	7.22	-	-
UDM [39]	7.16	-	-
LDM [65]	5.11	0.72	0.49
CONPREDIFF	**3.22**	**0.83**	**0.57**
LSUN-Churches 256×256			
Method	FID ↓	Prec. ↑	Recall ↑
DDPM [28]	7.89	-	-
ImageBART [16]	7.32	-	-
PGGAN [43]	6.42	-	-
StyleGAN [36]	4.21	-	-
StyleGAN2 [37]	3.86	-	-
ProjectedGAN [71]	**1.59**	0.61	0.44
LDM [65]	4.02	0.64	0.52
CONPREDIFF	**1.78**	**0.74**	**0.61**

A.4 More Synthesis Results

We visualize more text-to-image synthesis results on MS-COCO dataset in Figure 7. We observe that compared with previous powerful LDM and DALL-E 2, our CONPREDIFF generates more natural and smooth images that preserve local continuity.

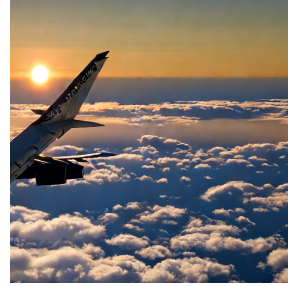"A photo of a dark Goth house" "A teddy bear sitting on a chair. " "A person holding a bunch of bananas on a table."
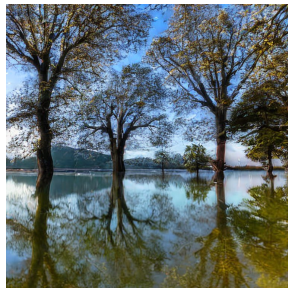
"A group of elephants walking in muddy water. " "Green frog on green grass" "The plane wing above the clouds. "

" A big round hole in brick wall " "Reflection of tree in lake" "An orange ball is put on the ground "

" Trees on African grassland " " Cat fell asleep on the owner's bed " "A red hydrant sitting in the snow."

" Pancakes with ketchup " "A photo of an adult lion." "A photo of an white garlic ice cream"

Figure 7: Synthesis examples demonstrating text-to-image capabilities of for various text prompts.

A.5 Human Evaluations

As demonstrated in qualitative results, our CONPREDIFF is able to synthesize realistic diverse, context-coherent images. However, using FID to estimate the sample quality is not always consistent with human judgment. Therefore, we follow the protocol of previous works [104, 68, 62], and conduct systematic human evaluations to better assess the generation capacities of our CONPREDIFF from the aspects of image photorealism and image-text alignment. We conduct side-by-side human evaluations, in which well-trained users are presented with two generated images for the same prompt and need to choose which image is of higher quality and more realistic (image photorealism) and which image better matches the input prompt (image-text alignment). For evaluating the coherence of local context, we propose a new evaluation protocol, in which users are presented with 1000 pairs of images and must choose which image better preserves local pixel/semantic continuity. The evaluation results are in Tab. 4, CONPREDIFF performs better in pairwise comparisons against both Improved VQ-Diffusion and Imagen. We find that CONPREDIFF is preferred in terms of all three evaluations, and CONPREDIFF is strongly preferred regarding context coherence, demonstrating that preserving local neighborhood context advances sample quality and semantic alignment.

Table 4: Human evaluation comparing CONPREDIFF to Improved VQ-Diffusion and Imagen.

	Improved VQ-Diffusion	Imagen
Image Photorealism	72%	65%
Image-Text Alignment	68%	63%
Context Coherence	84%	78%