

---

# LGDM: Latent Guidance in Diffusion Models for Perceptual Evaluations

---

Anonymous Authors<sup>1</sup>

## Abstract

Despite recent advancements in latent diffusion models that generate high-dimensional image data and perform various downstream tasks, there has been little exploration into perceptual consistency within these models on the task of No-Reference Image Quality Assessment (NR-IQA). In this paper, we hypothesize that latent diffusion models implicitly exhibit perceptually consistent local regions within the data manifold. We leverage this insight to guide on-manifold sampling using perceptual features and input measurements. Specifically, we propose **P**erceptual **M**anifold **G**uidance (PMG), an algorithm that utilizes pretrained latent diffusion models and perceptual quality features to obtain perceptually consistent multi-scale and multi-timestep feature maps from the denoising U-Net. We empirically demonstrate that these hyperfeatures exhibit high correlation with human perception in IQA tasks. Our method can be applied to any existing pretrained latent diffusion model and is straightforward to integrate. To the best of our knowledge, this paper is the first work on guiding diffusion model with perceptual features for NR-IQA. Extensive experiments on IQA datasets show that our method, LGDM, achieves state-of-the-art performance, underscoring the superior zero-shot generalization capabilities of diffusion models for NR-IQA tasks. The source code will be made publicly available upon publication.

## 1 Introduction

Score-based diffusion models have advanced significantly in recent years and have achieved remarkable success at synthesizing high-quality images across diverse scenes, views, and lighting conditions ((Ho et al., 2020; Song & Ermon, 2019; Song et al., 2020b; Zhang et al., 2023)). Latent Diffu-

<sup>1</sup>Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

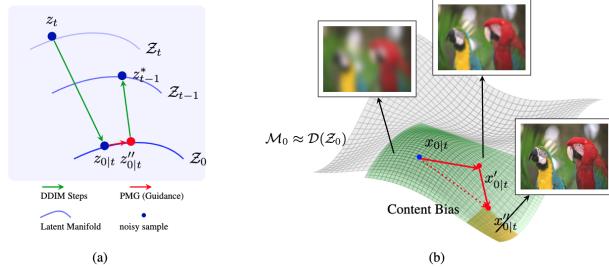


Figure 1: An overview of our proposed approach: (a) shows the transition of latent samples across latent manifolds, highlighting the steps of DDIM and our PMG. (b) depicts the content bias (green) on the manifold  $\mathcal{M}_0 \approx \mathcal{D}(Z_0)$ , showing that the guidance term in red (PMG) pushes a data sample ( $x'_{0|t} \sim \mathcal{D}(z'_{0|t})$ ) towards the perceptually consistent region (yellow) of the manifold. Here  $\mathcal{D}$  is decoder of VAE.

sion Models (LDMs), which embed data into a compressed latent space, enhance computational efficiency ((Rombach et al., 2022)). Diffusion models provide strong data priors that effectively capture the intricacies of high-dimensional data distributions, making them powerful for generative tasks. Conditional generation using posterior sampling has become crucial for solving various real-world low-level vision problems ((Kawar et al., 2022; Chung et al., 2023; Rout et al., 2024; Song et al., 2023a)). Additionally, several methods leverage the rich internal representations of diffusion models by extracting either hand-selected single or subsets of features from a denoising U-Net for downstream tasks ((Tumanyan et al., 2023; Ye et al., 2023; Xu et al., 2023; Baranchuk et al., 2021)). Despite these advancements in addressing tasks like inverse problems, segmentation, and semantic keypoint correspondence, there has been little exploration into perceptual consistency of diffusion models for No-Reference Image Quality Assessment (NR-IQA).

NR-IQA aims to evaluate image quality in line with human perception without a high-quality reference image ((Wang & Bovik, 2006)). It plays a crucial role in optimizing parameters for image processing tasks, such as resizing, compression ((Feng et al., 2023; Liu et al., 2023)), and enhancement ((Hou et al., 2024; Fei et al., 2023; Zhang et al., 2024)). Fig. 2 illustrate a diverse spectrum of distortions frequently encountered in real-world datasets. Early NR-IQA methods



Figure 2: Exemplar Authentic and Synthetic distorted images. Here (a) shows Camera Motion Blur, (b) shows compression artifacts, (c) shows noise, and (d) shows a night scene with grains. Best viewed when zoomed in.

used hand-crafted natural scene statistics features ((Zhang et al., 2015; Mittal et al., 2012; Saad et al., 2012)), and have evolved into learning-based quality metrics ((Madhusudana et al., 2022; Tu et al., 2021; Ke et al., 2021; Saini et al., 2024)). While learning-based methods show promise, they often lack generalizability. With the advent of generative models, some authors have explored the use of pixel diffusion models for NR-IQA tasks ((Li et al., 2024b; Babnik et al., 2024; Wang et al., 2024)). While these approaches show impressive progress, they are often *ad hoc*, focusing on tasks like quality feature denoising and image restoration by converting NR-IQA problems into Full-Reference IQA (FR-IQA) ones. Additionally, training on specific IQA datasets limits their generalizability. By contrast, our goal is to utilize pretrained latent diffusion models without fine-tuning, leveraging perceptual guidance to extract intermediate multi-scale and multi-time features, termed diffusion hyperfeatures ((Luo et al., 2024)), for NR-IQA. Our proposed method robustly handles breadth of degradations, capturing both local texture artifacts and larger-scale structural inconsistencies as shown in Fig. 2.

At the core of our method is the manifold hypothesis: *real data does not occupy the entire pixel space but instead lies on a smaller underlying manifold*. Previous works ((Chung et al., 2022b; He et al., 2023; Sun et al., 2023)) have used the manifold concept for guided sample generation and solving inverse problems. In IQA, deep models aim to learn distortion manifolds that correlate highly with human perceptual quality ((Agnolucci et al., 2024; Su et al., 2023; Guan et al., 2018; Gao et al., 2024)). These manifolds represent regions within the data manifold that contain perceptually consistent samples, with content bias further narrowing these regions.

Towards further advancing progress in this direction, we propose **Perceptual Manifold Guidance (PMG)** to ensure perceptually consistent on-manifold sampling, conditioned on weak perceptual quality features. Fig. 1 provides an overview and conceptual visualization of our approach. Unlike previous state-of-the-art CNN or transformer-based IQA models that only utilize the final feature layer, we extract intermediate multi-scale and multi-time features, termed diffusion hyperfeatures ((Luo et al., 2024)), from a denoising U-Net for NR-IQA. As shown by (Ghildyal et al., 2024),

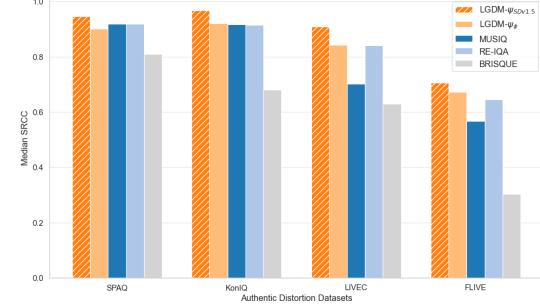


Figure 3: Median SRCC scores of NR-IQA methods across authentic distortion IQA datasets, demonstrating the superior performance of our method.

intermediate features of foundation models outperform state-of-the-art learned metrics based on final feature layers. Our method, **Latent Guidance in Diffusion Model (LGDM)**, is a framework for extracting perceptually consistent diffusion hyperfeatures from unconditionally pretrained latent diffusion models:

- We introduce a novel approach for leveraging unconditional latent diffusion models to tackle the challenging task of NR-IQA without any fine-tuning or additional training of feature extraction model.
- We design a manifold guidance scheme(PMG) that ensures the sampling process remains on the manifold and close to perceptually consistent region. We provide extensive theoretical proof that our perceptual guidance keeps gradient updates on the tangent spaces of the data manifold, maintaining proximity to the local perceptually consistent manifold. We also utilize intermediate multi-scale and multi-time features from the denoising U-Net, resulting in high correlation with human perceptual judgments.
- Extensive experiments on both authentic and synthetic IQA datasets demonstrate that our method achieves state-of-the-art performance. To the best of our knowledge, this is the first approach to introduce perceptual guidance in latent diffusion models for NR-IQA.

We evaluate LGDM against both supervised and unsuper-

110 vised state-of-the-art methods on ten IQA datasets, consistently achieving superior results (see Fig. 3).  
 111  
 112  
 113

## 2 Background

### 2.1 NR-IQA

117 With reference to NR-IQA, to better capture the complex  
 118 relationship between image content and perceived quality,  
 119 manifold learning techniques have been explored ((Agno-  
 120 nolucci et al., 2024; Su et al., 2023; Guan et al., 2018; Gao  
 121 et al., 2024)). These approaches aim to uncover intrinsic  
 122 low-dimensional structures within high-dimensional data,  
 123 thereby aligning more closely with human visual perception.  
 124 They generally rely on the following hypothesis:

125 **Assumption 1:** (Strong Manifold Hypothesis). *For a given*  
 126 *data distribution  $\mathbf{X} \in \mathbb{R}^D$ , the actual data points are con-*  
 127 *centrated on a  $k$ -dimensional locally linear subspace mani-*  
 128 *fold  $\mathcal{M} \subset \mathbb{R}^D$ , such that  $k \ll D$ .*

129 Latent diffusion models have demonstrated strong repre-  
 130 sentation learning capabilities when trained on large-scale  
 131 datasets containing a wide range of authentic and synthetic  
 132 distortions ((Rombach et al., 2022; Zhang et al., 2023)),  
 133 but their application to NR-IQA problems remains underex-  
 134 plored ((Li et al., 2024b; Babnik et al., 2024; Wang et al.,  
 135 2024)).

136 We demonstrate that latent diffusion models implicitly learn  
 137 perceptually consistent manifolds due to their extensive  
 138 training data and ability to capture data priors through  
 139 score matching. By leveraging the learned score function  
 140  $s_\theta$  with perceptual guidance from features of the  
 141 perceptual metric  $\psi_p$ , and extracting diffusion hyperfeatures  
 142  $\mathbf{H} = \bigcup_{t=0}^T \mathbf{h}_t = \bigcup_{t=0}^T \bigcup_{l=0}^L s_\theta(x_t, t)|_l$ , where  $T$  repre-  
 143 sents the total sampling steps and  $L$  is a subset of interme-  
 144 diate layers, we align features from the denoising network  $s_\theta$   
 145 at different time steps with human perceptual judgments.

### 2.2 Diffusion Models

146 We begin by reviewing the DDIM followed by conditional  
 147 diffusion models.

148 **Denoising Diffusion Implicit Models (DDIM).** To address  
 149 the slow generation of DDPM, Song et al. (2020a) proposed  
 150 Denoising Diffusion Implicit Models (DDIMs), which define a non-Markovian diffusion process for faster sampling.  
 151 The DDIM sampling update is:

$$x_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{x}_{0|t} + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} s_\theta(x_t, t) + \sigma_t \epsilon \quad t = T, \dots, 0, \quad (1)$$

152 where  $\alpha_t = 1 - \beta_t$ ,  $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$ ,  $\sigma_t =$   
 153  $\sqrt{(1 - \bar{\alpha}_{t-1})/(1 - \bar{\alpha}_t)} \sqrt{1 - \bar{\alpha}_t / \bar{\alpha}_{t-1}}$  corresponds

154 to DDPM sampling, and when  $\sigma_t = 0$  sampling becomes  
 155 deterministic, where  $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . The term  $\hat{x}_{0|t}$  is  
 156 direct estimation of the clean data  $x_0$  from noisy data  $x_t$ ,  
 157 calculated using Tweedie's formula ((Efron, 2011)):

$$\hat{x}_{0|t} = \frac{1}{\sqrt{\bar{\alpha}_t}} (x_t + \sqrt{1 - \bar{\alpha}_t} s_\theta(x_t, t)) \quad (2)$$

### Conditional Diffusion Models.

158 For conditional generation using unconditional diffusion  
 159 models ((Song et al., 2020b; Chung et al., 2022b; Yu et al.,  
 160 2023)), a common approach is to replace the score function  
 161 in DDPM equation A-7 with a conditional score function  
 162  $\nabla_{x_t} \log p(x_t|y)$ , where  $y$  is the conditioning variable. Using  
 163 Bayes' rule, the conditional score function can be decom-  
 164 posed into the unconditional score function and a likelihood  
 165 term:  $\nabla_{x_t} \log p(x_t|y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(y|x_t)$ ,  
 166 Incorporating this into the reverse SDE yields:

$$dx = \left[ -\frac{\beta_t}{2} x - \beta_t (\nabla_{x_t} \log p(x_t) \right. \\ \left. + \nabla_{x_t} \log p(y|x_t)) \right] dt + \sqrt{\beta_t} d\bar{\mathbf{w}} \quad (3)$$

167 The above SDE can be treated as a two-step process, the  
 168 first getting an unconditional denoised sample  $x_{t-1}$ , fol-  
 169 lowed by the gradient update with respect to  $x_t$ . Since, the  
 170 likelihood term  $\nabla_{x_t} \log p(y|x_t)$  is generally intractable, the  
 171 second term approximates a gradient update to minimizing  
 172 the guidance loss around the denoised sample  $x_{t-1}$ .

$$x'_{t-1} = \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0(x_t) + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} s_\theta(x_t, t) + \sigma_t \epsilon \quad (4)$$

$$x_{t-1} = x'_{t-1} - \zeta \nabla_{x_t} G(x_{0|t}, y) \quad (5)$$

173 where  $\zeta$  is a tunable step size. Here, Tweedie's estimate  $x_{0|t}$   
 174 is used since the guidance term is defined on the clean data  
 175  $x_0$ , i.e.,  $G_t(x_t, y) \approx \mathbb{E}_{p(x_0|x_t)}[G_t(x_0, y)] \sim G(x_{0|t}, y)$ .  
 176 The guidance term is optimized over a neighborhood around  
 177  $x_t \in \mathbb{R}^D$ .

178 Many methods use equation 5 for conditional generation  
 179 and various vision tasks ((Chung et al., 2022b; Kawar  
 180 et al., 2022; Yu et al., 2023; Song et al., 2023b)). For  
 181 example, Chung et al. (2022b) define an  $l_2$  loss term as  
 182  $|y - \mathcal{A}(x_{0|t})|_2^2$ , where  $\mathcal{A}$  represents a known differentiable  
 183 forward degradation model, effectively guiding the generated  
 184 sample to match the condition  $y$ .

## 3 Latent Guidance in Diffusion Model

185 As discussed in Section 1, diffusion models implicitly learn  
 186 perceptually consistent region. The diffusion process in  
 187 LDMs naturally generates on-manifold perceptually consist-  
 188 ent samples without requiring additional models to estimate

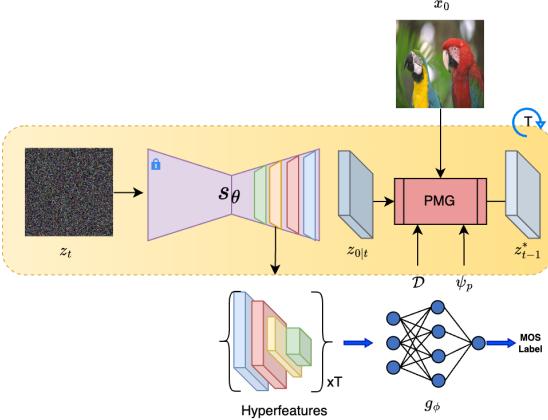


Figure 4: Illustrates aggregation of multi-scale and multi-timestep feature maps (Hyperfeatures) from denoising U-Net ( $s_\theta$ ) for NR-IQA. In the image,  $z_t$  is intermediate noisy latent,  $\mathcal{D}$  and  $\psi$  are decoder and perceptual quality features, respectively. PMG is our proposed algorithm for estimating the noisy sample  $z_{t-1}^*$ .

tangent spaces of the data manifold ((Srinivas et al., 2023; Bordt et al., 2023; He et al., 2023)), as we will demonstrate in this section. We exploit the same, and aim to generate perceptually consistent sample (input measurement) by guiding the sampling process on manifold. In the process, U-Net features align well with perceptual quality of measurement for relatively less noisy steps (discussed in 4).

In LDMs, the diffusion process operates within the latent space, training a score function  $s_\theta(z_t, t)$ . Let  $x \in \mathbb{R}^D$  represent the original high-dimensional data, and let  $\mathcal{E} : \mathbb{R}^D \rightarrow \mathbb{R}^k$  be an encoder and  $\mathcal{D} : \mathbb{R}^k \rightarrow \mathbb{R}^D$  be a decoder, where  $k \ll D$ . The embeddings in the latent space are given by  $z = \mathcal{E}(x) \in \mathbb{R}^k$ .

To guide the sampling process towards the perceptually consistent region on the manifold and ensure perceptually consistent hyperfeature extraction from the denoising score function, we propose the PMG framework. An overview of our proposed sampling process is depicted in Fig. 1, which illustrates the step-by-step guidance for extracting perceptually aligned features in the latent space. During the sampling process, our PMG guidance term push estimate on-manifold to reduce the error term.

### 3.1 Perceptual Manifold Guidance

We propose using perceptual features from an input measurement  $y$  derived via a perceptual quality metric  $\psi_p$  in the conditional score function, leading to  $\nabla_{z_t} \log p(z_t | \psi_p(y), y)$ . The choice of  $\psi_p$  is detailed in Section 4 and Appendix D. Before redefining the sampling steps, let's first consider the noisy sample manifolds.

Given Assumption 1, Chung et al. (2022a;b) show that noisy data  $x_t$  is probabilistically concentrated on a  $(D - 1)$ -dimensional manifold  $\mathcal{M}_t$ , which encapsulates the clean data manifold  $\mathcal{M}$ . Formally (see Appendix B for a detailed proof):

**Proposition 1** (Noisy Data Manifold) Let the distance function be defined as  $d(x, \mathcal{M}) := \inf_{y \in \mathcal{M}} \|x - y\|_2$ , and define the neighborhood around the manifold  $\mathcal{M}$  as  $B(\mathcal{M}; r) := \{x \in \mathbb{R}^D \mid d(x, \mathcal{M}) < r\}$ . Consider the distribution of noisy data given by  $p(x_t) = \int p(x_t | x_0) p(x_0) dx_0$ ,  $p(x_t | x_0) := \mathcal{N}(\sqrt{\alpha_t} x_0, (1 - \bar{\alpha}_t) \mathbf{I})$  represents the Gaussian perturbation of the data at time  $t$ , and  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$  is the cumulative product of the noise schedule  $\alpha_t$ . Under the Assumption 1, the distribution  $p_t(x_t)$  is concentrated on a  $(D - 1)$ -dim manifold  $\mathcal{M}_t := y \in \mathbb{R}^D : d(y, \sqrt{x_t} \mathcal{M}) = r_t := \sqrt{(1 - \bar{\alpha}_t)(D - k)}$ .

Most posterior sampling methods ((Chung et al., 2022a)) optimize the guidance term  $G(x_{0|t}, y)$  over  $x_t \in \mathbb{R}^D$ , whereas the score function  $s_\theta$  is trained only with samples on  $\mathcal{M}_t$ , as indicated by Proposition 1. This discrepancy implies that the solution  $x_t^*$  (leading to  $x_{0|t}^*$  via Tweedie's formula 2) may not reside on  $\mathcal{M}_t$ , resulting in a suboptimal solution ((Yu et al., 2023)). To overcome this limitation, we propose a solution over  $\mathcal{M}_t$ . From Assumption 1, the manifold  $\mathcal{M}_t$  coincides with its tangent space  $\mathcal{T}_{x_t} \mathcal{M}_t$ , i.e.,  $\mathcal{T}_{x_t} \mathcal{M}_t \simeq \mathbb{R}^k$  with  $k \ll D$  ((Park et al., 2023)). Practically, we optimize the guidance term  $G(x_{0|t}, y)$  over  $x_t \in \mathcal{T}_{x_t} \mathcal{M}_t$ . This new compact solution space ensures consistent on-manifold sampling throughout the process.

The latent space of a well-trained autoencoder implicitly captures the lower-dimensional structure of the data manifold, which can be leveraged for tangent space projection ((Srinivas et al., 2023; Bordt et al., 2023)). The latent processing of LDMs aids this as the samples already lie in the lower-dimensional space  $\mathbb{R}^k$ . Formally (proof follows He et al. (2023), see Appendix B):

**Proposition 2** (On-manifold sample with LDM) Given a perfect autoencoder, i.e.  $x = \mathcal{D}(\mathcal{E}(x))$ , and a gradient  $\nabla_{z_{0|t}} G(z_{0|t}, y) \in \mathcal{T}_{z_0} \mathcal{Z}$  then  $\mathcal{D}(\nabla_{z_{0|t}} G(z_{0|t}, y)) \in \mathcal{T}_{x_0} \mathcal{M}$ .

For LDMs, the minimization of the guidance term occurs within the tangent space of the clean data manifold. This guarantees that the generated sample remains close to the real data, without deviations. Although authors of Rout et al. (2024) do not explicitly discuss on-manifold sampling in LDMs, their results empirically suggest the inherent manifold consistency of LDMs.

Having defined consistent on-manifold sampling, we finally present our Perceptual Manifold Guidance (PMG) for perceptually consistent on-manifold sampling. Using Bayes' theorem on our new conditional score function

220  $\nabla_{z_t} \log p(z_t | \psi_p(y), y)$  (see Appendix B for details):  
 221  
 222  $\nabla_{z_t} \log p(z_t | \psi_p(y), y) \approx \nabla_{z_t} \log p(z_t)$   
 223  
 224  $+ \nabla_{z_t} \log p(\psi_p(y) | z_t) + \nabla_{z_t} \log p(y | z_t)$  (6)  
 225

226 From ((Rout et al., 2024)), LDM’s intractable terms can be  
 227 approximated as:

228  $\nabla_{z_t} \log p(\psi_p(y) | z_t) = \nabla_{z_t} \log p(\psi_p(y) | x_{0|t} = \mathcal{D}(z_{0|t}))$   
 229 (7)  
 230

231  $\nabla_{z_t} \log p(y | z_t) = \nabla_{z_t} \log p(y | x_{0|t} = \mathcal{D}(z_{0|t})).$  (8)  
 232

233 Based on Assumption 1, Propositions 1 and 2, Equations  
 234 6-8, and Lemma 2 in Appendix B, we derive the following  
 235 theorem for Perceptual Manifold Guidance (proof in  
 236 Appendix B):  
 237

238 **Theorem 1** (Perceptual Manifold Guidance) Given As-  
 239 sumption 1, given a perfect encoder  $\mathcal{E}$ , decoder  $\mathcal{D}$ ,  
 240 and an efficient score function  $s_\theta(z_t, t)$ , let the gradient  
 241  $\nabla_{z_{0|t}} G_1(\mathcal{D}(z_{0|t}), y)$  and  $\nabla_{z_{0|t}} G_2(\psi_p(\mathcal{D}(z_{0|t})), \psi_p(y))$  re-  
 242 side on the tangent space  $\mathcal{T}_{z_{0|t}} \mathcal{Z}$  of the latent manifold  $\mathcal{Z}$ .  
 243 Throughout the diffusion process, all update terms  $z_t$  remain  
 244 on noisy latent manifolds  $\mathcal{Z}_t$ , with  $z''_{0|t}$  lying in a perceptually  
 245 consistent manifold locality.  
 246

247 Discretized steps based on Theorem 1 can be written as:  
 248

249  $z'_{0|t} \leftarrow z_{0|t} - \zeta_1 \nabla_{z_{0|t}} G_1(\mathcal{D}(z_{0|t}), y)$  (9)  
 250

251  $z''_{0|t} \leftarrow z'_{0|t} - \zeta_2 \nabla_{z_{0|t}} G_2(\psi_p(\mathcal{D}(z_{0|t})), \psi_p(y))$  (10)  
 252

253  $z^*_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} z''_{0|t} - \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} s_\theta(z_t, t) + \sigma_t \epsilon$  (11)  
 254

255 Equation 9, 10, 11 are posterior sampling, perceptual con-  
 256 sistency step, and DDIM update step respectively. We use  
 257  $G_i$  as  $l_2$  functions. The perceptual consistency step in PMG  
 258 (Equation 10), guides the sample to be close to the percep-  
 259 tual quality of the input measurement. Specifically, during  
 260 sampling, the perceptual guidance term adjusts the earlier  
 261 estimate of the clean latent sample  $z'_{0|t}$  toward a perceptually  
 262 consistent locality on the tangent space of the clean  
 263 latent manifold,  $\mathcal{T}_{z_{0|t}} \mathcal{Z}$ . From Theorem 1, all update terms  
 264  $z_t$ , including  $z_{0|t}$ , are on the manifold  $\mathcal{Z}$ . This ensures the  
 265 sampling process remains close to a perceptually consistent  
 266 region on the manifold, with  $\mathcal{D}(z''_{0|t})$  closely aligned with  
 267 the perceptual quality of the input measurement (see Fig. 1).  
 268 We use internal representations from the denoising U-Net  
 269  $s_\theta$  to measure this perceptual consistency, detailed in Sec-  
 270 tion 3.2. The effectiveness of this approach is demonstrated  
 271 empirically in Section 4, where the absence of the percep-  
 272 tual guidance term ( $\psi_\phi$ ) in LGDM results in suboptimal  
 273 performance.  
 274

### Algorithm 1 LGDM: Latent Guidance in Diffusion Models

**Require:** Input image  $x$ , encoder  $\mathcal{E}(\cdot)$ , decoder  $\mathcal{D}(\cdot)$ , score function  $s_\theta(\cdot, t)$ , perceptual metric  $\psi_p(\cdot)$ , regression model  $g_\phi$ , time steps  $T$ , guidance weights  $\zeta_1, \zeta_2$

**Output:** Predicted quality score  $q_p$

```

1:  $z_0 \leftarrow \mathcal{E}(x)$ 
2:  $\mathbf{H} \leftarrow \emptyset$ 
3: for  $t = T, T-1, \dots, 1$  do
4:    $\epsilon \sim \mathcal{N}(0, I)$ 
5:    $\epsilon_t, h_t = s_\theta(z_t, t)$ 
6:    $\mathbf{H} \leftarrow \mathbf{H} \cup h_t$ 
7:    $\hat{z}_{0|t} \leftarrow \frac{1}{\sqrt{\bar{\alpha}_t}} (z_t - \sqrt{1 - \bar{\alpha}_t} \cdot \epsilon_t)$ 
8:    $z'_{0|t} \leftarrow \hat{z}_{0|t} - \zeta_1 \nabla_{z_{0|t}} G_1(\mathcal{D}(\hat{z}_{0|t}), x)$ 
9:    $z''_{0|t} \leftarrow z'_{0|t} - \zeta_2 \nabla_{z_{0|t}} G_2(\psi_p(\mathcal{D}(\hat{z}_{0|t})), \psi_p(x))$ 
10:   $z^*_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \cdot z''_{0|t} - \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \cdot \epsilon_t + \sigma_t \epsilon$ 
11: end for
12:  $q_p \leftarrow g_\phi(\mathbf{H})$ 
13: return  $q_p$ 
```

## 3.2 Diffusion Hyperfeatures & NR-IQA

We propose to use *diffusion hyperfeatures*—multi-scale and multi-timestep feature maps extracted from the denoising U-Net ( $s_\theta$ ) of a pretrained latent diffusion model.

Previous NR-IQA methods typically rely on features extracted from fine-tuned models ((Ke et al., 2021; Madhusudana et al., 2022; Liu et al., 2022; Saini et al., 2024)). However, these methods often use features from the final layer or a single scale, limiting their ability to capture the complete spectrum of image characteristics. In contrast, we harness the rich hierarchical representations available in the intermediate layers of the denoising U-Net across multiple diffusion timesteps. This enables us to capture both coarse and fine-grained image features crucial for assessing perceptual quality ((Ghildyal et al., 2024)). Recent studies ((Xu et al., 2023; Wu et al., 2023; Luo et al., 2024)) have shown that intermediate representations within diffusion models exhibit reliable semantic correspondences, although they have mostly been used for tasks such as data augmentation, generation, and segmentation. We hypothesize that these intermediate features also correlate strongly with human perceptual judgments of image quality, motivated by the diffusion models’ ability to generate perceptually appealing images and their robust representational capabilities for various downstream tasks ((Zhao et al., 2023)).

To extract these diffusion hyperfeatures, we gather intermediate feature maps from all upsampling layers of the denoising U-Net across multiple diffusion timesteps during the sampling process, see Fig. 4. These feature maps inherently contain shared representations that capture different image characteristics, such as semantic content, at

Table 1: Comparison of our proposed LGDM with SOTA NR-IQA methods on PLCC and SRCC Scores for authentic IQA datasets. The best results are in red, and the second-best results are in blue.

METHODS	LIVEC		KONIQ		FLIVE		SPAQ	
	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑	PLCC↑	SRCC↑
ILNIQE ((ZHANG ET AL., 2015))	0.508	0.508	0.537	0.523	-	-	0.712	0.713
BRISQUE ((MITTAL ET AL., 2012))	0.629	0.629	0.685	0.681	0.341	0.303	0.817	0.809
WADIQAM ((BOSSE ET AL., 2018))	0.671	0.682	0.807	0.804	0.467	0.455	-	-
DBCNN ((ZHANG ET AL., 2020))	0.869	0.851	0.884	0.875	0.551	0.545	0.915	0.911
TIQA ((STEPIEN & OSZUST, 2023))	0.861	0.845	0.903	0.892	0.581	0.541	-	-
METAIQA ((ZHU ET AL., 2020))	0.802	0.835	0.856	0.887	0.507	0.540	-	-
P2P-BM ((YING ET AL., 2020))	0.842	0.844	0.885	0.872	0.598	0.526	-	-
HYPERIQA ((SU ET AL., 2020))	0.882	0.859	0.917	0.906	0.602	0.544	0.915	0.911
TRES ((GOLESTANEH ET AL., 2022A))	0.877	0.846	0.928	0.915	0.625	0.554	-	-
MUSIQ ((KE ET AL., 2021))	0.746	0.702	0.928	0.916	0.661	0.566	0.921	0.918
RE-IQA ((SAHA ET AL., 2023))	0.854	0.840	0.923	0.914	0.733	0.645	0.925	0.918
LODA ((XU ET AL., 2024))	0.899	0.876	0.944	0.932	0.679	0.578	0.928	0.925
LGDM- $\psi_\phi$	0.853	0.842	0.929	0.921	0.751	0.672	0.912	0.901
LGDM- $\psi_{BRISQUE}$	0.852	0.840	0.924	0.919	0.691	0.598	0.917	0.908
LGDM- $\psi_{MUSIQ}$	0.869	0.858	0.939	0.928	0.747	0.672	0.922	0.920
LGDM- $\zeta_1 = 0, \psi_{SDv1.5}$	0.901	0.893	0.952	0.941	0.799	0.683	0.931	0.929
LGDM- $\psi_{RE-IQA}$	0.903	0.891	0.952	0.944	0.761	0.679	0.929	0.924
LGDM- $\psi_{SDv1.5}$	0.940	0.908	0.972	0.967	0.812	0.705	0.948	0.947

various scales and levels of abstraction. Since these features are distributed over both the network layers and diffusion timesteps, we aggregate  $s_\theta$  layers and timesteps as diffusion hyperfeatures for NR-IQA. Specifically, the set of all extracted features is denoted as:

$$\mathbf{H} = \bigcup_{t=1}^T \left\{ s_\theta^{(l)}(\mathbf{x}_t) \mid l \in \mathcal{L} \right\} \quad (12)$$

where  $s_\theta^{(l)}(\mathbf{x}_t)$  represents the feature map from layer  $l$  at timestep  $t$ ,  $\mathcal{L}$  is the set of layers from which we extract features, and  $T$  is the total number of timesteps considered. Our experiments show that perceptual quality is built progressively during reverse diffusion (later timesteps), making an appropriate range of sampling timestep to be in [0-100], where  $t = 0$  denotes a completely clean image. Finally, with the aggregated diffusion hyperfeatures  $\mathbf{H}$ , we employ a lightweight regression network  $g_\phi$  parameterized by  $\phi$  following standard NR-IQA practice (Madhusudana et al., 2022; Saha et al., 2023; Saini et al., 2024) to predict the perceptual quality score:

$$q_p = g_\phi(\mathbf{H}) \quad (13)$$

where  $q_p$  is the predicted quality score. Importantly, the diffusion model  $s_\theta$  remains fixed and is not fine-tuned at any stage, preserving its zero-shot generalization capabilities. The use of multi-scale and multi-timestep features enables the model to be sensitive to different types of distortions and image artifacts, which might not be captured when using single-scale feature. Our experiments show

that LGDM with PMG gives better performance across all NR-IQA benchmarks (see Section 4).

## 4 Experiments

### 4.1 Experimental Settings

To thoroughly evaluate the effectiveness of our proposed method, we conducted extensive experiments on ten publicly available and well-recognized IQA datasets, covering synthetic distortions, authentic distortions, and the latest AI-generated content (AIGC). These datasets are summarized in Table 6 (Appendix D). Many previous methods focused only on synthetic distortions, because of the difficulty of generalizing to real-world distortions. By contrast, our LDM is pretrained on a diverse dataset that includes both synthetic and authentic distortions, allowing for a fair comparison across all types of IQA datasets, including recent AIGC datasets.

For LDM, we use the widely adopted Stable Diffusion v1.5 (Rombach et al., 2022), pretrained on the LAION-5B dataset (Schuhmann et al., 2022). For text conditioning, we use an empty string “” as prompt. We run 10 DDIM steps, with  $t$  within the range (0, 100] and set the hyperparameters  $\zeta_1$  and  $\zeta_2$  to 1 and 0.2, respectively. For regression, we use a small neural network with two hidden layers. We use Pearson Linear Correlation Coefficient (PLCC) and Spearman’s Rank Order Correlation Coefficient (SRCC) as evaluation metrics. The impact of the choice of  $\psi_p$  is discussed in detail in the ablation study and Appendix D. All experiments were conducted on an NVIDIA A100 GPU using

Table 2: PLCC and SRCC comparison of LGDM on AI-Generated Datasets for IQA. The best results are in red, and the second-best results are in blue.

Method	AGIQA-1K		AGIQA-3K	
	PLCC↑	SRCC↑	PLCC↑	SRCC↑
CONTRIQUE (Madhusudana et al., 2022)	0.708	0.670	0.868	0.804
RE-IQA (Saha et al., 2023)	0.670	0.614	0.845	0.785
GenZIQA (De et al., 2024)	0.861	0.840	0.892	0.832
LGDM- $\psi_{SDv1.5}$	0.903	0.891	0.929	0.863

PyTorch. Additional implementation details are provided in the Appendix C.

## 4.2 Experimental Results & Comparisons

We evaluate LGDM on ten datasets. Table 1 presents the performance of LGDM on four authentic distortion (“In the Wild”) datasets, with LGDM- $\psi_{SDv1.5}$  achieving the best results across all datasets. Here  $\psi_{SDv1.5}$  denote that we use SDv1-5 as perceptual metric, and feed its intermediate feature during sampling. Specifically, on the LIVEC ((Ghadiallyaram & Bovik, 2015)) dataset, LGDM- $\psi_{SDv1.5}$  attained a PLCC of 0.940 and an SRCC of 0.908, significantly surpassing the previous best method, LoDA ((Xu et al., 2024)). On the FLIVE dataset ((Ying et al., 2020)), which contains the largest collection of human-labeled authentically distorted images emulating social media content (UGC), our method achieves a state-of-the-art PLCC of 0.812 and an SRCC of 0.705, demonstrating its robustness at handling diverse and complex real-world distortions. We also evaluated our method on AIGC datasets to assess its ability to handle AI-generated images, which often present unique challenges. As shown in Table 2, LGDM- $\psi_{SDv1.5}$  outperformed previous methods on both AGIQA-1K ((Li et al., 2023)) and AGIQA-3K ((Li et al., 2024a)) datasets, achieving PLCC scores of 0.903 and 0.929, respectively. As compared to GenZIQA ((De et al., 2024)), the previous best-performing method, our approach demonstrates significant improvements, highlighting its strong prior for AI-generated content, which is often lacking in previous methods. Additional results on synthetic distortion datasets are provided in Appendix D.

## 4.3 Ablation Study

**Cross-Dataset Generalization.** We conducted cross-dataset evaluations to assess the generalization capability of our method. Table 3 presents the results of inter-dataset evaluations. Our LGDM- $\psi_{SDv1.5}$  consistently achieved the highest SRCC scores across all cross-dataset combinations, demonstrating the robustness and strong generalization capabilities of LGDM’s perceptual feature maps. Interestingly, when trained on smaller datasets such as LIVEC and evaluated on x10 larger dataset KonIQ, LGDM- $\psi_{SDv1.5}$  still outperforms other methods by large margin. This demon-

Table 3: SRCC Scores for Cross Dataset Evaluations. The best results are in red, and the second-best results are in blue.

Train	Test	Methods			
		REIQA	DEIQT	LoDA	LGDM- $\psi_{SDv1.5}$
FLIVE	KonIQ	0.764	0.733	0.763	0.802
FLIVE	LIVEC	0.699	0.781	0.805	0.849
KonIQ	LIVEC	0.791	0.794	0.811	0.853
LIVEC	KonIQ	0.769	0.744	0.745	0.794

strates the strong perceptual features of LGDM and less dependence on regression model.

**Different Choices of  $\psi_p$ .** We evaluated different perceptual model features  $\psi_p$  to analyze their impact on perceptual guidance during the sampling process. We list experimental results for all variants of  $\psi_p$  in Table 1 and 7. Specifically, we use  $\phi$ (no perceptual guidance), BRISQUE, MUSIQ, RE-IQA, and SDv1-5.  $\zeta_1 = 0$  implies that during sampling we do not perform update in Equation 9, i.e. gradient for l2 loss of actual image, and just rely on perceptual features’ loss (Equation 10). Given large pretrained diffusion model, they have strong bias for perceptual features that provide the best guidance, hence  $\psi_{SDv1.5}$  gives the best performance across datasets. It should be noted that LGDM with perceptual features generally boosts the performance of underlying perceptual metric with strong priors from Diffusion models. For instance, RE-IQA demonstrating a strong perceptual correlation, when used with LGDM gives around 3%-5% performance boost. While models with worse human judgment correlation (e.g.,  $\psi_{BRISQUE}$ ) tend to reduce performance by deviating samples away from the perceptually consistent regions as compared when no perceptual guidance is applied ( $\psi_\phi$ ). Similar experiments were also conducted on synthetic datasets, please see Appendix D.

Table 4: SRCC and time taken for different number of timesteps on FLIVE ((Ying et al., 2020)) by LGDM- $\psi_{SDv1.5}$ .

Time Steps	SRCC↑	Time Taken (s) ↓
1	0.624	3.27
5	0.673	9.89
10	0.705	21.30
50	0.711	110.45

**Time Step Variations and Performance.** Table 4 shows the impact of varying the number of timesteps during sampling on FLIVE ((Ying et al., 2020)) using LGDM- $\psi_{SDv1.5}$ . We observe a convex trend in the SRCC scores—performance improves with an increase in the number of timesteps up to 50, but further increments result in diminishing returns, with increased computational cost (time taken). For practical use,

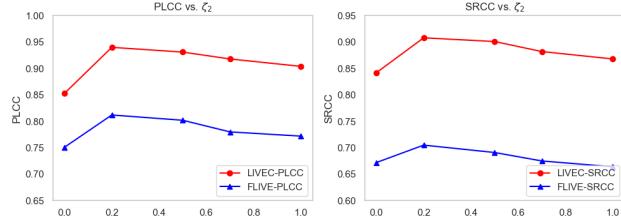


Figure 5: Effect of  $\zeta_2$  on the hyperfeatures. Figure on the left and right, shows the PLCC and SRCC values respectively for FLIVE and LIVEC datasets. Hyperfeatures were collected for single time step.

a trade-off between SRCC and computational efficiency is required. We choose the optimum value of 10 time steps that gives us very close to highest performance with five times less compute time.

Table 5: PLCC and SRCC for Different Versions of SD on LIVEC ((Ghadiyaram & Bovik, 2015)) and FLIVE ((Ying et al., 2020)). The best results are highlighted in red, and the second-best results are highlighted in blue.

SD Version	LIVEC		FLIVE	
	PLCC↑	SRCC↑	PLCC↑	SRCC↑
1.3	0.932	0.901	0.804	0.695
1.4	0.938	0.903	0.807	0.698
1.5	0.940	0.908	0.812	0.705
2	0.910	0.882	0.781	0.674
2.1	0.917	0.886	0.788	0.681

**Effect of Different Versions of SD.** We also compared the effect of using different versions of Stable Diffusion. Table 5 shows that SD v1.5 consistently outperforms other versions, achieving the highest PLCC and SRCC scores. Specifically, SD v1.5 reaches an SRCC of 0.908 on LIVEC and 0.705 on FLIVE, outperforming newer versions like v2.0 and v2.1, which exhibit lower correlation values. The decline in performance in newer versions may be attributed to architectural changes or training modifications that diverge from the characteristics required for effective NR-IQA, i.e. focus on the generation of high quality aesthetic image, rather than a broader coverage of image quality.

**Impact of Weights of Perceptual Guidance Term ( $\zeta$ ).** We conducted experiments to evaluate the effect of different values for the hyperparameters  $\zeta_1$  and  $\zeta_2$ . LGDM- $\psi_\phi$  in Table 1 represents the case where only the first term in PMG is used. The first term provides a strong baseline due to content bias effects from data consistency. We set  $\zeta_1 = 1$  following Rout et al. (2024). In Fig. 5, we show SRCC scores for different values of  $\zeta_2$  on FLIVE. We observe that values too small or large for  $\zeta$  lead to poor perceptual features by pushing the samples away from the perceptually

consistent region on  $\mathcal{M}$ .

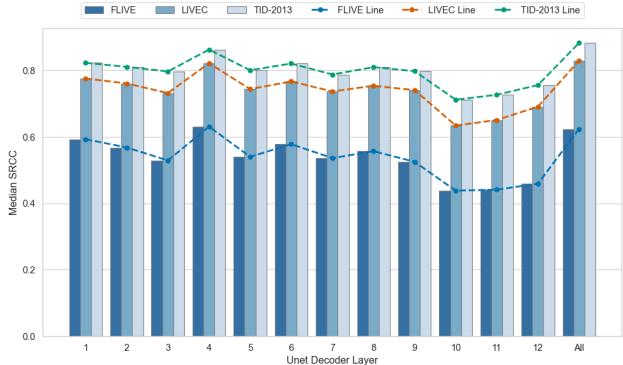


Figure 6: Contribution of individual layer of SDv1.5 towards SRCC for FLIVE, LIVEC, and TID-2013. A consistent pattern of contribution of layers across synthetic and authentic datasets.

**Contribution of Individual SD Layers** We also investigate the contribution of individual layer in SDv1.5 U-Net. It may be observed that different layers contribute differently to the overall performance. Given the high computational cost of running SD, we limited this experiment to single time sampling. Notably, layer 4 contributes significantly more to final quality prediction than other layers. Initial layers seem to develop strong perceptual correlation, we have reported the corresponding SRCC values for each decoder layer in Fig. 6.

## 5 Conclusion

In this work, we introduced the Latent Guidance in Diffusion Models (LGDM) for No-Reference Image Quality Assessment (NR-IQA). Leveraging the strong representation capabilities of pretrained latent diffusion models (LDMs), we proposed Perceptual Manifold Guidance (PMG) to direct the sampling process toward perceptually consistent regions on the data manifold. We demonstrated the value of extracting multi-scale and multi-timestep features—diffusion hyperfeatures from the denoising U-Net, providing a rich representation for quality assessment. To our knowledge, this is the first work utilizing pretrained LDMs directly for NR-IQA without any fine-tuning of diffusion backbone.

## References

- Agnolucci, L., Galteri, L., Bertini, M., and Del Bimbo, A. Arniqa: Learning distortion manifold for image quality assessment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 189–198, 2024.

Babnik, Ž., Peer, P., and Štruc, V. ediffiqa: Towards effi-

- 440 client face image quality assessment based on denoising  
 441 diffusion probabilistic models. *IEEE Transactions on*  
 442 *Biometrics, Behavior, and Identity Science*, 2024.
- 443 Baranchuk, D., Rubachev, I., Voynov, A., Khrulkov, V.,  
 444 and Babenko, A. Label-efficient semantic segmentation  
 445 with diffusion models. *arXiv preprint arXiv:2112.03126*,  
 446 2021.
- 447 Bordt, S., Upadhyay, U., Akata, Z., and von Luxburg, U.  
 448 The manifold hypothesis for gradient-based explanations.  
 449 In *Proceedings of the IEEE/CVF Conference on Computer*  
 450 *Vision and Pattern Recognition*, pp. 3697–3702,  
 451 2023.
- 452 Bosse, S., Maniry, D., Müller, K.-R., Wiegand, T., and  
 453 Samek, W. Deep neural networks for no-reference and  
 454 full-reference image quality assessment. *IEEE Transactions*  
 455 *on Image Processing*, 27(1):206–219, 2018. doi:  
 456 10.1109/TIP.2017.2760518.
- 457 Chung, H., Kim, J., Mccann, M. T., Klasky, M. L., and Ye,  
 458 J. C. Diffusion posterior sampling for general noisy inverse  
 459 problems. *arXiv preprint arXiv:2209.14687*, 2022a.
- 460 Chung, H., Sim, B., Ryu, D., and Ye, J. C. Improving  
 461 diffusion models for inverse problems using manifold  
 462 constraints. *Advances in Neural Information Processing*  
 463 *Systems*, 35:25683–25696, 2022b.
- 464 Chung, H., Lee, S., and Ye, J. C. Decomposed diffusion  
 465 sampler for accelerating large-scale inverse problems.  
 466 *arXiv preprint arXiv:2303.05754*, 2023.
- 467 De, D., Mitra, S., and Soundararajan, R. Genziqa: Generalized  
 468 image quality assessment using prompt-guided  
 469 latent diffusion models. *arXiv preprint arXiv:2406.04654*,  
 470 2024.
- 471 Efron, B. Tweedie’s formula and selection bias. *Journal*  
 472 *of the American Statistical Association*, 106(496):1602–  
 473 1614, 2011.
- 474 Fang, Y., Zhu, H., Zeng, Y., Ma, K., and Wang, Z. Perceptual  
 475 quality assessment of smartphone photography. In *Proceedings*  
 476 *of the IEEE/CVF conference on computer*  
 477 *vision and pattern recognition*, pp. 3677–3686, 2020.
- 478 Fei, B., Lyu, Z., Pan, L., Zhang, J., Yang, W., Luo, T., Zhang,  
 479 B., and Dai, B. Generative diffusion prior for unified  
 480 image restoration and enhancement. In *Proceedings of the*  
 481 *IEEE/CVF Conference on Computer Vision and Pattern*  
 482 *Recognition*, pp. 9935–9946, 2023.
- 483 Feng, R., Gao, Y., Jin, X., Feng, R., and Chen, Z. Semantically  
 484 structured image compression via irregular group-based  
 485 decoupling. In *Proceedings of the IEEE/CVF International*  
 486 *Conference on Computer Vision*, pp. 17237–  
 487 17247, 2023.
- 488 Gao, T., Pan, W., Zhang, Y., Zhao, S., Zhang, S., Zheng, X.,  
 489 Li, K., Cao, L., and Ji, R. Local manifold learning for  
 490 no-reference image quality assessment. *arXiv preprint*  
 491 *arXiv:2406.19247*, 2024.
- 492 Ghadiyaram, D. and Bovik, A. C. Massive online crowd-  
 493 sourced study of subjective and objective picture quality.  
 494 *IEEE Transactions on Image Processing*, 25(1):372–387,  
 495 2015.
- 496 Ghildyal, A., Barman, N., and Zadtootaghaj, S. Foundation  
 497 models boost low-level perceptual similarity metrics.  
 498 *arXiv preprint arXiv:2409.07650*, 2024.
- 499 Golestaneh, S. A., Dadsetan, S., and Kitani, K. M. No-  
 500 reference image quality assessment via transformers, rel-  
 501 ative ranking, and self-consistency. In *Proceedings of the*  
 502 *IEEE/CVF Winter Conference on Applications of Computer*  
 503 *Vision*, pp. 3209–3218, 2022a.
- 504 Golestaneh, S. A., Dadsetan, S., and Kitani, K. M. No-  
 505 reference image quality assessment via transformers, rel-  
 506 ative ranking, and self-consistency. In *Proceedings of the*  
 507 *IEEE/CVF winter conference on applications of computer*  
 508 *vision*, pp. 1220–1230, 2022b.
- 509 Guan, F., Jiang, G., Song, Y., Yu, M., Peng, Z., and Chen,  
 510 F. No-reference high-dynamic-range image quality as-  
 511 sessment based on tensor decomposition and manifold  
 512 learning. *Applied Optics*, 57(4):839–848, 2018.
- 513 Guan, J., Yi, S., Zeng, X., Cham, W.-K., and Wang, X. Vi-  
 514 sual importance and distortion guided deep image quality  
 515 assessment framework. *IEEE Transactions on Multi媒*  
 516 *dia*, 19(11):2505–2520, 2017.
- 517 He, Y., Murata, N., Lai, C.-H., Takida, Y., Uesaka, T., Kim,  
 518 D., Liao, W.-H., Mitsufuji, Y., Kolter, J. Z., Salakhutdinov,  
 519 R., et al. Manifold preserving guided diffusion.  
 520 *arXiv preprint arXiv:2311.16424*, 2023.
- 521 Ho, J., Jain, A., and Abbeel, P. Denoising diffusion proba-  
 522 bilistic models. *Advances in neural information process-*  
 523 *ing systems*, 33:6840–6851, 2020.
- 524 Hosu, V., Lin, H., Sziranyi, T., and Saupe, D. Koniq-10k:  
 525 An ecologically valid database for deep learning of blind  
 526 image quality assessment. *IEEE Transactions on Image*  
 527 *Processing*, 29:4041–4056, 2020.
- 528 Hou, J., Zhu, Z., Hou, J., Liu, H., Zeng, H., and Yuan, H.  
 529 Global structure-aware diffusion process for low-light  
 530 image enhancement. *Advances in Neural Information*  
 531 *Processing Systems*, 36, 2024.
- 532 ITU-T RECOMMENDATION, P. Subjective video quality  
 533 assessment methods for multimedia applications. 1999.

- 495 Jiang, Q., Shao, F., Gao, W., Chen, Z., Jiang, G., and Ho,  
 496 Y.-S. Unified no-reference quality assessment of singly  
 497 and multiply distorted stereoscopic images. *IEEE Trans-*  
 498 *actions on Image Processing*, 28(4):1866–1881, 2018.  
 499
- 500 Kawar, B., Elad, M., Ermon, S., and Song, J. Denoising  
 501 diffusion restoration models. *Advances in Neural Infor-*  
 502 *mation Processing Systems*, 35:23593–23606, 2022.
- 503 Ke, J., Wang, Q., Wang, Y., Milanfar, P., and Yang, F. Musiq:  
 504 Multi-scale image quality transformer. In *Proceedings*  
 505 *of the IEEE/CVF international conference on computer*  
 506 *vision*, pp. 5148–5157, 2021.
- 507 Larson, E. C. and Chandler, D. M. Most apparent distortion:  
 508 full-reference image quality assessment and the role of  
 509 strategy. *Journal of electronic imaging*, 19(1):0111006–  
 510 0111006, 2010.
- 511 Li, C., Zhang, Z., Wu, H., Sun, W., Min, X., Liu, X., Zhai,  
 512 G., and Lin, W. Agiqa-3k: An open database for ai-  
 513 generated image quality assessment. *IEEE Transactions*  
 514 *on Circuits and Systems for Video Technology*, 2023.
- 515 Li, C., Kou, T., Gao, Y., Cao, Y., Sun, W., Zhang, Z., Zhou,  
 516 Y., Zhang, Z., Zhang, W., Wu, H., et al. Aqiqa-20k: A  
 517 large database for ai-generated image quality assessment.  
 518 *arXiv preprint arXiv:2404.03407*, 2(3):5, 2024a.
- 519 Li, X., Zheng, J., Hu, R., Zhang, Y., Li, K., Shen, Y., Zheng,  
 520 X., Liu, Y., Zhang, S., Dai, P., et al. Feature denoising  
 521 diffusion model for blind image quality assessment.  
 522 *arXiv preprint arXiv:2401.11949*, 2024b.
- 523 Lin, H., Hosu, V., and Saupe, D. Kadid-10k: A large-  
 524 scale artificially distorted iqa database. In *2019 Eleventh*  
 525 *International Conference on Quality of Multimedia Experi-*  
 526 *ence (QoMEX)*, pp. 1–3. IEEE, 2019.
- 527 Liu, J., Zhou, W., Li, X., Xu, J., and Chen, Z. Liqa: Lifelong  
 528 blind image quality assessment. *IEEE Transactions on*  
 529 *Multimedia*, 25:5358–5373, 2022.
- 530 Liu, J., Sun, H., and Katto, J. Learned image compression  
 531 with mixed transformer-cnn architectures. In *Proceedings*  
 532 *of the IEEE/CVF conference on computer vision and*  
 533 *pattern recognition*, pp. 14388–14397, 2023.
- 534 Luo, G., Dunlap, L., Park, D. H., Holynski, A., and Darrell,  
 535 T. Diffusion hyperfeatures: Searching through time and  
 536 space for semantic correspondence. *Advances in Neural*  
 537 *Information Processing Systems*, 36, 2024.
- 538 Madhusudana, P. C., Birkbeck, N., Wang, Y., Adsumilli,  
 539 B., and Bovik, A. C. Image quality assessment using  
 540 contrastive learning. *IEEE Transactions on Image Pro-*  
 541 *cessing*, 31:4149–4161, 2022.
- 542 Mittal, A., Soundararajan, R., and Bovik, A. C. Making a  
 543 “completely blind” image quality analyzer. *IEEE Signal*  
 544 *processing letters*, 20(3):209–212, 2012.
- 545 Moorthy, A. K. and Bovik, A. C. Blind image quality  
 546 assessment: From natural scene statistics to perceptual  
 547 quality. *IEEE transactions on Image Processing*, 20(12):  
 548 3350–3364, 2011.
- 549 Park, Y.-H., Kwon, M., Choi, J., Jo, J., and Uh, Y. Under-  
 550 standing the latent space of diffusion models through  
 551 the lens of riemannian geometry. *Advances in Neural*  
 552 *Information Processing Systems*, 36:24129–24142, 2023.
- 553 Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian,  
 554 K., Jin, L., Astola, J., Vozel, B., Chehdi, K., Carli, M.,  
 555 Battisti, F., et al. Color image database tid2013: Pecu-  
 556 liarities and preliminary results. In *European workshop*  
 557 *on visual information processing (EUVIP)*, pp. 106–111.  
 558 IEEE, 2013.
- 559 Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and  
 560 Ommer, B. High-resolution image synthesis with latent  
 561 diffusion models. In *Proceedings of the IEEE/CVF con-*  
 562 *ference on computer vision and pattern recognition*, pp.  
 563 10684–10695, 2022.
- 564 Rout, L., Raoof, N., Daras, G., Caramanis, C., Dimakis, A.,  
 565 and Shakkottai, S. Solving linear inverse problems prov-  
 566 ably via posterior sampling with latent diffusion models.  
 567 *Advances in Neural Information Processing Systems*, 36,  
 568 2024.
- 569 Saad, M. A., Bovik, A. C., and Charrier, C. Blind image  
 570 quality assessment: A natural scene statistics approach in  
 571 the dct domain. *IEEE transactions on Image Processing*,  
 572 21(8):3339–3352, 2012.
- 573 Saha, A., Mishra, S., and Bovik, A. C. Re-iqa: Unsupervised  
 574 learning for image quality assessment in the wild. In  
 575 *Proceedings of the IEEE/CVF conference on computer*  
 576 *vision and pattern recognition*, pp. 5846–5855, 2023.
- 577 Saini, S., Saha, A., and Bovik, A. C. Hidro-vqa: High  
 578 dynamic range oracle for video quality assessment. In  
 579 *Proceedings of the IEEE/CVF Winter Conference on Ap-*  
 580 *plications of Computer Vision*, pp. 469–479, 2024.
- 581 Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C.,  
 582 Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis,  
 583 C., Wortsman, M., et al. Laion-5b: An open large-scale  
 584 dataset for training next generation image-text models.  
 585 *Advances in Neural Information Processing Systems*, 35:  
 586 25278–25294, 2022.
- 587 Shao, H., Kumar, A., and Thomas Fletcher, P. The rieman-  
 588 nian geometry of deep generative models. In *Proceedings*  
 589 *of the IEEE Conference on Computer Vision and Pattern*  
 590 *Recognition Workshops*, pp. 315–323, 2018.

- 550 Sheikh, H. R., Sabir, M. F., and Bovik, A. C. A statistical  
 551 evaluation of recent full reference image quality assess-  
 552 ment algorithms. *IEEE Transactions on image processing*,  
 553 15(11):3440–3451, 2006.
- 554
- 555 Song, B., Kwon, S. M., Zhang, Z., Hu, X., Qu, Q., and  
 556 Shen, L. Solving inverse problems with latent diffusion  
 557 models via hard data consistency. *arXiv preprint arXiv:2307.08123*, 2023a.
- 558
- 559 Song, J., Meng, C., and Ermon, S. Denoising diffusion im-  
 560 plicit models. *arXiv preprint arXiv:2010.02502*, 2020a.
- 561
- 562 Song, J., Zhang, Q., Yin, H., Mardani, M., Liu, M.-Y.,  
 563 Kautz, J., Chen, Y., and Vahdat, A. Loss-guided diffusion  
 564 models for plug-and-play controllable generation.  
 565 In *International Conference on Machine Learning*, pp.  
 566 32483–32498. PMLR, 2023b.
- 567
- 568 Song, Y. and Ermon, S. Generative modeling by estimating  
 569 gradients of the data distribution. *Advances in neural  
 570 information processing systems*, 32, 2019.
- 571
- 572 Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Er-  
 573 mon, S., and Poole, B. Score-based generative modeling  
 574 through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.
- 575
- 576 Srinivas, S., Bordt, S., and Lakkaraju, H. Which models  
 577 have perceptually-aligned gradients? an explanation via  
 578 off-manifold robustness. *Advances in neural information  
 579 processing systems*, 36:21172–21195, 2023.
- 580
- 581 Stepien, I. and Oszust, M. Tiqa-psi: Toolbox for percep-  
 582 tual image quality assessment of pan-sharpened images.  
 583 *SoftwareX*, 23:101494, 2023.
- 584
- 585 Su, S., Yan, Q., Zhu, Y., Zhang, C., Ge, X., Sun, J., and  
 586 Zhang, Y. Blindly assess image quality in the wild guided  
 587 by a self-adaptive hyper network. In *Proceedings of the  
 588 IEEE/CVF conference on computer vision and pattern  
 589 recognition*, pp. 3667–3676, 2020.
- 590
- 591 Su, S., Yan, Q., Zhu, Y., Sun, J., and Zhang, Y. From distor-  
 592 tion manifold to perceptual quality: a data efficient blind  
 593 image quality assessment approach. *Pattern Recognition*,  
 594 133:109047, 2023.
- 595
- 596 Sun, S., Wei, L., Xing, J., Jia, J., and Tian, Q. Sddm: score-  
 597 decomposed diffusion models on manifolds for unpaired  
 598 image-to-image translation. In *International Conference  
 599 on Machine Learning*, pp. 33115–33134. PMLR, 2023.
- 600
- 601 Tu, Z., Yu, X., Wang, Y., Birkbeck, N., Adsumilli, B., and  
 602 Bovik, A. C. Rapique: Rapid and accurate video quality  
 603 prediction of user generated content. *IEEE Open Journal of Signal Processing*, 2:425–440, 2021.
- 604
- Tumanyan, N., Geyer, M., Bagon, S., and Dekel, T. Plug-  
 and-play diffusion features for text-driven image-to-  
 image translation. In *Proceedings of the IEEE/CVF Con-  
 ference on Computer Vision and Pattern Recognition*, pp.  
 1921–1930, 2023.
- Vincent, P. A connection between score matching and de-  
 noising autoencoders. *Neural computation*, 23(7):1661–  
 1674, 2011.
- Wang, Z. and Bovik, A. C. *Modern image quality assess-  
 ment*. PhD thesis, Springer, 2006.
- Wang, Z., Hu, B., Zhang, M., Li, J., Li, L., Gong, M., and  
 Gao, X. Diffusion model based visual compensation guid-  
 ance and visual difference analysis for no-reference im-  
 age quality assessment. *arXiv preprint arXiv:2402.14401*,  
 2024.
- Wu, W., Zhao, Y., Chen, H., Gu, Y., Zhao, R., He, Y., Zhou,  
 H., Shou, M. Z., and Shen, C. Datasetdm: Synthesizing  
 data with perception annotations using diffusion models.  
*Advances in Neural Information Processing Systems*, 36:  
 54683–54695, 2023.
- Xu, J., Liu, S., Vahdat, A., Byeon, W., Wang, X., and  
 De Mello, S. Open-vocabulary panoptic segmentation  
 with text-to-image diffusion models. In *Proceedings of  
 the IEEE/CVF Conference on Computer Vision and Pat-  
 tern Recognition*, pp. 2955–2966, 2023.
- Xu, K., Liao, L., Xiao, J., Chen, C., Wu, H., Yan, Q., and  
 Lin, W. Boosting image quality assessment through  
 efficient transformer adaptation with local feature en-  
 hancement. In *Proceedings of the IEEE/CVF Conference  
 on Computer Vision and Pattern Recognition*, pp. 2662–  
 2672, 2024.
- Ye, J., Wang, N., and Wang, X. Featurenerf: Learning  
 generalizable nerfs by distilling foundation models. In  
*Proceedings of the IEEE/CVF International Conference  
 on Computer Vision*, pp. 8962–8973, 2023.
- Ying, Z., Niu, H., Gupta, P., Mahajan, D., Ghadiyaram,  
 D., and Bovik, A. From patches to pictures (paq-2-piq):  
 Mapping the perceptual space of picture quality. In *Pro-  
 ceedings of the IEEE/CVF conference on computer vision  
 and pattern recognition*, pp. 3575–3585, 2020.
- You, J. and Korhonen, J. Transformer for image quality  
 assessment. In *2021 IEEE international conference on  
 image processing (ICIP)*, pp. 1389–1393. IEEE, 2021.
- Yu, J., Wang, Y., Zhao, C., Ghanem, B., and Zhang, J. Free-  
 dom: Training-free energy-guided conditional diffusion  
 model. In *Proceedings of the IEEE/CVF International  
 Conference on Computer Vision*, pp. 23174–23184, 2023.

605 Zhang, C., Zhang, C., Zhang, M., and Kweon, I. S. Text-to-  
606 image diffusion models in generative ai: A survey. *arXiv*  
607 preprint arXiv:2303.07909, 2023.

608 Zhang, L., Zhang, L., and Bovik, A. C. A feature-enriched  
609 completely blind image quality evaluator. *IEEE Transac-*  
610 *tions on Image Processing*, 24(8):2579–2591, 2015.

612 Zhang, W., Ma, K., Yan, J., Deng, D., and Wang, Z. Blind  
613 image quality assessment using a deep bilinear convolu-  
614 tional neural network. *IEEE Transactions on Circuits and*  
615 *Systems for Video Technology*, 30(1):36–47, 2020.

617 Zhang, Y., Shi, X., Li, D., Wang, X., Wang, J., and Li, H. A  
618 unified conditional framework for diffusion-based image  
619 restoration. *Advances in Neural Information Processing*  
620 *Systems*, 36, 2024.

621 Zhao, W., Rao, Y., Liu, Z., Liu, B., Zhou, J., and Lu, J.  
622 Unleashing text-to-image diffusion models for visual per-  
623 ception. In *Proceedings of the IEEE/CVF International*  
624 *Conference on Computer Vision*, pp. 5729–5739, 2023.

626 Zhu, H., Li, L., Wu, J., Dong, W., and Shi, G. Metaiqa: Deep  
627 meta-learning for no-reference image quality assessment.  
628 In *Proceedings of the IEEE/CVF conference on computer*  
629 *vision and pattern recognition*, pp. 14143–14152, 2020.

630

631

632

633

634

635

636

637

638

639

640

641

642

643

644

645

646

647

648

649

650

651

652

653

654

655

656

657

658

659

---

## Appendix

---

660  
661  
662  
663  
664  
665 Here, we provide additional theoretical proof, implementation details, and experimental results to complement those in  
666 the main paper. Specifically, Section A discusses more related work and background on diffusion models and NR-IQA,  
667 Section B presents detailed theoretical proofs and supporting discussionn, Section C describes the implementation details,  
668 Section D includes further quantitative analyses to demonstrate the performance of LGDM, Finally in Section E we discuss  
669 the main limitation of our proposed method and possible extensions.  
670

## A Related Work

### A.1 NR-IQA

671 No-Reference Image Quality Assessment (NR-IQA) has been a focal point of research over the past two decades, aiming to  
672 evaluate image quality based on human perception without relying on reference images. Early approaches predominantly  
673 utilized handcrafted features derived from natural scene statistics (NSS), with models such as BRISQUE ((Mittal et al.,  
674 2012), DIIVINE (Moorthy & Bovik, 2011), BLIINDS (Saad et al., 2012), and NIQE (Mittal et al., 2012)). While these  
675 methods effectively leveraged statistical regularities in natural images, their performance often suffered when dealing with  
676 complex or unseen distortions due to their reliance on specific statistical models.  
677

678 The emergence of deep learning introduced convolutional neural networks (CNNs) into NR-IQA, enabling models to  
679 learn hierarchical feature representations directly from data. Transformer-based architectures further advanced the field  
680 by capturing long-range dependencies and contextual information, with models such as MUSIQ ((Ke et al., 2021), TReS  
681 (Golestaneh et al., 2022b), and TRIQ (You & Korhonen, 2021)) demonstrating significant improvements in performance.  
682 Despite these advancements, a major limitation persists: the lack of large-scale, diverse datasets encompassing the full  
683 spectrum of real-world distortions. This scarcity hampers the generalization capabilities of NR-IQA models, as they are  
684 trained on datasets that do not adequately represent all possible image degradation scenarios.  
685

686 To tackle the complexity of image distortions, the concept of perceptual or distortion manifolds has been explored in image  
687 quality assessment models. Manifold learning techniques aim to uncover the intrinsic low-dimensional structures within  
688 high-dimensional data, which better align with human visual perception. For instance, Jiang et al. (Jiang et al., 2018)  
689 applied manifold learning to reduce the dimensionality of RGB images, constructing low-dimensional representations for  
690 stereoscopic image quality assessment. Similarly, Guan et al. (Guan et al., 2017) employed manifold learning on feature  
691 maps to capture the intrinsic geometric structures of high-dimensional data in a low-dimensional space, thereby enhancing  
692 prediction accuracy for High-Dynamic-Range (HDR) images. These approaches highlight the potential of manifold learning  
693 in modeling the complex relationships between image content and perceived quality.  
694

695 Although diffusion models (DMs) have demonstrated remarkable efficacy in generating high-dimensional data and capturing  
696 rich feature representations within their intermediate layers ((Ho et al., 2020; Song et al., 2020a)), their application to  
697 NR-IQA has been minimal. Existing works incorporating DMs often use them for specific tasks, such as quality feature  
698 denoising or image restoration, effectively converting NR-IQA into full-reference IQA (FR-IQA) problems ((Li et al., 2024b;  
699 Babnik et al., 2024)). Typically, these methods involve training on specific IQA datasets, limiting their generalizability to  
700 diverse distortions.  
701

702 In our work, we demonstrate that since diffusion models are trained on large-scale datasets containing user-generated content  
703 (UGC) images—with a wide range of authentic and synthetic distortions—they inherently learn perceptually consistent  
704 manifolds. Although these models are not specifically trained for IQA tasks, they are designed to capture data priors by  
705 learning score functions, enabling them to model complex data distributions and capture both high-level and low-level  
706 features. This capability allows them to generate a diverse set of images with fine details. We believe that, with appropriate  
707 perceptual guidance, it is possible to extract features from diffusion models that correlate highly with human perception, in a  
708 zero-shot setting.  
709

710 Diffusion models have also demonstrated the ability to learn meaningful representations within their U-Net architectures,  
711 as evidenced by studies that leverage intermediate features for various vision tasks ((Zhao et al., 2023; Wu et al., 2023)).  
712 This suggests an opportunity to harness these models for NR-IQA, which has so far remained underexplored. Our  
713

work aims to address this gap by utilizing pretrained diffusion models without any fine-tuning, thereby preserving their inherent generalization capabilities. By extracting multi-scale and multi-time-step features—referred to as diffusion hyperfeatures—and incorporating perceptual guidance, we propose a method that overcomes the limitations of current NR-IQA approaches. This strategy leverages the rich representations within diffusion models to improve generalization across diverse image distortions, aligning more closely with human perceptual judgments.

## A.2 Diffusion Models

Diffusion models consist of a forward noise process and a backward denoising process. In the discrete formulation (Song et al., 2020b; Ho et al., 2020), the forward process manifests as a Markov chain described by:

$$q(\mathbf{x}_{1:N} \mid \mathbf{x}_0) = \prod_{k=1}^N q(\mathbf{x}_k \mid \mathbf{x}_{k-1}), \quad q(\mathbf{x}_k \mid \mathbf{x}_{k-1}) = \mathcal{N}(A_k \mathbf{x}_{k-1}, b_k^2 I). \quad (\text{A-1})$$

The coefficients  $\{a_k\}_{k=1}^N$  and  $\{b_k\}_{k=1}^N$  are manually set and may differ depending on various diffusion formulations (Song et al., 2020b). Given that each Markov step  $q(\mathbf{x}_k \mid \mathbf{x}_{k-1})$  is a linear Gaussian model, the resultant marginal distribution  $q(\mathbf{x}_k \mid \mathbf{x}_0)$  assumes a Gaussian form,  $\mathcal{N}(c_k \mathbf{x}_0, d_k^2 I)$ . The parameters  $\{c_k\}_{k=1}^N$  and  $\{d_k\}_{k=1}^N$  can be derived from  $\{a_k\}_{k=1}^N$  and  $\{b_k\}_{k=1}^N$ . For sample generation, we train a neural network,  $s_\theta(\mathbf{x}_k, t_k)$ , to estimate the score function  $\nabla_{\mathbf{x}_k} \log q(\mathbf{x}_k \mid \mathbf{x}_0)$ . The backward process, which we assume to be a Markov chain, is typically represented as:

$$p_\theta(\mathbf{x}_{k-1} \mid \mathbf{x}_k) = \mathcal{N}(u_k \hat{\mathbf{x}}_0(\mathbf{x}_k) + v_k s_\theta(\mathbf{x}_k, t_k), w_k^2 I) \quad (\text{A-2})$$

where  $\hat{\mathbf{x}}_0(\mathbf{x}_k) := \mathbf{x}_k + d_k^2 s_\theta(\mathbf{x}_k, t_k)/c_k$  is the predicted  $\mathbf{x}_0$  obtained from the Tweedie's formula. Here  $\{u_k\}_{k=1}^N$ ,  $\{v_k\}_{k=1}^N$ , and  $\{w_k\}_{k=1}^N$  can be computed from the forward process coefficients  $\{a_k\}_{k=1}^N$  and  $\{b_k\}_{k=1}^N$ . The formulation in Equation A-2 encompasses many stochastic samplers of diffusion models, including the ancestral sampler in DDPM ((Ho et al., 2020)), and the DDIM sampler in (Song et al., 2020b). For variance-preserving diffusion models (Ho et al., 2020), we have:

$$a_k = \sqrt{\alpha_k}, \quad b_k = \sqrt{\beta_k}, \quad c_k = \sqrt{\alpha_k}, \quad d_k = \sqrt{1 - \alpha_k}, \quad (\text{A-3})$$

where  $\alpha_k := 1 - \beta_k$ ,  $\bar{\alpha}_k := \prod_{j=1}^k \alpha_j$ , and  $\alpha_k, \beta_k$  follow the notations in (Ho et al., 2020). DDPM sampling:

$$u_k = \sqrt{\alpha_{k-1}}, \quad v_k = -\sqrt{\alpha_k}(1 - \bar{\alpha}_{k-1}), \quad w_k = \sqrt{\beta_k} \cdot \sqrt{\frac{1 - \bar{\alpha}_{k-1}}{1 - \bar{\alpha}_k}}, \quad (\text{A-4})$$

and for DDIM sampling (Song et al., 2020b), we have:

$$u_k = \sqrt{\alpha_k}, \quad v_k = \sqrt{1 - \bar{\alpha}_{k-1} - \sigma_k^2} \cdot \sqrt{1 - \bar{\alpha}_k}, \quad w_k = \sigma_k, \quad (\text{A-5})$$

where the conditional variance sequence  $\{\sigma_k\}_{k=1}^N$  can be arbitrary. And depending on the value of  $\sigma_k^2$ , it can become DDPM or DDIM sampling, i.e. With  $\beta_k \cdot \frac{(1 - \bar{\alpha}_{k-1})}{(1 - \bar{\alpha}_k)}$  it becomes DDPM.

**Score Based Diffusion Models.** Let  $x_0 \sim p(\mathbf{X})$  represent samples from the data distribution. Diffusion models define the generative process as the reverse of a noising process, which can be represented by the variance-preserving stochastic differential equation (VP-SDE) ((Song et al., 2020b))  $x(t)$ ,  $t \in [0, T]$ :

$$dx = -\frac{\beta_t}{2} x dt + \sqrt{\beta_t} d\mathbf{w} \quad (\text{A-6})$$

where  $\beta_t \in (0, 1)$  is the noise schedule of the process, a monotonically increasing function of  $t$ , and  $\mathbf{w}$  is a  $d$ -dimensional standard Wiener process. This SDE is defined such that  $x_0 \sim p(\mathbf{X})$  when  $t = 0$ , and as  $t \rightarrow T$ , the distribution approaches a standard Gaussian, i.e.,  $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ . Our goal is to learn the reverse-time SDE corresponding to equation (A-6):

$$dx = \left[ -\frac{\beta_t}{2} x - \beta_t \nabla_{x_t} \log p(x_t) \right] dt + \sqrt{\beta_t} d\bar{\mathbf{w}} \quad (\text{A-7})$$

770 where  $d\bar{w}$  is a reverse-time Wiener process and  $dt$  runs backward, and  $\nabla_{x_t} \log p(x_t)$  is the score function (Song et al.,  
 771 2020b). We approximate the score function using a neural network  $s_\theta(x_t, t)$  parameterized by  $\theta$ , trained via denoising score  
 772 matching (Vincent, 2011):

$$\begin{aligned} 773 \quad \theta^* &= \arg \min_{\theta} \mathbb{E}_{t \in [0, T], x_t \sim p(x_t | x_0), x_0 \sim p(\mathbf{X})} \\ 774 \quad &\left[ \|s_\theta(x_t, t) - \nabla_{x_t} \log p(x_t | x_0)\|_2^2 \right] \end{aligned} \quad (\text{A-8})$$

778 Once  $s_\theta$  is learned, we approximate the reverse-time SDE and generate clean data by iteratively solving Equation A-7 from  
 779 noisy samples ((Song & Ermon, 2019)).

780

## 781 B Theoretical Proofs

### 782 B.1 Lemma 1 (Tweedie's Formula for Exponential Family)

783 Let  $p(z|\eta)$  belong to the exponential family distribution:

$$784 \quad p(z | \eta) = p_0(z) \exp(\eta^\top T(z) - \Phi(\eta)) \quad (\text{B-1})$$

785 where  $\eta$  is the natural or canonical parameter of the family,  $\Phi(\eta)$  is the cumulant generating function (cfg) (which makes  
 786  $p_\eta(z)$  integrate to 1), and  $p_0(z)$  is the density when  $\eta = 0$ . Then, the posterior mean  $\hat{\eta} := \mathbb{E}[\eta | z]$  should satisfy:

$$787 \quad (\nabla_z T(z))^\top \hat{\eta} = \nabla_z \log p(z) - \nabla_z \log p_0(z) \quad (\text{B-2})$$

788 **Proof.** The marginal distribution  $p(z)$  can be expressed as:

$$789 \quad p(z) = \int_{\mathcal{Z}} p_\eta(z) p(\eta) d\eta \quad (\text{B-3})$$

790 which, using the form of  $p_\eta(z)$ , becomes:

$$791 \quad p(z) = p_0(z) \int_{\mathcal{Z}} \exp(\eta^\top T(z) - \Phi(\eta)) p(\eta) d\eta \quad (\text{B-4})$$

792 Taking the derivative of  $p(z)$  with respect to  $z$ :

$$\begin{aligned} 793 \quad \nabla_z p(z) &= \nabla_z p_0(z) \int_{\mathcal{Z}} \exp(\eta^\top T(z) - \Phi(\eta)) p(\eta) d\eta + \\ 794 \quad &\int_{\mathcal{Z}} (\nabla_z T(z))^\top \eta p_0(z) \exp(\eta^\top T(z) - \Phi(\eta)) p(\eta) d\eta \end{aligned} \quad (\text{B-5})$$

800 Rearranging, we get:

$$801 \quad \nabla_z p(z) = \frac{\nabla_z p_0(z)}{p_0(z)} p(z) + (\nabla_z T(z))^\top \int_{\mathcal{Z}} \eta p_\eta(z) p(\eta) d\eta \quad (\text{B-6})$$

802 which simplifies to:

$$803 \quad \nabla_z p(z) = \frac{\nabla_z p_0(z)}{p_0(z)} p(z) + (\nabla_z T(z))^\top \int_{\mathcal{Z}} \eta p_z(\eta) d\eta. \quad (\text{B-7})$$

804 Thus:

$$805 \quad \frac{\nabla_z p(z)}{p(z)} = \frac{\nabla_z p_0(z)}{p_0(z)} + (\nabla_z T(z))^\top \mathbb{E}[\eta | z] \quad (\text{B-8})$$

806 Finally:

$$807 \quad (\nabla_z T(z))^\top \mathbb{E}[\eta | z] = \nabla_z \log p(z) - \nabla_z \log p_0(z). \quad (\text{B-9})$$

808 This concludes the proof.

**B.2 Proposition 3 (Tweedie's formula for SDE)**

For the case of VP-SDE, we can estimate  $p(z_0|z_t)$  as:

$$z_{0|t} := \mathbb{E}[z_0 | z_t] = \frac{1}{\sqrt{\bar{\alpha}(t)}} (z_t + (1 - \bar{\alpha}(t)) \nabla_{z_t} \log p_t(z_t)) \quad (\text{B-10})$$

**Proof.** For the case of VP-SDE, we have

$$p(z_t|z_0) = \frac{1}{(2\pi(1 - \bar{\alpha}(t)))^{d/2}} \exp\left(-\frac{\|z_t - \sqrt{\bar{\alpha}(t)}z_0\|^2}{2(1 - \bar{\alpha}(t))}\right) \quad (\text{B-11})$$

A Gaussian distribution. We can get the canonical decomposition as:

$$p(z_t|z_0) = p_0(z_t) \exp(z_0^\top T(z_t) - \Phi(z_0)), \quad (\text{B-12})$$

And,

$$p_0(z_t) := \frac{1}{(2\pi(1 - \bar{\alpha}(t)))^{d/2}} \exp\left(-\frac{\|z_t\|^2}{2(1 - \bar{\alpha}(t))}\right) \quad (\text{B-13})$$

$$T(z_t) := \frac{\sqrt{\bar{\alpha}(t)}}{1 - \bar{\alpha}(t)} z_t \quad (\text{B-14})$$

$$\Phi(z_0) := \frac{\bar{\alpha}(t)\|z_0\|^2}{2(1 - \bar{\alpha}(t))} \quad (\text{B-15})$$

Therefore, from **Lemma 1**:

$$\frac{\sqrt{\bar{\alpha}(t)}}{1 - \bar{\alpha}(t)} \hat{z}_0 = \nabla_{z_t} \log p_t(z_t) + \frac{1}{1 - \bar{\alpha}(t)} z_t \quad (\text{B-16})$$

Giving us:

$$z_{0|t} = \frac{1}{\sqrt{\bar{\alpha}(t)}} (z_t + (1 - \bar{\alpha}(t)) \nabla_{z_t} \log p_t(z_t)) \quad (\text{B-17})$$

This concludes the proof.

**B.3 Conditional Score functions**

As mentioned in the main paper, conditional score function can be written as (Equation 6):

$$\begin{aligned} \nabla_{z_t} \log p(z_t|\psi_p(y), y) &\approx \nabla_{z_t} \log p(z_t) \\ &+ \nabla_{z_t} \log p(y|x_{0|t} = \mathcal{D}(z_{0|t})) \\ &+ \nabla_{z_t} \log p(\psi_p(y)|x_{0|t} = \mathcal{D}(z_{0|t})) \end{aligned} \quad (\text{B-18})$$

Proof. From Baye's theorem we can write the conditional distribution as:

$$p(z|\psi(y), y) = p(\psi(y)|z_t)p(y|z_t, \psi(y))p(z_t) \quad (\text{B-19})$$

Note,  $y$  is conditionally independent of  $\psi(y)$  given  $z_t$  for later timesteps in diffusion process as  $z_t$  gives more structural information for image. Therefore:

$$p(z|\psi(y), y) = p(\psi(y)|z_t)p(y|z_t)p(z_t) \quad (\text{B-20})$$

Our score function becomes:

$$\nabla_{z_t} \log p(z_t|\psi_p(y), y) \approx \nabla_{z_t} \log p(z_t) + \nabla_{z_t} \log p(\psi_p(y)|z_t) + \nabla_{z_t} \log p(y|z_t) \quad (\text{B-21})$$

We can write the posterior as:

$$p(y|z_t) = \int p(y|z_0)p(z_0|z_t)dz_0 \quad (\text{B-22})$$

Following ((Chung et al., 2022a)) and Proposition 3, we can have the posterior as:

$$p(y|z_t) \approx p(y|z_{0|t}) \quad (\text{B-23})$$

Therefore:

$$\nabla_{z_t} \log p(z_t|\psi_p(y), y) \approx \nabla_{z_t} \log p(z_t) + \nabla_{z_t} \log p(\psi_p(y)|z_{0|t}) + \nabla_{z_t} \log p(y|z_{0|t}) \quad (\text{B-24})$$

Following ((Rout et al., 2024)), we can approximately write the conditional probability for LDM given decoder  $\mathcal{D}$ :

$$p(y|z_t) \approx p(y|x_0 = \mathcal{D}(z_{0|t})) \quad (\text{B-25})$$

Note that we ignore the gluing term proposed by ((Rout et al., 2024)) as it depends on the forward degradation model only valid for inverse problems. Our final conditional score function becomes:

$$\nabla_{z_t} \log p(z_t|\psi_p(y), y) \approx \nabla_{z_t} \log p(z_t) + \nabla_{z_t} \log p(y|x_{0|t} = \mathcal{D}(z_{0|t})) + \nabla_{z_t} \log p(\psi_p(y)|x_{0|t} = \mathcal{D}(z_{0|t})) \quad (\text{B-26})$$

#### B.4 Proposition 1 (Noisy Data Manifold)

Let the distance function be defined as  $d(x, \mathcal{M}) := \inf_{y \in \mathcal{M}} \|x - y\|_2$ , and define the neighborhood around the manifold  $\mathcal{M}$  as  $B(\mathcal{M}; r) := \{x \in \mathbb{R}^D \mid d(x, \mathcal{M}) < r\}$ . Consider the distribution of noisy data given by  $p(x_t) = \int p(x_t|x_0)p(x_0)dx_0$ ,  $p(x_t|x_0) := \mathcal{N}(\sqrt{\bar{\alpha}_t} \mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I})$  represents the Gaussian perturbation of the data at time  $t$ , and  $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$  is the cumulative product of the noise schedule  $\alpha_t$ . Under the Assumption 1, the distribution  $p_t(x_t)$  is concentrated on a  $(D - 1)$ -dim manifold  $\mathcal{M}_t := y \in \mathbb{R}^D : d(y, \sqrt{\bar{\alpha}_t}\mathcal{M}) = r_t := \sqrt{(1 - \bar{\alpha}_t)(D - k)}$ .

Proof. (Mainly follow (Chung et al., 2022b)):

We begin by defining the manifold  $\mathcal{M}$  as  $\mathcal{M} := \{x \in \mathbb{R}^D : x_{k+1:D} = 0\}$  which represents a subspace where the last  $D - k$  coordinates are zero. Essentially, this means that  $\mathcal{M}$  lies within a lower-dimensional subspace of  $\mathbb{R}^D$ . Let  $X$  be a  $\chi^2$  random variable with  $n$  degrees of freedom. We use the following concentration bounds:

$$\mathbb{P}(X - n \geq 2\sqrt{n\tau} + 2\tau) \leq e^{-\tau}, \quad (\text{B-27})$$

$$\mathbb{P}(X - n \leq -2\sqrt{n\tau}) \leq e^{-\tau}. \quad (\text{B-28})$$

Now, consider the quantity  $\sum_{i=k+1}^D \frac{x_{t,i}^2}{1 - \bar{\alpha}_t}$ , which follows a  $\chi^2$  distribution with  $D - k$  degrees of freedom. Using the concentration bounds and setting  $\tau = (D - k)\epsilon'$ , we can express the following bound:

$$\mathbb{P}\left(-2(D - k)\sqrt{\epsilon'} \leq \sum_{i=k+1}^D \frac{x_{t,i}^2}{1 - \bar{\alpha}_t} - (D - k) \leq 2(D - k)(\sqrt{\epsilon'} + \epsilon')\right) \geq 1 - \delta. \quad (\text{B-29})$$

The above inequality gives us a range for the summation of the squared components of  $x_t$  beyond the first  $k$  dimensions. We can now rewrite this in terms of the Euclidean norm of these components:

$$\mathbb{P}\left(\sqrt{\sum_{i=k+1}^D x_{t,i}^2} \in \left(r_t \sqrt{\max\{0, 1 - 2\sqrt{\epsilon'}\}}, r_t \sqrt{1 + 2\sqrt{\epsilon'} + 2\epsilon'}\right)\right) \geq 1 - \delta, \quad (\text{B-30})$$

where we have defined:

$$r_t := \sqrt{(1 - \bar{\alpha}_t)(D - k)}. \quad (\text{B-31})$$

To ensure that the probability holds for a given confidence level  $1 - \delta$ , we define:

$$\epsilon'_{t,D-k} = -\frac{1}{D - k} \log \frac{\delta}{2}. \quad (\text{B-32})$$

We then use  $\epsilon'_{t,D-k}$  to define:

$$\epsilon_{t,D-k} = \min \left\{ 1, \sqrt{\max\{0, 1 - 2\sqrt{\epsilon'_{t,D-k}}\}} + \frac{1 + 2\sqrt{\epsilon'_{t,D-k}} + 2\epsilon'_{t,D-k} - 1}{\sqrt{1 - \bar{\alpha}_t}(D - k)} \right\}, \quad (\text{B-33})$$

which ensures  $0 < \epsilon_{t,D-k} \leq 1$ . This value  $\epsilon_{t,D-k}$  helps in determining the size of the neighborhood around the manifold  $\mathcal{M}_t$ , such that:

$$\mathbb{P}\left(x_t \in B(\mathcal{M}_t; \epsilon_{t,D-k} \cdot \sqrt{(1 - \bar{\alpha}_t)(D - k)})\right) \geq 1 - \delta. \quad (\text{B-34})$$

Thus, we have shown that the noisy data distribution  $p(x_t)$  is concentrated within a certain neighborhood around the manifold  $\mathcal{M}_t$ , with high probability. The parameter  $\epsilon_{t,D-k}$  is decreasing with respect to  $\delta$  and  $D - k$ , because  $\epsilon'_{t,D-k}$  is also decreasing in these parameters, and  $\epsilon_{t,D-k}$  is an increasing function of  $\epsilon'_{t,D-k}$ .

This concludes the proof.

In pixel space, as discussed by (Chung et al., 2022a), the optimization of the guidance term occurs in the entire space  $\mathbb{R}^D$ . However, from Proposition 1, we know that  $x_t$  actually lies in a much smaller subspace of  $\mathbb{R}^D$ , specifically in  $\mathbb{R}^k$ . To prevent sampling from deviating from the content-bias region on the manifold, one obvious way to improve the sampling process is to restrict the optimization space to  $M_t$ , specifically to the tangent space  $\mathcal{T}_{x_t}\mathcal{M}_t$ . Previous literature has suggested using autoencoders to approximate this tangent space  $\mathcal{T}_{x_t}\mathcal{M}_t$  (Shao et al., 2018)). However, since autoencoders are not trained on intermediate noisy samples, their practical effectiveness is limited. We instead use Latent Diffusion Models (LDM), where the entire sampling process occurs in the latent space  $\mathcal{M}$ . This approach ensures overall data consistency, but it may still not fully achieve perceptual consistency within the content-bias region on the manifold  $\mathcal{M}$  (see Fig. 1 for an intuitive illustration).

## B.5 Proposition 2 (On-manifold sample with LDM)

Given a perfect autoencoder, i.e.  $x = \mathcal{D}(\mathcal{E}(x))$ , and a gradient  $\nabla_{z_0|t} G(z_0|t, y) \in \mathcal{T}_{z_0}\mathcal{Z}$  then  $\mathcal{D}(\nabla_{z_0|t} G(z_0|t, y)) \in \mathcal{T}_{x_0}\mathcal{M}$ .

**Proof.** We begin by considering a perfect autoencoder, consisting of an encoder  $\mathcal{E}$  and a decoder  $\mathcal{D}$ , which satisfies the property  $x = \mathcal{D}(\mathcal{E}(x))$ . for any data point  $x \in X \subset \mathcal{M}$ . Let  $z_0 = \mathcal{E}(x_0)$  be the latent representation of  $x_0$ . Since the autoencoder is perfect, we have  $x_0 = \mathcal{D}(z_0)$ .

To understand how the encoder and decoder interact in terms of their mappings, we consider their Jacobians. The Jacobian of the encoder  $\frac{\partial \mathcal{E}}{\partial x_0}$  maps changes in the data space  $\mathbb{R}^D$  to changes in the latent space  $\mathbb{R}^k$ . The Jacobian of the decoder  $\frac{\partial \mathcal{D}}{\partial z_0}$  maps changes in the latent space  $\mathbb{R}^k$  back to the data space  $\mathbb{R}^D$ . Since the autoencoder is perfect, encoding and then decoding must recover the original input exactly. This implies that the composition of the encoder and decoder Jacobians must yield the identity mapping:

$$\frac{\partial \mathcal{E}}{\partial x_0} \frac{\partial \mathcal{D}}{\partial z_0} = I, \quad (\text{B-35})$$

where  $I$  is the identity matrix. This property ensures that the encoder and decoder are exact inverses of each other in terms of their linear mappings at  $x_0$  and  $z_0$ .

Consider a gradient  $\nabla_{z_0|t} G(z_0|t, y) \in \mathcal{T}_{z_0}\mathcal{Z}$ , where  $\mathcal{T}_{z_0}\mathcal{Z}$  is the tangent space of the latent space  $\mathcal{Z}$  at  $z_0$ . We want to determine the behavior of this gradient when mapped back to the data space using the decoder. The decoder Jacobian  $\frac{\partial \mathcal{D}}{\partial z_0}$  maps vectors from the latent space to the data space. Since  $\nabla_{z_0|t} G(z_0|t, y)$  is in the tangent space  $\mathcal{T}_{z_0}\mathcal{Z}$ , applying the decoder Jacobian gives:

$$\mathcal{D}(\nabla_{z_0|t} G(z_0|t, y)) = \frac{\partial \mathcal{D}}{\partial z_0} \nabla_{z_0|t} G(z_0|t, y). \quad (\text{B-36})$$

Since the Jacobian  $\frac{\partial \mathcal{D}}{\partial z_0}$  maps changes in the latent space to corresponding changes in the data space, and the latent space  $\mathcal{Z}$  is designed to represent the underlying data manifold  $\mathcal{M}$ , it follows that:

$$\frac{\partial \mathcal{D}}{\partial z_0} : \mathcal{T}_{z_0}\mathcal{Z} \rightarrow \mathcal{T}_{x_0}\mathcal{M}. \quad (\text{B-37})$$

Thus, the vector  $\mathcal{D}(\nabla_{z_0|t} G(z_0|t, y))$  lies in the tangent space  $\mathcal{T}_{x_0}\mathcal{M}$  of the data manifold at  $x_0$ . This implies that the gradient update, when mapped back to the data space, remains on the data manifold, ensuring consistency in the sampling process.

This concludes the proof.

## B.6 Lemma 2 (Distribution Concentration)

Consider the optimality of the diffusion model, i.e.,  $\epsilon_\theta(\sqrt{\alpha_t}z + \sqrt{1 - \alpha_t}\epsilon_t, t) = \epsilon_t$  for  $z \in \mathcal{Z}$ . For some  $\epsilon \sim \mathcal{N}(0, I)$ , the sum of noise components  $\sqrt{1 - \bar{\alpha}_{t-1}} - \sigma_t^2 \epsilon_\theta(z_t, t) + \sigma_t \epsilon_t$  in DDIM sampling can be expressed as:

$$\sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(z_t, t) + \sigma_t \epsilon_t = \sqrt{1 - \bar{\alpha}_{t-1}} \tilde{\epsilon}, \quad (\text{B-38})$$

where  $\tilde{\epsilon} \sim \mathcal{N}(0, I)$ . Since  $\sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(z_t, t)$  and  $\sigma_t \epsilon_t$  are independent, their sum is also a Gaussian random variable with a mean of 0 and a variance of  $(1 - \bar{\alpha}_{t-1} - \sigma_t^2) + \sigma_t^2 = (1 - \bar{\alpha}_{t-1})$ .

Furthermore, let the latent data distribution  $p(z)$  be a probability distribution with support on the linear manifold  $\mathcal{M}$  that satisfies Assumption 1. For any  $z \sim p(z)$ , consider

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}} z + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(z_t, t) + \sigma_t \epsilon_t. \quad (\text{B-39})$$

Then, the marginal distribution  $\hat{p}_{t-1}(z_{t-1})$ , which is defined as:

$$\hat{p}_{t-1}(z_{t-1}) = \int \mathcal{N}\left(z_{t-1}; \sqrt{\bar{\alpha}_{t-1}} z + \sqrt{1 - \bar{\alpha}_{t-1} - \sigma_t^2} \epsilon_\theta(z_t, t), \sigma_t^2 I\right) p(z|z) p(z) dz dz_t, \quad (\text{B-40})$$

is probabilistically concentrated on  $\mathcal{Z}_{t-1}$  for  $\epsilon_t \sim \mathcal{N}(0, I)$ .

**Proof.** Since  $\epsilon_\theta(z_t, t)$  is independent of  $\epsilon_t$ , their sum is the sum of independent Gaussian random variables, resulting in a Gaussian distribution with a variance  $(1 - \bar{\alpha}_{t-1})$ . By this result, the multivariate normal distribution has a mean  $\sqrt{\bar{\alpha}_{t-1}} z$  and a covariance matrix  $(1 - \bar{\alpha}_{t-1})I$ . Consequently, the marginal distribution of the target can be represented as:

$$\hat{p}_{t-1}(z_{t-1}) = \int \mathcal{N}(z_{t-1}; \sqrt{\bar{\alpha}_{t-1}} z, (1 - \bar{\alpha}_{t-1})I) p(z) dz, \quad (\text{B-41})$$

which matches the marginal distribution defined in Proposition 1. Therefore, in accordance with Proposition 1, the probability distribution  $\hat{p}_{t-1}(z_{t-1})$  probabilistically concentrates on  $\mathcal{M}_{t-1}$ .  $\square$

## B.7 Theorem 1 (Perceptual Manifold Guidance)

Given Assumption 1, for perfect encoder  $\mathcal{E}$ , decoder  $\mathcal{D}$ , and an efficient score function  $s_\theta(z_t, t)$ , let gradient  $\nabla_{z_{0|t}} G_1(\mathcal{D}(z_{0|t}), y)$  and  $\nabla_{z_{0|t}} G_2(\psi_p(\mathcal{D}(z_{0|t})), \psi_p(y))$  reside on the tangent space  $\mathcal{T}_{z_{0|t}} \mathcal{Z}$  of latent manifold  $\mathcal{Z}$ . Throughout the diffusion process, all update terms  $z_t$  remain on noisy latent manifolds  $\mathcal{Z}_t$ , with  $z''_{0|t}$  in perceptually consistent manifold locality.

**Proof.** We begin by establishing that both the gradients for data and perceptual consistency are constrained to the tangent space of the latent manifold, ensuring that updates remain on the manifold during the diffusion process. At  $t = T$ , we consider the noisy sample  $z_T$  generated from a Gaussian distribution. Noisy sample is expressed as:

$$z_T = \sqrt{\bar{\alpha}_T} z_0 + \sqrt{1 - \bar{\alpha}_T} \epsilon_T, \quad \epsilon_T \sim \mathcal{N}(0, I) \quad (\text{B-42})$$

where  $z_0 = \mathcal{E}(x_0)$  represents the latent variable corresponding to the clean sample  $x_0$ . The support of the distribution  $p(z_0)$  lies on the manifold  $\mathcal{Z}$ , ensuring that  $z_0 \in \mathcal{Z}$ . Assume that for all  $t \geq T_1$ , there exists a  $z_0 \in \mathcal{Z}$  such that:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \quad (\text{B-43})$$

We aim to prove that this also holds for  $t = T_1 - 1$ . At timestep  $t = T_1$ , the two gradients  $\nabla_{z_{0|t}} G_1$  (for data consistency) and  $\nabla_{z_{0|t}} G_2$  (for perceptual consistency) lie in the tangent space  $\mathcal{T}_{z_0} \mathcal{Z}$ . These gradients contribute to the update of the latent representation. The overall gradient update becomes:

$$z'_{0|T_1} = z_{0|T_1} - (\zeta_1 \nabla_{z_{0|T_1}} G_1 + \zeta_2 \nabla_{z_{0|T_1}} G_2) \quad (\text{B-44})$$

Where  $\zeta$  are scalars. Since both gradients reside in the tangent space  $\mathcal{T}_{z_{0|T_1}} \mathcal{Z}$ , and Assumption 1, the updated term  $z'_{0|T_1}$  remains on the latent manifold  $\mathcal{Z}$ . Using the update rule for  $z_{T_1-1}$ , similar to the diffusion update step, we have:

$$z_{T_1-1} = \sqrt{\bar{\alpha}_{T_1-1}} z'_{0|T_1} + \sqrt{1 - \bar{\alpha}_{T_1-1}} \epsilon', \quad \epsilon' \sim \mathcal{N}(0, I) \quad (\text{B-45})$$

Table 6: Summary of the IQA datasets used in our experiments.

Dataset	Type	Images	Description
LIVE IQA ((Sheikh et al., 2006))	Synthetic	779	29 reference images; 5 distortions at 4 levels
CSIQ-IQA ((Larson & Chandler, 2010))	Synthetic	866	30 reference images; 6 distortions
TID2013 ((Ponomarenko et al., 2013))	Synthetic	3,000	25 reference images; 24 distortions at 5 levels
KADID-10k ((Lin et al., 2019))	Synthetic	10,125	81 reference images; 25 distortions at 5 levels
LIVEC ((Ghadiyaram & Bovik, 2015))	Authentic	1,162	Mobile images with real-world distortions
KonIQ-10k ((Hosu et al., 2020))	Authentic	10,073	Diverse images from YFCC100M dataset
FLIVE ((Ying et al., 2020))	Authentic	39,810	Emulates social media content
SPAQ ((Fang et al., 2020))	Authentic	11,000	Mobile images with annotations
AGIQA-3K ((Li et al., 2023))	AIGC	3,000	AI-generated images for IQA
AGIQA-1K ((Li et al., 2023))	AIGC	1,000	AI-generated images for IQA

Thus, the updated latent variable remains on the manifold, as the noise component  $\epsilon'$  is Gaussian, and the mean update is based on  $z'_{0|T_1} \in \mathcal{Z}$ . Applying Lemma 2, give us  $p(z_{T_1-1})$ , that is probabilistically concentrated on  $Z_{T-1}$ .

The perceptual manifold, a subspace of the content-bias manifold defined by the data-consistency gradient  $\nabla_{z_{0|t}} G_1$ . In other words, data consistency keeps the sample within a region where the structural content is retained, and perceptual consistency term ensures that the sample moves toward regions of the manifold  $\mathcal{Z}$  that are perceptually meaningful. The second gradient term  $\nabla_{z_{0|t}} G_2(\psi_p(\mathcal{D}(z_{0|t})), \psi_p(y))$  represents a movement within this subspace to align with human perception. Since  $\nabla_{z_{0|t}} G_2$  resides in the tangent space  $T_{z_0} \mathcal{Z}$  and is also influenced by the perceptual features  $\psi_p$ , the update ensures that the latent variable moves towards a more perceptually consistent locality within the overall content-bias region.

Formally, let  $\mathcal{M}_{\text{content}}$  be the subspace of manifold corresponding to content consistency based on  $G_1$ , and let  $\mathcal{M}_{\text{perceptual}} \subset \mathcal{M}_{\text{content}}$  be the sub-manifold that represents regions of perceptual consistency. The update using  $\nabla_{z_{0|t}} G_2$  effectively ensures:

$$z''_{0|t} \in \mathcal{M}_{\text{perceptual}}, \quad (\text{B-46})$$

where  $\mathcal{M}_{\text{perceptual}}$  is a more constrained subspace within the content-consistent manifold, ensuring perceptual quality. By induction, we have shown that for all  $t$ , there exists a  $z_0 \in \mathcal{Z}$  such that  $z_t$  remains on the latent manifold throughout the diffusion process. Furthermore, the inclusion of the perceptual consistency gradient ensures that  $z''_{0|t}$  is updated towards a perceptually consistent region on the manifold. Thus, the final updated latent variable  $z''_{0|t}$  is not only data-consistent but also perceptually consistent within the manifold  $\mathcal{Z}$ , as required.

This concludes the proof.

## C Implementation Details

### C.1 Datasets and Evaluation Protocol

The datasets used in our study (Table 6) contain images labeled with Mean Opinion Scores (MOS) following ITU-T P.910 guidelines (ITU-T RECOMMENDATION, 1999). We train the regressor  $g_\phi$  using  $l_2$  loss on MOS ground truth values. Evaluation metrics include Pearson Linear Correlation Coefficient (PLCC) and Spearman’s Rank Order Correlation Coefficient (SRCC), ranging from 0 to 1, with higher values indicating better correlation.

Following (Saha et al., 2023; Madhusudana et al., 2022), we split each dataset into training, validation, and test sets (70%, 10%, and 20%, respectively), using source image-based splits to prevent content overlap. The process is repeated 10 times, and median performance is reported to ensure robustness.

### C.2 Implementation Details

**Model Configuration** For text conditioning, we use an empty string “ ” as prompt. We adopt the SDv1.5 and VQ-VAE from the official Stable Diffusion v1.5, with default settings from GitHub<sup>1</sup> and Hugging Face<sup>2</sup>. VQ-VAE is used with 8x

<sup>1</sup><https://github.com/CompVis/stable-diffusion>

<sup>2</sup><https://huggingface.co/CompVis>,

<https://huggingface.co/stable-diffusion-v1-5/stable-diffusion-v1-5>

Table 7: Comparison of our proposed LGDM with SOTA NR-IQA methods on PLCC and SRCC Scores for Synthetic IQA datasets. The best results are in red.

Methods	LIVE		CSIQ		TID2013		KADID	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
ILNIQE ((Zhang et al., 2015))	0.906	0.902	0.865	0.822	0.648	0.521	0.558	0.534
BRISQUE ((Mittal et al., 2012))	0.944	0.929	0.748	0.812	0.571	0.626	0.567	0.528
WaDIQaM ((Bosse et al., 2018))	0.955	0.960	0.844	0.852	0.855	0.835	0.752	0.739
DBCNN ((Zhang et al., 2020))	0.971	0.968	0.959	0.946	0.865	0.816	0.856	0.851
TIQA ((Stepien & Oszust, 2023))	0.965	0.949	0.838	0.825	0.858	0.846	0.855	0.850
MetalIQA ((Zhu et al., 2020))	0.959	0.960	0.908	0.899	0.868	0.856	0.775	0.762
P2P-BM ((Ying et al., 2020))	0.958	0.959	0.902	0.899	0.856	0.862	0.849	0.840
HyperIQA ((Su et al., 2020))	0.966	0.962	0.942	0.923	0.858	0.840	0.845	0.852
TReS ((Golestaneh et al., 2022a))	0.968	0.969	0.942	0.922	0.883	0.863	0.859	0.859
MUSIQ ((Ke et al., 2021))	0.911	0.940	0.893	0.871	0.815	0.773	0.872	0.875
RE-IQA ((Saha et al., 2023))	0.971	0.970	0.960	0.947	0.861	0.804	0.885	0.872
LoDA ((Xu et al., 2024))	0.979	0.975	-	-	0.901	0.869	0.936	0.931
LGDM- $\psi_\phi$	0.980	0.978	0.971	0.952	0.871	0.814	0.892	0.885
LGDM- $\psi_{BRISQUE}$	0.980	0.977	0.970	0.951	0.814	0.771	0.841	0.838
LGDM- $\psi_{MUSIQ}$	0.981	0.979	0.969	0.950	0.869	0.811	0.898	0.887
LGDM- $\psi_{Re-IQA}$	0.983	0.981	0.972	0.952	0.904	0.882	0.932	0.930
LGDM- $\zeta_1 = 0 \psi_{SDv1.5}$	0.981	0.978	0.971	0.958	0.908	0.876	0.935	0.931
<b>LGDM-<math>\psi_{SDv1.5}</math></b>	<b>0.988</b>	<b>0.986</b>	<b>0.981</b>	<b>0.964</b>	<b>0.921</b>	<b>0.883</b>	<b>0.961</b>	<b>0.958</b>

downsampling for  $512 \times 512$  resolution, which matches the typical resolution of IQA datasets like LIVE ((Sheikh et al., 2006)) and CSIQ ((Larson & Chandler, 2010)).

**Sampling and Perceptual Features** We use 10 DDIM steps for sampling, balancing efficiency and quality. The choice of perceptual metric  $\psi_p$  is crucial—using well-correlated metrics such as RE-IQA or MUSIQ improves model performance. Poor metrics can degrade results, as shown in our ablation studies.

**Guidance Weights** The weights for perceptual guidance,  $\zeta_1$  and  $\zeta_2$ , are set to 1 and 0.2 based on empirical evaluations. This setup provides sufficient guidance, which enhances prediction quality.

### C.3 Autoencoder

Though a perfect autoencoder is ideal for maintaining samples on the manifold  $\mathcal{M}$ , the Stable Diffusion v1.5 VAE yields effective results despite minor imperfections. As shown in Table 5, it provides the best performance across configurations.

## D Additional Results and Ablation Study

In this section, we provide further empirical evaluations of our proposed Latent Guidance in Diffusion Models (LGDM) by presenting additional experimental results, ablation studies, and analyses to supplement the findings in the main paper. We also evaluate the impact of model hyperparameters and different configurations, including the version of Stable Diffusion (SD), the number of time steps, and the weighting of perceptual guidance terms. Lastly, we discuss the impact of various layers of UNet towards NR-IQA performance.

### D.1 Synthetic IQA Datasets

Table 7 reports the Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank Order Correlation Coefficient (SRCC) scores of LGDM and existing NR-IQA methods on four synthetic datasets: LIVE ((Sheikh et al., 2006)), CSIQ ((Larson & Chandler, 2010)), TID2013 ((Ponomarenko et al., 2013)), and KADID ((Lin et al., 2019)). The results demonstrate that our proposed approach outperforms other methods across almost all datasets. Similar to the performance on authentic distortion dataset in Table 1, LGDM- $\psi_{SDv1.5}$  achieves the best performance across all synthetic datasets, indicating a strong alignment with human perceptual judgments. The results also suggest that perceptual guidance using  $\psi_{Re-IQA}$  and  $\psi_{SDv1.5}$  consistently enhances the model’s generalization capabilities. The superior performance of LGDM- $\psi_{SDv1.5}$  on the TID2013

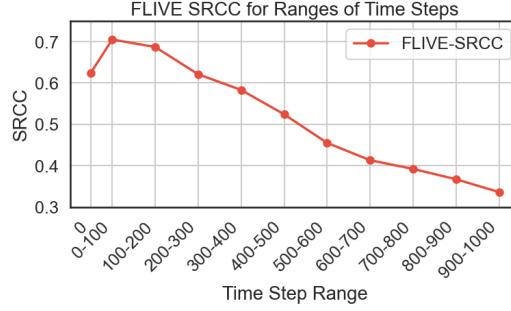


Figure 7: We report the behaviour of LGDM as we change the range of timesteps in the sampling process. As we move towards larger timestep range buckets, where there are more noisy samples, SRCC on FLIVE ((Ying et al., 2020)) decreases.

((Ponomarenko et al., 2013))and KADID ((Lin et al., 2019)) datasets, with PLCC and SRCC scores of 0.921/0.883 and 0.961/0.958, respectively, underscores the value of utilizing diffusion hyperfeatures.

## D.2 Impact of Time Step Range on Sampling Process

Figure 7 shows the effect of varying the range of time steps used during the sampling process. We observe a general decline in SRCC as we increase the time range, specifically when more noisy samples are involved. For larger time step ranges, the model relies on noisier intermediate representations, which reduces its ability to accurately predict image quality. This is an expected behaviour since, details in the images are generated towards the later time steps in diffusion process,i.e. less noise. This suggests that optimizing the range of time steps used for feature extraction is critical to maintaining high-quality predictions.

## D.3 Effect of Time Steps on Quality and Computation Time

Table 4 provides an analysis of the SRCC scores and the computation time for different numbers of time steps on the FLIVE ((Ying et al., 2020)) dataset using LGDM- $\psi_{SDv1.5}$ . As expected, increasing the number of time steps improves the SRCC score, with the highest value of 0.711 obtained at 50 time steps. However, this comes at the cost of increased computational time, with a significant jump from 21.30 seconds for 10 time steps to 110.45 seconds for 50 time steps. This trade-off suggests that while more time steps can yield better performance, it is essential to balance quality with computational efficiency, especially for real-time applications.

Table 8: PLCC and SRCC Scores for varying the weights of second term in PMG (equation 10) on LIVEC ((Ghadiyaram & Bovik, 2015)) and FLIVE ((Ying et al., 2020)). The best results are highlighted in bold, and the second-best results are underlined.

$\zeta_2$	LIVEC		FLIVE	
	PLCC↑	SRCC↑	PLCC↑	SRCC↑
0	0.853	0.842	0.751	0.672
0.2	<b>0.940</b>	<b>0.908</b>	<b>0.812</b>	<b>0.705</b>
0.5	<u>0.931</u>	<u>0.901</u>	<u>0.802</u>	<u>0.691</u>
0.7	0.918	0.882	0.780	0.675
1	0.904	0.868	0.772	0.664

## D.4 Effect of Perceptual Guidance Weighting

Table 8 presents the PLCC and SRCC scores for varying the perceptual guidance weight,  $\zeta_2$ , in the Perceptual Manifold Guidance (PMG) term for the LIVEC ((Ghadiyaram & Bovik, 2015)) and FLIVE ((Ying et al., 2020)) datasets. The results indicate that an optimal value of  $\zeta_2 = 0.2$  yields the best PLCC and SRCC scores all ten datasets. Specifically, an SRCC of 0.908 is achieved on LIVEC ((Ghadiyaram & Bovik, 2015)) and 0.705 on FLIVE ((Ying et al., 2020)) at this weight. When  $\zeta_2$  is set too high (e.g.,  $\zeta_2 = 1$ ), the model’s performance deteriorates, suggesting that samples move away from

1210 the perceptually consistent region on manifold. Conversely, setting  $\zeta_2$  too low results in underutilization of the perceptual  
1211 guidance, which leads to suboptimal quality predictions. With moderate weighting, superior performance can be achieved  
1212 across all benchmarks.

1213

## 1214 D.5 Summary and Insights

1215 Our extended experiments validate the effectiveness of LGDM across various IQA datasets, demonstrating its robustness  
1216 against both synthetic and real-world distortions. The ablation studies provide valuable insights into the factors that impact  
1217 model performance:

1218

- 1220 • Model Version Selection: SD v1.5 emerged as the best-performing version for NR-IQA, emphasizing the importance  
1221 of selecting the appropriate diffusion model.
- 1222 • Time Step Range: If sampling time steps from higher ranges are taken, it diminishes the performance significantly.
- 1223 • Total Time steps: While increasing time steps improves prediction quality, it also significantly increases computation  
1224 time, highlighting a trade-off between accuracy and efficiency.
- 1225 • Layer Importance: Intermediate layers of the diffusion model were found to contribute the most towards perceptual  
1226 quality, suggesting the possibility of optimizing feature extraction by focusing on specific layers.
- 1227 • Perceptual Guidance Weighting: The optimal perceptual guidance weight strikes a balance between content and  
1228 perceptual terms, which is crucial for maintaining high-quality predictions.

1229

1230 Overall, these results underscore the capability of pretrained diffusion models to serve as effective feature extractors for  
1231 NR-IQA tasks, provided that the appropriate guidance is provided. Our method, which exploits the inherent generalization  
1232 capabilities of diffusion models, successfully advances the state-of-the-art in NR-IQA, offering a promising approach for  
1233 future developments in perceptually consistent no-reference image quality assessment.

1234

## 1235 E Limitations & Extension

1236

1237 Our proposed LGDM framework shows strong performance in NR-IQA; however, there are a few limitations and potential  
1238 extensions worth noting.

1239

1240 **Limitations:** The computational cost of our approach, particularly due to the iterative diffusion model sampling, can be  
1241 high, which might limit real-time or resource-constrained applications. Also, the scope of our work is limited to NR-IQA.  
1242 We did not extend our evaluation to other low-level vision tasks or explore the use of perceptual control in image generation  
1243 due to time constraints and the focus on a single task in this paper. Throughout the experiments and proofs, we assumed that  
1244 VAE is ideal, whereas in real world they introduce notable reconstruction errors.

1245

1246 **Extensions:** The general framework of LGDM, particularly its ability to extract perceptual features, has potential applications  
1247 beyond NR-IQA. It can be extended to other low-level vision tasks that are sensitive to perceptual features, such as image  
1248 denoising, super-resolution, and enhancement, where quality assessment is crucial. Moreover, since posterior sampling  
1249 is known to be limiting due to its clean sample estimation from intermediate time step  $t$ , one can explore using more  
1250 advanced techniques such as Bayesian filtering to directly approximate posterior sampling in intermediate time step  $t$ , that  
1251 is computationally more efficient as it avoids taking derivatives on the estimated score function. Our method can also  
1252 be adapted for perceptually controllable image generation. Exploring these directions can help expand the impact of our  
1253 approach and leverage its strengths across a broader range of vision tasks.

1254

1255 In future work, we plan to explore these extensions, allowing LGDM to contribute more broadly to the field of low-level  
1256 computer vision.

1257

1258

1259

1260

1261

1262

1263

1264