DiffNat : Exploiting the Kurtosis Concentration Property for Image Quality Improvement

Anonymous authors
Paper under double-blind review

Abstract

Diffusion models have significantly advanced generative AI in terms of creating and editing naturalistic images. However, improving the image quality of generated images is still of paramount interest. In this context, we propose a generic kurtosis concentration (KC) loss that can be readily applied to any standard diffusion model pipeline to improve image quality. Our motivation stems from the *projected kurtosis concentration property* of natural images, which states that natural images have nearly constant kurtosis values across different band-pass filtered versions of the image. To improve the image quality of generated images, we reduce the gap between the highest and lowest kurtosis values across the band-pass filtered versions (e.g., Discrete Wavelet Transform (DWT)) of images. In addition, we also propose a novel condition-agnostic perceptual guidance strategy during inference to further improve the quality. We validate the proposed approach for four diverse tasks, viz., (1) personalized few-shot finetuning using text guidance, (2) unconditional image generation, (3) image super-resolution, and (4) blind face-restoration. Integrating the proposed KC loss and perceptual guidance has improved the perceptual quality in all these tasks in terms of FID, MUSIQ score, and user evaluation. Code is provided in the supplementary.

1 Introduction

Multi-modal generative AI has advanced by leaps and bounds with the advent of the diffusion model. Large-scale text-to-image diffusion models, e.g., DALLE Ramesh et al. (2022), Stable-diffusion Rombach et al. (2022) synthesize high-quality images in diverse scenes, views, and lighting conditions from text prompts. These models generate high-quality and diverse images since they have been trained on a large collection of image-text pairs and can capture the visual-semantic correspondence effectively. While diffusion models generate images that appear highly realistic, recent studies have demonstrated that these images can still be distinguished from natural ones using advanced image forensic tools Corvi et al. (2023). This suggests that although state-of-the-art generative models excel at tasks like image editing, they often leave behind subtle, unnatural artifacts. Ensuring high image quality is therefore critical for various generative tasks, such as personalized few-shot finetuning Ruiz et al. (2022); Kumari et al. (2022), super-resolution Karras et al. (2022); Dhariwal & Nichol (2021), image restoration, and unconditional image generation.

Our goal is to improve the image quality using natural image statistics by exploiting the well-known kurtosis concentration property of natural images Zhang & Lyu (2014); Zoran & Weiss (2009); Wainwright & Simoncelli (1999). This property states that natural images have nearly constant kurtosis (fourth order moment) values across different band-pass (e.g., Discrete Cosine Transform (DCT), Discrete Wavelet Transform (DWT)) versions of the images Zhang & Lyu (2014). Inspired by this property, we propose a novel kurtosis concentration (KC) loss, which is generic and applicable to any diffusion-based pipeline. More specifically, this loss minimizes the gap in the kurtosis of an image across band-pass filtered versions and improves the quality of the generated images. We also propose a novel perceptual guidance (PG) strategy during inference which is agnostic to conditioning (e.g., text/class) and further improves image quality. Both KC loss and PG strategies are general-purpose and do not require any labels. It can be adapted to various generative tasks with minimal effort. In this work, we experiment with diverse tasks: (1) personalized few-shot finetuning of

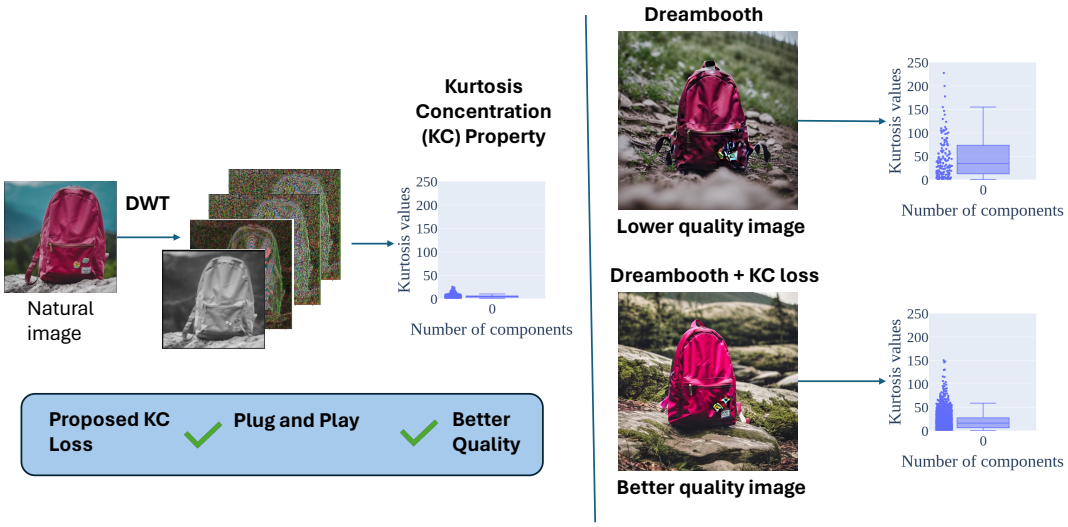


Figure 1: Overview of DiffNat: Natural images exhibit consistent kurtosis values across bandpass-filtered versions, shown by the tight blue box in the left figure. Diffusion-generated images (top right) show greater kurtosis spread, indicating lower quality. Adding KC loss (bottom right) reduces kurtosis variance and improves image quality.

text-to-image diffusion model, (2) unconditional image generation, (3) image super-resolution, and (4) blind face-restoration.

Our major contributions are as follows:

- We introduce DiffNat - a framework for improving the image quality of diffusion models using the kurtosis concentration property of natural images.
- We provide insights on how reducing kurtosis improves image quality. This is the primary motivation for the proposed loss function.
- A novel condition-agnostic perceptual guidance strategy is proposed which further improves image quality.
- We validate the proposed KC loss and PG strategy on diverse generative tasks, e.g., (1) personalized few-shot finetuning of text-to-image diffusion model using text guidance, (2) unconditional image generation, (3) image super-resolution, and (4) blind face-restoration. Experiments indicate that incorporating the proposed KC loss and PG enhances perceptual quality across various tasks and benchmarks, and this improvement has been validated through a user study.

2 Related Work

Generative models. Recent progress in generating high-fidelity, diverse images from text inputs has been remarkable. Initially, GAN-based methods dominated text-to-image generation Qiao et al. (2019); Tao et al. (2022); Liao et al. (2022); Zhu et al. (2019); Ruan et al. (2021), but recent advances have shifted towards diffusion models like Stable Diffusion Rombach et al. (2022) and Imagen Saharia et al. (2022), which leverage large datasets for training. Text-based image editing has also advanced significantly; GAN-based approaches have improved with CLIP Radford et al. (2021), while diffusion-based methods offer better control and impressive results Ruiz et al. (2022); Kumari et al. (2022); Gal et al. (2022). Personalization techniques such as Textual Inversion Gal et al. (2022), DreamBooth Ruiz et al. (2022), and Custom Diffusion Kumari et al. (2022) allow for the creation of unique images by embedding subjects or concepts into the model's output. In unconditional image generation, the Denoised Diffusion Probabilistic Model (DDPM) Ho et al. (2020) is a leading method, providing superior image quality through variational inference and image-space denoising. For conditional tasks like image super-resolution, guided diffusion Dhariwal & Nichol (2021)

and latent diffusion models Karras et al. (2022) are highly effective, producing high-resolution images from low-resolution inputs.

Natural Image Statistics. Natural images have interesting scale-invariance and noise properties Zoran & Weiss (2009), which have been used for image restoration problems. Projected KC property of natural images, i.e., natural images tend to have constant kurtosis values across different band-pass (DCT, DWT) filtered versions has been used for blind forgery detection Zhang & Lyu (2014). Inspired by these observations, we propose a novel loss function based on natural image statistics for generating better quality images.

3 Method

In this section, we present the concept of KC loss (Sec. 3.2) which can be applied to various generative tasks for enhancing the quality of generated images. We start by providing a basic understanding of the KC property of natural images.

3.1 Kurtosis Concentration Property

Definition 1 *Kurtosis: It is a measure of the "peakedness" of the probability distribution of a random variable Zhang & Lyu (2014). For a random variable x, its kurtosis is defined as,*

$$\kappa(x) = \frac{\mu_4(x)}{(\sigma^2(x))^2} - 3. \tag{1}$$

where $\sigma^2(x) = \mathbb{E}_x[(x - \mathbb{E}_x(x))^2]$ and $\mu_4(x) = \mathbb{E}_x[(x - \mathbb{E}_x(x))^4]$ are the second order and fourth order moment of x. For example, the Gaussian random variable has a kurtosis value of 0.

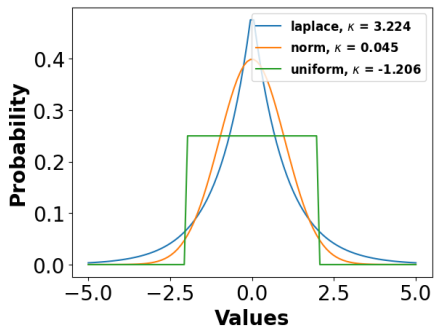


Figure 2: Kurtosis of various distributions where kurtosis captures the peakedness of distributions.

Kurtosis of well-known distributions is shown in Fig. 2. We can observe that a positive kurtosis indicates that the distribution is more peaked than the normal distribution and negative kurtosis indicates it to be less peaked than normal distribution Zhang & Lyu (2014). Kurtosis is a useful statistic used for blind source separation Naik et al. (2014) and independent component analysis (ICA) Stone (2002).

For a random vector x, we define the kurtosis of the 1D projection of x onto a unit vector w as a projection kurtosis, i.e., $\kappa(w^T x)$. Projection kurtosis is an effective measure of the statistical properties of high-dimensional data. E.g., if x is a Gaussian, its projection over any w has a 1D Gaussian distribution. Therefore, its projection kurtosis is always zero, which exhibits the kurtosis concentration (to a single value, i.e., zero) of Gaussian.

It is well-known that natural images can be modeled using zero-mean Gaussian Scale Mixture (GSM) vector Zoran & Weiss (2009); Zhang & Lyu (2014); Lyu et al. (2014); Wainwright & Simoncelli (1999). Next, we analyze an interesting property of the GSM vector.

Lemma 1 *A GSM vector x with zero mean has the following probability density function:*

$$p(x) = \int_0^\infty \mathcal{N}(x; 0, z\Sigma_x) p_z(z) dz \tag{2}$$

and its projection kurtosis is <u>constant</u> with respect to the projection direction w, i.e.,

$$\kappa(w^T x) = \frac{3 var_z\{z\}}{\mathcal{E}_z\{z\}^2} \tag{3}$$

where $\mathcal{N}(x; 0, z\Sigma_x)$ denotes a Gaussian distribution with zero mean and covariance matrix $z\Sigma_x$, with z a positive random variable with density $p_z(z)$. $\mathcal{E}_z\{z\}$ and $var_z\{z\}$ are the mean and variance of latent variable z respectively.

Proof. The proof is provided in the Appendix.

This result by Zhang & Lyu (2014) shows that projection kurtosis is constant across projection directions (e.g., wavelet basis), which provides theoretical insights into the kurtosis concentration property, which we will discuss next.

Kurtosis Concentration Property: It has been observed that for natural images, projection kurtosis values across different bandpass-filtered channels tend to be close to a constant value. This is termed as kurtosis concentration property of natural images Zhang & Lyu (2014); Zoran & Weiss (2009); Lyu et al. (2014); Bethge (2006); Lyu & Simoncelli (2009); Wainwright & Simoncelli (1999). It can also be interpreted as an implication of Lemma 1, if we consider patches of natural images as zero-mean GSM vector (Zoran & Weiss (2009), Wainwright & Simoncelli (1999)) and projection directions correspond to bandpass filters, e.g., DWT.

One intuitive reasoning of the projected kurtosis concentration property is given as follows. It is observed that the distribution ($p(x, \alpha, \beta)$) of different bandpass (DWT) filtered versions of natural images follows a generalized Gaussian density of the form Zhang & Lyu (2014); Zoran & Weiss (2009).

$$p(x, \alpha, \beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} exp(-\frac{|x|}{\alpha})^{\beta} \tag{4}$$

where α, β are scaling parameters and $\Gamma(.)$ is the Gamma function. The kurtosis of this function is given by Zoran & Weiss (2009),

$$\kappa = \frac{\Gamma(1/\beta)\Gamma(5/\beta)}{\Gamma(3/\beta)^2} \tag{5}$$

Empirically, it has been shown that for natural images, β takes relatively small values, ranging from 0.5 to 1 Zoran & Weiss (2009), and this kurtosis value tends to be constant Zhang & Lyu (2014); Zoran & Weiss (2009); Lyu et al. (2014); Wainwright & Simoncelli (1999), independent of α or x.

We investigate and experimentally verify this property for natural images on large datasets, e.g., FFHQ dataset (Fig. 9(c)), Dreambooth dataset, Oxford-flowers dataset (in Appendix). We conclude that this property actually holds for both object datasets (Dreambooth dataset, Oxford flowers), face dataset (FFHQ) with sufficient variations in viewpoint, scale, illumination, color, objects, pose, lighting condition etc. Analysis of kurtosis difference has been shown in Fig. 9 and also in the Appendix, which clearly shows that the difference of kurtosis values are higher in diffusion-generated images compared to natural images in these datasets.

3.2 Kurtosis Concentration (KC) Loss

In this work, we introduce a novel KC loss function for training deep generative models, leveraging the KC property of natural images to improve perceptual quality. Unlike previous approaches Zhang & Lyu (2014) that used the KC property for tasks like noise estimation and source separation, our KC loss can be integrated into any diffusion pipeline, and we validate its effectiveness with state-of-the-art diffusion models.

Suppose we need to train or finetune a diffusion model f_θ using input training images (x) with or without a conditioning vector c. The conditioning vector could be text, image, or none (in the case of the unconditional diffusion model). Given an initial noise map $\epsilon \sim \mathcal{N}(0, I)$, and a conditioning vector c, the generated image obtained from f_θ is given by $x_{gen} = f_\theta(x, \epsilon, c)$. Typically, the diffusion model is trained to minimize the $l2$ distance between the ground truth image (x) and the noisy image (x_{gen}) Dhariwal & Nichol (2021) or their corresponding latent in case of Latent Diffusion Model (LDM) Rombach et al. (2022). Without loss of generality, we are referring to that as reconstruction loss (L_{recon}) between the ground-truth image (x) and the generated image (x_{gen}), denoted by,

$$L_{recon} = \mathbb{E}_{x,c,\epsilon}[\ ||x_{gen} - x||_2^2] \tag{6}$$

Note that for diffusion models trained to predict the added noise, we could deterministically obtain the intermediate clean image from the predicted noise and apply the loss to that. Next, we will describe the KC

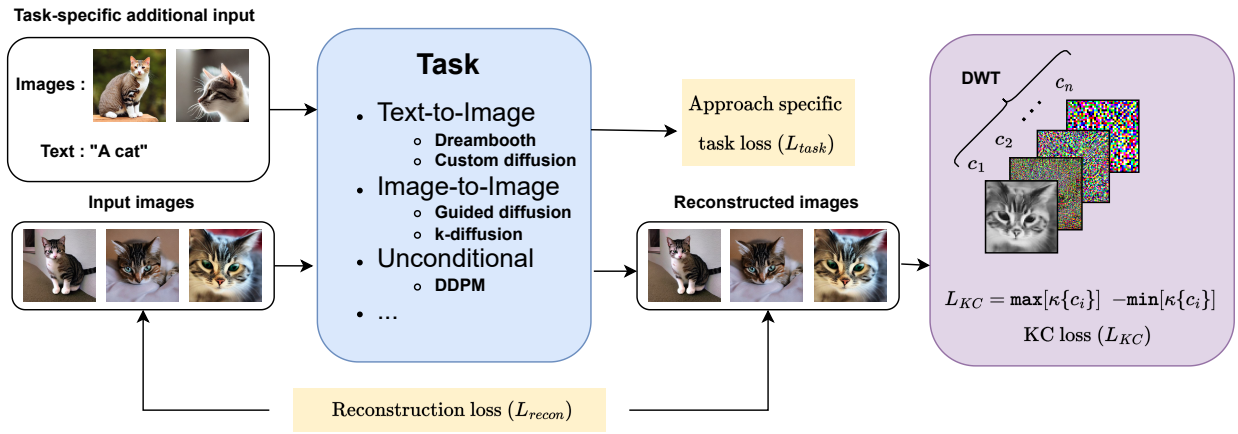


Figure 3: Overview of DiffNat. The proposed kurtosis concentration loss can be integrated into any diffusion-based approach for various tasks (e.g., text-to-image generation (DreamBooth, Custom diffusion), super-resolution image-to-image generation (Guided diffusion, k-diffusion), unconditional image generation (DDPM)). In addition to the task-specific losses, and general reconstruction loss, we incorporate the kurtosis concentration loss (L_{KC}), which operates on the reconstructed images and minimizes the kurtosis deviation (i.e., max[$\kappa\{c_i\}$] - min[$\kappa\{c_i\}$]) across Discrete Wavelet Transform (DWT) filtered version of the reconstructed image, Here, c_1, c_2 .. are DWT filtered version of the reconstructed image and $\kappa(x)$ denote kurtosis of x.

loss. Note, that the KC property holds across different bandpass transformed domains (DCT, DWT, fastICA) and we choose DWT because it is widely used due to its hierarchical structure and energy compaction properties E Woods & C Gonzalez (2008). Typically, DWT transforms images into LL (low-low), LH (low-high), HL (high-low), HH (high-high) frequency bands and each of the sub-bands contains several sparse details of the image. E.g., the LL and HH subbands contain a low-pass and high-pass filtered version of the image, respectively Zhang & Lyu (2014). The generated image x_{gen} is then transformed using Discrete Wavelet Transform (DWT) with kernels $k_1, k_2, .., k_n$ producing filtered images $g_{gen,1}, g_{gen,2}, .., g_{gen,n}$ respectively, such that, $g_{gen,i} = F_{k_i}(x_{gen})$. Here, F_l denotes the discrete wavelet transform with kernel l.

Now, kurtosis values of these $g_{gen,i}$ should be constant by the KC property, therefore, we minimize the difference between the maximum and minimum values of the kurtosis of $g_{gen,i}$'s to finetune the model using the loss,

$$L_{KC} = \mathbb{E}_{x,c,\epsilon}[\max(\kappa\{g_{gen,i}\}) - \min(\kappa\{g_{gen,i}\})] \tag{7}$$

Here, $\kappa(x)$ is kurtosis of x. Note that, this loss (Algorithm. 1) is quite generic and can be applied to both image or latent diffusion models for training. In the case of latent diffusion models, we need to transform the latent to image space (via a pretrained VQVAE), before applying this loss, since this prior holds for image space only. At each timestep (t), we extract the clean image (x_0) from the noisy latents (ϵ_t) ($x_0 = (x_t - \sigma_t.\epsilon_t)/\alpha_t$) and apply KC loss exclusively to the clean image, not the noisy one. In case of applying this loss to any task T (DreamBooth, super-resolution, unconditional image generation), the overall loss (L) function would be, $L = L_{task} + L_{recon} + L_{KC}$, where L_{task} is the task-specific loss.

Algorithm 1: Kurtosis Concentration loss

Input: Diffusion model (f_θ), training images (x), condition vector (c)
Output: KC loss L_{KC}

1. $\epsilon \sim \mathcal{N}(0, I)$; `// Sample random noise`
2. $x_{gen} = f_\theta(x, \epsilon, c)$; `// Generate image`
3. $g_{gen,1}, g_{gen,2}, g_{gen,3}, .. = DWT(x_{gen})$; `// Wavelet decomposed images`
4. $L_{KC} = \mathbb{E}_{x,c,\epsilon}[\max(\kappa\{g_{gen,i}\}) - \min(\kappa\{g_{gen,i}\})]$; `// Compute the KC loss`

3.3 Perceptual Guidance during Inference

Applying the KC loss during diffusion training enhances sample quality compared to vanilla diffusion. Additionally, we propose a novel Perceptual Guidance (PG) mechanism during inference to further improve

Algorithm 2: Perceptual Guidance

Input: Base diffusion model (θ_B), Diffusion model trained with KC (θ_P), prompt (c), guidance scale (γ)
Output: output image (x_0)
$x_T = \mathcal{N}(0, I)$
for *t in T, T-1, .., 1* **do**
  $\hat{\epsilon}_{\theta_B,\theta_P}(x_t, c, t) = \epsilon_{\theta_B}(x_t, c, t) + \gamma\Big(\epsilon_{\theta_P}(x_t, c, t) - \epsilon_{\theta_B}(x_t, c, t)\Big)$
  if *t > 1* **then**
    $x_{t-1} \sim \mathcal{N}(\mu(x_t), \Sigma_t)$
  else
    $x_0 = (x_1 - \sigma_1 \hat{\epsilon_1})/\alpha_1$
  end
end
return x_0

perceptual quality. Conceptually, this is analogous to classifier-free guidance, where the diffusion model is evaluated both conditionally and unconditionally at each step, and the difference in their outputs serves as a gradient direction toward the condition. However, while classifier-free guidance can indirectly improve perceptual quality, it has notable limitations, i.e., (1) it is restricted to conditional models, and (2) its primary goal is alignment with the condition, which does not necessarily guarantee improved perceptual quality. In contrast, our PG mechanism generates two outputs with differing perceptual qualities—using two models, one trained with KC loss and the other without—and amplifies the intermediate output at each step toward the model with better perceptual quality. This approach offers key advantages - (1) it is condition-agnostic, making it more versatile, (2) it can operate alongside traditional classifier-free guidance, complementing its functionality (Tab. 11).

Diffusion models typically generate images unconditionally by approximating the true data distribution $q(x_0)$ with $p_\theta(x_0)$. However, many applications require conditioning on labels or text prompts c Dhariwal & Nichol (2021). From the Bayesian perspective, the conditional score can be expressed as:

$$\nabla_x \log p(x \mid c) = \nabla_x \log p(x) + \nabla_x \log p(c \mid x) \tag{8}$$

Given $\epsilon_\theta(x_t, t)$ ($\approx -\sigma_t \nabla_{x_t} \log p(x_t)$) is a model trained to predict the noise added to a sample, the above equation can be approximated as,

$$\hat{\epsilon}_\theta(x_t, c, t) = \epsilon_\theta(x_t, t) - \sigma_t \nabla_{x_t} \log p_\phi(c \mid x_t) \tag{9}$$

Classifier guidance Dhariwal & Nichol (2021) meets this need by introducing an auxiliary model $p_\phi(c \mid x_t)$ and uses the modified score function $\hat{\epsilon}$, which, while effective, requires training a robust classifier for all denoising steps—a challenging and resource-intensive task.

To avoid this, classifier-free guidance (CFG) Ho & Salimans (2022) uses a single neural network to parameterize both the conditional and unconditional model. Ho & Salimans (2022) incorporates c directly into the denoising network, training it to predict both the conditional score $\epsilon_\theta(x_t, c, t)$ and the unconditional score $\epsilon_\theta(x_t, t)$ by randomly dropping the prompt during training. The resulting conditional gradient,

$$-\frac{1}{\sigma_t}\Big(\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, t)\Big) \tag{10}$$

leads to an updated diffusion score (putting Eq. 10 in Eq. 9):

$$\hat{\epsilon}_\theta(x_t, c, t) = \epsilon_\theta(x_t, t) + \gamma\Big(\epsilon_\theta(x_t, c, t) - \epsilon_\theta(x_t, t)\Big). \tag{11}$$

where γ is a constant guidance scale.

With a similar intuition, Perceptual Guidance (PG) refines the denoising trajectory in diffusion models by integrating a perceptual prior during inference. Instead of calculating class/text-conditional or unconditional

scores, we calculate scores corresponding to high or low perceptual quality. In PG, we redefine the condition as an indicator function based on the presence of KC loss, $c' = \mathbf{1}_{\text{KC}}$. We can reformulate Eq. 8 as,

$$\nabla_x \log p_\gamma(x \mid c' = \mathbf{1}_{\text{KC}}, c) = \nabla_x \log p(x) + \gamma(\nabla_x \log p(x \mid c' = \mathbf{1}_{\text{KC}}, c) - \nabla_x \log p(x)) \tag{12}$$

We derive the guiding gradient from the difference between models trained with and w/o KC loss (representing conditional and unconditional predictions). Suppose the baseline diffusion model is denoted by θ_B, and the model trained with KC loss is denoted by θ_P (i.e., if $c' = True$, then $\theta = \theta_P$, else $\theta = \theta_B$). During perceptual guidance, we guide the diffusion process through the difference of output predicted by θ_B and θ_P as follows,

$$\hat{\epsilon}_{\theta_B,\theta_P}(x_t, c, t) = \epsilon_{\theta_B}(x_t, c, t) + \gamma\Big(\epsilon_{\theta_P}(x_t, c, t) - \epsilon_{\theta_B}(x_t, c, t)\Big).$$

As a result, PG introduces a perceptual-optimized (using KC) denoiser where $\gamma > 1$ enhances perceptual quality. Note that, here c could be a text prompt or NULL. This process is iterated T times to generate the final sample as shown in Algorithm 2. A concurrent work on auto-guidance Karras et al. (2024) improves image quality by guiding diffusion models using a weaker version of itself. Instead, PG leverages the difference between models trained with and without KC loss for guidance (Tab. 10).

3.4 Intuitive Justification

Why do diffusion-generated images have higher kurtosis values? Natural images typically exhibit smooth transitions and structured patterns, leading to a pixel intensity distribution with fewer outliers and, consequently, lower kurtosis Zoran & Weiss (2009); Lyu et al. (2014); Wainwright & Simoncelli (1999); Bethge (2006). In contrast, diffusion models generate images through iterative refinement of pure noise into coherent structures. Due to imperfections in the trained UNet and finite denoising steps, residual high-frequency noise may persist in the final output, leading to more extreme pixel intensity values and contributing to heavier tails in the distribution Zhang et al. (2023). This phenomenon is more effectively characterized in the frequency domain (wavelet transform), which generally correspond to higher kurtosis values Zhang & Lyu (2014) (Fig. 9).

How does KC loss improve image quality? Intuitively, minimizing the KC loss can be seen as a locality-aware smoothing of the data distribution to enhance perceptual quality. For instance, in Fig. 6, the output of GD without KC loss exhibits undesirable abrupt changes near the eye region. Ideally, we should first detect such regions and then apply suitable operations to improve perceptual quality. However, performing a spatially varying refinement is challenging, and globally applying such filters might be sub-optimal for other regions. It has been observed that the wavelet coefficients of natural images follow a Generalized Gaussian density Moulin & Liu (1999); Sharifi & Leon-Garcia (1995); Mallat (1989), and Kurtosis quantifies the heaviness of the tails and peakedness of a distribution compared to the Gaussian distribution Zhang & Lyu (2014); Maier (2021). Therefore, kurtosis across bandpass wavelet filtered versions of the image automatically provides locality of the abrupt changes and minimizing KC loss performs locality-aware smoothing of the data distribution. In Fig. 6, incorporating KC loss enhances the generated eye region.

4 Experiments

We evaluate the efficacy of the proposed loss for four tasks - (1) personalized few-shot finetuning of diffusion model using text guidance, (2) unconditional image generation, (3) image super-resolution, and (4) blind face-restoration.

4.1 Task 1: Personalized few-shot finetuning using text guidance

In this section, we address the problem of finetuning the text-to-image diffusion model from a few examples for text-guided image generation in a subject-driven manner. Specifically, given only a few images (e.g., 3-5) of a particular subject without any textual description, our task is to learn the subject-specific details and generate new images of that particular subject in different conditions specified by the text prompt. To evaluate the efficacy of KC loss in this task, we build upon methods, (1) DreamBooth Ruiz et al. (2022), (2) Custom diffusion Kumari et al. (2022), and (3) AttnDreambooth (AttnDB) Pang et al. (2024). We evaluate all approaches with/without KC loss on the DreamBooth dataset with same setup for a fair comparison. We

Figure 4: Qualitative comparison of with/without KC loss and PG in DreamBooth (DB). The quality of teapot image (top row) improves while adding KC and PG. In the bottom row, the background of DB generated images looks blurry and unnatural (highlighted in red), while adding KC and PG improves image quality, and the background look more natural.

have also compared with another naturalness loss, i.e., LPIPS loss Zhang et al. (2018) as a baseline. For KC loss, we decompose the reconstructed images using 27 'Daubechies' filter banks and get the average difference of the kurtosis values as a loss function. When adding the proposed KC loss to these approaches, we obtain performance improvements in visual quality, i.e., FID Lucic et al. (2018), MUSIQ score Ke et al. (2021), HPSv2 Wu et al. (2023), subject/prompt fidelity metrics (DINO, CLIP-I, CLIP-T) and LPIPS-diversity Zhang et al. (2018) as shown in Tab. 1. The qualitative results are shown in Fig. 4. In practice, better perceptual quality often correlates with higher semantic fidelity Dhariwal & Nichol (2021); Ramesh et al. (2022), as both improvements are driven by more accurate and coherent image generation that captures the intended high-level semantics of the prompt, verified in Tab. 1. More training details and ablation are provided in the Appendix.

Table 1: Comparison on personalized few-shot finetuning task

Method	Image quality			Subject fidelity		Prompt fidelity	Image diversity
	FID ↓	MUSIQ ↑	HPSv2 ↑	DINO ↑	CLIP-I ↑	CLIP-T ↑	LPIPS-div ↑
DB Ruiz et al. (2022)	111.76	68.31	25.12	0.65	0.81	0.31	0.38
DB Ruiz et al. (2022) + LPIPS	108.23	68.39	25.43	0.65	0.80	0.32	0.40
DB + KC loss (Ours)	100.08	69.78	25.82	0.68	0.84	0.34	0.42
DB + KC loss + PG (Ours)	**93.45**	**70.82**	**26.04**	**0.70**	**0.86**	**0.35**	**0.43**
CD Kumari et al. (2022)	84.65	70.15	26.12	0.71	0.87	0.38	0.40
CD Kumari et al. (2022) + LPIPS	80.12	70.56	26.33	0.71	0.87	0.37	0.43
CD + KC loss (Ours)	75.68	72.22	26.64	0.73	0.88	0.40	0.44
CD + KC loss + PG (Ours)	**66.27**	**73.77**	**27.10**	**0.77**	**0.89**	**0.43**	**0.46**
AttnDB Pang et al. (2024)	80.59	70.23	26.42	0.72	0.85	0.35	0.41
AttnDB Pang et al. (2024) + LPIPS	80.06	70.50	26.66	0.73	0.87	0.35	0.43
AttnDB + KC loss (Ours)	73.02	71.10	26.83	0.75	0.89	0.37	0.45
AttnDB + KC loss + PG (Ours)	**64.78**	**72.89**	**27.35**	**0.78**	**0.90**	**0.38**	**0.47**

Human evaluation. Since perceptual metrics are not always reliable, we also conducted a human preference study using Amazon Mechanical Turk (AMT) for (1) subject fidelity assessment and (2) image quality ranking. For the subject fidelity assessment, we evaluated the visual similarity of real and generated images, both with and without KC loss, to the actual subject. We asked around 5000 visual similarity questions to 50 unbiased users (age 20-50, randomized gender, AMT). The average rating was 5.8 (on a scale where 0 is "extremely unlikely" and 10 is "extremely likely"), indicating our proposed loss retains subject fidelity in most cases. We also had 50 unbiased users rank our method against baselines (i.e., "DiffNat", "DreamBooth",

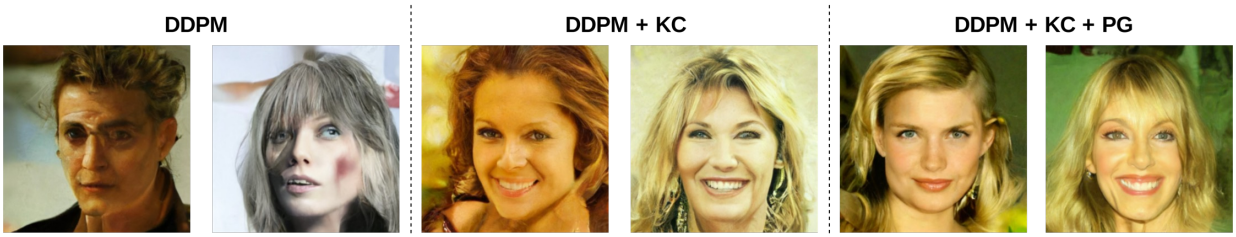


Figure 5: Comparison of unconditional image generation (DDPM) with/without KC loss. Integrating KC loss significantly improve image quality, whereas DDPM generated images have unnatural image artifacts.

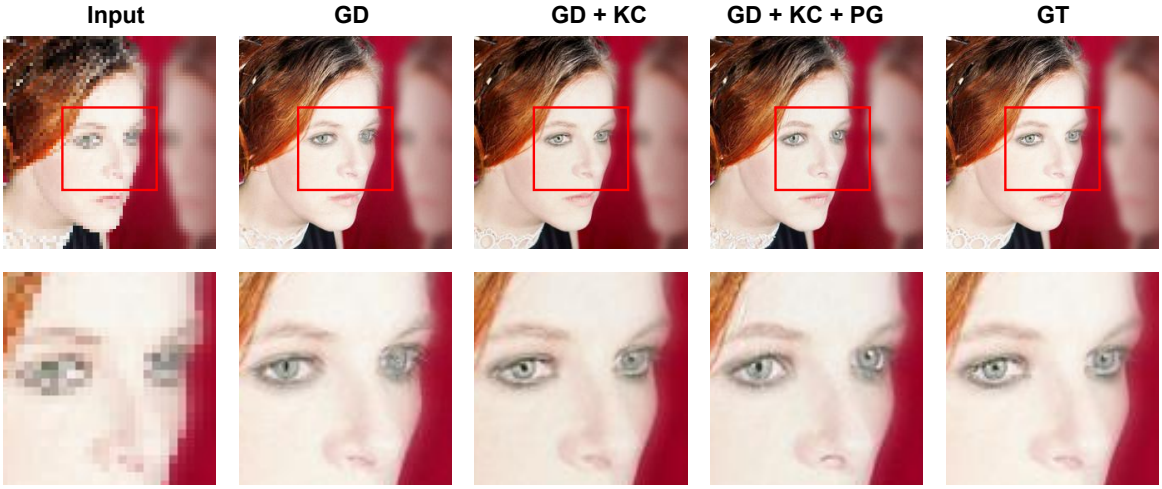


Figure 6: Qualitative comparison of with/without KC loss, PG in guided diffusion (GD). The bottom image (with KC loss) has better eye details (best viewed in color).

"Custom diffusion", "None is satisfactory"), totaling 1500 questionnaires. The aggregate responses showed that DiffNat-generated images significantly outperformed the baselines by a large margin (50.4%). Further details are provided in the Appendix.

4.2 Task 2: Unconditional image generation

Unconditional image generation operates without the need for text or image guidance. It aims to learn the training data distribution through a generative model (in this case, a diffusion model) and produce samples that resemble the training data distribution. We opted for the well-known unconditional image generation pipeline, the DDPM Ho et al. (2020), to test the efficacy of KC loss. PG is especially effective here because classifier-free guidance cannot be applied.

In DDPM, we directly integrate the KC loss into the image space, demonstrating the flexibility of our proposed loss. We experimented with the Oxford-flowers Nilsback & Zisserman (2006), CelebA-faces Zhang et al. (2020), CelebAHQ Karras et al. (2017), Stanford-Dogs Khosla et al. (2011) and Stanford-Cars Krause et al. (2013) datasets, achieving consistent improvements in image quality, as shown in Tab. 2 and Figure 5. Additionally, PG further enhances image quality, as indicated in Tab. 2. Human evaluation is not feasible for unconditional image generation due to the lack of one-to-one correspondence between training and generated images, but quantitative and qualitative analyses demonstrate the effectiveness of our approach.

4.3 Task 3: Image super-resolution

Image super-resolution typically takes the form of a conditional generation task, leveraging a low-resolution image as an additional condition for the diffusion model. In this study, we use two state-of-the-art diffusion pipelines as baselines for comparison. Guided diffusion (GD) Dhariwal & Nichol (2021) directly takes the

Table 2: Comparison of unconditional image generation task

Method	Oxford flowers		Celeb-faces		CelebAHQ		Stanford-Dogs		Stanford-Cars	
	FID ↓	MUSIQ ↑	FID ↓	MUSIQ ↑	FID ↓	MUSIQ ↑	FID ↓	MUSIQ ↑	FID ↓	MUSIQ ↑
DDPM Ho et al. (2020)	243.43	20.67	202.67	19.07	199.77	46.05	129.91	50.12	143.71	53.77
DDPM Ho et al. (2020) + LPIPS	242.62	20.80	201.55	19.21	197.17	46.15	115.72	50.86	137.22	53.98
DDPM + KC loss (Ours)	237.73	21.13	198.23	19.52	190.59	46.83	105.45	51.53	125.85	54.21
DDPM + KC loss + PG (Ours)	**200.12**	**22.45**	**188.49**	**20.82**	**175.12**	**48.32**	**98.77**	**52.17**	**115.93**	**54.85**

Table 3: Comparison of image super-resolution (x4) task

Method	Image quality				Image Diversity
	FID ↓	PSNR ↑	SSIM ↑	MUSIQ ↑	LPIPS-div ↑
GD Dhariwal & Nichol (2021)	121.23	18.13	0.54	57.31	0.37
GD Dhariwal & Nichol (2021) + LPIPS	119.81	18.22	0.54	57.42	0.40
GD + KC loss (Ours)	103.19	18.92	0.55	58.69	0.42
GD + KC loss + PG (Ours)	**93.45**	**20.17**	**0.58**	**60.13**	**0.45**
LD Karras et al. (2022)	95.83	19.16	0.56	59.57	0.38
LD Karras et al. (2022) + LPIPS	92.77	19.42	0.57	59.82	0.40
LD + KC loss (Ours)	83.34	20.25	0.58	61.20	0.42
LD + KC loss + PG (Ours)	**71.33**	**21.92**	**0.60**	**62.85**	**0.44**

low-resolution image as a condition and performs the diffusion operation in the pixel space. Additionally, we also explore the latent diffusion model (LD) Karras et al. (2022) that operates in the latent space of a pre-trained VQVAE Esser et al. (2021).

Note that, as GD operates in the pixel space, we directly add the proposed KC loss to the output of the denoising UNet. Conversely, for LD, we initially convert the latent embedding to image space using the pre-trained decoder and integrate the KC loss on the output of the decoder. For training, we use the standard FFHQ dataset Karras et al. (2017), which contains 70k high-quality images. We address the task of $\times 2$, $\times 4$, and $\times 8$ super-resolution where the GT images are of resolution 256×256. We evaluate randomly sampled 3000 images from CelebA-Test dataset Karras et al. (2017) under the same $\times 2$, $\times 4$ and $\times 8$-SR setting in Tab. 4, Tab. 3 and Tab. 5 respectively. In the qualitative results shown in Fig. 6, we observe that adding KC loss improves the image quality and finer details, e.g., eye structure, texture, and lighting.

Human evaluation. We conduct a human evaluation of the image super-resolution task to compare GD and LD with the addition of KC loss to each counterpart (DiffNat). The aggregate response of choices (corresponding to best quality images w.r.t methods) from 50 unbiased users (age 20-50, randomized gender, AMT) across 1000 questionnaires, shown in Fig. 8, indicates that DiffNat-generated images have superior quality compared to the GD and LD baselines.

4.4 Task 4: Blind face restoration

We also verify the efficacy of KC loss and PG for blind face restoration task. For blind face restoration task Suin et al. (2024), we train on FFHQ dataset with and without KC loss on IPC baseline Suin et al. (2024), and evaluate on a subset of Celeb-A test set with a resolution of 256x256. Average LPIPS, FID, IDS, PSNR, SSIM are reported in Tab. 18. Qualitative results (Fig. 28) also verify that adding KC loss and PG improves the image quality.

Table 6: Face restoration

Method	LPIPS-div↑	FID↓	IDS ↑	PSNR ↑	SSIM ↑
DifFace	0.20	70.69	0.48	22.82	0.61
RestoreFormer	0.29	60.98	0.39	21.77	0.53
IPC (WACV'24)	0.32	55.42	0.54	22.34	0.60
IPC + KC	0.34	43.21	0.61	24.19	0.64
IPC + KC + PG	**0.36**	**38.23**	**0.65**	**25.71**	**0.69**

5 Ablation and analysis

Ablation of loss and guidance. Here we perform ablation studies of diffusion backbone (SD-1.5, SDXL), KC loss and PG as shown in Tab. 7 & 8 on DreamBooth dataset. We observe that PG is complementary to classifier-free guidance (CFG) and both KC loss and PG improve image quality as shown in Fig. 4, and

Table 4: Image super-resolution (x2) task

Method	Image quality				
	FID ↓	PSNR ↑	SSIM ↑	LPIPS-div ↑	MUSIQ ↑
GD Dhariwal & Nichol (2021)	100.2	19.4	0.62	0.25	58.12
GD + KC loss (Ours)	80.9	20.2	0.66	0.28	59.91
GD + KC loss + PG (Ours)	**71.3**	**21.7**	**0.69**	**0.33**	**60.32**
LD. Karras et al. (2022)	82.45	21.2	0.64	0.26	60.23
LD + KC loss (Ours)	70.12	22.3	0.70	0.29	62.15
LD + KC loss + PG (Ours)	**59.32**	**23.6**	**0.73**	**0.31**	**63.72**

Table 5: Image super-resolution (x8) task

Method	Image quality				
	FID ↓	PSNR ↑	SSIM ↑	LPIPS-div ↑	MUSIQ ↑
GD Dhariwal & Nichol (2021)	140.3	17.5	0.52	0.33	55.26
GD + KC loss (Ours)	125.5	18.7	0.56	0.35	57.33
GD + KC loss + PG (Ours)	**108.6**	**19.1**	**0.58**	**0.38**	**58.62**
LD. Karras et al. (2022)	103.2	18.7	0.59	0.40	58.62
LD + KC loss (Ours)	80.1	19.5	0.67	0.43	60.31
LD + KC loss + PG (Ours)	**67.3**	**20.8**	**0.69**	**0.44**	**61.87**

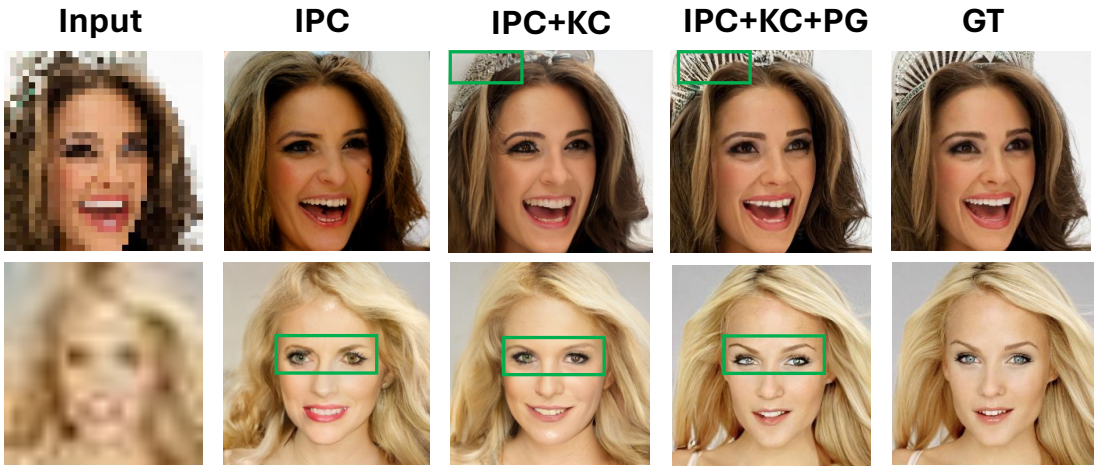


Figure 7: Restoration comparison with IPC. KC and PG improves image quality (highlighted in green).

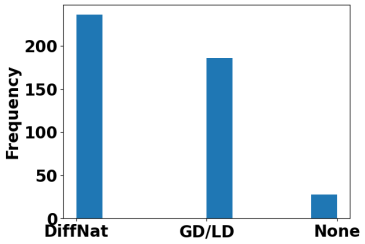


Figure 8: Human evaluation for image super-resolution task.

Table 7: Loss & guidance ablation on DB (SD-1.5)

KC	CFG	PG	FID ↓	MUSIQ ↑
✗	✗	✗	125.56	67.12
✗	✗	✓	114.38	68.05
✗	✓	✗	111.76	68.31
✗	✓	✓	109.12	68.92
✓	✗	✗	105.33	69.34
✓	✓	✗	100.08	69.78
✓	✗	✓	98.21	70.02
✓	✓	✓	**93.45**	**70.82**

Table 8: Loss & guidance ablation on DB (SDXL)

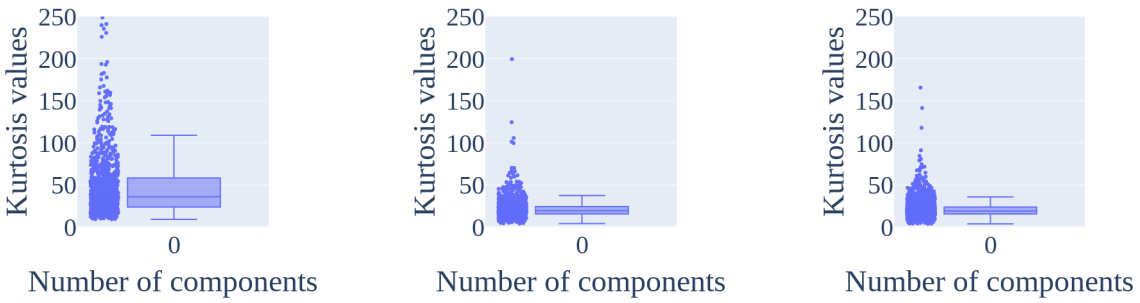
KC	CFG	PG	FID ↓	MUSIQ ↑
✗	✗	✗	102.32	70.18
✗	✗	✓	96.17	70.92
✗	✓	✗	95.32	71.35
✗	✓	✓	89.72	71.87
✓	✗	✗	85.11	72.05
✓	✓	✗	80.75	72.32
✓	✗	✓	73.33	72.88
✓	✓	✓	**70.18**	**73.02**

Fig. 6. We have also performed ablations w.r.t transforms (DCT, DWT) and the results are provided in the Appendix.

Kurtosis analysis. To verify the efficacy of the proposed KC loss, we performed an average kurtosis analysis by computing the average kurtosis deviation of DWT-filtered images from the FFHQ dataset and plotting the results in Fig. 9. The analysis showed that images generated with GD had the highest kurtosis deviation (Fig. 9 (a)), while natural images had the least deviation (Fig. 9 (c)), and adding KC loss reduced the kurtosis deviation (Fig. 9 (b)), thus improving image quality as demonstrated both qualitatively and quantitatively.

Table 9: Comparison of real vs synthetic detection

Method	Accuracy
DB	93.33%
DB + KC loss	66.66%
CD	94.16 %
CD + KC loss	92.50%



(a) Avg. κ of GD generated images (b) Avg. κ of images with GD + KC (c) Avg. κ of Natural images

Figure 9: Average kurtosis (κ) analysis of guided diffusion (GD) framework trained on FFHQ dataset. From this analysis, it is evident that GD-generated images have higher kurtosis deviation. Integrating KC loss reduces the kurtosis deviation to preserve the naturalness of the generated images. Natural images have more concentrated kurtosis values.

(a) GD (b) GD + KC (c) LD (d) LD + KC

Figure 10: Perceptual artifact ratio analysis. Green boundaries localizes perceptual artifacts. Adding KC loss reduces such artifacts

Comparison of real vs synthetic detection. To analyze the robustness of the proposed KC loss, we train a classifier to distinguish real images from synthetic ones generated by diffusion models, including those with and without KC loss. The results in Tab. 9 show that adding KC loss decreased the real vs synthetic classification accuracy, indicating that the generated images with KC loss have higher perceptual quality and appear more natural to both human viewers and machine algorithms.

Perceptual artifact analysis. Zhang et al. (2023) identified perceptual artifacts in diffusion-generated images, which adversely impact image quality, and developed a dataset and metric (Perceptual Artifacts Ratio, PAR) to automate artifact localization/editing. Our analysis shows that incorporating KC loss reduces these perceptual artifacts (Fig. 10), as evidenced by a decrease in average PAR (Table. 12), demonstrating that KC loss inherently enhances image quality by minimizing artifacts.

Computational cost. The time and space complexity of CFG and PG for SDXL in A5000 machine (single image inference) are presented in Tab. 10. We observe that while PG incurs a bit higher compute overhead than CFG, as both require two forward passes, it achieves greater quality improvement (on Dreambooth (DB) task, DB dataset) with minimal memory overhead due to optimized attention in Diffusers.

Comparison with recent SD model. We ablate KC and PG on SOTA models like SDXL (Tab. 11) and FLUX (Tab. 11) for DB task, observing consistent gains. However, the improvements are more pronounced in pixel-space models like GD Dhariwal & Nichol (2021) (Tab. 3, 4, 5).

Table 10: Complexity

Metrics	w/o guidance	CFG	PG	PG + CFG
GPU mem. (MB)	12167	12218	12200	12500
Inference time (s)	7.2	14.5	14.9	29.3
FID	102.32	95.32	73.33	70.08
MUSIQ	70.18	71.35	72.88	73.02

Table 11: Ablation on DB task, DB dataset for FLUX and SDXL

KC	CFG	PG	FLUX				SDXL			
			FID ↓	MUSIQ ↑	LIQE ↑	Q-align ↑	FID ↓	MUSIQ ↑	LIQE ↑	Q-align ↑
✗	✗	✗	60.13	75.24	7.20	4.37	102.32	70.18	4.32	1.85
✗	✗	✓	55.43	75.98	7.54	4.52	96.17	70.92	4.81	2.01
✗	✓	✗	52.12	76.33	7.83	4.73	95.32	71.35	5.15	2.32
✗	✓	✓	50.42	76.72	7.97	4.96	89.72	71.87	5.37	2.41
✓	✗	✗	45.23	77.50	8.32	5.21	85.11	72.05	5.62	2.52
✓	✓	✗	42.74	77.93	8.61	5.63	80.75	72.32	5.98	2.83
✓	✗	✓	40.21	78.34	8.90	5.82	73.33	72.88	6.35	3.12
✓	✓	✓	**36.96**	**78.85**	**9.13**	**6.04**	**70.08**	**73.02**	**6.82**	**3.55**

Limitations. The proposed PG strategy necessitates two forward passes through the diffusion model to obtain the guidance direction, which is time-consuming. We aim to address this issue in future work.

6 Conclusion

While diffusion models have made significant strides in generating naturalistic images, enhancing image quality remains a key focus. We introduce a novel and generic KC loss, leveraging the KC property of natural images, which minimizes the gap between maximum and minimum kurtosis values across different DWT-filtered versions of the image. Additionally, we propose a condition-agnostic PG strategy to further improve image quality. Our experiments show that KC loss and PG improve image quality in various generative tasks, including personalized few-shot fine-tuning of text-to-image models, unconditional image generation, image super-resolution, and blind face-restoration. Human studies validate the effectiveness of our approach.

Table 12: PAR analysis of tasks - DB, CD, DDPM on Oxford flowers (OF), DDPM on CelebFaces (CF), DDPM on CelebAHQ (CelebHQ), GD on FFHQ, LD on FFHQ has been reported. (in %)

Setting	DB	CD	DDPM (OF)	DDPM (CF)	DDPM (CelebHQ)	GD	LD
w/o KC loss	1.64	0.63	3.02	7.09	3.20	0.89	1.11
w KC loss	**0.75**	**0.36**	**2.99**	**6.97**	**2.63**	**0.51**	**1.07**

References

Matthias Bethge. Factorial coding of natural images: how effective are linear models in removing higher-order dependencies? *JOSA A*, 23(6):1253–1268, 2006.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Richard E Woods and Rafael C Gonzalez. Digital image processing, 2008.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. *arXiv preprint arXiv:2406.02507*, 2024.

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

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

Jonathan Krause, Jia Deng, Michael Stark, and Li Fei-Fei. Collecting a large-scale dataset of fine-grained cars. 2013.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *arXiv preprint arXiv:2212.04488*, 2022.

Wentong Liao, Kai Hu, Michael Ying Yang, and Bodo Rosenhahn. Text to image generation with semantic-spatial aware gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18187–18196, 2022.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *Advances in neural information processing systems*, 31, 2018.

Siwei Lyu and Eero P Simoncelli. Nonlinear extraction of independent components of natural images using radial gaussianization. *Neural computation*, 21(6):1485–1519, 2009.

Siwei Lyu, Xunyu Pan, and Xing Zhang. Exposing region splicing forgeries with blind local noise estimation. *International journal of computer vision*, 110:202–221, 2014.

Andreas Maier. *Lecture Notes in Pattern Recognition: Episode 34 – Measures of Non-Gaussianity.* 2021.

Stephane G Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE transactions on pattern analysis and machine intelligence*, 11(7):674–693, 1989.

Pierre Moulin and Juan Liu. Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors. *IEEE transactions on Information Theory*, 45(3):909–919, 1999.

Ganesh R Naik, Wenwu Wang, et al. Blind source separation. *Berlin: Springer*, 10:978–3, 2014.

M-E Nilsback and Andrew Zisserman. A visual vocabulary for flower classification. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pp. 1447–1454. IEEE, 2006.

Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. Attndreambooth: Towards text-aligned personalized text-to-image generation. *Advances in Neural Information Processing Systems*, 37:39869–39900, 2024.

Tingting Qiao, Jing Zhang, Duanqing Xu, and Dacheng Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1505–1514, 2019.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

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

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

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

Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13960–13969, 2021.

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

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494, 2022.

Karnran Sharifi and Alberto Leon-Garcia. Estimation of shape parameter for generalized gaussian distributions in subband decompositions of video. *IEEE Transactions on Circuits and Systems for Video Technology*, 5 (1):52–56, 1995.

James V Stone. Independent component analysis: an introduction. *Trends in cognitive sciences*, 6(2):59–64, 2002.

Maitreya Suin, Nithin Gopalakrishnan Nair, Chun Pong Lau, Vishal M Patel, and Rama Chellappa. Diffuse and restore: A region-adaptive diffusion model for identity-preserving blind face restoration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6343–6352, 2024.

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

Martin J Wainwright and Eero Simoncelli. Scale mixtures of gaussians and the statistics of natural images. *Advances in neural information processing systems*, 12, 1999.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023.

Guiyu Zhang, Huan-ang Gao, Zijian Jiang, Hao Zhao, and Zhedong Zheng. Ctrl-u: Robust conditional image generation via uncertainty-aware reward modeling. *arXiv preprint arXiv:2410.11236*, 2024.

Lingzhi Zhang, Zhengjie Xu, Connelly Barnes, Yuqian Zhou, Qing Liu, He Zhang, Sohrab Amirghodsi, Zhe Lin, Eli Shechtman, and Jianbo Shi. Perceptual artifacts localization for image synthesis tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7579–7590, 2023.

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

Xing Zhang and Siwei Lyu. Using projection kurtosis concentration of natural images for blind noise covariance matrix estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2870–2876, 2014.

Yuanhan Zhang, ZhenFei Yin, Yidong Li, Guojun Yin, Junjie Yan, Jing Shao, and Ziwei Liu. Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pp. 70–85. Springer, 2020.

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

Daniel Zoran and Yair Weiss. Scale invariance and noise in natural images. In *2009 IEEE 12th International Conference on Computer Vision*, pp. 2209–2216. IEEE, 2009.

A Appendix

In this supplementary material, we will provide the following details.

1. Training details.
2. Theoretical justification.
3. KC loss added as a regularizer.
4. Additional Ablations.
5. Failure cases.
6. Kurtosis analysis.
7. Computational complexity.
8. Convergence analysis.
9. Qualitative analysis.
10. Experiments on image super-resolution.
11. Experiments on other tasks.

B Training Details

The training details of finetuning the diffusion model for various tasks have been provided here. For personalized few-shot finetuning, we consider two methods - Dreambooth Ruiz et al. (2022) and Custom diffusion Kumari et al. (2022). For fair comparison, we applied both the approaches on the dataset and setting introduced by Dreambooth. The dataset contains 30 subjects (e.g., backpack, stuffed animal, dogs, cats, sunglasses, cartoons etc) and 25 prompts including 20 re-contextualization prompts and 5 property modification prompts. DINO, which is the average pairwise cosine similarity between the ViT-S/16 DINO embeddings Caron et al. (2021) of the generated and real images. (2) CLIP-I, i.e., the average pairwise cosine similarity between CLIP Radford et al. (2015) embeddings of the generated and real images. To measure the prompt fidelity, we use CLIP-T, which is the average cosine similarity between prompt and image CLIP embeddings.

For unconditional image generation, we have experimented on oxford flowers, CelebAfaces and CelebAHQ datasets. Image quality has been measured by FID and MUSIQ score.

In case of image super-resolution, we experimented with guided diffusion Dhariwal & Nichol (2021) and latent diffusion Karras et al. (2022) pipelines. We use FFHQ dataset for training, and test on a subset of 1000 images from CelebAHQ test set for x4 super-resolution task. The hyperparameter details are given in Tab. 13.

C Theoretical Justification

Here we provide theoretical analysis of the Lemmas mentioned in the main paper.

Definition 2 *Lipschitz Continuity : A function f is said to be Lipschitz continuous if there exists a constant L (called the Lipschitz constant) such that for all x and y in the domain of f:*

$$|f(x) - f(y)| \leq L|x - y| \tag{13}$$

Definition 3 *Max-Min Difference : Consider the function f :*

$$f(\kappa_1, \kappa_2, \ldots, \kappa_n) = \max(\kappa_i) - \min(\kappa_i) \tag{14}$$

Table 13: Hyperparameters

Hyperparameter	Values
Coefficient of L_{recon}	1
Coefficient of L_{prior}	1
Coefficient of L_{KC}	1
Perceptual guidance scale	1.001
Learning rate	10^{-5}
Batch size (Dreambooth, Custom diffusion)	8
Batch size (DDPM)	125
Batch size (GD)	16
Batch size (LD)	9
Text-to-image diffusion model	Stable Diffusion-v1 Rombach et al. (2022)
Number of class prior images (Dreambooth, Custom diffusion)	10
Number of DWT components	25
DWT filter	Daubechies

Definition 4 *Lipschitz Condition: We need to show that there exists a constant L such that for any two sets of kurtosis values $(\kappa_1, \kappa_2, \ldots, \kappa_n)$ and $(\kappa'_1, \kappa'_2, \ldots, \kappa'_n)$,*

$$|f(\kappa_1, \kappa_2, \ldots, \kappa_n) - f(\kappa'_1, \kappa'_2, \ldots, \kappa'_n)| \leq L \sum_{i=1}^{n} |\kappa_i - \kappa'_i| \tag{15}$$

Lemma 2 *KC loss is differentiable and Lipschitz continuous with Lipschitz constant 2.*

Proof. We are taking maximum and minimum across kurtosis values, therefore KC loss is differentiable and so as the combined loss, since differentiability preserves over addition.

Next, we proof the Lipschitz continuity. Note, the function $\max(\kappa_i)$ is 1-Lipschitz because:

$$|\max(\kappa_i) - \max(\kappa'_i)| \leq \max_i |\kappa_i - \kappa'_i| \tag{16}$$

Similarly, the function $\min(\kappa_i)$ is also 1-Lipschitz because:

$$|\min(\kappa_i) - \min(\kappa'_i)| \leq \max_i |\kappa_i - \kappa'_i| \tag{17}$$

Since both the maximum and minimum functions are 1-Lipschitz, their difference is also Lipschitz continuous with a constant of 2:

$$|f(\kappa_1, \kappa_2, \ldots, \kappa_n) - f(\kappa'_1, \kappa'_2, \ldots, \kappa'_n)| \leq 2 \max_i |\kappa_i - \kappa'_i| \tag{18}$$

For simplicity, if we consider the l_1 norm of the differences, we get:

$$|f(\kappa_1, \kappa_2, \ldots, \kappa_n) - f(\kappa'_1, \kappa'_2, \ldots, \kappa'_n)| \leq 2 \sum_{i=1}^{n} |\kappa_i - \kappa'_i| \tag{19}$$

Thus, we have shown that the KC loss function, i.e., the difference between the maximum and minimum kurtosis values of wavelet-transformed coefficients of natural images is Lipschitz continuous with a Lipschitz constant of 2 when considering the l_1 norm.

Lemma 3 *A Gaussian scale mixture (GSM) vector x with zero mean has the following probability density function:*

$$p(x) = \int_0^\infty \mathcal{N}(x; 0, z\Sigma_x) p_z(z) dz \tag{20}$$

and its projection kurtosis is <u>constant</u> with respect to the projection direction w, i.e.,

$$\kappa(w^T x) = \frac{3var_z\{z\}}{\mathcal{E}_z\{z\}^2} \tag{21}$$

where $\mathcal{E}_z\{z\}$ and $var_z\{z\}$ are the mean and variance of latent variable z respectively.

Proof. Marginal distribution of the projection of x on non-zero vector w is given by Zhang & Lyu (2014),

$$\begin{aligned} p_w(t) &= \int_{x:w^T x=t} p(x)dx \\ &= \int_z p_z(z)dz. \int_{x:w^T x=t} \frac{1}{\sqrt{(2\pi z)^d}|det(\Sigma_x)|} exp(-\frac{x^T \Sigma_x^{-1} x}{2z})dx \\ &= \int_z \mathcal{N}_t(0, zw^T \Sigma_x w) p_z(z)dz \end{aligned}$$

Note that, the last equality holds from the marginalization property of Gaussian, i.e., $X \approx \mathcal{N}(\mu, \Sigma)$, then, $AX \approx \mathcal{N}(A\mu, A\Sigma A^T)$.

The variance of $w^T x$,

$$\begin{aligned} \mathcal{E}_t\{t^2\} &= \int_z p_z dz \int_t t^2 \mathcal{N}_t(0, zw^T \Sigma_x w)dz \\ &= w^T \Sigma_x w \int_z z p_z dz \\ &= w^T \Sigma_x w \mathcal{E}_z\{z\} \end{aligned}$$

The fourth order moment of $w^T x$,

$$\begin{aligned} \mathcal{E}_t\{t^4\} &= \int_z p_z dz \int_t t^4 \mathcal{N}_t(0, zw^T \Sigma_x w)dz \\ &= 3(w^T \Sigma_x w)^2 \int_z z^2 p_z dz \\ &= 3(w^T \Sigma_x w)^2 \mathcal{E}_z\{z^2\} \end{aligned}$$

We utilize the property that $\mathcal{N}_t(0, \sigma^2)$ has a fourth order moment of $3\sigma^4$.

Finally, the kurtosis becomes,

$$\begin{aligned} \kappa(w^T x) &= \frac{\mathcal{E}_t\{t\}^4}{\mathcal{E}_t\{t\}^2} - 3 \\ &= \frac{3\mathcal{E}_z\{z\}^2}{\mathcal{E}_z\{z\}^2} - 3 \\ &= \frac{3(\mathcal{E}_z\{z^2\} - \mathcal{E}_z\{z\}^2)}{\mathcal{E}_z\{z\}^2} \\ &= \frac{3var_z\{z\}}{\mathcal{E}_z\{z\}^2} \end{aligned}$$

D KC Loss Added as a Regularizer

We would like to highlight that in our work, the underlying theoretical framework behind the forward and reverse diffusion processes remains unchanged; rather, we focus on improving the performance of the denoising neural network used to approximate the reverse diffusion trajectory.

Suppose, we have the input training images (x) and conditioning vector c. The conditioning vector could be text (text-to-image model), image (image-to-image model), or none (in case of the unconditional diffusion model). In the forward process, the noisy versions of image x at timestep t is generated as $x_t = \alpha_t x + \sigma_t \epsilon$, where $\epsilon \sim N(0, I)$.

In the reverse process, a denoised autoencoder (f_θ) is trained to predict the denoised version of the image ($x_{t,gen}$) at each timestep t from the noisy images x_t, i.e., $x_{t,gen} = f_\theta(x_t, c, t)$. Typically, the denoised autoencoder (f_θ) is trained by minimzing the Mean Squared Error between the real image (x) and the generated denoised version of the image at time step t ($x_{t,gen}$) averaged over timesteps and noise variances as denoted by,

$$L_{recon} = \mathbb{E}_{x,c,\epsilon,t}[\ ||x_{t,gen} - x||_2^2] \tag{22}$$

The kurtosis concentration loss is applied on the generated images ($x_{t,gen}$), and therefore can be considered as a function (f') of x_{gen} as follows:

$$L_{KC} = \mathbb{E}_{x,c,\epsilon,t}[f'(x_{t,gen})] \tag{23}$$

Note the function f' is difference between the maximum and minimum values of the DWT filtered version of input $x_{t,gen}$.

Therefore, the total loss function can be written as,

$$L_{total} = \mathbb{E}_{x,c,\epsilon,t}[\ ||x_{t,gen} - x||_2^2] + \mathbb{E}_{x,c,\epsilon,t}[f'(x_{t,gen})] \tag{24}$$

In our work, the above-mentioned framework remains the same. Instead, the proposed KC loss acts as an additional regularizer to the training of the denoising neural network, which helps it to denoise x_t better (Lemma 2, main paper), ultimately improving the approximation of x, i.e., $x_{t,gen}$ at each time step t.

E Additional Ablations

In Fig. 11, we visualize some of the DiffNat generated images using various text-prompts. The generated images capture the context of the text-prompt and also retain naturalness. We have also provided qualitative comparison w.r.t Dreambooth in Fig. 12.

We also provide ablations for using DCT transforms and analyse the performance with respect to other tasks and methods. Experiments in Table. 14 suggests DWT performs better than DCT for different methods across datasets.

F Failure Cases

We also present some of the failure cases of DiffNat in Fig. 13. E.g., our model fails to generate images of "A [V] berry bowl with the Eiffel Tower in the background", but actually generates images with "the Eiffel Tower" in the berry bowl. Similarly, the model fails to generate "A cube shaped [V] can", since these object do not appear in the training set. The model also fails to generate "A [V] cat on top of a purple rug in a forest" and instead generated some version of purple cat.

G Kurtosis Analysis

To verify the efficacy of the proposed KC loss, we perform average kurtosis analysis in this section. we compute the average kurtosis deviation of DWT filtered version of images from the dataset and plot them in Fig. 15, Fig. 16 and Fig. 17. E.g., in case of dreambooth task, we compute the kurtosis statistics of bandpass filtered version of natural images from Dreambooth dataset, images generated by Dreambooth and images generated by DiffNat (i.e., adding KC loss) and plot it in Fig. 15. We observe that the Dreambooth generated images (Fig. 15 (a)) have highest kurtosis deviation. The average deviation is least for natural images (Fig. 15 (c)) and adding KC loss reduces the kurtosis deviation (Fig. 15 (b)). Similar trends can be observed for

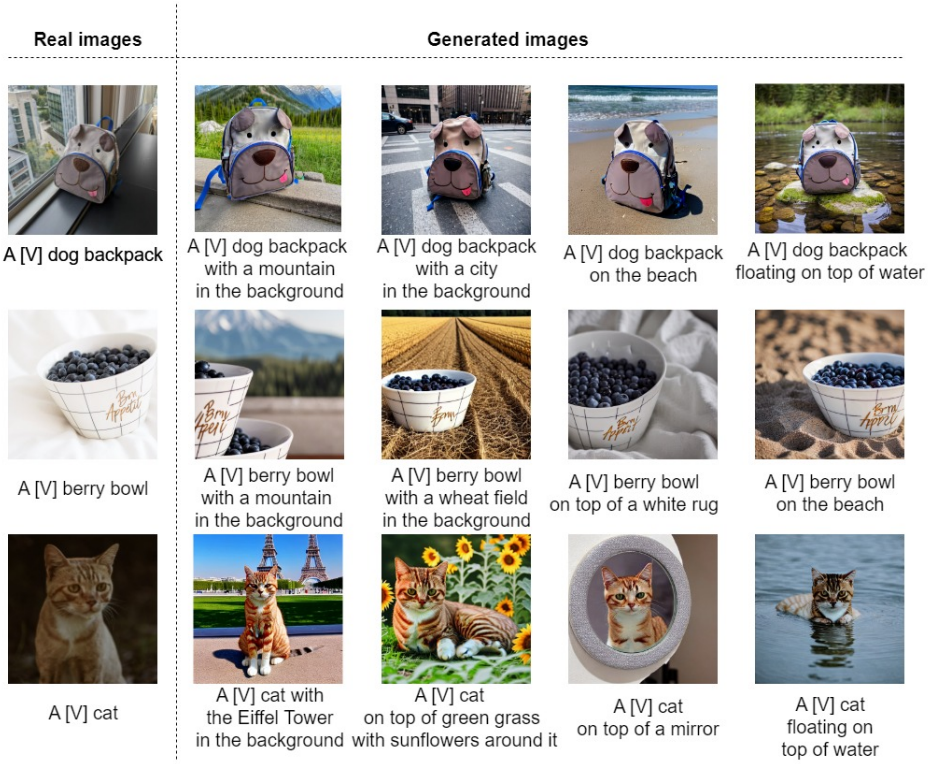


Figure 11: DiffNat generated images. The task is to learn a unique identifier ("A [V] dog backpack") of the training images and generate variations w.r.t. background, lighting conditions etc. The generated images look natural in different background context, e.g., "A [V] dog backpack on the beach/ with a city in the background etc". The generated images are of high quality.

Table 14: Comparison of DCT vs DWT

Method	**FID score** ↓	**MUSIQ score** ↑
DB (Dreambooth dataset)	111.76	68.31
DB + KC (DCT)	106.23	68.72
DB + KC (DWT)	**100.08**	**69.78**
CD (Dreambooth dataset)	84.65	70.15
CD + KC (DCT)	80.33	70.67
CD + KC (DWT)	**75.68**	**72.22**
DDPM (Oxford flowers)	243.43	20.67
DDPM + KC (DCT)	240.12	20.98
DDPM + KC (DWT)	**237.73**	**21.13**
GD (FFHQ)	121.23	57.31
GD + KC (DCT)	112.66	58.12
GD + KC (DWT)	**103.19**	**58.69**
LD (FFHQ)	95.83	59.57
LD + KC (DCT)	88.52	60.37
LD + KC (DWT)	**83.34**	**61.20**

Figure 12: Comparison of DreamBooth and DiffNat. DiffNat generated images have better visual quality.

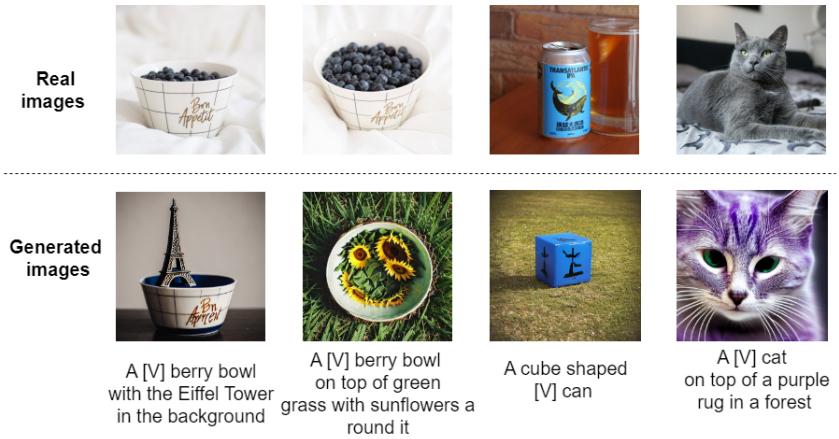


Figure 13: Failure cases of DiffNat. Instead of generating "A [V] berry bowl with the Eiffel Tower in the background", our method generates image with the Eiffel Tower in the berry bowl. Also, while generating "A [V] cat on top of a purple rug in a forest", it generates a purple [V] cat, which shows the color bias w.r.t the text-prompt of the model.

DDPM (Fig. 16), guided diffusion (Fig. 17) as well. Adding KC loss improves image quality has been verified both qualitatively and quantitatively in the paper. This analysis verifies minimizing kurtosis loss improves diffusion image quality.

H Computational Complexity

Here we analyze the computational complexity of the proposed KC loss. Suppose, given a batch of N images. We need to perform DWT of each images using k different filters. Since, DWT for 'Haar'/'Daubechis' wavelet can be done in linear time, the complexity of performing DWT with k filters can be done in $\mathcal{O}(Nk)$ time. Now, calculating the difference between maximum and minimum kurtosis can be done in linear time, therefore, the computational complexity of calculating KC loss is $\mathcal{O}(Nk)$. This minimal overhead of computing KC loss can be observed in the training time analysis provided next. The run time analysis has been provided in Table. 15. Note that the experiments for Dreambooth, Custom diffusion, DDPM have been performed on a single A5000 machine with 24GB GPU. We have performed guided diffusion (GD) and latent diffusion (LD) experiments on a server of 8 24GB A5000 GPUs. The experimental results in Table. 15 show that incorporating KC loss induces minimum training overhead.

Table 15: Training time analysis

Method	**dataset**	**Training time**
DreamBooth Ruiz et al. (2022)	5-shot finetuning	10 min 21s
DreamBooth Ruiz et al. (2022) + KC loss	5-shot finetuning	11 min 30s
Custom Diffusion Kumari et al. (2022)	5-shot finetuning	6m 43s
Custom Diffusion Kumari et al. (2022) + KC loss	5-shot finetuning	7m 11s
DDPM Ho et al. (2020)	CelebAfaces	2d 8h 21m
DDPM Ho et al. (2020) + KC loss	CelebAfaces	2d 9h 19m
GD Dhariwal & Nichol (2021)	FFHQ	23h 10m
GD Dhariwal & Nichol (2021) + KC loss	FFHQ	1d 1h 29m
LD Karras et al. (2022)	FFHQ	20h 15m
LD Karras et al. (2022) + KC loss	FFHQ	22h 40m

I Convergence Analysis

The main idea of the diffusion model is to train a UNet, which learns to denoise from a random noise to a specific image distribution. More denoising steps ensure a better denoised version of the image, e.g., DDPM Ho et al. (2020), LDM Karras et al. (2022). In proposition 1 (main paper), we show that minimizing projection kurtosis further denoise input signals. Therefore, KC loss helps in the denoising process and improves the convergence speed. We have shown that adding KC loss improves the loss to converge faster for Dreambooth task in Fig. 14.

J Qualitative Analysis

In this section, we provide more qualitative analysis to show that adding KC loss improves image quality. Zoomed view of the generated images are shown to compare w.r.t the baselines in Fig. 19, Fig. 20, Fig. 21, Fig. 22, Fig. 23, Fig. 24, Fig. 25, Fig. 26. Details are provided in the caption.

K Experiments on Image Super-resolution

In this section, we provide more experimental results for image super-resolution task. This includes quantitative results and human evaluation.

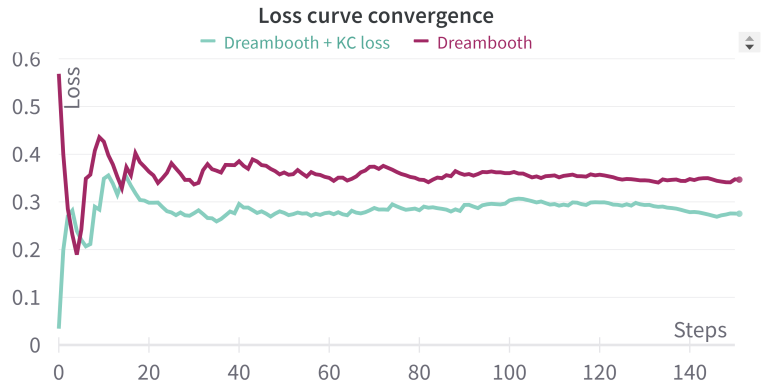
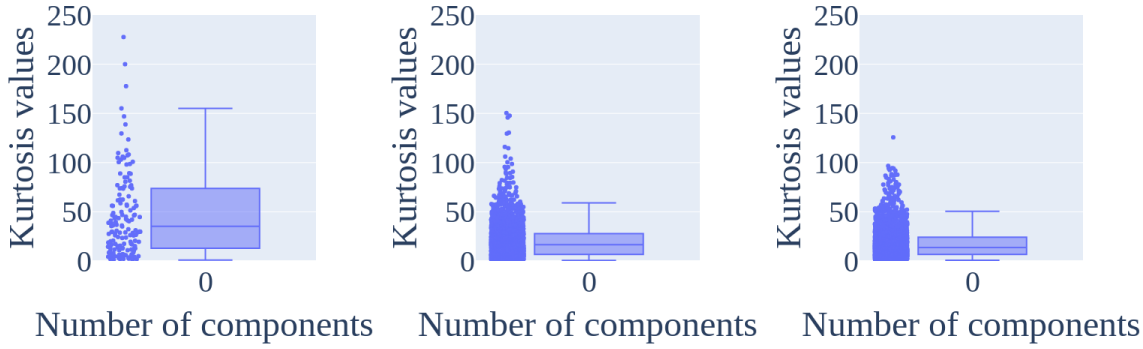


Figure 14: Loss curve convergence of Dreambooth.

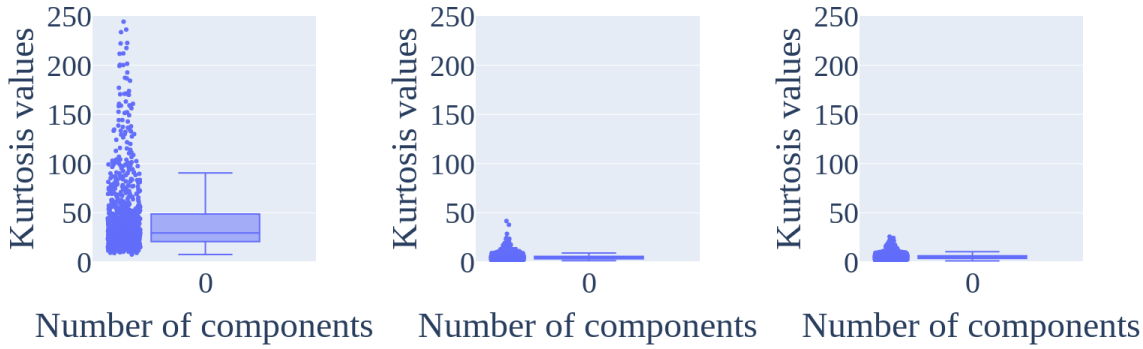


(a) Average kurtosis of Dreambooth images

(b) Avg. kurtosis (trained with Dreambooth + KC loss)

(c) Average kurtosis of Natural images

Figure 15: Average kurtosis analysis of Dreambooth, DiffNat and natural images over the dataset used in Dreambooth. From this analysis, it is evident that Dreambooth generated images have higher kurtosis deviation. Integrating KC loss reduces the kurtosis deviation to preserve the naturalness of the generated images. Natural images have more concentrated kurtosis values.

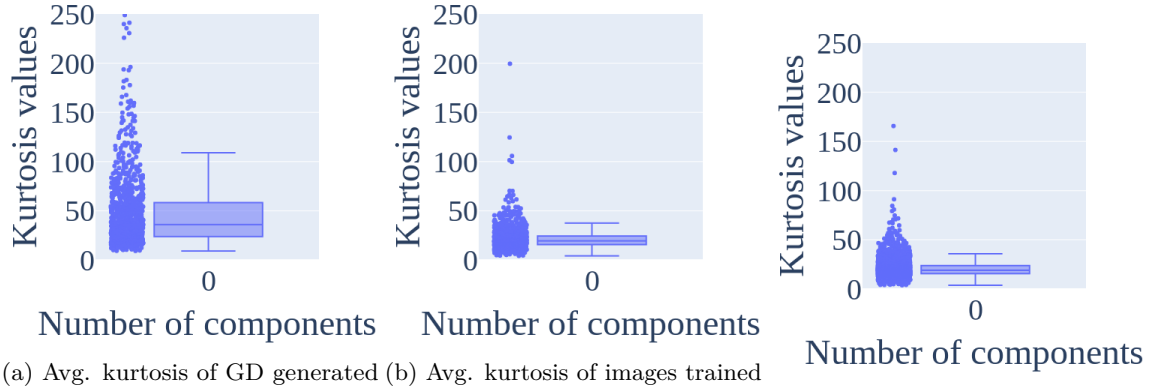


(a) Average kurtosis of DDPM images

(b) Average kurtosis of images trained with DDPM + KC loss

(c) Average kurtosis of Natural images

Figure 16: Average kurtosis analysis of DDPM framework trained on Oxford flowers dataset. From this analysis, it is evident that DDPM generated images have higher kurtosis deviation. Integrating KC loss reduces the kurtosis deviation to preserve the naturalness of the generated images. Natural images have more concentrated kurtosis values.



(a) Avg. kurtosis of GD generated images

(b) Avg. kurtosis of images trained with GD + KC loss

(c) Avg. kurtosis of Natural images

Figure 17: Average kurtosis analysis of guided diffusion (GD) framework trained on FFHQ dataset. From this analysis, it is evident that GD generated images have higher kurtosis deviation. Integrating KC loss reduces the kurtosis deviation to preserve the naturalness of the generated images. Natural images have more concentrated kurtosis values.

Table 16: Comparison of image super-resolution (x2) task

Method	Image quality			
	FID score ↓	**PSNR ↑**	**SSIM ↑**	**MUSIQ score ↑**
GD Dhariwal & Nichol (2021)	100.2	19.4	0.62	58.12
GD + KC loss(Ours)	**80.9**	**20.2**	**0.66**	**59.91**
LD. Karras et al. (2022)	82.45	21.2	0.64	60.23
LD + KC loss(Ours)	**70.12**	**22.3**	**0.70**	**62.15**

K.1 Quantitative Results

In addition to the super resolution task (x4) shown in the main paper, we conduct experiments for x2 and x8 tasks as well in the same setting. The ground-truth images are of size 256 X 256. Therefore, x2 task performs image super-resolution from 128 X 128 → 256 X 256 and x8 task performs image super-resolution from 32 X 32 → 256 X 256 and the corresponding experiments are shown in Table 16 and Table 17 respectively. For training, we use standard FFHQ dataset Karras et al. (2017), and evaluation is performed on CelebA-Test dataset Karras et al. (2017). We observe that adding KC loss improves image quality quantitatively both for guided diffusion (GD) and latent diffusion (LD). Qualitative results are shown in Fig. 23, Fig. 24, Fig. 25 and Fig. 26. Next, we also perform human study to validate our approach.

K.2 Human Evaluation

We conduct human evaluation of image super-resolution task to compare guided diffusion (GD)/ latent diffusion (LD) and adding KC loss to the corresponding counterpart (DiffNat). We provide 20 examples of natural images and corresponding generated images using GD, LD and our method DiffNat (i.e., adding KC loss) and asked the following question to amazon mechanical turks: "which of the generated images is of best visual quality considering factors include image quality and preserving the identity of the original image?" Similar to Dreambooth task, we evaluate this by 50 users, totalling 1000 questionnaires. The available options are { 'DiffNat', 'GD/LD', 'None is satisfactory' }. The aggregate response shows that DiffNat generated images are of better image quality compared to the baselines, as shown in Fig. 27. Therefore, we verified the improved image quality quantitatively, qualitatively and through human evaluation as well. Note that, human evaluation is not applicable for unconditional image generation task since there is no one-to-one correspondence between the training images and the generated images. It will be ambiguous for the human observers to compare quality between approaches. Therefore, we abstain ourselves from performing human evaluation for this task. However, the quantitative and qualitative analysis exhibit the efficacy of our approach.

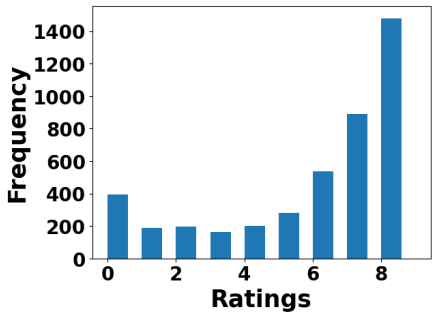


Figure 18: Subject fidelity assessment by user study. The ratings ranges from "0" being "extremely unlikely" to 10 being "extremely likely". We observe from the plot that most of the users find DiffNat preserves subject fidelity. The average rating is 5.8, which is "moderately likely" to "highly likely".

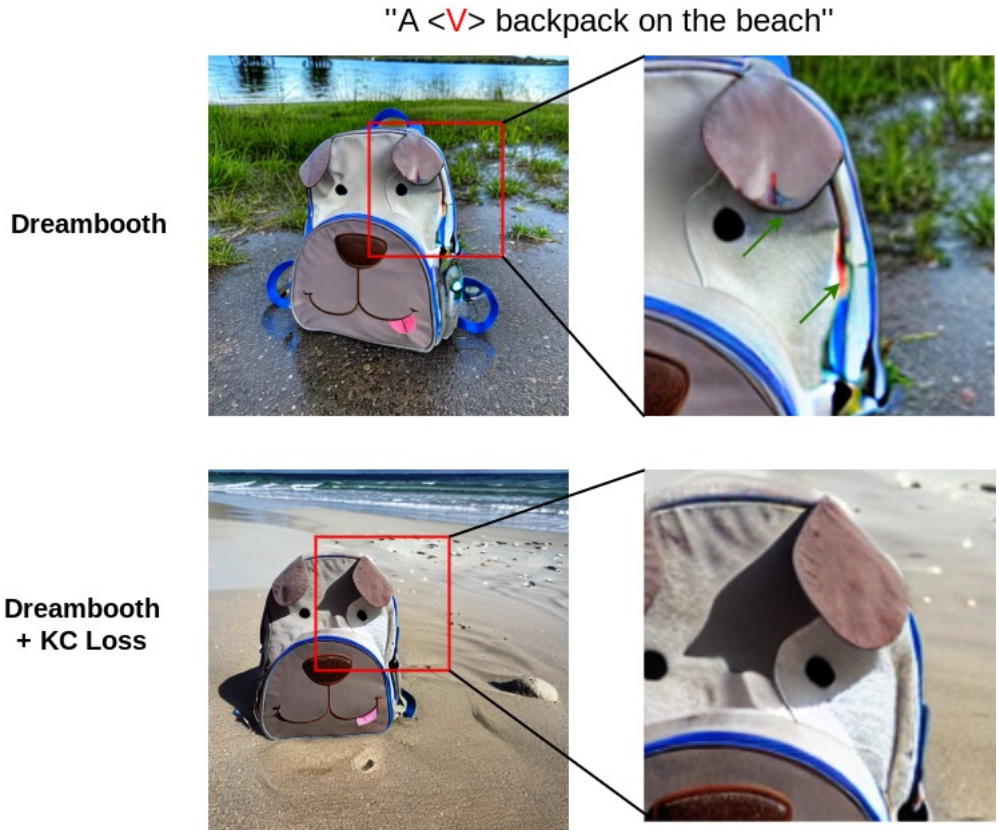


Figure 19: Qualitative comparison of with/without KC loss in Dreambooth. The bottom image (with KC loss) shows better image quality and shadows (best viewed in color).

Table 17: Comparison of image super-resolution (x8) task

Method	Image quality			
	FID score ↓	**PSNR ↑**	**SSIM ↑**	**MUSIQ score ↑**
GD Dhariwal & Nichol (2021)	140.3	17.5	0.52	55.26
GD + KC loss(Ours)	**125.5**	**18.7**	**0.56**	**57.33**
LD. Karras et al. (2022)	103.2	18.7	0.59	58.62
LD + KC loss(Ours)	**80.1**	**19.5**	**0.67**	**60.31**

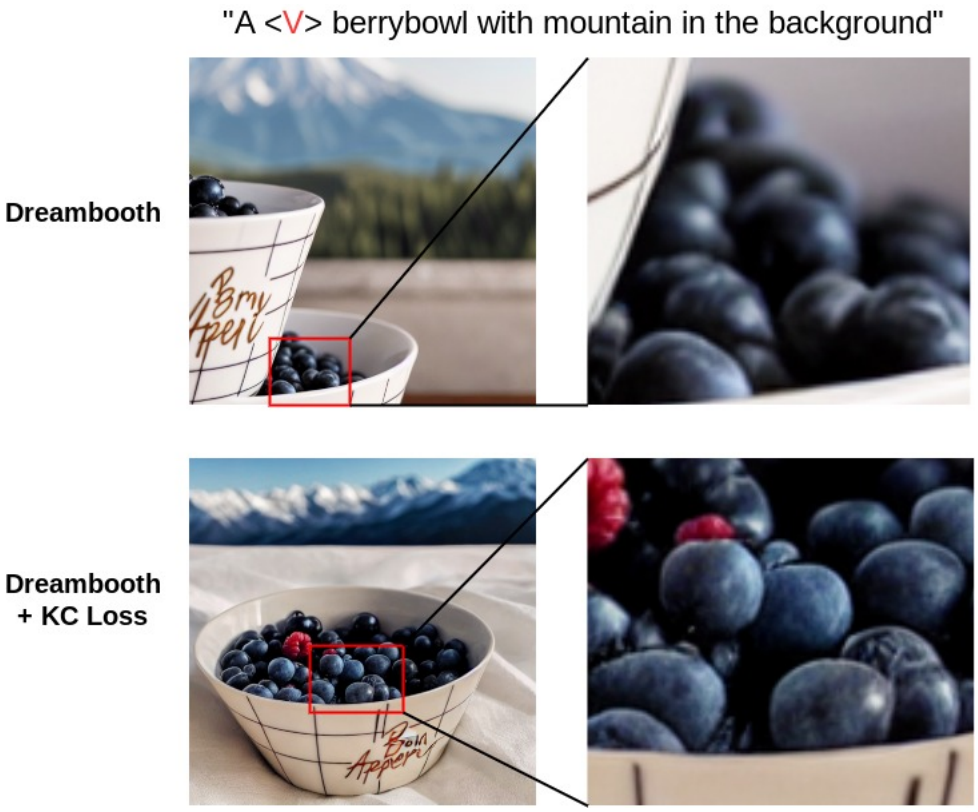


Figure 20: Qualitative comparison of with/without KC loss in Dreambooth. The bottom image (with KC loss) shows better image quality and reflections on the bowl full of berries (best viewed in color).

Figure 21: Qualitative comparison of with/without KC loss in Custom diffusion. The bottom image (with KC loss) shows better image quality in terms of color vividness and contrast (best viewed in color).

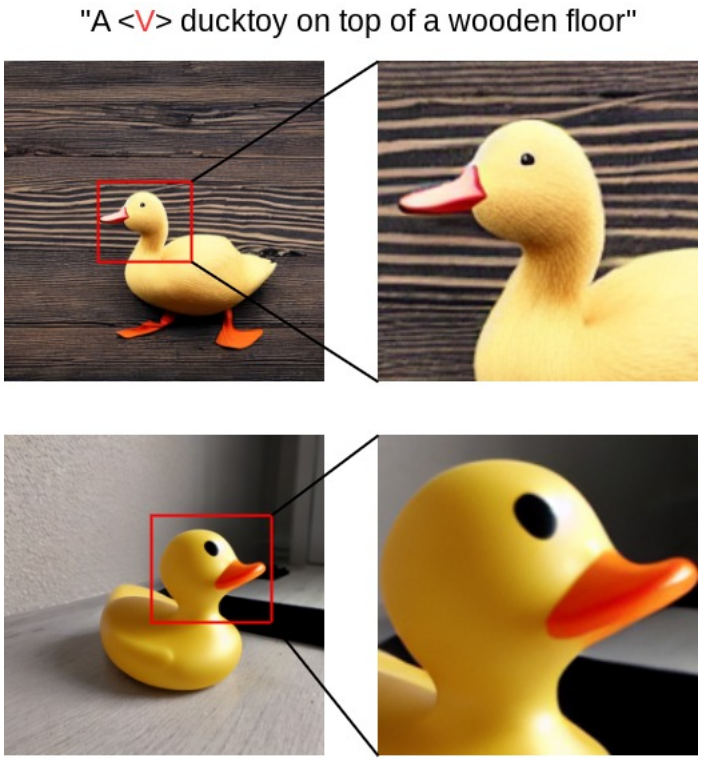


Figure 22: Qualitative comparison of with/without KC loss in Custom diffusion. The bottom image (with KC loss) shows better image quality in terms of detail and smoothness (best viewed in color).

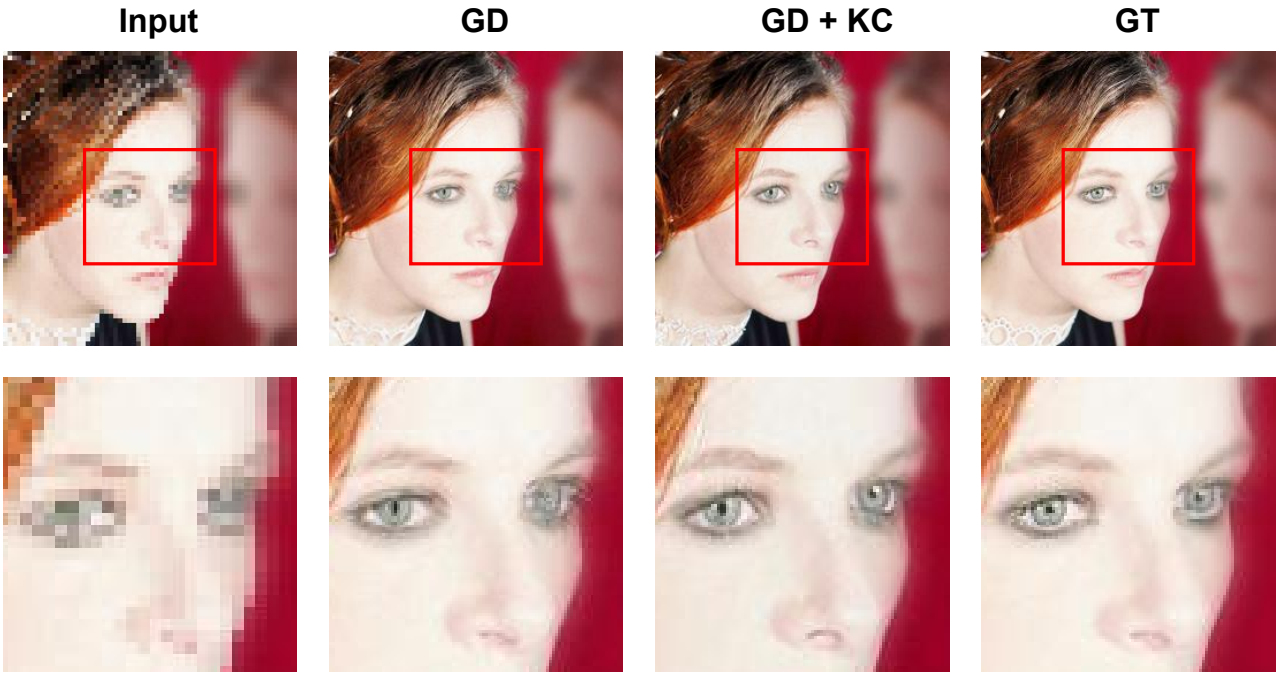


Figure 23: Qualitative comparison of with/without KC loss in guided diffusion (GD). The bottom image (with KC loss) has better eye and hair details (best viewed in color).

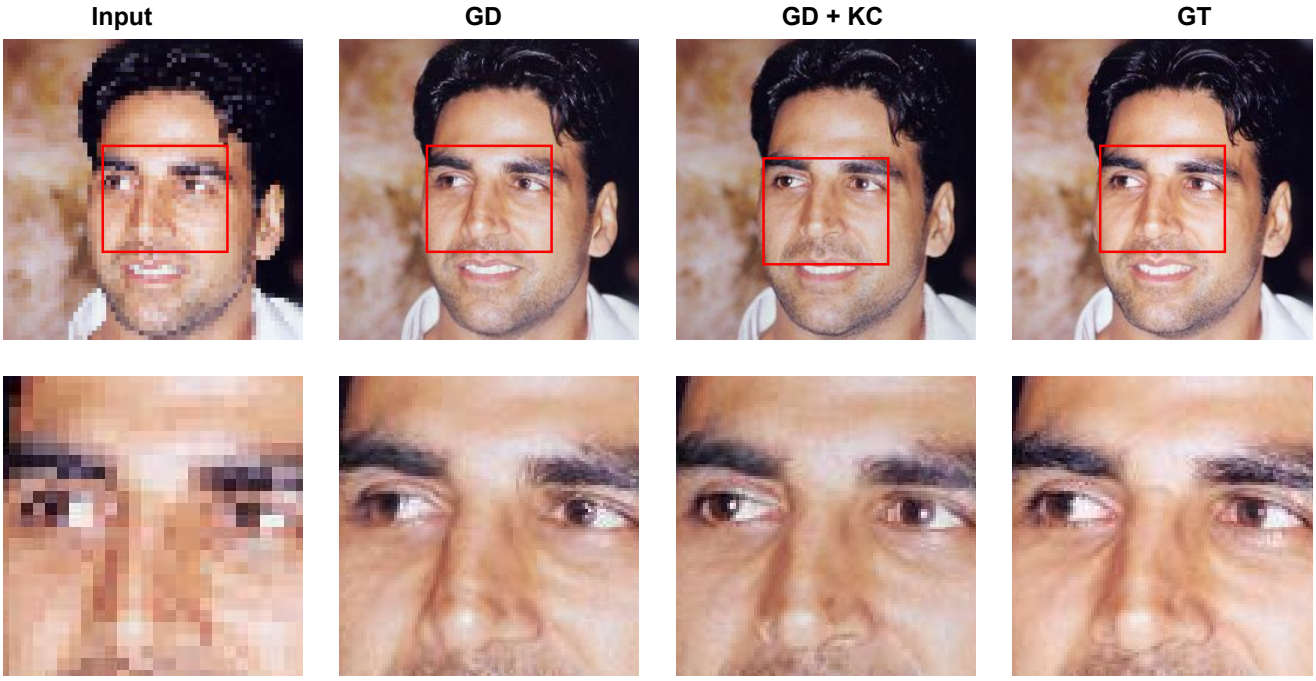


Figure 24: Qualitative comparison of with/without KC loss in guided diffusion (GD). The bottom image (with KC loss) has better eye details and skin smoothness (best viewed in color).

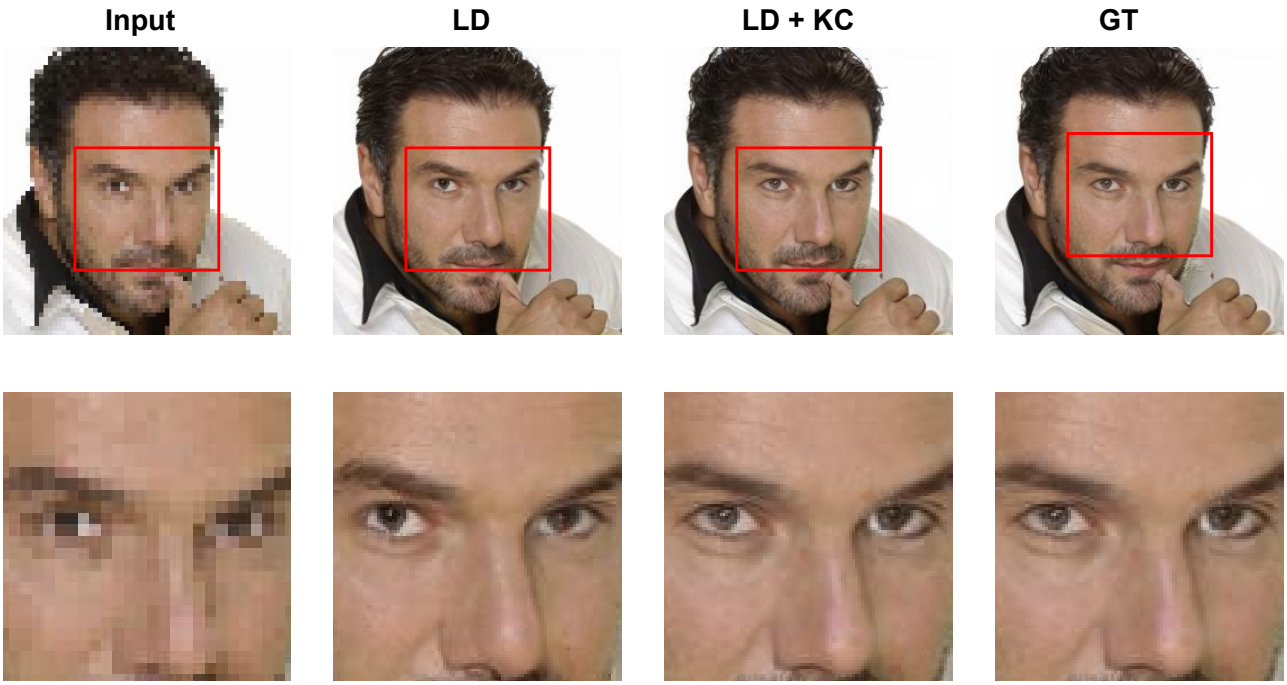


Figure 25: Qualitative comparison of with/without KC loss in Latent diffusion (LD). The bottom image (with KC loss) has higher similarity w.r.t the ground truth in terms of left eye and skin color (best viewed in color).

Figure 26: Qualitative comparison of with/without KC loss in Latent diffusion (LD). The bottom image (with KC loss) has higher similarity w.r.t the ground truth in terms of left eye and skin color (best viewed in color).

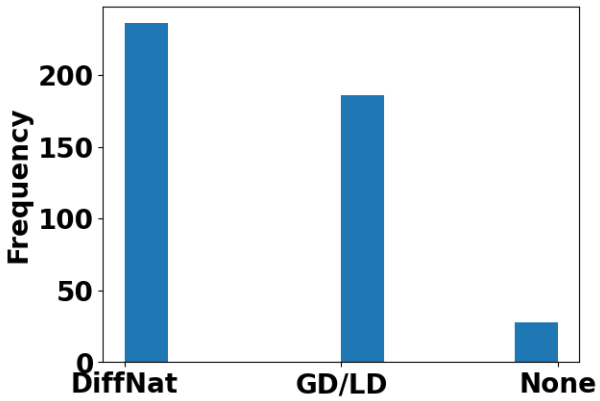


Figure 27: Human evaluation for image super-resolution task. DiffNat performs better than guided diffusion (GD), latent diffusion (LD) in user study as well.

Table 18: Blind Face restoration

Method	**LPIPS-div↑**	**FID↓**	**IDS ↑**	**PSNR ↑**	**SSIM ↑**
DifFace	0.20	70.69	0.48	22.82	0.61
RestoreFormer	0.29	60.98	0.39	21.77	0.53
IPC Suin et al. (2024)	0.32	55.42	0.54	22.34	0.60
IPC + KC	**0.34**	**43.21**	**0.61**	**24.19**	**0.64**

L Experiments on other Tasks

We analyze the effectiveness of KC loss on other visual recognition tasks, e.g., inverse problem like blind face restoration.

L.1 Blind Face Restoration

For blind face restoration task Suin et al. (2024), we train on FFHQ dataset with and without KC loss on IPC baseline Suin et al. (2024), and evaluate on 3000 images on Celeb-A test set with a resolution of 256x256. Average LPIPS, FID, IDS, PSNR, SSIM are reported in Tab. 18. Qualitative results (Fig. 28) also verify that adding KC loss improves image quality.

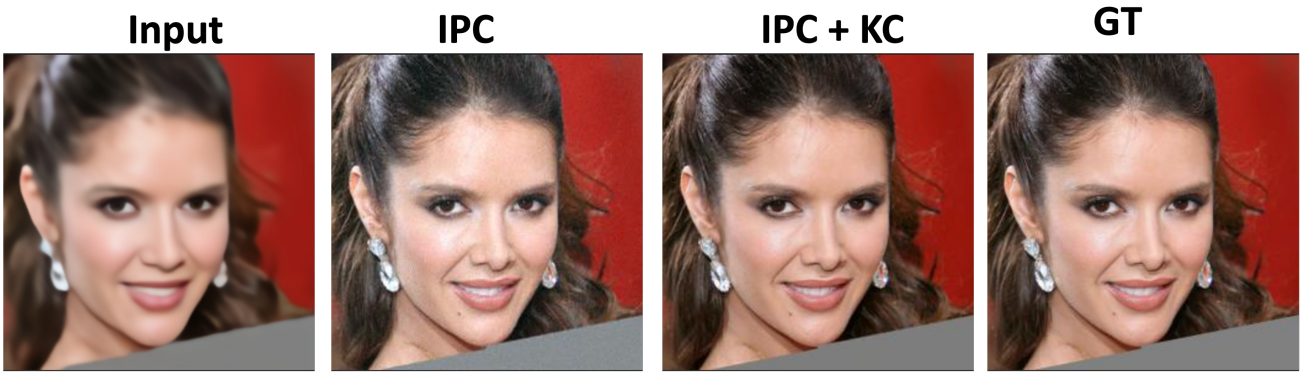


Figure 28: Restoration comparison with IPC

L.2 Conditional Image Generation with Structural Constraints.

We also experiment with more challenging conditional image generation with structural constraint. In this context, we experiment with Crtl-U Zhang et al. (2024), to verify the efficacy of KC loss and PG strategy. Ctrl-U Zhang et al. (2024) proposes uncertainty-aware reward modeling for conditional image generation. It estimates reward uncertainty by running two forward passes of the diffusion model at different timesteps. The framework uses this uncertainty to adaptively regularize a pixel-wise consistency loss. They have expreimented on ADE20K, COCO-Stuff datasets for segmentation, edge, depth conditions.

We integrate KC loss additionally to the existing losses of Ctrl-U for training and use PG strategy during inference for segmentation mask as condition. We experiment on ADE20K and COCO-stuff dataset and observe improvements in both segmentation performance and image quality. We use the same setting used by Crtl-U. Experimental results are shown in Tab. 19 and Fig. 29, Fig. 30.

L.3 Comparison with AttnDreambooth

For the personalized few-shot finetuning task, we also compare with a recent baseline - AttnDreamBooth (AttnDB) Pang et al. (2024). It introduces a novel method to enhance personalized text-to-image synthesis by addressing limitations in existing approaches like Textual Inversion and DreamBooth, which often suffer from overfitting or neglecting the target concept. AttnDreamBooth decomposes the personalization process into three distinct training stages: (1) learning embedding alignment to integrate new concepts into the model's vocabulary, (2) refining the attention map through fine-tuning cross-attention layers with a regularization term that encourages similarity between the attention maps of the new concept and its super-category, and (3) acquiring the subject identity by fine-tuning the entire U-Net. We add both KC loss and PG strategy to this, and obtain performance improvement both in image quality, and diversity as shown in Tab. 20. Qualitative results are shown in Fig. 31.

Table 19: Performance Comparison w.r.t Ctrl-U on ADE20K and COCO-stuff Datasets

Method	ADE20K			COCO-stuff		
	mIoU ↑	FID ↓	CLIPscore ↑	mIoU ↑	FID ↓	CLIPscore ↑
Ctrl-U (ICLR'25) Zhang et al. (2024)	46.49	28.01	32.26	49.91	25.79	31.23
Ctrl-U + KC	46.72	24.32	29.32	50.12	14.08	31.75
Ctrl-U + KC + PG	**46.91**	**20.12**	**30.12**	**50.47**	**11.06**	**32.18**

Table 20: Comparison on personalized few-shot finetuning task

Method	**Image quality**			**Subject fidelity**		**Prompt fidelity**	**Image diversity**
	FID ↓	**MUSIQ ↑**	**HPSv2 ↑**	**DINO ↑**	**CLIP-I ↑**	**CLIP-T ↑**	**LPIPS-div ↑**
DB Ruiz et al. (2022)	111.76	68.31	25.12	0.65	0.81	0.31	0.38
DB Ruiz et al. (2022) + LPIPS	108.23	68.39	25.43	0.65	0.80	0.32	0.40
DB + KC loss (Ours)	100.08	69.78	25.82	0.68	0.84	0.34	0.42
DB + KC loss + PG (Ours)	**93.45**	**70.82**	**26.04**	**0.70**	**0.86**	**0.35**	**0.43**
CD Kumari et al. (2022)	84.65	70.15	26.12	0.71	0.87	0.38	0.40
CD Kumari et al. (2022) + LPIPS	80.12	70.56	26.33	0.71	0.87	0.37	0.43
CD + KC loss (Ours)	75.68	72.22	26.64	0.73	0.88	0.40	0.44
CD + KC loss + PG (Ours)	**66.27**	**73.77**	**27.10**	**0.77**	**0.89**	**0.43**	**0.46**
AttnDB Pang et al. (2024)	80.59	70.23	26.42	0.72	0.85	0.35	0.41
AttnDB Pang et al. (2024) + LPIPS	80.06	70.50	26.66	0.73	0.87	0.35	0.43
AttnDB + KC loss (Ours)	73.02	71.10	26.83	0.75	0.89	0.37	0.45
AttnDB + KC loss + PG (Ours)	**64.78**	**72.89**	**27.35**	**0.78**	**0.90**	**0.38**	**0.47**

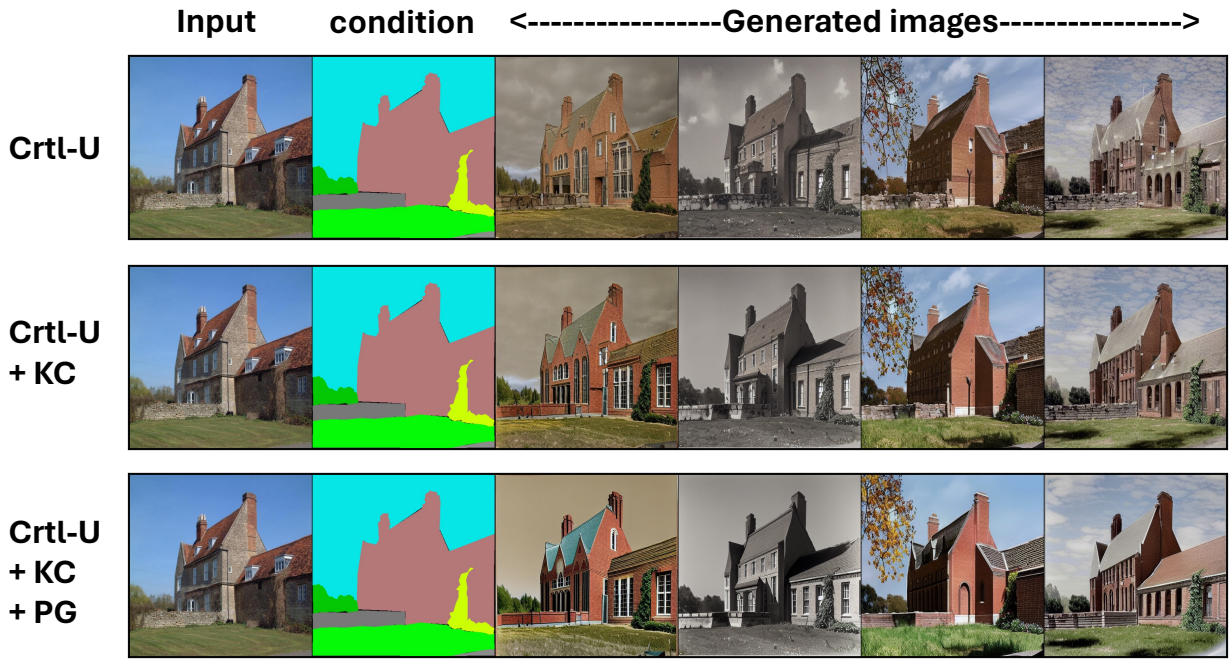


Figure 29: Qualitative comparison with Ctrl-U

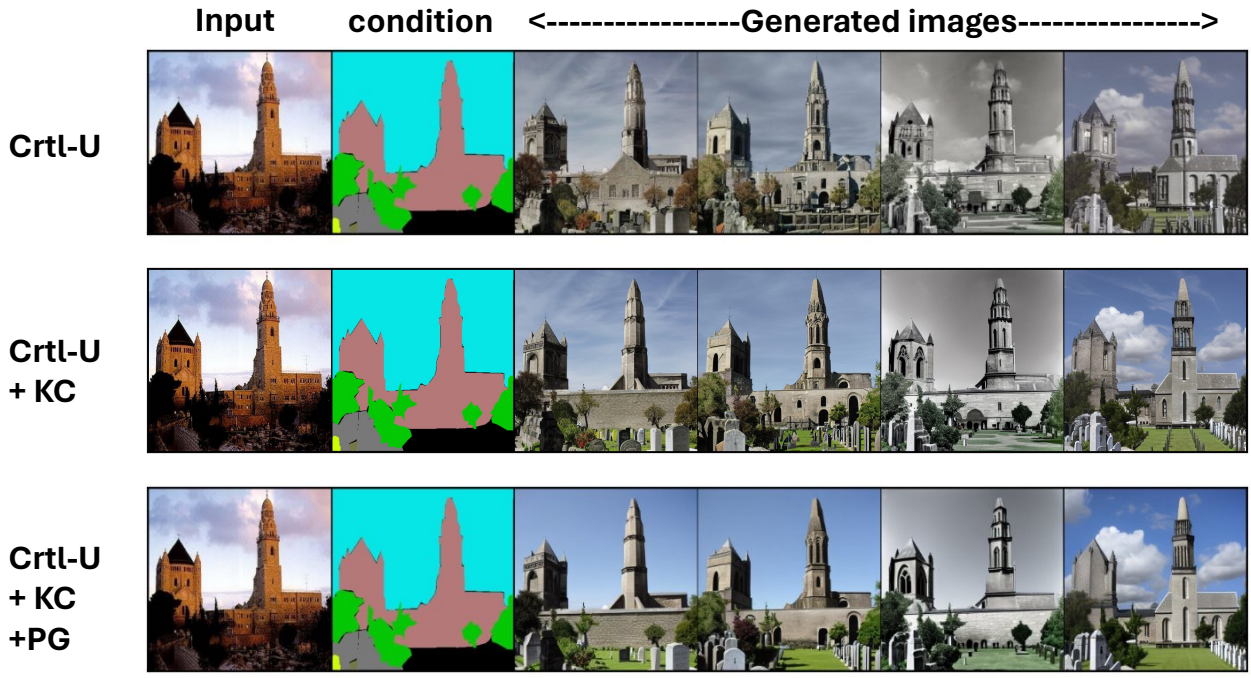


Figure 30: Qualitative comparison with Ctrl-U

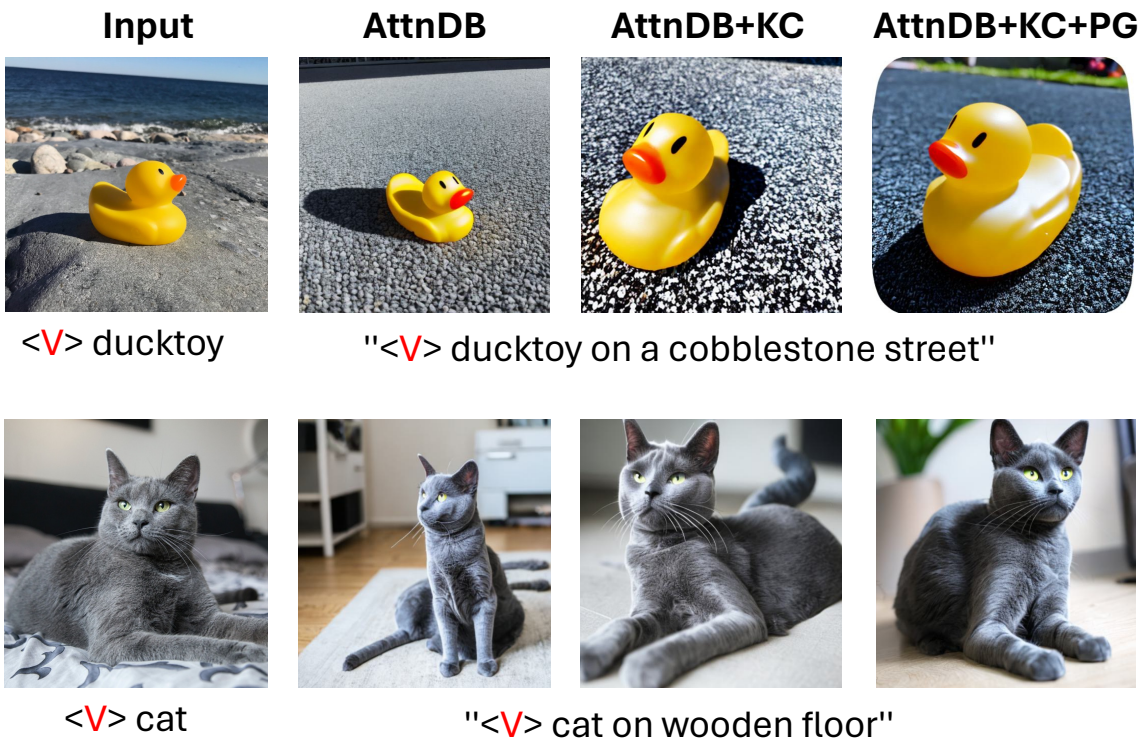


Figure 31: Qualitative comparison with AttnDreamBooth