ImageRAGTurbo: Towards One-step Text-to-Image Generation with Retrieval-Augmented Diffusion Models

Peijie Qiu[1]*, Hariharan Ramshankar[2], Arnau Ramisa[2], Amit Kumar K C[2],
René Vidal[2], Vamsi Salaka[2], Rahul Bhagat[2]
[1]Washington University in St. Louis, [2]Amazon Core Search
peijie,qiu@wustl.edu {dsthari, aramisay, amitkrkc, rvidal, vsalaka, rbhagat}@amazon.com

Abstract

Diffusion models have emerged as the leading approach for text-to-image generation. However, their iterative sampling process, which gradually morphs random noise into coherent images, introduces significant latency that limits their applicability. While recent few-step diffusion models reduce the number of sampling steps to as few as one to four steps, they often compromise image quality and prompt alignment, especially in one-step generation. Additionally, these models require computationally expensive training procedures. To address these limitations, we propose ImageRAGTurbo, a novel approach to efficiently finetune few-step diffusion models via retrieval augmentation. Given a text prompt, we retrieve relevant text-image pairs from a database and use them to condition the generation process. We argue that such retrieved examples provide rich contextual information to the UNet denoiser that helps reduce the number of denoising steps without compromising image quality. Indeed, our initial investigations show that using the retrieved content to edit the denoiser's latent space ($\mathcal{H}$-space) without additional finetuning already improves prompt fidelity. To further improve the quality of the generated images, we augment the UNet denoiser with a trainable adapter in the $\mathcal{H}$-space, which efficiently blends the retrieved content with the target prompt using a cross-attention mechanism. Experimental results on fast text-to-image generation demonstrate that our approach produces high-fidelity images without compromising latency compared to existing methods.

1. Introduction

Diffusion models [14, 39–41] have demonstrated remarkable capabilities in generating high-quality images. However, their iterative sampling process introduces a significant computational bottleneck that limits their practical applications. In particular, standard diffusion models typically require $25 - 50$ denoising steps to generate a single image, with each step involving an expensive network function evaluation of the denoiser. This sequential sampling process leads to high latency, limiting the use of these diffusion models in real-time and interactive applications.

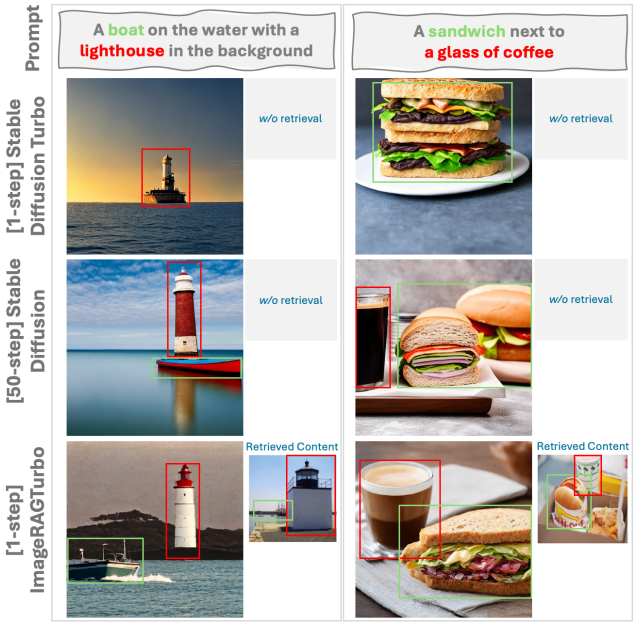


Figure 1. Illustrative examples of ImageRAGTurbo, where the visual concepts are highlighted by colored boxes. ImageRAGTurbo outperforms Stable Diffusion Turbo (adversarially distilled Stable Diffusion without retrieval) in generating accurate visual concepts. Even with a single step, ImageRAGTurbo performs comparably to Stable Diffusion with 50 steps. Please zoom in for better quality.

Classical approaches to improving the efficiency of diffusion models include (1) latent diffusion models that operate in a lower-resolution latent space to gain efficiency in the spatial dimension [40], and (2) fast ordinary differential equation (ODE) samplers that directly reduce the number of sampling steps [27, 28, 51, 59]. However, despite remark-

*This work was done during an internship with Amazon

able progress in developing fast ODE samplers, they still require more than 10 steps to generate images with satisfactory quality.

Recent approaches to improving the efficiency of diffusion models include model quantization [24, 25], cached sampling [32, 62], and few-step distilled models [30, 43, 44, 53]. Although the first two families of methods accelerate sampling, they still require a large number of steps. In contrast, few-step models distill the full sampling trajectory of the teacher diffusion models to a short trajectory with just one to four steps. In the extreme one-step case, the distilled model learns to directly map noise to the target distribution. However, this poses the significant challenge of achieving high image and prompt fidelity[1] [43, 61]. For example, given the prompt *"A boat on the water with a lighthouse in the background."*, the distilled few-step model may fail to generate the visual concept of *"boat"* in the resulting images (see illustration in Fig. 1). Additionally, distilling these few-step models from teacher diffusion models typically requires substantial computational resources, making them less accessible for easy adoption.

One potential solution to improve prompt fidelity is to train preference models [21, 55]. However, this requires human annotated preference datasets. Another potential solution, drawing inspiration from retrieval-augmented generation (RAG) in large language models (LLMs) [5, 23, 38], is to adapt RAG to enhance prompt fidelity in image generation [1, 3, 47, 48, 58]. These models first query a database for relevant text-image pairs and then condition the diffusion models on the retrieved pairs to enhance prompt fidelity. Although prior retrieval-augmented image generation (RAIG) models have investigated various conditioning strategies [1, 3, 48] and fine-grained retrieval content using LLMs [47, 58], RAIG with few-step diffusion models has not yet been explored. In addition, most previous methods do not focus on improving efficiency in both training and inference [1, 3, 48, 58].

In this paper, we propose to use retrieval augmentation to improve the generation quality of few-step diffusion models while maintaining efficiency. The key intuition is that injecting semantically relevant retrieved information into the diffusion model can simplify the task of mapping random noise to target distributions. Preliminary results on using the retrieved content to directly edit the latent space of the diffusion model's denoiser ($\mathcal{H}$-space) without any finetuning already show improved image and prompt fidelity (see Fig. 1). However, this simple approach requires an expensive hyperparameter search at inference time to maximize performance, compromising the latency benefits of few-step image generation. To mitigate this challenge, we propose to augment the denoiser's $\mathcal{H}$-space with a trainable adapter that blends the retrieved content with the target prompt using a cross-attention mechanism (Sec. 3.3). Experiments show that the proposed ImageRAGTurbo method produces high-fidelity images without compromising efficiency.

Our main contributions can be summarized as follows:

- We introduce a novel retrieval-augmented framework for few-step image generation with diffusion models, which significantly improves prompt fidelity and image quality while maintaining fast generation speeds.
- We develop an efficient finetuning approach using a lightweight adapter network in the $\mathcal{H}$-space, which reduces computational requirements compared to traditional few-step model training methods.
- We conduct extensive experiments demonstrating that our method outperforms existing few-step approaches in terms of both generation quality and prompt alignment, while maintaining comparable inference speed.

The remainder of the paper is organized as follows: Section 2 reviews relevant literature, Section 3 details our method including our preliminary investigation in Section 3.2 and the proposed $\mathcal{H}$-adapter finetuning in Section 3.3. Section 4 presents our experimental validation and results, and Section 5 concludes the paper.

2. Related work

Few-step Text-to-Image Generation. Standard diffusion models [13] generate images via an iterative denoising process that aims to reverse the forward noising process step by step. This iterative reverse sampling process occurs sequentially and cannot be parallelized to speed it up. Although fast ODE samplers [27, 28, 51, 59] can skip some steps by enabling a large sampling step size, these models cannot directly map noise to images like GANs [7, 19]. To address this challenge, consistency models [30, 53] enforce the self-consistency of the denoising networks along the sampling trajectory such that they can map any point in an ODE trajectory to its starting point. This enables one-step or few-step generation by directly mapping random noise to images, or through intermediate steps. However, despite being theoretically grounded, consistency models distilled from pretrained diffusion models, such as latent consistency models [30], typically result in low-quality samples when using few steps [44]. In contrast, adversarial diffusion distillation [44] and its variant in latent space [43] show strong empirical performance in few-step generation. There are also hybrid methods that combine consistency models and adversarial training [4, 60]. However, all these methods share a common flaw: they often struggle to generate faithful images that align well with the target prompts.

Retrieval-Augmented Image Generation. RAG has been extensively explored in natural language processing [5, 23, 38], but has received comparatively less attention in the context of image generation. An early example is the retrieval-

[1] We use *prompt fidelity* and *text-to-image alignment* interchangeably.

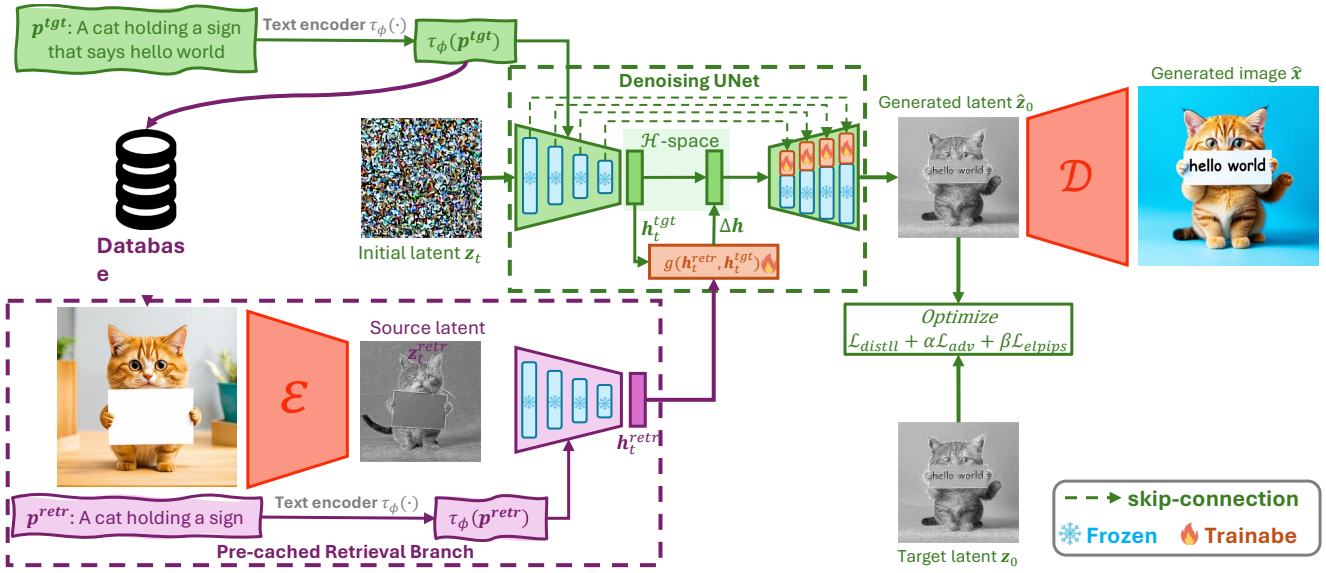


Figure 2. Overview of the proposed ImageRAGTurbo framework for efficiently finetuning the few-step diffusion models with retrieval-augmented generation. The framework involves two main branches: *i)* a standard denoising branch (highlighted by green), and *ii)* a retrieval branch (highlighted by purple). For a target prompt $\mathrm{p}^{\boldsymbol{tgt}}$, it will first be converted to embeddings with a pretrained text encoder $\tau_\phi(\cdot)$. Then we query a database based on the target text embeddings to obtain the retrieved text-latent embeddings $(\tau_\phi(\mathrm{p}^{\boldsymbol{retr}}), \boldsymbol{z}_t^{retr})$, which is then fed into the encoder of the denoiser to obtain the retrieved $\mathcal{H}$-space feature $\boldsymbol{h}_t^{retr}$. Finally, the $\mathcal{H}$-space feature $\boldsymbol{h}_t^{retr}$ in the retrieved branch is injected into the denoising branch by the proposed $\mathcal{H}$-space adapter to guide the generation. See details in Sec. 3.

augmented diffusion model (RDM) [1], which conditions the denoiser of a latent diffusion model [40] on CLIP embeddings [37] of image neighbors retrieved from an external database. Similarly, Re-Imagen [3] and its variant [16] extend the pixel-space training method of Imagen [41] to condition the denoiser on the retrieved text-image pairs. In addition, RealRAG [31] aims to enhance the retrieval branch of retrieval-augmented image generation by training a reflective retriever that selects retrieved images to complement the model's missing knowledge. However, all these methods are not optimized for few-step image generation. Recent approaches like ImageRAG [47] and FineRAG [58] leverage Multimodal LLMs (MLLMs) for dynamic image retrieval based on text prompts, and use these MLLMs to assess and progressively improve the generated images through iterative refinement. However, the repeated calls to MLLMs in these methods introduce substantial computational overhead, which defeats the primary goal of fast generation in this paper.

3. Method

In this work, we propose ImageRAGTurbo, a framework for boosting the performance of few-step generation models. Specifically, given a target prompt, $\boldsymbol{p}^{\text{tgt}}$, we retrieve relevant text-to-image pairs $(\boldsymbol{p}^{\text{retr}}, \boldsymbol{x}^{\text{retr}})$ from a database. The retrieved information is then injected into the text-to-image diffusion models to guide the generation process. Motivated by the recent finding that the $\mathcal{H}$-space of the UNET denoiser already contains semantically meaningful representations, our investigation starts from a training-free direct $\mathcal{H}$-space injection (Sec. 3.2) and then extends to the proposed efficient $\mathcal{H}$-space adapter tuning (Sec. 3.3). The overview of the proposed framework is shown in Fig. 2.

3.1. Preliminary

Latent diffusion models. In this work we focus on latent diffusion models (LDMs) [40] and their popular successor stable diffusion models [6, 36], rather than pixel-space alternatives like Imagen [41], due to their popularity and state-of-the-art performance for text-to-image generation. Specifically, LDMs have two stages: first, a pretrained VAE that transforms images $\boldsymbol{x}$ to their latent representations $\boldsymbol{z}$ via the encoder $\mathcal{E} : \boldsymbol{x} \mapsto \boldsymbol{z}$ and then back to the image space via the decoder $\mathcal{D} : \boldsymbol{z} \mapsto \boldsymbol{x}$; second, a diffusion model applied to the resulting latent space $\mathcal{Z}$. LDMs for text-to-image generation henceforth learn to progressively transform a standard Gaussian distribution $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ along with a target prompt $\boldsymbol{p}^{\text{tgt}}$ into the target latent $\boldsymbol{z}_0 = \mathcal{E}(\boldsymbol{x})$ through a denoiser f_θ:

$$\hat{\boldsymbol{z}}_0 = f_\theta(\boldsymbol{z}_t, t, \tau_\phi(\boldsymbol{p}^{\text{tgt}})), \boldsymbol{z}_t = \alpha_t \boldsymbol{z}_0 + \sigma_t \boldsymbol{\epsilon}, t \in \mathcal{U}(0, 1), \quad (1)$$

where α_t and σ_t control the noise scheduling, and $\tau_\phi(\cdot)$ is a text encoder (*e.g.,* a pretrained CLIP text encoder [37]).

Although there are different parameterizations of the denoiser [20], we opt for the one that predicts $\hat{\boldsymbol{z}}_0$ in Eq. (1), as it is widely used in few-step diffusion models [30, 43, 44, 53]. The denoiser can be trained by optimizing the *scoring matching* objective [52, 54]:

$$\min_{\theta} \mathbb{E}_{\boldsymbol{x}\sim p(\boldsymbol{x}),\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I}),t\sim\mathcal{U}(0,1)} \left[\lambda(t)\|\hat{\boldsymbol{z}}_0 - \boldsymbol{z}_0\|_2^2\right], \quad (2)$$

where $p(\boldsymbol{x})$ is the data distribution, and $\lambda(t)$ is a weighting function. After training, we can generate the target latent $\hat{\boldsymbol{z}}_0$ by solving an SDE/ODE using a variety of solvers [13, 27, 28, 51, 59], starting from a random Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The resulting latent $\hat{\boldsymbol{z}}_0$ is then projected back to the image space via the VAE decoder $\mathcal{D}$ to produce the final generated image $\hat{\boldsymbol{x}} = \mathcal{D}(\hat{\boldsymbol{z}}_0)$.

$\mathcal{H}$-space of the UNet denoiser. We define the $\mathcal{H}$-space of the denoiser [22], which plays a crucial role in our method, as the space of feature maps $\boldsymbol{h}_t$ corresponding to the deepest layer of the UNet denoiser (see Fig. 2). High-level concepts of a generated image, such as image class or object presence and attributes, are largely determined by $\mathcal{H}$-space features, which consequently impact the prompt fidelity [46]. In contrast, low-level features corresponding to earlier layers of the denoiser determine the details of a generated image.

3.2. Retrieval-Augmented Direct $\mathcal{H}$-space Injection

Recent explorations have revealed that the denoising UNet already has a semantically meaningful $\mathcal{H}$-space [22, 34]. In light of this, we first investigate a simple training-free mechanism to directly inject the retrieved information into the $\mathcal{H}$-space of a UNet denoiser to demonstrate the effectiveness of the retrieval augmented generation. Specifically, we first project the retrieved text-image pairs $(\boldsymbol{p}^{\text{retr}}, \boldsymbol{x}^{\text{retr}})$ to their corresponding embeddings $(\tau_\phi(\boldsymbol{p}^{\text{retr}}), \boldsymbol{z}_0^{\text{retr}})$, where $\boldsymbol{z}_0^{\text{retr}} = \mathcal{E}(\boldsymbol{x}^{\text{retr}})$ (see illustration in Fig. 2). Subsequently, the retrieved embeddings $(\tau_\phi(\text{p}^{\text{retr}}), \boldsymbol{z}_0^{\text{retr}})$ are fed to the encoder of the UNet denoiser f_θ^{enc} to extract the $\mathcal{H}$-space features: $\boldsymbol{h}_t^{\text{retr}} = f_\theta^{enc}(\tau_\phi(\text{p}^{\text{retr}}), t, \boldsymbol{z}_0^{\text{retr}})$. In practice, we set $t = 0$, as the retrieved images are noise-free. Likewise, we can obtain the target $\mathcal{H}$-space features $\boldsymbol{h}_t^{\text{tgt}} = f_\theta^{enc}(\tau_\phi(\text{p}^{\text{tgt}}), t, \boldsymbol{z}_t)$. We then enrich the target $\mathcal{H}$-space features by blending it with the retrieved $\mathcal{H}$-space features via spherical normalized interpolation:

$$\boldsymbol{h}_t^{\text{blend}} = \frac{\sin[(1-w)\Omega_t]}{\sin\Omega_t}\boldsymbol{h}_t^{\text{retr}} + \frac{\sin[w\Omega_t]}{\sin\Omega_t}\boldsymbol{h}_t^{\text{tgt}}, \quad (3)$$

where $\Omega_t = \arccos(\langle \boldsymbol{h}_t^{\text{tgt}}, \boldsymbol{h}_t^{\text{retr}} \rangle)$, and $w \in (0, 1)$ controls the strength of the blending. We employ spherical normalized interpolation rather than linear interpolation, as geodesic interpolation paths provide smoother semantic transitions [26], effectively preventing the abrupt changes typically caused by phase transitions [46]. The blended $\mathcal{H}$-space features along with the target prompt embeddings are then fed to the decoder of the UNet denoiser f_θ^{dec} to produce the denoised latent

$$\hat{\boldsymbol{z}}_{t-1} = f_\theta^{dec}(\tau_\phi(\text{p}^{\text{tgt}}), t, \boldsymbol{h}_t^{\text{retr}}).$$

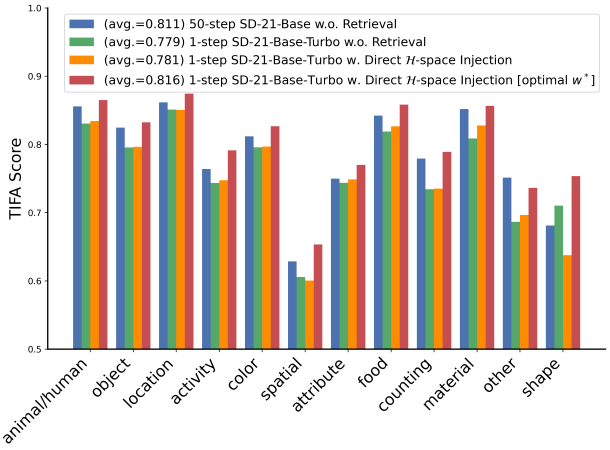


Figure 3. Performance of direct $\mathcal{H}$-space injection, shown as a histogram of TIFA scores across various categories.

Our initial investigation has revealed that the direct $\mathcal{H}$-space injection without any further tuning yields a modest improvement in prompt fidelity (see Fig. 3), increasing the TIFA score [17] from 0.779 to 0.781 with a fixed $w = 0.8$ for all prompts. Observing that different prompts require different optimal blending strengths, we determine the optimal w^* that maximizes the TIFA score for each prompt within a set of predefined values $\{0.1, 0.2, \ldots, 0.9\}$ via brute force search/exhaustive search. This further improves the TIFA score to 0.816, even surpassing the 50-step stable diffusion model without retrieval augmentation. However, determining the optimal blending strength w^* of each prompt is an ill-posed problem, as the optimal blending depends on latent factors such as retrieval relevance, prompt semantics, generation difficulty, *etc*.

A more principled solution to this problem is to edit the $\mathcal{H}$-space representations along multiple directions and compute the edit strength automatically. For example, recent work for LLMs [29] constructs a large dictionary of latent editing directions (*e.g.,* 40,000) and uses sparse coding to automatically select a small, relevant subset of directions along with their corresponding strengths. However, solving a sparse coding problem during inference can largely increase the inference time. Next, we propose an efficient method based on training an adapter that blends retrieved and target $\mathcal{H}$-space features.

3.3. Retrieval-Augmented Efficient $\mathcal{H}$-space Tuning

Although directly injecting content into $\mathcal{H}$-space improves prompt fidelity, it requires exhaustive search for the optimal

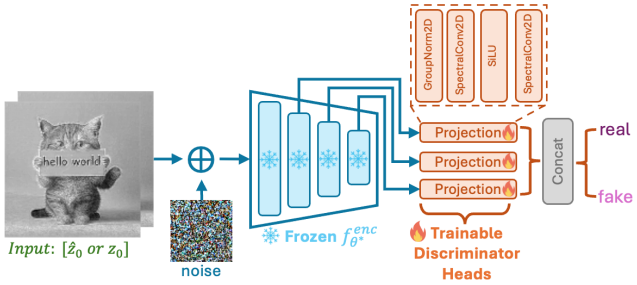


Figure 4. The architecture of the discriminator used for latent adversarial training, which adds noise to the samples $\hat{\boldsymbol{z}}_0$ from the student model and $\boldsymbol{z}_0$ from the teacher model, and differentiate them.

blending strength, making it impractical for real-world applications. This motivates us to automatically learn the correlations between the retrieved and target $\mathcal{H}$-space features. For this purpose, we introduce a trainable adapter $g_\varphi(\cdot,\cdot)$ to the $\mathcal{H}$-space of the UNet denoiser while freezing the other parts of the denoiser. The adapter $g_\varphi(\cdot,\cdot)$ leverages a cross-attention mechanism to automatically determine the correlations between the retrieved and target $\mathcal{H}$-space features:

$$g_\varphi(\boldsymbol{h}_t^{\text{tgt}}, \boldsymbol{h}_t^{\text{retr}}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}}\right)\boldsymbol{V}, \\ \boldsymbol{Q} = \boldsymbol{W}_Q \cdot \boldsymbol{h}_t^{\text{tgt}}, \boldsymbol{K} = \boldsymbol{W}_K \cdot \boldsymbol{h}_t^{\text{retr}}, \boldsymbol{V} = \boldsymbol{W}_V \cdot \boldsymbol{h}_t^{\text{retr}}. \tag{4}$$

The outputs from the adapter are then added to the raw target $\mathcal{H}$-space features as follows:

$$\boldsymbol{h}_t^{\text{retr}} = \boldsymbol{h}_t^{\text{tgt}} + \lambda \underbrace{g_\varphi(\boldsymbol{h}_t^{\text{tgt}}, \boldsymbol{h}_t^{\text{retr}})}_{\Delta \boldsymbol{h}_t}, \tag{5}$$

where λ is a weighting parameter. We empirically find that the results are not sensitive to the choice of λ, and set it to be the cosine similarity between the retrieved and target text CLIP embeddings. The intuition is that if the retrieved content has a high correlation to the target prompt, it should contribute more to the final generated content.

3.4. Training Procedure

Few-step diffusion models, trained through either self-consistency [30, 53] or adversarial training [43, 44], operate on a subset of n timesteps $T_n = \{t_1, t_2, \cdots, t_n\}$ drawn from the complete N timesteps in the teacher model. For example, a 4-step student model uses $T_n = \{1.0, 0.75, 0.5, 0.25\}$. The predicted $\hat{\boldsymbol{z}}_0$ at timestep $t_i, i \in \{1, \cdots, n\}$ of the proposed denoiser reads

$$\hat{\boldsymbol{z}}_0 = f_\theta(\boldsymbol{z}_{t_i}, t_i, \tau_\phi(\text{p}^{\text{tgt}}), \boldsymbol{h}_0^{\text{retr}}).$$

In this work, we focus on latent adversarial training [43], as it empirically shows better performance than self-consistency training, especially in the one-step scenario. In addition, latent adversarial training, which operates on the latent $\mathcal{Z}$ space, is more efficient than its predecessor that performs adversarial training in the pixel space [44].

Without loss of generality, we denote the teacher and few-step student denoisers by f_{θ^*} and f_θ, respectively, both frozen in our training schema. The trainable module for the few-step denoiser is the $\mathcal{H}$-space adapter. To further help translate the adapted features into the final generation, we also finetune the decoder of the student denoiser using the parameter-efficient low-rank adaption tuning [15].

In our latent adversarial training, the discriminator (denoted by D) consists of the frozen UNet encoder of the teacher denoiser ($f_{\theta^*}^{enc}$), followed by a few trainable projection layers that project multi-stage encoder features to output embeddings (see Fig. 4). Unlike standard adversarial training, latent adversarial training leverages the idea of DiffusionGAN [56] that differentiates a pair of noisy latent variables ($\hat{\boldsymbol{z}}_s$, $\boldsymbol{z}_s$), where $\hat{\boldsymbol{z}}_s = \alpha_s \hat{\boldsymbol{z}}_0 + \sigma_s \boldsymbol{\epsilon}$ and $\boldsymbol{z}_s = \alpha_s \boldsymbol{z}_0 + \sigma_s \boldsymbol{\epsilon}$, $s \sim \{t_1, \cdots, t_s, \cdots, t_N\}$. It is worth noting that the timesteps used to train the discriminator are sampled uniformly from the full set of N timesteps such that it can learn to distinguish samples across different noise levels throughout the entire diffusion process.

Following [42–44], the hinge loss is used as the adversarial objective function. The objective function for training the discriminator is given as:

$$\mathcal{L}_{\text{adv}}^D = \sum_k \mathbb{E}_{\boldsymbol{z}_0}\left[\max(0, 1 - D_k(\boldsymbol{z}_s, \tau_\phi(P^{\text{tgt}}), t_s))\right] + \\ \mathbb{E}_{\hat{\boldsymbol{z}}_0}[\max(0, 1 + D_k(\hat{\boldsymbol{z}}_s, \tau_\phi(P^{\text{tgt}}), t_s))], \tag{6}$$

where D stands for discriminator including the frozen teacher encoder and trainable projection layers, and k denotes the k-th stage features. Similarly, the objective function for training the few-step denoiser, denoted by G following the convention of adversarial training, reads:

$$\mathcal{L}_{\text{adv}}^G = -\sum_k \mathbb{E}_{\boldsymbol{z}_0}[D_k(\hat{\boldsymbol{z}}_s, \tau_\phi(\text{p}^{\text{tgt}})]. \tag{7}$$

To stabilize the adversarial training procedure, we apply spectral normalization [33] to the convolutional layers in the trainable projection heads of the discriminator (see Fig. 4). Compared to the gradient penalty method used in [18, 44], spectral normalization does not require double backpropagation, and hence it saves computations.

In addition to the adversarial training objective, we use two auxiliary objectives to stabilize training and improve image quality: *i)* score distillation loss, and *ii)* latent LPIPS loss [18]. The score distillation loss is computed as the smoothed L_1 loss between $\hat{\boldsymbol{z}}_0$ and $\boldsymbol{z}_0$:

$$\mathcal{L}_{\text{distill}} = \begin{cases} 0.5 * \|\hat{\boldsymbol{z}}_0 - \boldsymbol{z}_0\|_2^2, & \text{if } |\hat{\boldsymbol{z}}_0 - \boldsymbol{z}_0| < 1 \\ |\hat{\boldsymbol{z}}_0 - \boldsymbol{z}_0| - 0.5, & \text{otherwise.} \end{cases} \tag{8}$$

Similar to the LPIPS loss, the latent LPIPS loss is computed in the latent space between $\hat{z}_0$ and z_0:

$$\mathcal{L}_{\text{elpips}} = \mathbb{E}_{\mathcal{T}}\left[\|F(\mathcal{T}(\hat{z}_0)) - F(\mathcal{T}(z_0)\|_2^2\right], \tag{9}$$

where $\mathcal{T}(\cdot)$ stands for a set of random differentiable augmentations, including general geometric transformations and cutoff, and $F(\cdot)$ denotes the embedding network (*e.g.,* VGG16 [50]).

The final training objective is the weighted sum of the aforementioned three objectives:

$$\mathcal{L} = \mathcal{L}_{\text{adv}} + \alpha\mathcal{L}_{\text{distill}} + \beta\mathcal{L}_{\text{latentLPIPS}}, \tag{10}$$

where α and β are weight-balance parameters. However, the extensive search of these parameters is expensive in practice. Following [18, 44], we empirically set $\alpha = 2.5$ and $\beta = 1.0$ in this paper.

4. Experiments

4.1. Experimental Setup

Datasets. We finetune the proposed method using a mix of synthetic and real text-image pairs, as prior studies have shown that synthetic data can help improve prompt fidelity [43]. We generate synthetic data via the teacher model (*i.e.,* Stable Diffusion v2-1-base [40]) at a constant classifier-free guidance [12] value of 7.5, using around 3 million prompts from LAION-Aesthetic 6.25+ dataset [45], which is commonly used for few-step diffusion distillation [57]. For the real data, we select 500K images from the LAION-Aesthetic 5.5+ dataset [45] after filtering out those with resolutions smaller than 1024×1024. Both the real and synthetic images are subsequently resized to 512×512 and fed into the VAE encoder to obtain 64×64 latent embeddings.

Evaluation benchmarks. We evaluate our proposed method on two widely used benchmarks for evaluating text-to-image generation: *i)* MS-COCO [2] and *ii)* TIFA. Specifically, the MS-COCO benchmark comprises $5,000$ text-image pairs from the validation set of the MS-COCO 2017 dataset, while the TIFA benchmark consists of $4,081$ text prompts that contain elements from 12 categories (*e.g.,* object, spatial, activity, counting).

Evaluation metrics. Our quantitative evaluation employs a set of metrics to assess different aspects of the generated images, including photorealism and faithfulness. For measuring the photorealism of the generated images, we consider the widely-adopted FID score [11] on the COCO benchmark, which measures if the generated images follow the same distribution of the real images. In contrast, due to the absence of reference images, we use Aesthetics (AES) score to measure the visual appeal on the TIFA benchmark. To assess the faithfulness of the generated images to the prompts, we adapt the widely used CLIP score [10] on the COCO benchmark. We compute CLIP score using the `OpenCLIP-ViT-bigG-14` model. Although the CLIP score provides an overall assessment of text-to-image alignment, it may fail to provide an accurate evaluation of specific object attributes, quantities, and spatial relationships in the generated images. Therefore, we use the TIFA score [17] to measure if the generated images can truthfully reflect the target prompts by visual question answering.

Baselines. To ensure a fair comparison, we compare the proposed method with several popular text-to-image LDM baselines with similar model sizes. In particular, we consider stable diffusion models [40], such as Stable Diffusion v1-5 and Stable Diffusion v2-1-base. In addition, we compare our model with few-step diffusion models, including the Latent Consistency Model (LCM) [30] and Stable Diffusion Turbo v2-1-base trained by latent adversarial distillation [43] rather than pixel-space adversarial distillation as in [44]. We also compare our method to RDM [1], which integrates retrieval augmentation into stable diffusion models. To ensure a fair comparison, our method and all baselines have almost the same model size. For the distilled few-step models, the classifier-free guidance [12] is disabled, as it will easily overshoot [43]. For the other baseline models, we use a constant classifier-free guidance value of 7.5.

Implementation details. To obtain retrieved text-image pairs we apply the ScaNN search algorithm [8] in the text feature space of a pretrained CLIP text encoder [37]. Here, we reuse the same CLIP text features used by the stable diffusion model (*i.e.,* `OpenCLIP-ViT-H-14`) without introducing any additional computational burden during inference. Other retrieval methods, such as the lexical prompt retriever (*e.g.,* BM25 retriever) can be applied in different applications without introducing expensive computational overhead. We use the 0.63M text-image pairs from the OpenImage database as the retrieval database. Our finetuning is performed on 64× NVIDIA L40S GPUs with a total batch size of $2,048$. We finetune the 1-step stable diffusion turbo model for 20K iterations using the AdamW optimizer with a constant learning rate of 1×10^{-5} for both few-step student model and the discriminator. The overall finetuning process takes approximately one week.

4.2. Results on the MS-COCO Benchmark

Our evaluation on the MS-COCO benchmark demonstrates ImageRAGTurbo's superior performance (see Table 1). Compared to the baseline Stable Diffusion Turbo v2-1-base without retrieval augmentation, our approach achieves better text-image alignment with a CLIP score that is 1.37% higher, and improved image quality as indicated by a lower FID score. Our 1-step ImageRAGTurbo even surpasses the 4-step latent consistency model by a large margin in both CLIP and FID scores. Furthermore, our 1-step model

Table 1. Quantitative comparison of text-to-image generation models on the MS-COCO benchmark measured by the FID and CLIP scores, as well as the number of function evaluations (NFE) in the denoising process.

Models	NFE	FID↓	CLIP↑
w/o Retrieval			
Stable Diffusion v1-5	50	24.38	0.319
Stable Diffusion v2-1	50	25.33	0.330
Latent Consistency Model	4	36.52	0.307
Stable Diffusion Turbo v2-1	1	26.04	0.319
w/ Retrieval			
RDM	50	27.60	0.293
ImageRAGTurbo (Ours)	1	25.59	0.323

Table 2. Quantitative comparison of text-to-image generation models on the TIFA benchmark as measured by the TIFA and AES scores, and the number of function evaluations (NFE) in denoising.

Models	NFE	AES↑	TIFA↑
w/o Retrieval			
Stable Diffusion v1-5	50	5.79	0.768
Stable Diffusion v2-1	50	6.04	0.811
Latent Consistency Model	4	5.80	0.764
Stable Diffusion Turbo v2-1	1	5.85	0.779
w/ Retrieval			
RDM	50	5.40	0.725
ImageRAGTurbo (Ours)	1	5.88	0.801

achieves comparable results to the 50-step teacher model (*i.e.,* Stable Diffusion v2-1-base), showing a drop of 1.03% and 2.1% in FID and CLIP scores, respectively. These improvements can be attributed to ImageRAGTurbo's enhanced ability to retrieve and leverage fine-grained, high-quality reference images for more precise image synthesis. In comparison to other retrieval-augmented approaches, ImageRAGTurbo shows substantial improvement over RDM with a CLIP score that is 10% higher, which highlights the effectiveness of the proposed $\mathcal{H}$-space adapter tuning.

4.3. Results on the TIFA Benchmark

As shown in Table 2, ImageRAGTurbo also demonstrates superior performance on the TIFA benchmark. Compared to the baseline Stable Diffusion Turbo v2-1-base without retrieval augmentation, our approach achieves better text-image fidelity with a TIFA score that is 2.2% higher, and a slightly better aesthetic quality as indicated by the aesthetic score (5.88 vs 5.83). Additionally, our 1-step ImageRAGTurbo outperforms the 4-step latent consistency model by a large margin, showing a 3.7% improvement in TIFA score. Although our 1-step model shows an overall drop of 1.2% in TIFA score compared to 50-step Stable Diffusion v2-1-base model, it achieves a comparable or even slightly higher TIFA score in certain categories such as *object, location, activity, and material* (see Fig. 5). Compared to other retrieval-augmented approaches, ImageRAGTurbo substantially outperforms RDM with a TIFA score 7.6% higher, and a better aesthetic quality (5.88 vs 5.40), while requiring only a single inference step instead of 50 steps.

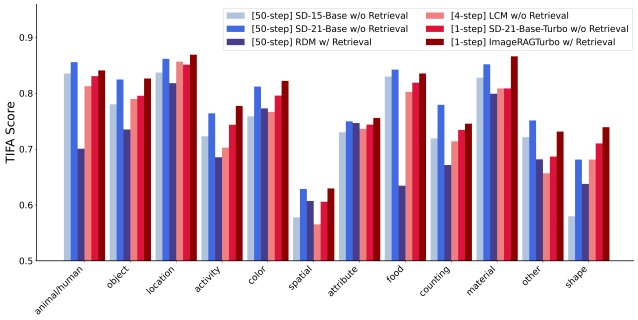


Figure 5. Detailed histogram of TIFA scores across various categories. Despite still lagging behind 50-step Stable Diffusion v2-1-base model, our 1-step ImageRAGTurbo achieves comparable or even slightly higher TIFA score in certain categories such as *object, activity, and material*.

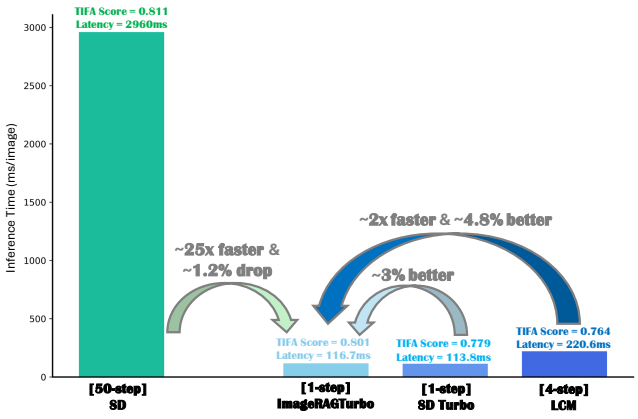
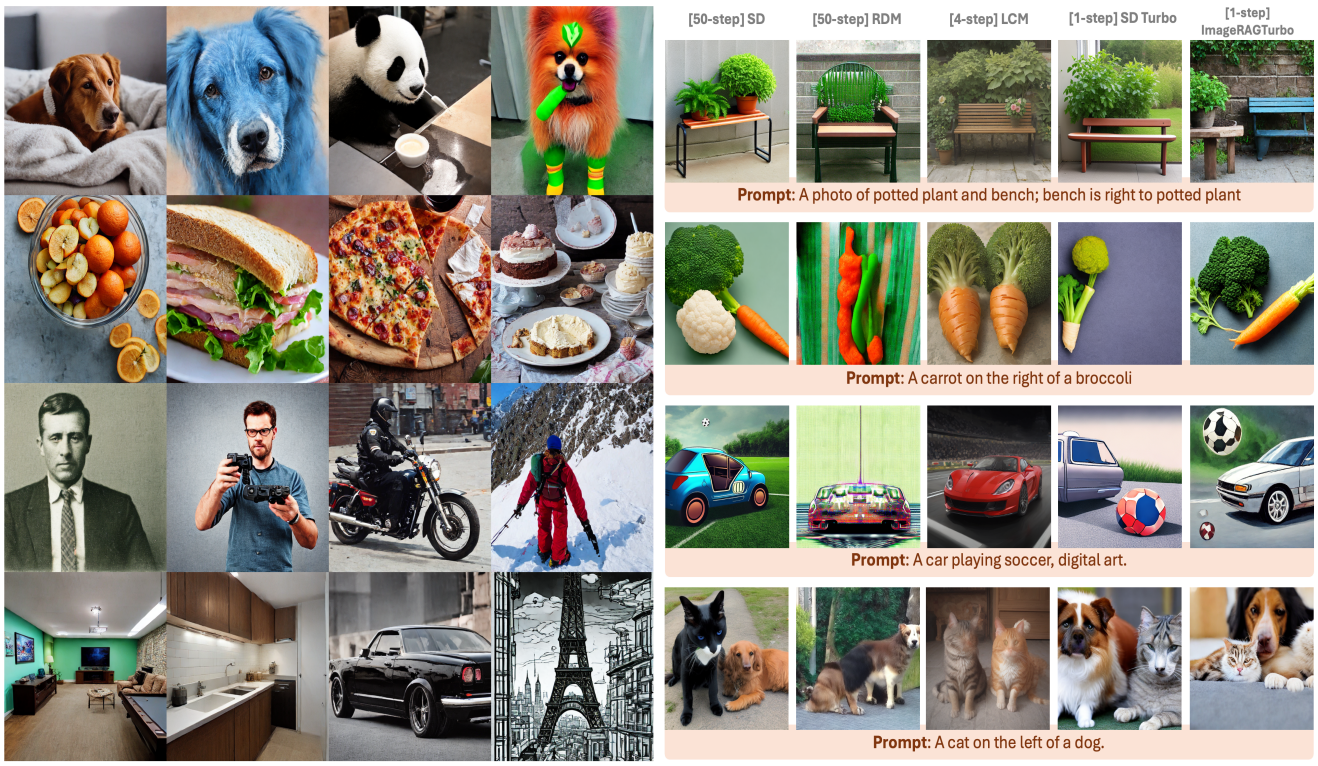


Figure 6. Comparison of inference time (ms/image). We report the average inference time calculated over the same set of 100 prompts at 512×512 resolution on a single NVIDIA L40S GPU.

4.4. Efficiency Analysis

Fig. 6 shows the computational efficiency of our method compared to baselines. The total inference time of our method (116.7ms) is on par with that of the Stable Diffusion (SD) Turbo (113.8ms). The inference time of our method can be broken down into two components: retrieval (1.9ms/image) and denoising (114.8ms/image). The 1 ms increase in denoising latency compared to SD Turbo is attributable to the $\mathcal{H}$-space adapter, with 36M additional parameters, representing only 4% of the total model parameters. Our method introduces only a minimal overhead of 2.5% in latency compared to the SD Turbo, due to the added



(a) Representative samples generated by ImageRAGTurbo.

(b) Qualitative comparison with baseline methods.

Figure 7. Qualitative results of ImageRAGTurbo. (a) ImageRAGTurbo produces high-quality samples across diverse prompts and visual concepts with a single denoising step. (b) ImageRAGTurbo achieves better text-to-image alignment than other competing few-step methods on a variety of scenes, objects, and compositions.

retrieval mechanism and adapter, while delivering a significant 3% improvement in TIFA score. Additionally, the proposed method is approximately 25× faster than the SD teacher model, with only a 1.2% drop in TIFA score. Compared to the latent consistency model, our method demonstrates a 3.7% higher TIFA score with around 47% more latency (116.7ms vs 220.6ms).

4.5. Qualitative Analysis

To highlight the effectiveness of our proposed model and its potential applications, we present qualitative examples in Fig. 7. Specifically, our approach can generate high-quality images in a single step (see Fig. 7a). Notably, our approach outperforms other few-step baselines in text-to-image alignment (see Fig. 7b). In particular, our RAG-based proposed method can generate accurate objects (demonstrated in *examples/rows 2 and 3*) and proper spatial relationship (shown in *examples/rows 1 and 4*).

5. Conclusion and Future Work

In this paper, we have presented ImageRAGTurbo, a novel retrieval-augmented framework designed to enhance few-step diffusion models through an efficient $\mathcal{H}$-space adapter tuning. Our approach tackles a key challenge in few-step diffusion models (insufficient text-to-image alignment) by effectively leveraging relevant information from a large retrieval database. Through this efficient retrieval-augmented mechanism, ImageRAGTurbo can improve the model's ability to generate images that faithfully reflect the input prompts without compromising the inference latency. Experimental results show that our method significantly improves text-to-image alignment without compromising image quality.

While our framework opens up a new direction for retrieval-augmented few-step diffusion models, several avenues remain for future investigation. First, our current framework relies on CLIP-based retrieval, while incorporating more fine-grained strategies, such as compositional retrieval, is a promising direction for future work. Second, we evaluate the current framework only on UNet-based architectures, and extending the proposed method to diffusion transformers (DiTs) [35] is worth exploring. Similar to the $\mathcal{H}$-space in UNets, recent studies suggest that DiTs also learn semantically structured latent representations in deeper layers, where structured and disentangled features emerge in latent patch spaces and attention maps [9, 49].

References

[1] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. *arXiv preprint arXiv:2204.11824*, 2022. 2, 3, 6

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context, 2018. 6

[3] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 2, 3

[4] Quan Dao, Hao Phung, Trung Tuan Dao, Dimitris N Metaxas, and Anh Tran. Self-corrected flow distillation for consistent one-step and few-step image generation. In *AAAI Conference on Artificial Intelligence*, pages 2654–2662, 2025. 2

[5] Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. 2

[6] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Int. Conf. Machine Learning*, 2024. 3

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[8] Ruiqi Guo, Philip Sun, Erik Lindgren, Quan Geng, David Simcha, Felix Chern, and Sanjiv Kumar. Accelerating large-scale inference with anisotropic vector quantization, 2020. 6

[9] Alec Helbling, Tuna Han Salih Meral, Ben Hoover, Pinar Yanardag, and Duen Horng Chau. Conceptattention: Diffusion transformers learn highly interpretable features. *arXiv preprint arXiv:2502.04320*, 2025. 8

[10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 6

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. 6

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2, 4

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 1

[15] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 5

[16] Hexiang Hu, Kelvin CK Chan, Yu-Chuan Su, Wenhu Chen, Yandong Li, Kihyuk Sohn, Yang Zhao, Xue Ben, Boqing Gong, William Cohen, et al. Instruct-imagen: Image generation with multi-modal instruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4754–4763, 2024. 3

[17] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20406–20417, 2023. 4, 6

[18] Minguk Kang, Richard Zhang, Connelly Barnes, Sylvain Paris, Suha Kwak, Jaesik Park, Eli Shechtman, Jun-Yan Zhu, and Taesung Park. Distilling Diffusion Models into Conditional GANs. In *European Conference on Computer Vision (ECCV)*, 2024. 5, 6

[19] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2

[20] 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. 4

[21] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation, 2023. 2

[22] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022. 4

[23] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Adv. Neural Inform. Process. Syst.*, pages 9459–9474, 2020. 2

[24] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17535–17545, 2023. 2

[25] Yanjing Li, Sheng Xu, Xianbin Cao, Xiao Sun, and Baochang Zhang. Q-dm: An efficient low-bit quantized diffusion model. In *Adv. Neural Inform. Process. Syst.*, pages 76680–76691, 2023. 2

[26] Alexander Lobashev, Dmitry Guskov, Maria Larchenko, and Mikhail Tamm. Hessian geometry of latent space in generative models. In *Forty-second International Conference on Machine Learning*, 2025. 4

[27] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022. 1, 2, 4

[28] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sam-

pling of diffusion probabilistic models. *Machine Intelligence Research*, pages 1–22, 2025. 1, 2, 4

[29] Jinqi Luo, Tianjiao Ding, Kwan Ho Ryan Chan, Darshan Thaker, Aditya Chattopadhyay, Chris Callison-Burch, and René Vidal. Pace: Parsimonious concept engineering for large language models. *Advances in Neural Information Processing Systems*, 37:99347–99381, 2024. 4

[30] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2, 4, 5, 6

[31] Yuanhuiyi Lyu, Xu Zheng, Lutao Jiang, Yibo Yan, Xin Zou, Huiyu Zhou, Linfeng Zhang, and Xuming Hu. RealRAG: Retrieval-augmented realistic image generation via self-reflective contrastive learning. In *Forty-second International Conference on Machine Learning*, 2025. 3

[32] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 15762–15772, 2024. 2

[33] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 5

[34] Yong-Hyun Park, Mingi Kwon, Jaewoong Choi, Junghyo Jo, and Youngjung Uh. Understanding the latent space of diffusion models through the lens of riemannian geometry. *Advances in Neural Information Processing Systems*, 36: 24129–24142, 2023. 4

[35] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 8

[36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3

[37] 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*, pages 8748–8763. PmLR, 2021. 3, 6

[38] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023. 2

[39] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 1

[40] 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*, pages 10684–10695, 2022. 1, 3, 6

[41] 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. 1, 3

[42] Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster, 2021. 5

[43] Axel Sauer, Frederic Boesel, Tim Dockhorn, Andreas Blattmann, Patrick Esser, and Robin Rombach. Fast high-resolution image synthesis with latent adversarial diffusion distillation. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2, 4, 5, 6

[44] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2024. 2, 4, 5, 6

[45] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 6

[46] Antonio Sclocchi, Alessandro Favero, and Matthieu Wyart. A phase transition in diffusion models reveals the hierarchical nature of data. *Proceedings of the National Academy of Sciences*, 122(1):e2408799121, 2025. 4

[47] Rotem Shalev-Arkushin, Rinon Gal, Amit H Bermano, and Ohad Fried. ImageRAG: Dynamic image retrieval for reference-guided image generation. *arXiv preprint arXiv:2502.09411*, 2025. 2, 3

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

[49] Zitao Shuai, Chenwei Wu, Zhengxu Tang, Bowen Song, and Liyue Shen. Latent space disentanglement in diffusion transformers enables precise zero-shot semantic editing. *arXiv preprint arXiv:2411.08196*, 2024. 8

[50] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. 6

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 4

[52] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 4

[53] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252. PMLR, 2023. 2, 4, 5

[54] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011. 4

[55] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization, 2023. 2

[56] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*, 2022. 5

[57] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and Bill Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024. 6

[58] Huaying Yuan, Ziliang Zhao, Shuting Wang, Shitao Xiao, Minheng Ni, Zheng Liu, and Zhicheng Dou. Finerag: Fine-grained retrieval-augmented text-to-image generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11196–11205, 2025. 2, 3

[59] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems*, 36:49842–49869, 2023. 1, 2, 4

[60] Zhenyu Zhou, Defang Chen, Can Wang, Chun Chen, and Siwei Lyu. Simple and fast distillation of diffusion models. *Advances in Neural Information Processing Systems*, 37:40831–40860, 2024. 2

[61] Zhenyu Zhou, Defang Chen, Can Wang, Chun Chen, and Siwei Lyu. Simple and fast distillation of diffusion models, 2024. 2

[62] Chang Zou, Xuyang Liu, Ting Liu, Siteng Huang, and Linfeng Zhang. Accelerating diffusion transformers with token-wise feature caching. In *The Thirteenth International Conference on Learning Representations*, 2025. 2