RECON: Training-Free Acceleration for Text-to-Image Synthesis with Retrieval of Concept Prompt Trajectories

Chen-Yi Lu[1,2](✉), Shubham Agarwal[2], Md Mehrab Tanjim[2], Kanak Mahadik[2], Anup Rao[2], Subrata Mitra[2], Shiv Kumar Saini[2], Saurabh Bagchi[1], and Somali Chaterji[1]

[1] Purdue University, West Lafayette, USA
lu842@purdue.edu
[2] Adobe Research, San Jose, USA

Abstract. Text-to-image diffusion models excel in generating photo-realistic images but are hampered by slow processing times. Training-free retrieval-based acceleration methods, which leverage pre-generated "trajectories," have been introduced to address this. Yet, these methods often lack diversity and fidelity as they depend heavily on similarities to stored prompts. To address this, we present ReCon (**Re**trieving **Con**cepts), an innovative retrieval-based diffusion acceleration method that extracts visual "concepts" from prompts, forming a knowledge base that facilitates the creation of adaptable trajectories. Consequently, ReCon surpasses existing retrieval-based methods, producing high-fidelity images and reducing required Neural Function Evaluations (NFEs) by up to 40%. Extensive testing on MS-COCO, Pick-a-pick, and DiffusionDB datasets confirms that ReCon consistently outperforms established methods across multiple metrics such as Pick Score, CLIP Score, and Aesthetics Score. A user study further indicates that 76% of images generated by ReCon are rated as the highest fidelity, outperforming two competing methods, a purely text-based retrieval and a noise similarity-based retrieval. Project URL: https://stevencylu.github.io/ReCon.

Keywords: Text-to-image Generation · Diffusion Model Acceleration · Retrieval-based Diffusion

1 Introduction

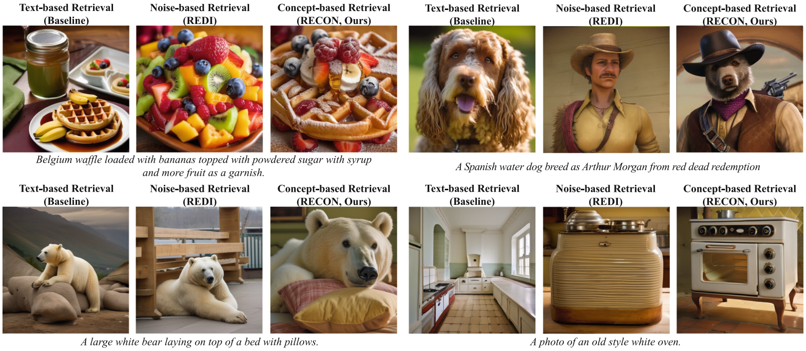
Diffusion Probabilistic Models (DMs) are a class of generative models that simulate a process of gradually reversing noise into structured outputs, drawing

C.-Y. Lu and S. Agarwal—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-73202-7_17.

A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15117, pp. 288–306, 2025.
https://doi.org/10.1007/978-3-031-73202-7_17

inspiration from thermodynamics. These models iteratively denoise an initial random noise distribution to the desired data distribution [46]. DMs have proven to be highly effective in a variety of applications, such as text-to-image generation (T2I) [11,17,38,40,41,43,56,65], speech synthesis [10,27,57], and 3D image generation [21,54,64]. In particular, DMs like Imagen [43], DALL-E 2 [40], and StableDiffusion XL (SDXL) [38] have excelled in T2I tasks. However, DMs are notably slower than other generative models due to the extensive number of neural function evaluations (NFEs), where each NFE entails a forward network pass. As AI art and related T2I applications gain popularity, enhancing user experience with DMs through faster model inference and reducing NFEs is crucial.


Text-based Retrieval (Baseline)
Noise-based Retrieval (REDI)
Concept-based Retrieval (RECON, Ours)
Text-based Retrieval (Baseline)
Noise-based Retrieval (REDI)
Concept-based Retrieval (RECON, Ours)

Belgium waffle loaded with bananas topped with powdered sugar with syrup and more fruit as a garnish.

A Spanish water dog breed as Arthur Morgan from red dead redemption

Text-based Retrieval (Baseline)
Noise-based Retrieval (REDI)
Concept-based Retrieval (RECON, Ours)
Text-based Retrieval (Baseline)
Noise-based Retrieval (REDI)
Concept-based Retrieval (RECON, Ours)

A large white bear laying on top of a bed with pillows.

A photo of an old style white oven.

Fig. 1. Examples of outputs from existing retrieval-based acceleration approaches, namely Text-based (Baseline) and Noise-based (ReDi [58]) Retrieval, show their difficulties in accurately representing the input prompts. On the other hand, our proposed Concept-based Retrieval framework (ReCon) is able to produce high-fidelity and faithful images compared to other techniques.

To address this issue, several approaches have been developed to accelerate the generation process by reducing the number of NFEs in DMs, including both training-based and training-free methods. Training-based methods, such as Distillation and Consistency Models [35,44,50], aim to distill a complex pre-trained model into a more efficient student model, capable of faster image generation with reduced NFEs. While these methods are promising, they often require a substantial amount of training time and data. This challenge becomes even more pronounced as SOTA diffusion models evolve to encompass over a billion parameters [2,38,43], intensifying the high training load.

To mitigate the demand for computing resources, training-free methods prioritize enhancing computational efficiency without necessitating additional model training. These strategies primarily concentrate on identifying more efficient numerical solvers, which reduce the number of denoising steps, thereby decreasing the NFEs required by DMs [33,59,62]. Our approach diverges from the more conventional focus on numerical solvers that primarily aim to streamline the

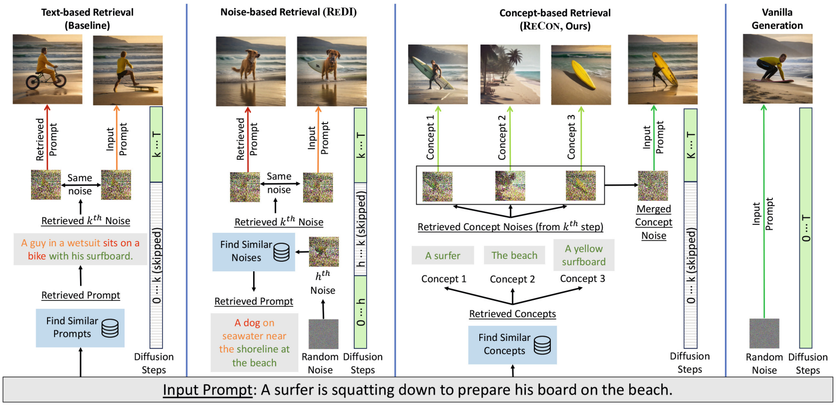


Fig. 2. Overview of different approaches. Text-based (Baseline) and noise-based retrieval (REDI [58]) often follow rigid trajectories with extraneous elements (highlighted in red). In contrast, our method (RECON) parses input prompts into visual concepts and selects closely related ones from a knowledge base (highlighted in green), enabling a malleable trajectory for incorporating input-specific additional details. Here, 0 = Start, h = Initial, k = Intermediate, and T = Final denote the diffusion steps. (Color figure online)

denoising process. Rather, it capitalizes on leveraging previously stored intermediate noisy outputs, or "noises," associated with a vast array of known prompts. For any new input prompt, retrieval-based methods retrieve a relevant noise vector from this repository of stored trajectories at a certain intermediate step. This enables the model to resume the denoising process from a later point, effectively skipping the initial steps and thus speeding up image generation. Additionally, this retrieval-based approach can be integrated with any diffusion model's numerical solvers, without altering the core denoising mechanics.

To retrieve the nearest trajectory for an input text, a straightforward approach can be based on text-based similarity [3]. Here, we can find the prompt that is most similar to the input prompt and retrieve its associated trajectory to get the noise at an intermediate step before starting the denoising process. In theory, semantically similar prompts (e.g., "An old man" vs. "A man who is old") should yield comparable trajectories in DMs, allowing the denoising process to resume from an intermediate noise level with analogous outcomes. Different from this simple textual similarity, [58] proposed REDI, a noise-based retrieval framework. This method initiates the denoising sequence for few steps, subsequently identifying the most closely matched trajectory based on noise-wise similarity.

The efficacy of text-based retrieval (Baseline) and noise-based retrieval (REDI) methods in generating images are illustrated in Fig. 1. The figure shows that these methods often struggle to produce images that are faithful to the input text. To understand this, we provide a high-level overview of each method in Fig. 2. As seen from the illustration, both methods rely on stored prompts, either directly for text similarity or indirectly for noise similarity. This strategy presupposes a diverse repository capable of reflecting the myriad of unique prompts encountered during inference. However, the reality often deviates, as truly analogous prompts with matching concepts are scarce in real-world scenarios. Moreover, even when a prompt is retrieved that seems closely matched in text (Baseline) or initial noise (ReDi), it might introduce extraneous concepts into the generated image. Key image attributes, such as layouts and color schemes, are predominantly determined during the early stages of the denoising process [15]. However, these stages are often skipped in retrieval-based methods, making it difficult to correct the trajectory using the input prompt within the constraints of reduced NFEs post-retrieval. This issue is depicted in Fig. 2, where red-colored texts like "on a bike" (Baseline) or "a dog" (ReDi) indicate unwanted details in the final image. Such trajectory deviations present a critical flaw in both techniques, making retrieval-based acceleration challenging.

To address these challenges, we present ReCon. The central idea of ReCon is **re**trieving **con**cepts from the input prompt, which can be thought of as "soft templates" for the image generation process. This method involves collecting intermediate noises corresponding to these concepts and then combining them. The combination results is what we call a "flexible trajectory," which incorporates some details from the retrieved prompts while still leaving scope for additional elements that are unique from the input prompt. This is illustrated in Fig. 2 under "Concept-based Retrieval," where the prompt is decomposed into three generic concepts (A surfer, The beach, and A yellow surfboard) retrieved from the knowledge base. This flexibility is central to ReCon's approach. Our strategy draws upon NLP research, which suggests that a sentence can be deconstructed into phrases and evaluated for various attributes such as named entities [26], abstractness [5,47], and visualness [52]. In our method, we primarily use a pre-trained language model [31] to extract essential noun phrases and determine their visualness using [52] to identify "concepts" (we use "visual concepts" and "concepts" interchangeably). As illustrated in Fig. 1, the adoption of visual concepts allows ReCon to generate images that are not only more accurate but also more faithful to the input prompts, surpassing previous methods. *To the best of our knowledge,* ReCon *is the first approach to extract visual concepts in T2I tasks for retrieval-based acceleration.* **Contributions. 1)** We introduce ReCon, first of its kind to extract concepts for accelerating T2I tasks without the need for additional training. By retrieving concepts similar to the input prompt, ReCon crafts a flexible trajectory that closely aligns with the input prompt, allowing for the incorporation of unique details. **2)** We conduct a comprehensive experimental study to compare ReCon with both text-based (Baseline) and noise-based (ReDi) [58] retrieval frameworks across three datasets: MS-COCO [7], Pick-a-Pick [23], and DiffusionDB [55], using various metrics, such as CLIP Score [39], PickScore [23], and Aesthetic Score [1]. Our findings reveal that while text-based retrieval is surprisingly robust relative to noise-based retrieval techniques (namely ReDi [58]), especially with SDXL's high-fidelity model, ReCon consistently excels in all measured dimensions. Specifically, ReCon excels in both fidelity and speed, generating high CLIP Score images in just 6–7 s on an A10G GPU—faster than REDI and Baseline, which require more than 10 s. **3)** Our contribution to the community extends beyond novel methodologies; we have curated a comprehensive concept knowledge base (KB), encompassing over 1.2M visual concepts and their trajectories derived from 1.6M unique prompts across varied datasets. This repository will be made publicly available (upon acceptance) to facilitate further research in the nascent field of retrieval-based T2I acceleration.

2 Related Work

Training-Based Acceleration. Several attempts have been made to accelerate the sampling process of diffusion models by introducing additional training beyond the base diffusion model. Distillation methods [14,30,34,35,44,63] aim to distill a teacher model into a more efficient student model with fewer denoising steps. Consistency Models [49,50], on the other hand, leverage the inherent properties of ODE samplers within diffusion models to train a new model by minimizing the difference between points along identical trajectories. Despite being able to drastically reduce generation steps (potentially to a single step), these acceleration methods require additional training. For larger base models with billions of parameters, such as SDXL [38], used in our paper, the requisite for training resources becomes even more pronounced. In contrast, our training-free, retrieval-based method can be easily implemented with a single A10G GPU.

Training-Free Acceleration. Since the introduction of interpreting diffusion models in continuous time [51], the sampling process can be executed by solving reverse SDEs or ODEs, given that the time-dependent distribution is equivalent. This enables a broad spectrum of research to investigate better numerical solvers for the reverse differential equation [22,24,33,60]. Consequently, the total number of required steps (and therefore, NFEs) can be significantly reduced, often to just 10–20. Our contribution, while complementary to the efforts in optimizing numerical solvers, introduces an additional layer of efficiency by exploiting the concept of caching and retrieving intermediate noises. This approach is orthogonal to the direct optimization of solvers because it does not alter the computational process of the samplers. Instead, it leverages the pre-generated data to bypass certain steps of the diffusion process. By caching intermediate noises we can quickly retrieve and resume the generation process from these points for new images. This method is compatible with any recent sampler, providing a versatile and powerful tool for acceleration.

Retrieval-Based Diffusion. Current retrieval-based T2I studies typically target generating images of rare entities [6] or adapting to new scenarios [45]. In the training-free acceleration direction, the closest work to ours is ReDi [58], where the authors introduced a training-free method for image generation that leverages initial noise to retrieve the most similar generation trajectory. Our approach differs fundamentally from ReDi by *concurrently* retrieving noises from *multiple* semantically similar concept prompts without the need for initial diffusion steps to obtain the noise vector, thus avoiding additional diffusion steps. Moreover, our study, using the enhanced SDXL model [38] alongside multiple datasets featuring complex prompts, demonstrates a surprising finding: simple text-based retrieval (Baseline in Fig. 2) outperform ReDi when tested across extended models and datasets. Nonetheless, both ReDi and the baseline approach fall short in fully capturing the intricacies of the input prompts, prompting the exploration of a more effective concept-based retrieval method.

Concept Retrieval for T2I. Several studies have explored synthesizing images with user-provided concepts. These concepts can be either textual concepts [29] or image concepts [8,25,42] to generate personalized or compositionally complex images. This exploration includes methods for customizing images with specific concept images [8,25,42] and composing multiple text-conditioned diffusion models for intricate image compositions [29]. Our work differs in two significant ways. *First*, we automatically extract visual concepts from text prompts instead of relying on user-provided concepts. *Second*, we cache and retrieve these automatically derived concepts alongside their corresponding noise vectors for accelerating DM. To achieve automatic extraction of visual concepts, we mainly draw inspiration from existing NLP research on word and phrase abstractness and imageability [5,19,32,47]. More recently, Verma *et al.* [52] proposed a method for evaluating prompt-level visualness in T2I, which is more relevant for our case. Inspired by this, we accomplish automatic extraction of visual concept prompts by adopting a fine-tuned CLIP [39] to gauge the text-visualness [52]. Using this, we mine around 1.2M unique concept texts, which we later use for retrieval-based acceleration. To the best of our knowledge, this is the first attempt to extract visual concepts specifically for accelerating T2I tasks.

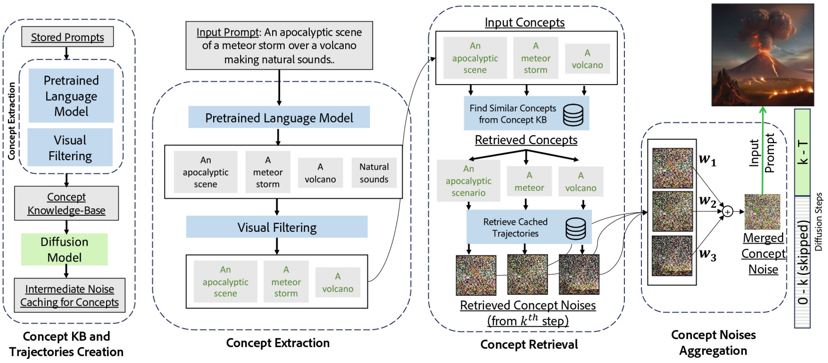


Fig. 3. Overview of ReCon. For a given input prompt, ReCon decomposes it into distinct visual concepts Our framework then stores unique concepts from known prompts as well as their trajectories for retrieval purpose. During inference, the same concept decomposition happens and the most similar concepts along with their linked noises are retrieved. These concept noise vectors are then aggregated at step k to resume the denoising process from $t = k$ to T, resulting in decreasing k NFEs.

3 Approach

3.1 Preliminaries

Diffusion Models. We approach diffusion models (DMs) through the lens of differential equations [22,51], focusing on the "trajectory" idea within the diffusion process. Let $\mathbf{x} \sim p_{data}(\mathbf{x})$ represent the data distribution. DMs perturb $\mathbf{x}$ by iteratively adding i.i.d. Gaussian noise with variance σ_t^2 over $t \in [T, 0]$, resulting in a noise distribution $p(\mathbf{x}_0; \sigma_0)$. When σ_0^2 reaches a sufficiently large value σ_{max}^2, the distribution $p(\mathbf{x}_0; \sigma_0^2)$ closely resembles isotropic Gaussian noise with variance σ_{max}^2. Image generation in DMs starts with a noise-saturated sample $\mathbf{x}_0$ and progressively denoises to approximate $p(\mathbf{x}_t; \sigma_t)$, where t ranges from 0 to T. In this progression, σ_{t+1} is smaller than σ_t, and σ_0 is equal to σ_{max}. The final output at $\sigma_T = 0$, should ideally reflect the original data distribution. This iterative denoising procedure is computationally represented by the Probability Flow ODE [51], which tracks the data sample trajectory over time:

$$d\mathbf{x} = -\bar{\sigma}(t)\sigma(t)\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma(t))dt, \tag{1}$$

where $\nabla_{\mathbf{x}} \log p(\mathbf{x}; \sigma)$ denotes the score function, and $\sigma(t)$ is a user-defined noise-level schedule with its time derivative represented by $\bar{\sigma}(t)$. In practice, the score function is estimated by training a deep neural network $\mathbf{s}_\theta(\mathbf{x}; \sigma)$ parameterized by θ. Note that we are using the same notation as Karras *et al.* [22], where $\mathbf{x}_0$ is the noisy input and $\mathbf{x}_T$ is the fully denoised output or generated final image. Throughout this paper, we refer to the trajectory $\{\mathbf{x}_t : t \in \{0 \cdots T\}\}$ as the one generated by a conditioned DM (typically using classifier-free guidance [18] for T2I), where $\mathbf{x}_t$ represents the noisy model outputs (noises) at each time step.

3.2 Retrieval-Based Acceleration

From Eq. (1), we can see that the sampling process becomes deterministic once $\mathbf{x}_0$ is sampled from the Gaussian distribution. This allows for the pre-generation and storage of trajectories corresponding to a diverse range of prompts within a database. During inference, the key challenge lies in retrieving the trajectory that most closely aligns with the trajectory of the input prompt, posing the critical question: *How do we select the optimal trajectory to resume the denoising process?* To this end, we explore two primary approaches from literature, which we describe below.

Text-Based Trajectory Retrieval. Since the input prompt is the initial element, it is intuitive to retrieve trajectories associated with the closest prompts. Specifically, for an input prompt $\mathcal{P}$, we encode it with a pre-trained text encoder to obtain its embedding $\phi(\mathcal{P})$. The nearest prompt $\mathcal{P}'$ is identified by finding the minimum cosine distance between text embeddings, and its trajectory $\{\mathbf{x}_t^{P'}\}_{t=0}^T$ is obtained. The intermediate output of $\mathcal{P}'$ at step k is looked up in the database and then, the diffusion process is resumed from the intermediate step $k \in [0, T]$

with the original prompt $\mathcal{P}$, effectively skipping k NFEs, as illustrated in Fig. 2. We refer to this method as Baseline.

Noise-Based Trajectory Retrieval. The alternative method, ReDi, retrieves trajectories by searching for the closest noise vector [58]. Specifically, for a given input prompt $\mathcal{P}$, ReDi first computes h steps in the denoising process and uses the noise $\mathbf{x}_h^{\mathcal{P}}$ as a query. It then searches for the nearest noise $\mathbf{x}_h^{P'}$ in the knowledge base and resumes the denoising process from step $k > h$, effectively skipping $k - h$ steps. However, this method adds h initial steps to the number of NFEs, as illustrated in Fig. 2.

3.3 Our Proposed Framework: ReCon

Both Baseline and ReDi rely on stored prompts $\mathcal{P}'$, either directly through text similarity or indirectly through noise similarity. However, they face challenges in generating images that are faithful to the input prompt $\mathcal{P}$. As shown in Fig. 2, both the baseline method and ReDi introduce extraneous components into the final images, compromising its fidelity. Crucial image attributes, such as layout and color schemes, are defined in the initial denoising stages [15], which are frequently bypassed in retrieval-based approaches, leading to a misalignment from the intended prompt details.

Concept Extraction and KB Creation. To address the challenge of incorporating specific details from input prompts while avoiding the rigidity seen in Baseline and ReDi, we introduce a novel approach: decomposing input prompts into visual concepts. However, the challenge lies in distinguishing visual concepts from abstract terms. For example, terms like "belief," "essentialness," and "spirituality," while conceptually significant, often rank high in abstractness [5], but do not translate into visual representations for T2I generation. This distinction is crucial as language models like RoBERTa [9,31] are proficient in extracting and understanding text, including abstract concepts, the challenge lies in ensuring that these concepts can be visually represented. To tackle this challenge, we adopt a fine-tuned CLIP model to evaluate the visual potential of extracted phrases, ensuring that only those with a clear visual representation are selected [52]. In the process of ReCon, each text prompt $\mathcal{P}$ is decomposed into N visual concepts $\{c_i\}_{i=1}^N$, using a pre-trained RoBERTa [31]. Subsequently, a visualness filtering model $h(\cdot)$, fine-tuned to distinguish between visually representable and non-visual text segments, evaluates these extracted noun phrases. representations. The visualness of a concept c is quantified through cosine similarity $\mathcal{S} = 1 - \langle h(c), h(\mathcal{I}_{null}) \rangle$, where $\mathcal{I}_{null}$ denotes an image of random RGB values, serving as a sample for non-visual content. Concepts with visualness scores $\mathcal{S}$ below a predefined threshold are deemed non-visual and excluded from further consideration. These visually relevant concepts and their trajectories from DMs are then cached in a vector database for retrieval purpose. Figure 3 illustrates these steps in this stage of ReCon under the section "Concept KB and Trajectories Creation."

Retrieving and Aggregating Concepts. During inference, given an input prompt $\mathcal{P}$, a concept knowledge base, and a pre-trained diffusion model f_θ, we first decompose the prompt into a set of visual concepts $\{c_i\}_{i=1}^N$. For each concept c_i, we compute its text embedding $\phi(c_i)$, which serves as a query for retrieving the closest matching concepts c'_i along with their associated trajectories $\mathbf{x}_t^{c'_i}$. Subsequently, at a pre-determined step $k \in [0, T]$, we aggregate these trajectories by combining their noise vectors $\mathbf{x}_k^{c'_i}$ at step k, thus forming a new starting point for the denoising process:

$$\hat{\mathbf{x}}_t = \sum_{i=1}^{N} w_i \cdot \mathbf{x}_k^{c'_i}, \tag{2}$$

$$\mathbf{x}_{t+1} = f_\theta(\hat{\mathbf{x}}_t; t, \mathcal{P}); t = k, \cdots, T-1, \tag{3}$$

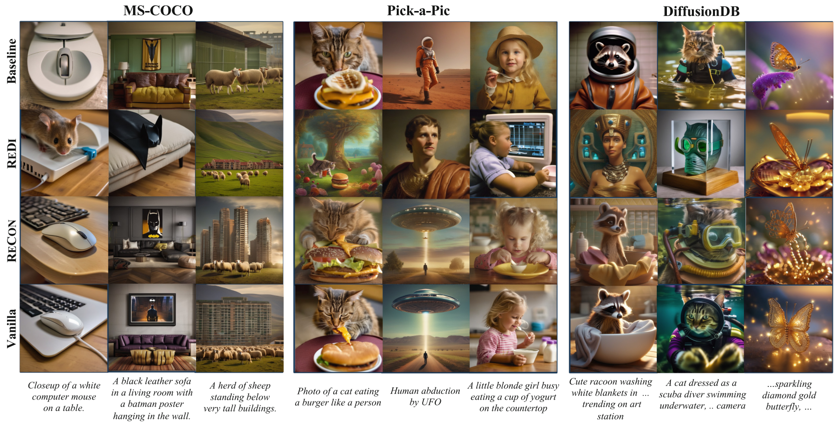

MS-COCO
Pick-a-Pic
DiffusionDB
Baseline
REDI
RECON
Vanilla
Closeup of a white computer mouse on a table.
A black leather sofa in a living room with a batman poster hanging in the wall.
A herd of sheep standing below very tall buildings.
Photo of a cat eating a burger like a person
Human abduction by UFO
A little blonde girl busy eating a cup of yogurt on the countertop
Cute racoon washing white blankets in ... trending on art station
A cat dressed as a scuba diver swimming underwater, .. camera
...sparkling diamond gold butterfly, ...

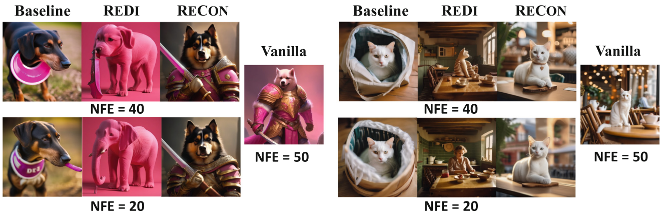
Fig. 4. Qualitative comparisons reveal that ReCon consistently produces images of quality comparable to those generated by Vanilla, with no retrieval-based acceleration. Here, the numbers of NFEs are: Vanilla = 50, Others = 35. Full prompts for DiffusionDB appear in *Supplementary.*

where $\{w_i : i \in [1, N]\}$ corresponds to the weights for each concept. Our method defines weights w_i using the cosine similarity between text embeddings of the query and retrieved concepts, normalized across all concepts: $w_i = \frac{\langle \phi(c_i), \phi(c'_i) \rangle}{\sum_{i=1}^{N} \langle \phi(c_i), \phi(c'_i) \rangle}$. This approach is grounded in the principle that diffusion models can be regarded as energy-based models (EBMs), facilitating the compositional aggregation of noise vectors [12, 28, 29] For detailed insights into noise vector compositionality in EBMs, please refer to our *Supplementary material.* The "Concept Retrieval" and "Concept Noises Aggregation" stages precede the continuation of the denoising process with the aggregated noise $\hat{\mathbf{x}}_t$ and the original prompt $\mathcal{P}$ to generate the final image $\mathbf{x}_T$, as depicted in Fig. 3. While only the concept noun phrases are aggregated, the final diffusion steps (steps $k \cdots T$ in Fig. 3) are guided by the original prompt. Thus, essential prompt details beyond nouns still get to influence the resulting image $\mathbf{x}_T$. For instance, in col. 3 of Fig. 4, although the concepts extracted from the original prompt "A herd of sheep standing below very tall buildings" are "a herd of sheep" and "buildings" without including "tall," the final image still captures this key attribute.

4 Experiments

4.1 Setup

Datasets. For a fair comparison, we first assess the performance of ReCon using the same dataset as ReDi [58]: MS-COCO [7] (containing 82K and 42K captions in train and validation). However, captions in MS-COCO may not always reflect real-world T2I prompts, as noted in [23]. Therefore, we also use Pick-a-Pic [23] (35K unique prompts) and DiffusionDB [55] (1.8M unique prompts). These datasets offer a broader range of prompts, more reflective of actual T2I user preferences. For evaluation, we randomly select 10K captions from MS-COCO, 500 prompts from Pick-a-Pic, and 10K from DiffusionDB from their validation/test set (detailed in *Supplementary*). To our knowledge, such a comprehensive study has not been done before for retrieval-based acceleration.

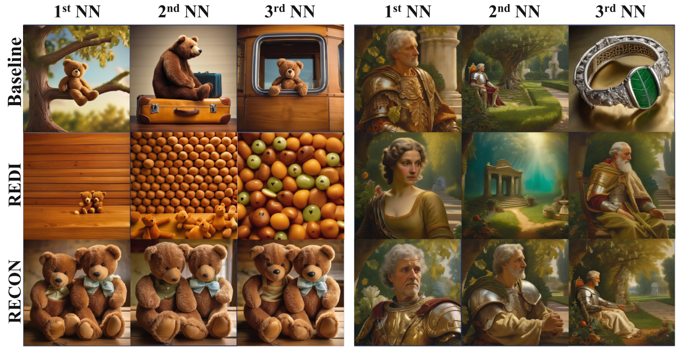


A dog warrior holding a sword in pink armor. *A cat is sitting on a wooden café table with décor.*

Fig. 5. While other methods struggle to generate faithful images with relatively higher NFEs, ours (ReCon) can generate faithful images at NFEs as low as 20.

Metrics. In unconditional settings, a popular metric to test photorealism is FID [16], which compares feature distributions from real and generated images. However, FID falls short in conditional settings such as T2I. Studies have pointed out FID is negatively correlated with human preferences [23,38]. Therefore, for a comprehensive evaluation of T2I models, we employ metrics that specifically address the conditional nature of the task. PickScore [23] measures the alignment between text descriptions and generated images. "CLIP Score" [39], on the other hand, provides a broader measure of semantic congruence between the text and generated image, useful for understanding how well the generated image captures the gist of the text. "Aesthetics Score" [1] complements these by assessing the visual appeal of images. Finally, we employed NFE to quantify ReCon's time efficiency, independent of hardware configurations.

Implementation Details. For building Baseline and ReDi (described in Sect. 3.2), we follow the steps mentioned in [58], but replace the base model with the improved SDXL [38]. We denote the default sampling process in SDXL without any acceleration as "Vanilla Generation." For ReCon, we first build a concept knowledge base (KB) from the unique captions and prompts in the training sets of all considered datasets. We extract noun phrases using a pre-trained language model, RoBERTa [31] and apply visual filtering with a CLIP model [39] from [52]. To efficiently store and retrieve these extracted concepts and their associated trajectories, we implement a vector database with FAISS [20]. Remarkably, the concept retrieval part of our pipeline merely takes 1.15% of the total inference time on an A10G GPU. We build the rest of our pipeline based on PyTorch and the HuggingFace Diffusers library [36,37]. We determine the weights for merging concept noise vectors using the CLIP text similarity scores, as in Eq. (2). Unless otherwise mentioned, for all the experiments, we use the Euler sampler [22] as default, setting the number of NFEs to 35 for retrieval-based methods and 50 for Vanilla Generation. Further implementation details, including configurations of SDXL, the KB size, specifics of our vector database for caching and retrieval are provided in *Supplementary.*



"Two brown teddy bears sitting on top a wooden table."

"renaissance painting of elder in the garden, short silver hair, a wise face, emotions closeup, dressed in roman armour."

Fig. 6. Baseline and ReDi often struggle with $1st$ Nearest Neighbor (NN) results, leading to poorer outcomes for $2nd$ and $3rd$ NNs. Conversely, ReCon consistently produces accurate images across all NN concepts, addressing the diversity challenge.

4.2 Qualitative Results

Enhanced Fidelity. In Fig. 4, Images generated using ReCon demonstrate a higher degree of fidelity to the provided prompts compared to both Baseline and ReDi methods. ReCon images also closely mirror the results of Vanilla Generation across all datasets. In contrast, images from Baseline and ReDi images exhibit notable discrepancies from the input prompt, such as depicting inaccurate objects (animal mouse *vs.* computer mouse), or omitting key elements described in the prompts (e.g., “very tall buildings”). ReCon is able to generate uncommon concept combinations like “..cat eating a burger like a person.”, “..raccoon washing blankets..”, and “..cat dressed as a scuba diver..”. ReDi and Baseline often miss the mark with these intricate descriptions, occasionally generating completely irrelevant images, like those for the prompt “Human abduction by UFO.” ReCon also performs better with rare input prompts. For instance, the phrase “sparkling diamond gold butterfly” is absent from KB. Nevertheless, ReCon produced a relevant image of a butterfly based on the concept “a sparkling gold glitter dress,” thus highlighting the advantages of the flexible trajectory of ReCon’s design. Figure 5 showcases the outputs from three retrieval-based methods with both high and low number of NFEs. Both Baseline and ReDi methods struggle to faithfully capture the input prompts in their images at relatively high number of NFEs. Unsurprisingly, their performance further diminishes at lower number of NFEs. ReCon, however, consistently performs better even with NFEs as low as 20.

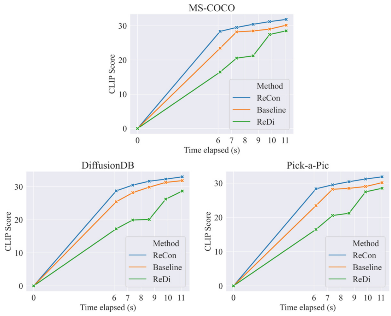


Fig. 7. ReCon can generate images with high CLIP scores in 6–7 s using an A10G GPU, whereas both ReDi and Baseline require more than 10 s to generate images of comparable quality.

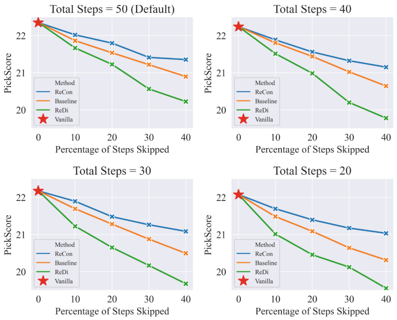


Fig. 8. ReCon consistently outperforms Baseline and ReDi on MS-COCO under varying total number of steps and different percentage of skipped steps w.r.t. total steps taken by Vanilla Generation.

Diversity. Diversity in T2I generation ensures that the model can produce a range of visually distinct images for a given prompt, enhancing user experience by offering creative variations. The traditional approach for diversity in T2I tasks involves caching different trajectories for multiple seeds. This method, while capable of producing diverse outputs, faces significant scalability challenges. Storing trajectories for a growing number of unique prompts, even for a single seed, quickly becomes prohibitively expensive. A cheaper solution to this problem is to pull the top-k Nearest Neighbor (NN) trajectories. However, Baseline and ReDi struggle with accuracy and faithfulness, particularly beyond the first NN, as shown in Fig. 6. ReCon effectively overcomes this by focusing on similar visual concepts rather than full-length prompts. This approach is more efficient as matching concepts is simpler than entire prompts. For instance, in one of the prompts showcased, the top-three retrieved NN concepts are: 1*st*: [‘two brown teddy bears’, ‘large wooden table’], 2*nd*: [‘two teddy bears’, ‘wooden kitchen table’], and 3*rd*: [‘two stuffed teddy bears’, ‘wooden counter’], all closely related to each other. This similarity among the top concepts enables ReCon to maintain diversity without sacrificing fidelity or facing scalability issues.

4.3 Quantitative Results

Comparison of All Methods. In Fig. 8, we show the performance with different percentage of skipped steps with varying the number of total steps (or NFEs) on MS-COCO (results from other datasets are shown in *Supplementary*). ReCon consistently surpasses both ReDi and Baseline across different NFE counts, demonstrating more notable improvement at lower number of NFEs. In Fig. 7, we demonstrate the CLIP Scores of ReCon in comparison to previous works. With just 6 s of computation time on an A10G GPU, ReCon is able to produce accurate images from prompts across various datasets, while both ReDi and Baseline require over 10 s to generate images of similar quality. A detailed study among the retrieval-based methods is presented in Table 1. It is evident that ReCon excels over Baseline and Redi in terms of all metrics. We also make a surprising discovery: *Baseline is quite competitive compared to the state-of-the-art* ReDi. This shows that when using an improved base model, a simple approach like the Baseline can perform quite well, especially on T2I alignment scores. For rest of the analysis, we primarily report these two scores (results for the rest are in the Supplementary).

Table 1. Our concept-based retrieval ReCon surpasses Baseline and ReDi across MS-COCO, Pick-a-Pic, and DiffusionDB datasets. NFE is fixed to 35.

	CLIP Score↑			PickScore↑			Aesthetics Score↑		
	COCO	Pick-a-Pic	DiffDB	COCO	Pick-a-Pic	DiffDB	COCO	Pick-a-Pic	DiffDB
Baseline	29.15	27.67	29.68	20.74	20.44	19.81	5.50	6.05	6.26
ReDi	18.17	21.80	20.43	19.51	19.03	18.21	5.51	6.06	6.29
ReCon	**30.19**	**29.93**	**31.46**	**21.31**	**20.76**	**20.08**	**5.52**	**6.10**	**6.31**

Table 2. ReCon's ablation study where 'Visual-filtering' leverages text-visualness scores [53], while 'weighting' evaluates the similarity between concepts. 'Top-1' prioritizes the highest scoring concept, while '>0.5' criterion includes only those concepts with similarities above this value.

Visual-filtering	✓	✓	✓	✗	✗	✓
Weighting	✗ (Top-1)	✗ (> 0.5)	✗	✗	✓	✓
CLIP Score ↑	27.46	29.79	29.33	28.98	29.06	**29.93**
PickScore ↑	19.88	20.59	20.54	20.12	20.36	**20.84**

Table 3. Comparison of all methods for different samplers. ReCon consistently outperforms Baseline and ReDi across different diffusion samplers, showcasing ReCon's ability to work with multiple efficient numerical solvers.

Scheduler	CLIP Score↑				PickScore↑			
	Euler	PNDM	DDIM	DPM	Euler	PNDM	DDIM	DPM
Baseline	27.67	25.96	26.77	26.73	20.44	19.82	20.10	20.00
ReDi	21.80	20.24	21.68	21.42	19.03	18.08	18.47	18.40
ReCon	**29.93**	**27.68**	**28.99**	**28.56**	**20.76**	**19.89**	**20.29**	**20.06**

Aggregation Ablation Study. Table 2 shows the importance of using both weighted aggregation Eq. (2) and visual-filtering in ReCon. Here, we compare with using only the most similar concept, denoted as Top-1, and excluding low-similarity concepts with a threshold of 0.5, denoted as (> 0.5). The results indicate that both the weighting and visualness filter are essential.

Concept Retrieval Analysis. Although both Baseline and ReCon rely on text similarity scores (either prompt-to-prompt in Baseline or concept-to-concept in ReCon), Table 4 shows that the latter consistently outperforms the former. At each similarity score interval, concept-based retrieval in ReCon leads to superior results, proving the effectiveness of our approach. The length and complexity of input prompts, often containing multiple concepts, poses the question: how many concepts to break a prompt into? Fig. 9 shows that consistently deconstructing prompts into concepts (ReCon) generally produces better outcomes than not doing so (Baseline). Also, it shows that there is a sweet spot for the number of concepts to decompose a text prompt into.

Table 4. ReCon consistently performs better than Baseline at every text similarity score interval and every NFE count (results shown for PickScore↑).

Text Similarity	NFE = 40		NFE = 35		NFE = 30		NFE = 25	
	Baseline	ReCon	Baseline	ReCon	Baseline	ReCon	Baseline	ReCon
0.5–0.6	20.38	**21.07**	20.11	**20.92**	19.95	**20.52**	19.62	**20.18**
0.6–0.7	20.51	**20.98**	20.07	**20.59**	19.86	**20.20**	19.53	**20.04**
0.7–0.8	20.77	**20.96**	20.44	**20.64**	20.25	**20.36**	19.92	**20.16**
0.8–0.9	21.26	**21.42**	20.97	**21.16**	20.83	**20.89**	20.43	**20.65**

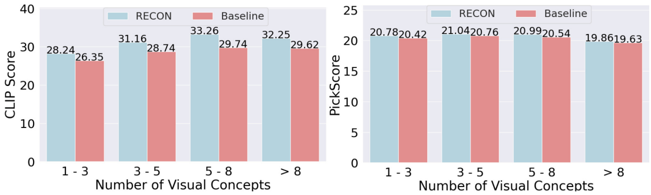


Fig. 9. Prompts can contain an arbitrary number of concepts; so, breaking down a prompt into its constituent concepts is better (ReCon) than not doing so (Baseline).

Comparison with Different Samplers. Table 3 showcases the results of incorporating our method into other diffusion samplers, including DDIM [48], DPM-Solver [33], PNDM [22], and Euler [22], with ReCon outperforming other retrieval-based techniques across different samplers.

Human Evaluation. We conducted a comprehensive user study involving 214 participants (details provided in *Supplementary*) to evaluate image generation using ReCon, Baseline, and ReDi. After filtering for consistency, we retained 200 valid participants. All methods used the same random seed, steps (NFE = 35), and Euler sampler. Results showed ReCon's superior performance, with **75.5%** of its generated images ranked first, 16% second, and 8.5% third. This preference for ReCon was statistically significant (p-value < 0.05).

5 Conclusion

We proposed ReCon, a training-free retrieval-based acceleration method for T2I tasks. We observed that previous approaches were hindered by the rigidity of their trajectories, which often resulted in images lacking fidelity and diversity. To address these challenges, ReCon retrieves trajectories for the noise by using concept similarity and then merges the noises derived from multiple semantically similar concepts. The visual concepts extracted for this process have broad applicability. While we focus on retrieval-based acceleration in this paper, the potential uses of these concepts can be applied to noise aesthetics classifications [13], noise inversion techniques [61], and latent-to-latent transformations within generative models [4]. We leave the exploration of our concept knowledge base as future work.

Acknowledgement. This material is based in part upon work supported by the National Science Foundation under Grant Number CNS-2146449 (NSF CAREER award), gift funding from Adobe Research, and by the Army Research Lab under contract number W911NF-2020-221. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

1. christophschuhmann/improved-aesthetic-predictor: CLIP+MLP aesthetic score predictor. https://github.com/christophschuhmann/improved-aesthetic-predictor
2. Dall-e 2 (2023). https://openai.com/dall-e-2
3. Agarwal, S., Mitra, S., Chakraborty, S., Karanam, S., Mukherjee, K., Saini, S.K.: Approximate caching for efficiently serving text-to-image diffusion models. In: 21st USENIX Symposium on Networked Systems Design and Implementation (NSDI 2024), pp. 1173–1189. USENIX Association, Santa Clara (2024). https://www.usenix.org/conference/nsdi24/presentation/agarwal-shubham
4. Bojanowski, P., Joulin, A., Lopez-Paz, D., Szlam, A.: Optimizing the latent space of generative networks. arXiv preprint arXiv:1707.05776 (2017)
5. Brysbaert, M., Warriner, A.B., Kuperman, V.: Concreteness ratings for 40 thousand generally known English word lemmas. Behav. Res. Methods **46**, 904–911 (2014)
6. Cai, D., Wang, Y., Liu, L., Shi, S.: Recent advances in retrieval-augmented text generation. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 3417–3419 (2022)
7. Chen, X., et al.: Microsoft COCO captions: data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
8. Choi, J., Choi, Y., Kim, Y., Kim, J., Yoon, S.: Custom-edit: text-guided image editing with customized diffusion models. arXiv preprint arXiv:2305.15779 (2023)
9. Chung, H.W., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022)
10. Deng, Y., Wu, N., Qiu, C., Luo, Y., Chen, Y.: MixGAN-TTS: efficient and stable speech synthesis based on diffusion model. IEEE Access **11**, 57674–57682 (2023)
11. Dhariwal, P., Nichol, A.: Diffusion models beat GANs on image synthesis. In: Advances in Neural Information Processing Systems 34, pp. 8780–8794 (2021)
12. Du, Y., Li, S., Mordatch, I.: Compositional visual generation with energy based models. In: Advances in Neural Information Processing Systems 33, pp. 6637–6647 (2020)
13. Gallego, V.: Personalizing text-to-image generation via aesthetic gradients. arXiv preprint arXiv:2209.12330 (2022)
14. Gu, J., Zhai, S., Zhang, Y., Liu, L., Susskind, J.M.: BOOT: data-free distillation of denoising diffusion models with bootstrapping. In: ICML 2023 Workshop on Structured Probabilistic Inference and Generative Modeling (2023)
15. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control (2022)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: Advances in Neural Information Processing Systems 30 (2017)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems 33, pp. 6840–6851 (2020)
18. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
19. Jeong, J.W., Wang, X.J., Lee, D.H.: Towards measuring the visualness of a concept. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 2415–2418 (2012)
20. Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Trans. Big Data **7**(3), 535–547 (2019)

21. Karnewar, A., Vedaldi, A., Novotny, D., Mitra, N.J.: HoloDiffusion: training a 3D diffusion model using 2D images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18423–18433 (2023)
22. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: Advances in Neural Information Processing Systems 35, pp. 26565–26577 (2022)
23. Kirstain, Y., Polyak, A., Singer, U., Matiana, S., Penna, J., Levy, O.: Pick-a-Pic: an open dataset of user preferences for text-to-image generation. arXiv preprint arXiv:2305.01569 (2023)
24. Kong, Z., Ping, W.: On fast sampling of diffusion probabilistic models. In: ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models (2021)
25. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1931–1941 (2023)
26. Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. IEEE Trans. Knowl. Data Eng. **34**(1), 50–70 (2020)
27. Liu, J., Li, C., Ren, Y., Chen, F., Zhao, Z.: DiffSinger: singing voice synthesis via shallow diffusion mechanism. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 11020–11028 (2022)
28. Liu, N., Li, S., Du, Y., Tenenbaum, J., Torralba, A.: Learning to compose visual relations. In: Advances in Neural Information Processing Systems 34, pp. 23166–23178 (2021)
29. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13677, pp. 423–439. Springer, Cham (2022). https://doi.org/10.1007/978-3-031-19790-1_26
30. Liu, X., Zhang, X., Ma, J., Peng, J., et al.: InstaFlow: one step is enough for high-quality diffusion-based text-to-image generation. In: The Twelfth International Conference on Learning Representations (2023)
31. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
32. Louis, A., Nenkova, A.: What makes writing great? First experiments on article quality prediction in the science journalism domain. Trans. Assoc. Comput. Linguist. **1**, 341–352 (2013)
33. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: DPM-Solver: a fast ode solver for diffusion probabilistic model sampling in around 10 steps. In: Advances in Neural Information Processing Systems 35, pp. 5775–5787 (2022)
34. Luhman, E., Luhman, T.: Knowledge distillation in iterative generative models for improved sampling speed. arXiv preprint arXiv:2101.02388 (2021)
35. Meng, C., et al.: On distillation of guided diffusion models. In: CVPR (2023)
36. Paszke, A., et al.: PyTorch: an imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32 (2019)
37. von Platen, P., et al.: Diffusers: state-of-the-art diffusion models (2022). https://github.com/huggingface/diffusers
38. Podell, D., et al.: SDXL: improving latent diffusion models for high-resolution image synthesis, pp. 1–13 (2024)
39. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)

40. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125 (2022)
41. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)
42. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: DreamBooth: fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22500–22510 (2023)
43. Saharia, C., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: Advances in Neural Information Processing Systems 35, pp. 36479–36494 (2022)
44. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models (2022)
45. Sheynin, S., et al.: KNN-Diffusion: image generation via large-scale retrieval. arXiv preprint arXiv:2204.02849 (2022)
46. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning, pp. 2256–2265. PMLR (2015)
47. Solovyev, V.: Concreteness/abstractness concept: state of the art. In: Velichkovsky, B.M., Balaban, P.M., Ushakov, V.L. (eds.) Intercognsci 2020. AISC, vol. 1358, pp. 275–283. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-71637-0_33
48. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (2020)
49. Song, Y., Dhariwal, P.: Improved techniques for training consistency models. In: The Twelfth International Conference on Learning Representations (2023)
50. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: Proceedings of Machine Learning Research, vol. 202. PMLR (2023). https://proceedings.mlr.press/v202/song23a.html
51. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (2020)
52. Verma, G., Rossi, R.A., Tensmeyer, C., Gu, J., Nenkova, A.: Learning the visualness of text using large vision-language models. In: The 2023 Conference on Empirical Methods in Natural Language Processing (2023)
53. Verma, G., Rossi, R.A., Tensmeyer, C., Gu, J., Nenkova, A.: Learning the visualness of text using large vision-language models. arXiv preprint arXiv:2305.10434 (2023)
54. Wang, T., et al.: RODIN: a generative model for sculpting 3D digital avatars using diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4563–4573 (2023)
55. Wang, Z.J., Montoya, E., Munechika, D., Yang, H., Hoover, B., Chau, D.H.: DiffusionDB: a large-scale prompt gallery dataset for text-to-image generative models. arXiv preprint arXiv:2210.14896 (2022)
56. Xu, X., Wang, Z., Zhang, G., Wang, K., Shi, H.: Versatile diffusion: text, images and variations all in one diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7754–7765 (2023)
57. Yasuda, Y., Toda, T.: Text-to-speech synthesis based on latent variable conversion using diffusion probabilistic model and variational autoencoder. In: ICASSP 2023-

2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5. IEEE (2023)

58. Zhang, K., Yang, X., Wang, W.Y., Li, L.: ReDi: efficient learning-free diffusion inference via trajectory retrieval. In: Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 41770–41785. PMLR (2023). https://proceedings.mlr.press/v202/zhang23as.html
59. Zhang, Q., Chen, Y.: Fast sampling of diffusion models with exponential integrator. In: NeurIPS 2022 Workshop on Score-Based Methods (2022)
60. Zhang, Q., Chen, Y.: Fast sampling of diffusion models with exponential integrator. In: The Eleventh International Conference on Learning Representations (2023). https://openreview.net/forum?id=Loek7hfb46P
61. Zhang, Y., et al.: Inversion-based style transfer with diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10146–10156 (2023)
62. Zhao, W., Bai, L., Rao, Y., Zhou, J., Lu, J.: UniPC: a unified predictor-corrector framework for fast sampling of diffusion models. arXiv preprint arXiv:2302.04867 (2023)
63. Zheng, H., Nie, W., Vahdat, A., Azizzadenesheli, K., Anandkumar, A.: Fast sampling of diffusion models via operator learning. In: International Conference on Machine Learning, pp. 42390–42402. PMLR (2023)
64. Zhou, L., Du, Y., Wu, J.: 3D shape generation and completion through point-voxel diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5826–5835 (2021)
65. Zhou, Y., Liu, B., Zhu, Y., Yang, X., Chen, C., Xu, J.: Shifted diffusion for text-to-image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10157–10166 (2023)