# ERUDIFF: Refactoring Knowledge in Diffusion Models for Advanced Text-to-Image Synthesis

Xiefan Guo[1,2,4] Xinzhu Ma[2,4] Haoxiang Ma[4]
Zihao Zhou[1,2] Di Huang[1,2,3⋆]

[1]State Key Laboratory of Complex and Critical Software Environment, Beihang University, Beijing, China
[2]School of Computer Science and Engineering, Beihang University, Beijing, China
[3]Shenzhen Loop Area Institute, Shenzhen, China
[4]Shanghai Artificial Intelligence Laboratory, Shanghai, China
{xfguo,dhuang}@buaa.edu.cn

**Abstract.** Text-to-image diffusion models have achieved remarkable fidelity in synthesizing images from explicit text prompts, yet exhibit a critical deficiency in processing implicit prompts that require deep-level world knowledge, ranging from natural sciences to cultural commonsense, resulting in counter-factual synthesis. This paper traces the root of this limitation to a fundamental dislocation of the underlying knowledge structures, manifesting as a chaotic organization of implicit prompts compared to their explicit counterparts. In this paper, we propose ERUDIFF, which aims to refactor the knowledge within diffusion models. Specifically, we develop the Diffusion Knowledge Distribution Matching (DK-DM) to register the knowledge distribution of intractable implicit prompts with that of well-defined explicit anchors. Furthermore, to rectify the inherent biases in explicit prompt rendering, we employ the Negative-Only Reinforcement Learning (NO-RL) strategy for fine-grained correction. Rigorous empirical evaluations demonstrate that our method significantly enhances the performance of leading diffusion models, including FLUX and Qwen-Image, across both the scientific knowledge benchmark (*i.e.*, SCIENCE-T2I) and the world knowledge benchmark (*i.e.*, WISE), underscoring the effectiveness and generalizability. Our code is available at `https://github.com/xiefan-guo/erudiff`.

**Keywords:** Diffusion models · Text-to-image synthesis · World knowledge informed image synthesis

## 1 Introduction

Text-to-image diffusion models [2, 9, 16, 42, 47, 60, 67, 81] have demonstrated substantial efficacy in synthesizing high-fidelity, visually-realistic, and diverse images conditioned on text prompts. Recent breakthroughs, characterized by

⋆ Corresponding author.


**Fig. 1: Visual results on WISE.** ERUDIFF exhibits a comprehensive integration of multifaceted world knowledge, spanning natural science to cultural commonsense.

complex compositional generation [6, 20, 27, 28, 33, 50, 56, 66, 85], enhanced aesthetic quality [11, 13, 18, 19, 25, 37, 38, 48, 71, 84, 88], ultra-high-resolution synthesis [8, 31, 63, 86, 94, 98], and efficient architectures and inference acceleration techniques [21, 24, 51, 52, 69, 70, 74, 92, 93], have further pushed the boundaries of the field. Despite these improvements, current models exhibit fundamental limitations in scenarios necessitating structured reasoning grounded in world knowledge [22, 44, 55, 58, 75]. Their proficiency remains largely confined to the precise rendering of explicit prompts involving superficial attributes, such as color, texture, and shape. Conversely, they demonstrate a restricted capacity to process implicit prompts that encompass a broad spectrum of world knowledge, ranging from natural sciences to cultural commonsense. This deficiency frequently results in counter-factual synthesis, as illustrated in Fig. 1.

Current mainstream research paradigms lean towards deep mining the intrinsic world knowledge and reasoning capabilities of Large Language Models (LLMs) to address the cognitive bottlenecks inherent in diffusion models when processing implicit prompts. A representative technical trajectory [41, 54, 62, 78, 80, 83, 91] involves leveraging LLMs to rewrite highly abstract, implicit instructions into granular, explicit descriptions, thereby providing concrete semantic guidance for text-to-image diffusion models to achieve precise synthesis of complex latent meanings. Another frontier seeks to establish unified multimodal understanding and generation frameworks [5, 7, 10, 12, 23, 35, 36, 59, 79, 87, 97], which involve the cross-modal transfer of deep reasoning capabilities from the textual domain, harnessing inferential power from linguistic modalities to enhance the parsing precision and alignment of visual generative models regarding implicit intentions. Despite these advancements, significant unexplored potential remains within the diffusion models.

This paper posits that a fundamental cause of this performance gap lies in the inherent disparity between training data distributions: while comprehensive world knowledge is predominantly distilled within text-only corpora, text-image

pairs are frequently restricted to superficial descriptions of visual appearances. Consequently, we pioneer a novel paradigm that leverages text-only corpora to fine-tune diffusion models, facilitating the internalization of world knowledge.

In this paper, we propose EruDiff (Erudite Diffusion) to facilitate a deep refactoring of the inherent knowledge systems within pre-trained diffusion models. This is achieved by leveraging structured text corpora rich in world knowledge, organized into pairs of *implicit prompts* and *explicit prompts*. Specifically, we develop the Diffusion Knowledge Distribution Matching (DK-DM), which achieves precise alignment of knowledge distributions across two semantic spaces by registering high-abstraction implicit prompts to well-defined explicit anchors. To mitigate catastrophic knowledge forgetting, we introduce an anti-forgetting knowledge consolidation mechanism, complemented by a timestep-aware curriculum learning strategy to effectively accelerate convergence. Furthermore, addressing the inherent representation biases and inaccurate visual rendering associated with explicit prompts, we employ the Negative-Only Reinforcement Learning (NO-RL) for fine-grained correction, leading to improved results.

To address the scarcity of structured training resources despite the abundance of world-knowledge evaluation benchmarks [22,55,58,75], we introduce the Knowledge-10K dataset to serve as the training data support for our method, which will be publicly released to facilitate community research. Extensive empirical evaluations demonstrate that the proposed method significantly enhances the performance of leading diffusion models, including FLUX and Qwen-Image, on the scientific knowledge benchmark (*i.e.*, Science-T2I) and the world knowledge benchmark (*i.e.*, WISE), which validates the superior efficacy and generalization capabilities of our approach.

Our research demonstrates that classical text-to-image diffusion models possess substantial latent cognitive capacity. Systematically refactoring the intrinsic knowledge frameworks unlocks high-dimensional world-knowledge comprehension and superior visual rendering capabilities. This study provides novel insights into the development of next-generation unified multimodal architectures that seamlessly integrate understanding, reasoning, and visual generation.

## 2 Related Work

**Text-to-image diffusion models.** Text-to-image synthesis aims to generate photorealistic images that align precisely with text prompts. Recently, Diffusion Models (DMs) [14, 32] have become a dominant paradigm in generative modeling, showing exceptional performance in diverse applications [26, 34, 39, 40, 61, 67, 72, 77, 95], most notably in text-to-image synthesis. Early landmarks such as GLIDE [57], DALL-E 2 [65], and Imagen [68] explored text-guided refinement, CLIP-based embedding spaces, and cascaded architectures, respectively. Notably, Stable Diffusion (SD) [67] significantly enhanced computational efficiency through latent-space diffusion. To further push the boundaries of generation quality, recent research has pivoted toward architectural innovations, including the construction of flow-based methods, the development of Diffu-

sion Transformers (DiT), the emergence of Multi-Modal Diffusion Transformer (MM-DiT) and the integration of large language models (LLMs). Representative state-of-the-art models include PixArt-$\alpha$ [9], Stable Diffusion 3 [16], FLUX.1 [42], HiDream-I1 [4], Qwen-Image [81], HunyuanImage [5] and Seedream [73].

Nevertheless, these models encounter fundamental limitations in scenarios requiring structured reasoning informed by world knowledge. Their efficacy is largely restricted to the high-fidelity rendering of explicit prompts concerning surface attributes. In contrast, they exhibit a diminished capacity for addressing implicit prompts rooted in extensive world knowledge, frequently yielding counter-factual synthesis.

**Prompt rewriting for diffusion models.** Prompt rewriting aims to augment initial instructions by leveraging the semantic modeling capabilities of Large Language Models (LLMs) before they are passed to a frozen text-to-image (T2I) model, thereby enhancing the quality of visual synthesis. Recent studies [62, 80, 83, 91] utilize LLMs to iteratively critique and refine prompts within a closed-loop system, while others [29, 41, 78, 82] introduced visual feedback signals as rewards to specifically fine-tune LLM-based prompt rewriters. Despite these advancements, the potential of diffusion models remains substantially under-explored. Our research demonstrates that classical text-to-image diffusion models harbor significant latent cognitive capacities that can be effectively unlocked through meticulously designed knowledge refactoring.

**RL for diffusion models.** As a direct intervention strategy to enhance the performance of pre-trained diffusion models, Reinforcement Learning (RL) approaches [3, 11, 15, 19, 25, 30, 43, 45, 46, 49, 76, 89, 90, 96] have demonstrated significant efficacy in improving visual quality and text-image alignment. However, its generalizability remains constrained by the scarcity of high-quality reward models and visual preference datasets enriched with extensive world knowledge. More critically, the RL fine-tuning process inevitably induces a distribution shift, posing an inherent risk of semantic degradation and the erosion of pre-trained knowledge. Our approach leverages text-only corpora supplemented by anti-forgetting knowledge consolidation mechanism to effectively mitigate these limitations.

## 3 EruDiff

EruDiff is dedicated to the deep refactoring of the inherent knowledge systems within pre-trained diffusion models. As illustrated in Fig. 2, EruDiff comprises the Diffusion Knowledge Distribution Matching (DK-DM) and the Negative-Only Reinforcement Learning (NO-RL). The former aligns the knowledge distribution of intractable implicit prompts with that of well-defined explicit anchors, while the latter rectifies inherent biases in the rendering of explicit prompts. Detailed expositions of these two components are provided in Sec. 3.1 and 3.2, respectively. Sec. 3.3 outlines our training strategy.

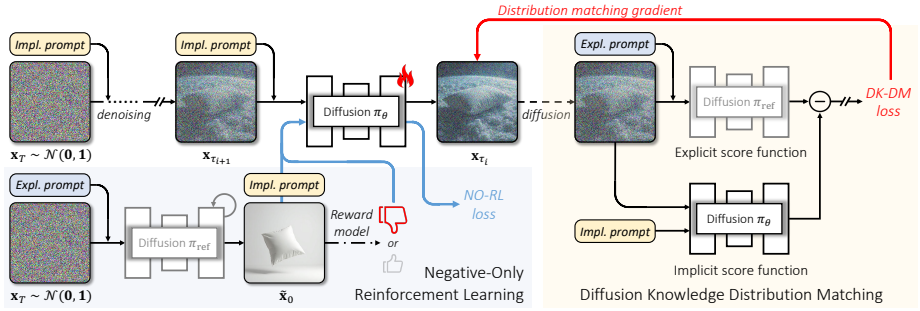


**Fig. 2: Illustration of EruDiff.** EruDiff focuses on the deep refactoring of the inherent knowledge systems within pre-trained text-to-image diffusion models, aiming to enhance factual fidelity when processing implicit prompts that necessitate a profound command of extensive world knowledge.

### 3.1 Diffusion Knowledge Distribution Matching

Inspired by the principles of Distribution Matching Distillation (DMD) [53, 92, 93], DK-DM is implemented by aligning the synthesis of implicit prompts $\mathbf{y}_{\text{impl}}$ to that of explicit prompts $\mathbf{y}_{\text{expl}}$ at the distribution level. This is achieved by minimizing the expectation of the Kullback-Leibler (KL) divergence between the diffused implicit image distribution $p_{\text{impl},t}$ and the diffused explicit image distribution $p_{\text{expl},t}$ over the timestep $t$, both of which are derived from implicit and explicit prompts, respectively:

$$\begin{aligned}\nabla_{\mathcal{L}_{\text{DK-DM}}} &= \mathbb{E}_t \left( \nabla_\theta \mathcal{D}_{\text{KL}} \left( p_{\text{impl},t} \| p_{\text{expl},t} \right) \right) \\ &= -\mathbb{E}_{t,\mathbf{z}} \left[ \left( s_{\text{expl}} \left( \mathcal{F} \left( g_\theta(\mathbf{z}), t \right), t \right) - s_{\text{impl}} \left( \mathcal{F} \left( g_\theta(\mathbf{z}), t \right), t \right) \right) \frac{\partial g_\theta(\mathbf{z})}{\partial \theta} \right],\end{aligned} \tag{1}$$

where $t \sim \mathcal{U}(T_{\min}, T_{\max})$. Following [93], $T_{\min} = 0.02T$ and $T_{\max} = 0.98T$, where $T$ denotes the total number of timesteps in the training phase. $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ represents the random Gaussian noise input. $\mathcal{F}(\cdot, t)$ signifies the forward diffusion process (*i.e.*, noise injection), with the noise level corresponding to the specific timestep $t$. In our implementation, $\mathcal{L}_{\text{DK-DM}}$ is computed by constructing a pseudo-MSE loss using the stop-gradient technique [92, 93]. Furthermore, as illustrated in Fig. 2, we have specifically designed the following components:

– **Image generator** $g_\theta(\cdot)$: The image generator $g_\theta(\cdot)$ is the text-to-image diffusion model undergoing fine-tuning, denoted as $\pi_\theta(\cdot)$, which takes implicit prompts as input and is optimized via gradient backpropagation. Given that $g_\theta(\cdot)$ is a multi-step diffusion model, straightforward backpropagation through the entire denoising trajectory would lead to excessive memory consumption and potential overflow. To mitigate this issue, we employ a gradient truncation strategy, restricting the backpropagation to only the final denoising step.

- **Explicit score function** $s_{\text{expl}}(\cdot)$: The explicit score $s_{\text{expl}}(\cdot)$ denotes the score function of the diffused explicit image distribution, which is provided by the fixed, pre-trained text-to-image diffusion model $\pi_{\text{ref}}(\cdot, \mathbf{y}_{\text{expl}})$ conditioned on explicit prompts. Notably, the image distribution corresponding to the explicit prompts serves as the target distribution in this work.
- **Implicit score function** $s_{\text{impl}}(\cdot)$: The implicit score $s_{\text{impl}}(\cdot)$ denotes the score function of the diffused implicit image distribution. Distinct from [53, 92, 93] that incur additional computational overhead to train a dynamically-learned denoiser, such a denoiser exists inherently within our framework in the form of the text-to-image diffusion model undergoing fine-tuning. Consequently, $s_{\text{impl}}(\cdot)$ is directly provided by the diffusion model undergoing fine-tuning $\pi_\theta(\cdot, \mathbf{y}_{\text{impl}})$ conditioned on implicit prompts.

**Full-timestep distribution matching.** In contrast to single-step [93] and few-step [53] image generators, the latter of which necessitate manual alignment of denoising timesteps with the target diffusion model. In this work, the image generator and the explicit score function share an identical timestep schedule. This inherent consistency naturally motivates the execution of full-timestep distribution matching, which emerges as a more efficient alternative. Specifically, given the inference timestep sequence $(T, \ldots, \tau_{i+1}, \tau_i, \tau_{i-1}, \ldots, 0)$, for each training iteration, rather than exclusively selecting timestep 0 to obtain a fully denoised clean image, we sample an intermediate timestep $\tau_i$ to generate a noisy image, upon which distribution matching is performed.

Since our image generator $g_\theta(\cdot)$ is a multi-step diffusion model, producing a fully denoised image at every iteration is computationally prohibitive. In addition, despite the application of a gradient truncation strategy, in scenarios with a high number of inference steps, imposing constraints solely on the final denoising step exerts a relatively negligible influence on the earlier stages of the denoising chain. Consequently, performing full-timestep distribution matching serves as a natural and effective choice to address these issues. Building upon this rationale, we reformulate $\mathcal{L}_{\text{DK-DM}}$ as follows:

$$\nabla_{\mathcal{L}_{\text{DK-DM}}} = -\mathbb{E}_{t,\tau_i,z}\left[\left(s_{\text{expl}}\left(\mathcal{F}\left(\mathbf{x}_{\tau_i}, t\right), t\right) - s_{\text{impl}}\left(\mathcal{F}\left(\mathbf{x}_{\tau_i}, t\right), t\right)\right)\frac{\partial \mathbf{x}_{\tau_i}}{\partial \theta}\right], \quad (2)$$

where $\mathbf{x}_{\tau_i} = g_\theta(\mathbf{z}, \tau_i)$, denoting that the denoising chain is truncated at the inference timestep $\tau_i$. Accordingly, $t \sim \mathcal{U}(T_{\min}, T_{\max})$, with $T_{\min} = \tau_i + 0.02 \times (\tau_{i+1} - \tau_i)$ and $T_{\max} = \tau_i + 0.98 \times (\tau_{i+1} - \tau_i)$.

**Anti-forgetting knowledge consolidation.** Distribution shift is an inevitable and challenging issue encountered during the post-training of text-to-image diffusion models, *e.g.*, reinforcement learning, posing a significant risk of pre-trained knowledge degradation. This risk is particularly pronounced in the context of knowledge refactoring. As illustrated in Fig. 6, a naive refactoring of knowledge within the diffusion model, aiming to enable it to comprehend the world-knowledge fact that "Albert Einstein's favorite musical instruments" is "Violin", unfortuantely results in the degradation of its inherent knowledge of "Albert

Einstein". To mitigate this issue, we introduce an anti-forgetting knowledge consolidation mechanism.

Specifically, we extend the vanilla Diffusion Knowledge Distribution Matching, denoted as DK-DM($\mathbf{y}_{\text{impl}}, \mathbf{y}_{\text{expl}}$), by incorporating an explicit knowledge consolidation term DK-DM($\mathbf{y}_{\text{expl}}, \mathbf{y}_{\text{expl}}$) and a foundational knowledge consolidation term DK-DM($\mathbf{y}_{\text{found}}, \mathbf{y}_{\text{found}}$), where $\mathbf{y}_{\text{found}}$ represents a sub-prompt extracted from the implicit prompt, specifically designed to eliminate requirements for world-knowledge reasoning. Furthermore, to enhance generalizability across diverse and complex prompts while reducing implementation difficulty, we extract noun phrases from the implicit prompts as $\mathbf{y}_{\text{found}}$, representing a more pragmatic and cost-effective alternative. For the aforementioned example, $\mathbf{y}_{\text{found}} \in$ {"Albert Einstein", "musical instruments"}. More details regarding the extraction of $\mathbf{y}_{\text{found}}$ are provided in the supplementary material.

Empirically, executing the anti-forgetting knowledge consolidation at a lower frequency is sufficient to effectively mitigate the loss of pre-trained knowledge. Specifically, during each training iteration, we execute DK-DM($\mathbf{y}_{\text{impl}}, \mathbf{y}_{\text{expl}}$), DK-DM($\mathbf{y}_{\text{expl}}, \mathbf{y}_{\text{expl}}$), and DK-DM($\mathbf{y}_{\text{found}}, \mathbf{y}_{\text{found}}$) with probabilities $p_{\text{impl}}$, $p_{\text{expl}}$, and $p_{\text{found}}$, respectively. We set $p_{\text{impl}} = 0.8$, $p_{\text{expl}} = 0.1$, and $p_{\text{found}} = 0.1$.

**Timestep-aware curriculum learning.** Within our framework, while uniformly sampling $\tau_i$ from the inference timestep set $\{T, \ldots, \tau_{i+1}, \tau_i, \tau_{i-1}, \ldots, 0\}$ is a viable approach for full-timestep distribution matching, it overlooks timestep-aware characteristic, leading to suboptimal convergence efficiency. In practice, we observe that the early stages of the denoising inference chain predominantly govern the overall distribution matching. This dominance arises not only from their inherently stronger prompt-dependency [1] but also from the cascading impact on the matching of subsequent timesteps. Consequently, we employ a timestep-aware curriculum learning strategy to prioritize these timesteps:

$$p(n) = \frac{\exp(-\lambda(n-1))}{\sum_{i=1}^{T_{\text{inference}}} \exp(-\lambda(i-1))}, \tag{3}$$

where $n$ is the number of inference steps performed, $n \in \{1, 2, \ldots, T_{\text{inference}}\}$, $T_{\text{inference}}$ represents the total number of timesteps in the inference phase, and $p(n)$ represents the probability that the image generator $g_\theta(\cdot)$ performs only the first $n$ denoising operations within the current training iteration. $\lambda$ represents the decay coefficient, which is set to $\lambda = 0.1$ in this work. Values in the range of 0.05 to 0.2 represent moderate and effective choices.

### 3.2 Negative-Only Reinforcement Learning

While state-of-the-art text-to-image diffusion models demonstrate a strong command of explicit prompts, they inevitably exhibit inherent representation biases and inaccurate visual rendering to some extent (see Fig. 5). To address these limitations, we employ Negative-Only Reinforcement Learning (NO-RL) for fine-grained correction, thereby achieving superior generative outcomes.

In this work, NO-RL is implemented as a variant of Kahneman-Tversky Optimization (KTO) [17], which offers the distinct advantage of bypassing the need

for costly paired preference data, instead, it aligns the model with human preferences using only binary feedback signals, making it inherently compatible with our problem formulation. As illustrated in Fig. 2, we exclusively leverage the failure sample set $\mathcal{D}_{\text{loss}}$, derived from explicit prompt synthesis, for the NO-RL process. Specifically, following the methodology in [46], the image generator $g_\theta(\cdot)$ is further optimized using the following objective function:

$$\max_{\pi_\theta} \mathbb{E}_{\tilde{\mathbf{x}}_0,t}\left[U\left(-\left(\beta \log \frac{\pi_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\pi_{\text{ref}}(\mathbf{x}_{t-1}|\mathbf{x}_t)} - Q_{\text{ref}}\right)\right)\right], \tag{4}$$

where $\tilde{\mathbf{x}}_0 \in \mathcal{D}_{\text{loss}}$, $t \sim \mathcal{U}[0,T]$, $\mathbf{x}_t = \mathcal{F}(\tilde{\mathbf{x}}_0, t)$. $\pi_*(x_{t-1}|x_t)$ represents a denoising sampling step. $U(\cdot)$ denotes a monotonically increasing value function that maps implicit rewards to subjective utilities, which is instantiated as a sigmoid function in this work. The reference point $Q_{\text{ref}}$ is computed by evaluating the term $\max(0, \frac{1}{m}\sum \beta \log \frac{\pi_\theta(\mathbf{x}'_{t-1}|\mathbf{x}'_t)}{\pi_{\text{ref}}(\mathbf{x}'_{t-1}|\mathbf{x}'_t)})$ over a batch of $m$ unrelated $(\mathbf{x}'_{t-1}, \mathbf{x}'_t)$ pairs.

NO-RL effectively reshapes the target distribution by excluding the distribution space associated with failure samples. Given that the positive distribution representation is already sufficiently captured by DK-DM, NO-RL adopts a more efficient strategy by focusing exclusively on constraints derived from negative samples, the integration of redundant positive sample learning into the existing NO-RL framework yields no further performance gains. Further methodological details, including the filtering strategy for the failure sample set $\mathcal{D}_{\text{loss}}$, are provided in the supplementary material.

### 3.3 Training strategy

As illustrated in Fig. 2, the optimization objective of our method comprises two components: DK-DM and NO-RL. Although both components aim to optimize the image generator $g_\theta(\cdot)$, they do not share the same network forward pass within the computational graph, distinguishing them from standard multi-objective joint training paradigms. Consequently, we perform two independent backpropagation steps to facilitate the learning

This strategy effectively reduces peak memory consumption without incurring additional computational overhead from redundant forward passes. Furthermore, to balance the two optimization objectives that operate at different scales, we implement an exponential moving average normalization (EMAN) mechanism with a decay coefficient of 0.99 for both loss functions to ensure training robustness and convergence stability. Algorithm 1 outlines the final training procedure.

## 4 Knowledge-10K

Despite the prevalence of world knowledge evaluation benchmarks, there remains a critical lack of structured training resources. To address this deficiency, we introduce the Knowledge-10K dataset for training support, which will be made available to the public to foster further investigation within the community.

**Algorithm 1** ERUDIFF

**Require:** Pre-trained text-to-image diffusion model weights $\theta$, prompt dataset $\mathcal{Y}$, sampling priors $\{p_{\text{impl}}, p_{\text{expl}}, p_{\text{found}}\}$, failure sample set $\mathcal{D}_{\text{loss}}$.

**while** not converged **do**
  *// Phase I: Diffusion Knowledge Distribution Matching (DK-DM)*
  Sample prompt category $c \in \{\text{impl, expl, found}\}$ with probability $p_c$;
  Sample prompt triplet $(\mathbf{y}_{\text{impl}}, \mathbf{y}_{\text{expl}}, \mathbf{y}_{\text{found}}) \in \mathcal{Y}$;
  Construct corresponding prompt pair;
    **if** $c$ **is** impl **then** $(\mathbf{y}_{\text{impl}}, \mathbf{y}_{\text{expl}}) \leftarrow (\mathbf{y}_{\text{impl}}, \mathbf{y}_{\text{expl}})$;
    **if** $c$ **is** expl **then** $(\mathbf{y}_{\text{impl}}, \mathbf{y}_{\text{expl}}) \leftarrow (\mathbf{y}_{\text{expl}}, \mathbf{y}_{\text{expl}})$;
    **if** $c$ **is** found **then** $(\mathbf{y}_{\text{impl}}, \mathbf{y}_{\text{expl}}) \leftarrow (\mathbf{y}_{\text{found}}, \mathbf{y}_{\text{found}})$;
  Sample Gaussian noise $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ and step count $n \sim p(n)$ using Eq. 3;
  Execute $n$ denoising steps to obtain truncated latent $\mathbf{x}_{\tau_i} = g_\theta(\mathbf{z}, \tau_i)$;
  Sample $t \sim \mathcal{U}(T_{\min}, T_{\max})$ relative to interval $[\tau_i, \tau_{i+1}]$;
  Calculate $\mathcal{L}_{\text{DK-DM}}$ using Eq. 2;
  Perform EMAN: $\mathcal{L}_{\text{DK-DM}} \leftarrow \text{EMAN}(\mathcal{L}_{\text{DK-DM}})$;
  Update $\theta \leftarrow \theta - \eta \cdot \nabla_{\mathcal{L}_{\text{DK-DM}}}$;
  *// Phase II: Negative-Only Reinforcement Learning (NO-RL)*
  Sample failure sample $\tilde{\mathbf{x}}_0 \sim \mathcal{D}_{\text{loss}}$ and timestep $t \sim \mathcal{U}[0, T]$;
  Compute $\mathcal{L}_{\text{NO-RL}}$ using Eq. 4;
  Perform EMAN: $\mathcal{L}_{\text{NO-RL}} \leftarrow \text{EMAN}(\mathcal{L}_{\text{NO-RL}})$;
  Update $\theta \leftarrow \theta - \eta \cdot \nabla_{\mathcal{L}_{\text{NO-RL}}}$;
**end while**

**Taxonomy.** Adhering to the design principle of prevailing evaluation benchmark [58], Knowledge-10K encompasses world knowledge across three primary domains: cultural commonsense, spatio-temporal reasoning, and natural science. Cultural commonsense spans diverse fields such as festivals, sports, and craftsmanship, it entails the recognition of traditional customs, ethnic handicrafts, and iconic landmarks, as well as events associated with prominent figures, among other cultural manifestations. Spatio-temporal reasoning integrates both temporal and spatial dimensions, necessitating reasoning regarding chronological relationships, positioning, and perspectives. Natural science incorporates domain-specific expertise across biology, physics, and chemistry.

**Format.** Each entry in the Knowledge-10K dataset primarily consists of an *implicit prompt* and its corresponding *explicit prompt*. The former necessitates the retrieval and reasoning of deep-seated world knowledge, while the latter serves as its intuitive counterpart, which circumvents logical complexity by directly describing the visual content. Samples are provided in the supplementary material.

**Data collection pipeline.** (1) Template customization. Initially, 100 seed samples were manually constructed to establish the specifications for target data characteristics and to define the taxonomy of world knowledge involved. (2) Scale expansion. We adopted a hybrid strategy combining information retrieval with Large Language Models (LLMs) synthesis. The former involves extracting raw information from the Internet and encyclopedic sources, followed by manual rewriting to ensure format alignment. The latter leverages the collabo-

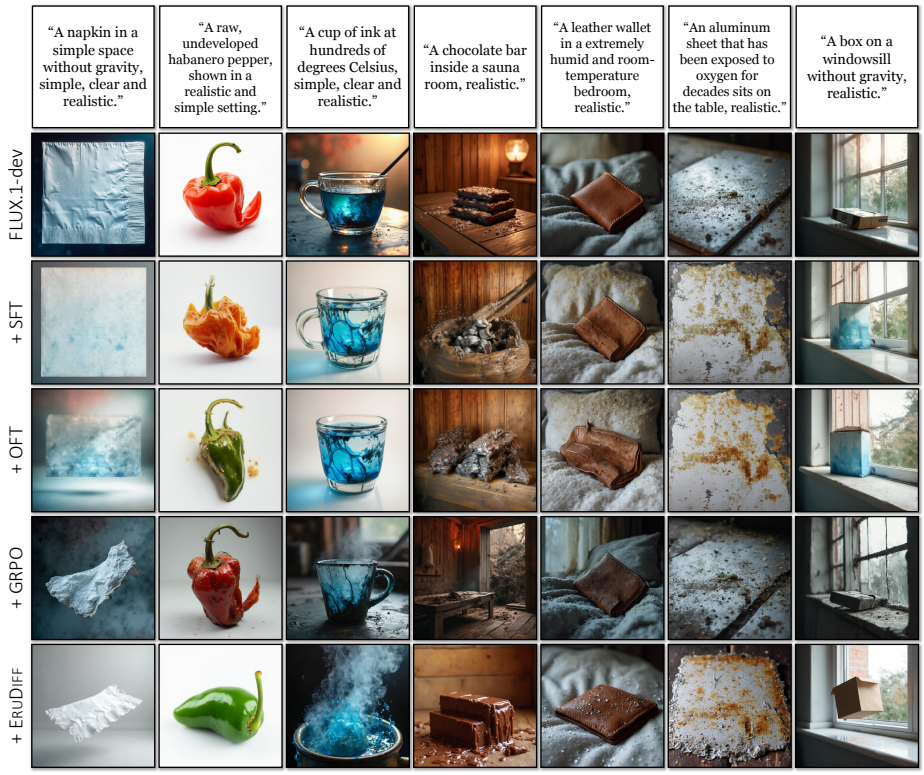


**Fig. 3: Qualitative comparison on SCIENCE-T2I based on FLUX.** Each image is generated with the same text prompt and random seed for all methods. ERUDIFF excels in precise scientific knowledge comprehension and rendering.

rative synergy of Gemini 3, GPT-5, and Grok 4 to synthesize new entries, which mitigates the redundancy and stylistic uniformity inherent in single-model generation. Subsequently, Gemini 3 was employed to perform semantic deduplication, filtering out redundant instances to yield an initial corpus of 10,000 entries. (3) Data review. To mitigate retrieval bias and LLM hallucinations, the more advanced Gemini 3 Pro was introduced for automated proofreading and correction. Finally, five volunteers holding at least a Bachelor's degree in Engineering were invited to conduct a final audit, thereby ensuring the high fidelity and rigorous quality of the dataset.

**Dataset Statistics.** The Knowledge-10K dataset comprises 10,000 data entries. The domains of cultural commonsense, spatio-temporal reasoning, and natural science consist of 4,000, 3,000, and 3,000 samples, respectively.

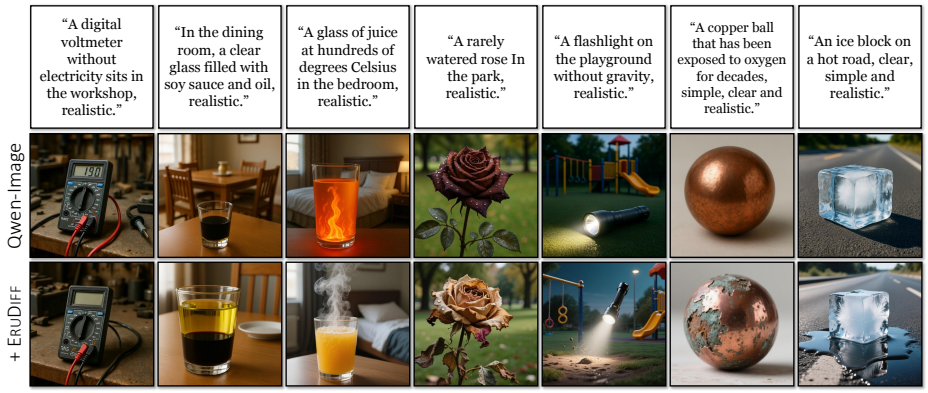


**Fig. 4: Qualitative comparison on SCIENCE-T2I based on Qwen-Image.** Each image is generated with the same text prompt and random seed for all methods. ERUDIFF excels in precise scientific knowledge comprehension and rendering.

## 5 Experiments

### 5.1 Experimental settings

**Implementation details.** ERUDIFF is designed as a model-agnostic training paradigm, engineered for seamless integration into diverse state-of-the-art text-to-image diffusion models. To rigorously evaluate its performance and cross-model generalizability, we instantiate ERUDIFF upon two representative high-capacity backbones: FLUX.1-dev [42] and Qwen-Image [81]. Our implementation is built upon the Diffusers library, utilizing Low-Rank Adaptation (LoRA) to efficiently fine-tune the attention layers within the denoising network. More parameter setting, training details are provided in the supplementary material.
**Benchmarks.** The performance of ERUDIFF is rigorously assessed across two widely recognized benchmarks: scientific knowledge benchmark (*i.e.*, SCIENCE-T2I [44]) and the world knowledge benchmark (*i.e.*, WISE [58]).

SCIENCE-T2I [44] spans multiple scientific domains, including physics, chemistry, and biology, encompassing 16 distinct scientific phenomena. It introduces SCISCORE, an end-to-end reward model integrated with expert-level scientific knowledge, designed to evaluate whether generated images precisely reflect the scientific phenomena described in the given prompts. The SCIENCE-T2I dataset comprises both training and testing sets, the latter is further divided into SCIENCE-T2I S and SCIENCE-T2I C. Specifically, SCIENCE-T2I S maintains the same stylistic attributes as the training set, whereas SCIENCE-T2I C introduces more diverse scene settings. We evaluate the performance of the proposed method using the corresponding dataset splits.

WISE [58] is a benchmark specifically designed for world knowledge-informed semantic evaluation. It utilizes 1,000 meticulously crafted prompts to assess the capability of text-to-image models in understanding and rendering world knowledge across 25 subdomains in cultural commonsense, spatio-temporal reasoning,

**Table 1: Objective evaluation on WISE.** ERUDIFF yields consistent performance gains across all categories, marking a significant advancement in the field of image synthesis guided by world knowledge. **Bold** numbers indicate the highest scores among text-to-image diffusion models without the aid of multimodal large language models.

| **Methods** | **Type** | Cultural | Time | Space | Biology | Physics | Chemistry | **Overall** |
|---|---|---|---|---|---|---|---|---|
| Janus-Pro-7B [10] | *unified* | 0.30 | 0.37 | 0.49 | 0.36 | 0.42 | 0.26 | 0.35 |
| Emu3 [79] | *unified* | 0.34 | 0.45 | 0.48 | 0.41 | 0.45 | 0.27 | 0.39 |
| BLIP3o-8B [7] | *unified* | 0.49 | 0.51 | 0.63 | 0.54 | 0.63 | 0.37 | 0.52 |
| BAGEL [12] | *unified* | 0.44 | 0.55 | 0.68 | 0.44 | 0.60 | 0.39 | 0.52 |
| BAGEL + CoT [12] | *unified* | 0.76 | 0.69 | 0.75 | 0.65 | 0.75 | 0.58 | 0.70 |
| Qwen-Image [81] | *MLLM + diffusion* | 0.62 | 0.63 | 0.77 | 0.57 | 0.75 | 0.40 | 0.62 |
| SD 1.5 [67] | *diffusion* | 0.34 | 0.35 | 0.32 | 0.28 | 0.29 | 0.21 | 0.32 |
| SD 2.1 [67] | *diffusion* | 0.30 | 0.38 | 0.35 | 0.33 | 0.34 | 0.21 | 0.32 |
| SD XL [60] | *diffusion* | 0.43 | 0.48 | 0.47 | 0.44 | 0.45 | 0.27 | 0.43 |
| PixArt-$\alpha$ [9] | *diffusion* | 0.45 | 0.50 | 0.48 | 0.49 | 0.56 | 0.34 | 0.47 |
| SD 3-medium [16] | *diffusion* | 0.42 | 0.44 | 0.48 | 0.39 | 0.47 | 0.29 | 0.42 |
| SD 3.5-medium [16] | *diffusion* | 0.43 | 0.50 | 0.52 | 0.41 | 0.53 | 0.33 | 0.45 |
| SD 3.5-large [16] | *diffusion* | 0.44 | 0.50 | 0.58 | 0.44 | 0.52 | 0.31 | 0.46 |
| FLUX.1-schnell [42] | *diffusion* | 0.39 | 0.44 | 0.50 | 0.31 | 0.44 | 0.26 | 0.40 |
| FLUX.1-dev [42] | *diffusion* | 0.48 | 0.58 | 0.62 | 0.42 | 0.51 | 0.35 | 0.50 |
| FLUX.1-dev + ERUDIFF | *diffusion* | **0.66** | **0.65** | **0.72** | **0.54** | **0.64** | **0.49** | **0.64** |

and natural science. Correspondingly, WISE introduces WiScore, a quantitative metric for evaluating world knowledge-image alignment. As WISE serves solely as an evaluation benchmark and does not provide associated training resources, We utilize the constructed Knowledge-10K dataset as the training support.

### 5.2 Qualitative comparison

Fig. 3, 4 present a comparative analysis of our results against state-of-the-art counterparts on the SCIENCE-T2I benchmark based on FLUX [42] and Qwen-Image [81], respectively. SFT involves supervised fine-tuning using high-quality text-image pairs precisely rendered with scientific knowledge from SCIENCE-T2I. OFT [44], an enhanced variant of DPO [64], and GRPO [49] are both reinforcement learning (RL) based strategies trained via the SCISCORE reward model provided by SCIENCE-T2I.

As shown in Fig. 3, although these methods marginally improve the performance of the pre-trained text-to-image diffusion model (*i.e.*, FLUX.1-dev), they continue to struggle with unnatural renderings (*e.g.*, "stiff floating of napkin without gravity"), as well as the omission of critical elements (*e.g.*, "chocolate") and unintended attribute leakage (*e.g.*, "erroneous attachment of blue tints to napkin and box"). In contrast, ERUDIFF yields more scientifically accurate and visually realistic synthesis, outperforming existing counterparts. Notably, ERUDIFF is trained exclusively on the text resources from SCIENCE-T2I, offering superior scalability compared to SFT and RL-based approaches, for which the requisite

**Table 2: Objective evaluation on Science-T2I.** EruDiff outperforms SFT and RL-based strategies and MLLM-dependent framework, demonstrating substantial improvements in scientific knowledge informed image synthesis. The best performance is highlighted with **bold** values.

| Methods | Science-T2I S SciScore | Science-T2I C SciScore |
|---|---|---|
| FLUX.1-dev [42] | 23.31 | 27.54 |
| FLUX.1-dev + prompt rewrite (GPT-4o) | 32.90 | 33.19 |
| FLUX.1-dev + SFT [44] | 28.64 | 30.01 |
| FLUX.1-dev + OFT [44] | 30.95 | 32.31 |
| FLUX.1-dev + GRPO [49] | 31.68 | 32.19 |
| FLUX.1-dev + EruDiff (Ours) | **33.70** | **34.54** |
| Qwen-Image [81] | 25.38 | 31.56 |
| Qwen-Image + prompt rewrite (GPT-4o) | 34.52 | 37.86 |
| Qwen-Image + EruDiff (Ours) | **35.43** | **38.48** |

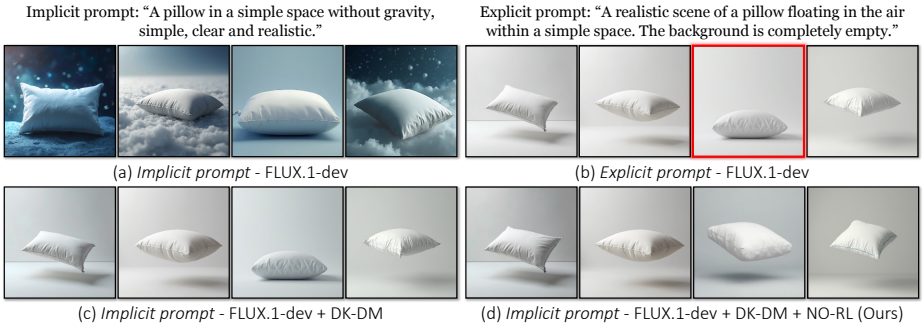


**Fig. 5: Ablation study** on negative-only reinforcement learning, which effectively rectifies inherent biases in explicit prompt rendering, as highlighted by the red box.

high-quality text-image pairs and reward models are often prohibitively expensive. Furthermore, as shown in Fig. 4, even though Qwen-Image incorporates a multimodal large language model (*i.e.*, Qwen2.5-VL) as a text encoder to derive superior representations, it still fails to capture complex phenomena, such as "immiscible liquid mixing", "wilting states", and "corrosion". EruDiff successfully masters these concepts, thereby enhancing overall performance.

Fig. 1 presents visual results on the WISE benchmark, EruDiff demonstrates a robust command of world knowledge, ranging from natural science to cultural commonsense, resulting in significant performance enhancements.

### 5.3 Quantitative comparison

**Objective evaluation.** Tables 1 and 2 present the objective evaluations conducted on the WISE and Science-T2I benchmarks, respectively. As indicated in Table 1, EruDiff fully unleashes the world knowledge informed image synthesis capabilities of FLUX, significantly enhancing performance across all eval-

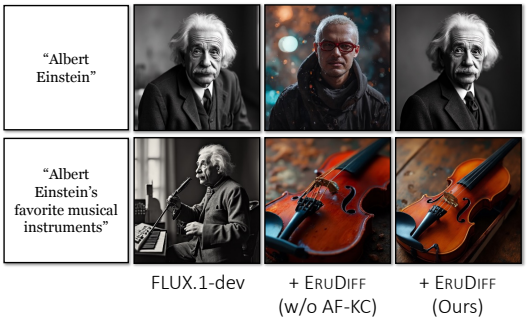


**Fig. 6: Ablation study** on AF-KC.

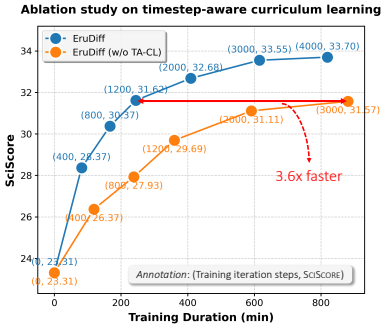


**Fig. 7: Ablation study** on TA-CL.

uation dimensions. It surpasses Qwen-Image, which leverages multimodal large language model (*i.e.*, Qwen2.5-VL) capabilities, and achieves state-of-the-art performance among diffusion-based models. Furthermore, the objective evaluation on the scientific knowledge benchmark in Table 2 demonstrates that our method outperforms both SFT and RL-based strategies. Notably, by effectively rectifying the inherent biases in explicit prompt rendering, our approach also demonstrates superior performance compared to prompt rewriting using the advanced commercial model GPT-4o, despite the capacity of latter for scientific knowledge comprehension.

**User study.** A subjective user study is provided in the supplementary material.

### 5.4 Ablation study

**On negative-only reinforcement learning.** As illustrated in Fig. 5, NO-RL focuses on circumventing the rendering inaccuracies inherently present in existing text-to-image diffusion models when processing explicit prompts, highlighted by the red box. These inaccuracies are otherwise propagated to the fine-tuned model following DK-DM. By leveraging these failure samples for single-preference reinforcement learning, NO-RL effectively mitigates this issue.

**On anti-forgetting knowledge consolidation.** As shown in Fig. 6, a naive implementation of knowledge refactoring inevitably leads to the degradation of pre-trained knowledge, such as the erosion of the "Albert Einstein" concept. By introducing AF-KC, we achieve effective preservation of these knowledge.

**On timestep-aware curriculum learning.** Given the timestep-aware characteristic of DK-DM, we design a customized timestep-aware curriculum learning strategy that prioritizes the early stages of the denoising inference chain. As shown in Fig. 7, our method achieves accelerated convergence compared to the standard uniform timestep sampling baseline.

## 6 Conclusion

This study focuses on the crucial shortcoming in text-to-image diffusion models concerning their inability to generate factually accurate images from implicit prompts requiring profound world knowledge, despite excelling in fidelity with explicit prompts. The proposed EruDiff mitigates this issue through Diffusion Knowledge Distribution Matching (DK-DM), which registers the distributions of implicit prompts with well-defined explicit anchors, and is further complemented by Negative-Only Reinforcement Learning (NO-RL) to precisely rectify inherent biases in explicit prompt rendering, ultimately leading to enhanced performance. Rigorous empirical evaluations on the Science-T2I and WISE benchmarks validate that EruDiff significantly enhances the performance of leading models, including FLUX and Qwen-Image, underscoring its efficacy in fostering more factually consistent and scientifically grounded generative AI.

## Acknowledgements

This work is supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM202).

## References

1. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Zhang, Q., Kreis, K., Aittala, M., Aila, T., Laine, S., et al.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324 (2022)
2. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn.openai.com/papers/dall-e-3.pdf **2**(3), 8 (2023)
3. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
4. Cai, Q., Chen, J., Chen, Y., Li, Y., Long, F., Pan, Y., Qiu, Z., Zhang, Y., Gao, F., Xu, P., et al.: Hidream-i1: A high-efficient image generative foundation model with sparse diffusion transformer. arXiv preprint arXiv:2505.22705 (2025)
5. Cao, S., Chen, H., Chen, P., Cheng, Y., Cui, Y., Deng, X., Dong, Y., Gong, K., Gu, T., Gu, X., et al.: Hunyuanimage 3.0 technical report. arXiv preprint arXiv:2509.23951 (2025)
6. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM Transactions on Graphics (TOG) **42**(4), 1–10 (2023)
7. Chen, J., Xu, Z., Pan, X., Hu, Y., Qin, C., Goldstein, T., Huang, L., Zhou, T., Xie, S., Savarese, S., et al.: Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset. arXiv preprint arXiv:2505.09568 (2025)
8. Chen, J., Ge, C., Xie, E., Wu, Y., Yao, L., Ren, X., Wang, Z., Luo, P., Lu, H., Li, Z.: Pixart-sigma: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In: European Conference on Computer Vision (ECCV) (2024)

9. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wang, Z., Kwok, J.T., Luo, P., Lu, H., Li, Z.: Pixart-alpha: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: International Conference on Learning Representations (ICLR) (2024)
10. Chen, X., Wu, Z., Liu, X., Pan, Z., Liu, W., Xie, Z., Yu, X., Ruan, C.: Janus-pro: Unified multimodal understanding and generation with data and model scaling. arXiv preprint arXiv:2501.17811 (2025)
11. Clark, K., Vicol, P., Swersky, K., Fleet, D.J.: Directly fine-tuning diffusion models on differentiable rewards. In: International Conference on Learning Representations (ICLR) (2024)
12. Deng, C., Zhu, D., Li, K., Gou, C., Li, F., Wang, Z., Zhong, S., Yu, W., Nie, X., Song, Z., et al.: Emerging properties in unified multimodal pretraining. arXiv preprint arXiv:2505.14683 (2025)
13. Deng, F., Wang, Q., Wei, W., Hou, T., Grundmann, M.: Prdp: Proximal reward difference prediction for large-scale reward finetuning of diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
14. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2021)
15. Dong, H., Xiong, W., Goyal, D., Zhang, Y., Chow, W., Pan, R., Diao, S., Zhang, J., Shum, K., Zhang, T.: Raft: Reward ranked finetuning for generative foundation model alignment. arXiv preprint arXiv:2304.06767 (2023)
16. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: International Conference on Machine Learning (ICML) (2024)
17. Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., Kiela, D.: Kto: Model alignment as prospect theoretic optimization. arXiv preprint arXiv:2402.01306 (2024)
18. Eyring, L., Karthik, S., Roth, K., Dosovitskiy, A., Akata, Z.: Reno: Enhancing one-step text-to-image models through reward-based noise optimization. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2024)
19. Fan, Y., Watkins, O., Du, Y., Liu, H., Ryu, M., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Lee, K., Lee, K.: Reinforcement learning for fine-tuning text-to-image diffusion models. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2023)
20. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: International Conference on Learning Representations (ICLR) (2023)
21. Frans, K., Hafner, D., Levine, S., Abbeel, P.: One step diffusion via shortcut models. In: International Conference on Learning Representations (ICLR) (2025)
22. Fu, X., He, M., Lu, Y., Wang, W.Y., Roth, D.: Commonsense-t2i challenge: Can text-to-image generation models understand commonsense? Conference on Language Modeling (COLM) (2024)
23. Ge, Y., Zhao, S., Zhu, J., Ge, Y., Yi, K., Song, L., Li, C., Ding, X., Shan, Y.: Seed-x: Multimodal models with unified multi-granularity comprehension and generation. arXiv preprint arXiv:2404.14396 (2024)
24. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)

25. Guo, X., Cui, M., Bo, L., Huang, D.: Shortft: Diffusion model alignment via shortcut-based fine-tuning. In: IEEE International Conference on Computer Vision (ICCV) (2025)
26. Guo, X., Liu, J., Cui, M., Bo, L., Huang, D.: I4vgen: Image as free stepping stone for text-to-video generation. arXiv preprint arXiv:2406.02230 (2024)
27. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
28. Guo, X., Ma, X., Zhang, H., Huang, D.: Ctcal: Rethinking text-to-image diffusion models via cross-timestep self-calibration. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2026)
29. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2023)
30. He, X., Fu, S., Zhao, Y., Li, W., Yang, J., Yin, D., Rao, F., Zhang, B.: Tempflow-grpo: When timing matters for grpo in flow models. In: International Conference on Learning Representations (ICLR) (2026)
31. He, Y., Yang, S., Chen, H., Cun, X., Xia, M., Zhang, Y., Wang, X., He, R., Chen, Q., Shan, Y.: Scalecrafter: Tuning-free higher-resolution visual generation with diffusion models. In: International Conference on Learning Representations (ICLR) (2024)
32. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2020)
33. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2023)
34. Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., Zhao, Z.: Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In: International Conference on Machine Learning (ICML) (2023)
35. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
36. Jiang, D., Guo, Z., Zhang, R., Zong, Z., Li, H., Zhuo, L., Yan, S., Heng, P.A., Li, H.: T2i-r1: Reinforcing image generation with collaborative semantic-level and token-level cot. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
37. Karthik, S., Coskun, H., Akata, Z., Tulyakov, S., Ren, J., Kag, A.: Scalable ranked preference optimization for text-to-image generation. In: IEEE International Conference on Computer Vision (ICCV) (2025)
38. 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. In: NeurIPS (2023)
39. Kong, W., Tian, Q., Zhang, Z., Min, R., Dai, Z., Zhou, J., Xiong, J., Li, X., Wu, B., Zhang, J., et al.: Hunyuanvideo: A systematic framework for large video generative models. arXiv preprint arXiv:2412.03603 (2024)
40. Kong, Z., Ping, W., Huang, J., Zhao, K., Catanzaro, B.: Diffwave: A versatile diffusion model for audio synthesis. In: International Conference on Learning Representations (ICLR) (2021)
41. Kou, S., Jin, J., Zhou, Z., Ma, Y., Wang, Y., Chen, Q., Jiang, P., Yang, X., Zhu, J., Yu, K., et al.: Think-then-generate: Reasoning-aware text-to-image diffusion with llm encoders. arXiv preprint arXiv:2601.10332 (2026)

42. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
43. Lee, K., Liu, H., Ryu, M., Watkins, O., Du, Y., Boutilier, C., Abbeel, P., Ghavamzadeh, M., Gu, S.S.: Aligning text-to-image models using human feedback. arXiv preprint arXiv:2302.12192 (2023)
44. Li, J., Chai, W., Fu, X., Xu, H., Xie, S.: Science-t2i: Addressing scientific illusions in image synthesis. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
45. Li, J., Cui, Y., Huang, T., Ma, Y., Fan, C., Yang, M., Zhong, Z.: Mixgrpo: Unlocking flow-based grpo efficiency with mixed ode-sde. arXiv preprint arXiv:2507.21802 (2025)
46. Li, S., Kallidromitis, K., Gokul, A., Kato, Y., Kozuka, K.: Aligning diffusion models by optimizing human utility. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2024)
47. Li, Z., Zhang, J., Lin, Q., Xiong, J., Long, Y., Deng, X., Zhang, Y., Liu, X., Huang, M., Xiao, Z., et al.: Hunyuan-dit: A powerful multi-resolution diffusion transformer with fine-grained chinese understanding. arXiv preprint arXiv:2405.08748 (2024)
48. Liang, Z., Yuan, Y., Gu, S., Chen, B., Hang, T., Cheng, M., Li, J., Zheng, L.: Aesthetic post-training diffusion models from generic preferences with step-by-step preference optimization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
49. Liu, J., Liu, G., Liang, J., Li, Y., Liu, J., Wang, X., Wan, P., Zhang, D., Ouyang, W.: Flow-grpo: Training flow matching models via online rl. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)
50. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European Conference on Computer Vision (ECCV) (2022)
51. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: International Conference on Learning Representations (ICLR) (2023)
52. Luo, S., Tan, Y., Huang, L., Li, J., Zhao, H.: Latent consistency models: Synthesizing high-resolution images with few-step inference. arXiv preprint arXiv:2310.04378 (2023)
53. Luo, Y., Hu, T., Sun, J., Cai, Y., Tang, J.: Learning few-step diffusion models by trajectory distribution matching. In: IEEE International Conference on Computer Vision (ICCV) (2025)
54. Mañas, O., Astolfi, P., Hall, M., Ross, C., Urbanek, J., Williams, A., Agrawal, A., Romero-Soriano, A., Drozdzal, M.: Improving text-to-image consistency via automatic prompt optimization. Transactions on Machine Learning Research (TMLR) (2024)
55. Meng, F., Shao, W., Luo, L., Wang, Y., Chen, Y., Lu, Q., Yang, Y., Yang, T., Zhang, K., Qiao, Y., et al.: Phybench: A physical commonsense benchmark for evaluating text-to-image models. arXiv preprint arXiv:2406.11802 (2024)
56. Meral, T.H.S., Simsar, E., Tombari, F., Yanardag, P.: Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
57. Nichol, A.Q., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., Mcgrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In: International Conference on Machine Learning (ICML) (2022)

58. Niu, Y., Ning, M., Zheng, M., Jin, W., Lin, B., Jin, P., Liao, J., Feng, C., Ning, K., Zhu, B., et al.: Wise: A world knowledge-informed semantic evaluation for text-to-image generation. arXiv preprint arXiv:2503.07265 (2025)
59. Pan, X., Shukla, S.N., Singh, A., Zhao, Z., Mishra, S.K., Wang, J., Xu, Z., Chen, J., Li, K., Juefei-Xu, F., et al.: Transfer between modalities with metaqueries. arXiv preprint arXiv:2504.06256 (2025)
60. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: International Conference on Learning Representations (ICLR) (2024)
61. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: International Conference on Learning Representations (ICLR) (2023)
62. Qin, J., Wu, J., Chen, W., Ren, Y., Li, H., Wu, H., Xiao, X., Wang, R., Wen, S.: Diffusiongpt: Llm-driven text-to-image generation system. arXiv preprint arXiv:2401.10061 (2024)
63. Qiu, H., Zhang, S., Wei, Y., Chu, R., Yuan, H., Wang, X., Zhang, Y., Liu, Z.: Freescale: Unleashing the resolution of diffusion models via tuning-free scale fusion. In: IEEE International Conference on Computer Vision (ICCV) (2025)
64. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2023)
65. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
66. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2023)
67. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2022)
68. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2022)
69. Salimans, T., Ho, J.: Progressive distillation for fast sampling of diffusion models. In: International Conference on Learning Representations (ICLR) (2022)
70. Sauer, A., Lorenz, D., Blattmann, A., Rombach, R.: Adversarial diffusion distillation. In: European Conference on Computer Vision (ECCV) (2024)
71. Schuhmann, C., Beaumont, R.: Laoin aesthetic predictor. `https://laion.ai/blog/laion-aesthetics/` (2022)
72. Seedance, T., Chen, H., Chen, S., Chen, X., Chen, Y., Chen, Y., Chen, Z., Cheng, F., Cheng, T., Cheng, X., et al.: Seedance 1.5 pro: A native audio-visual joint generation foundation model. arXiv preprint arXiv:2512.13507 (2025)
73. Seedream, T., Chen, Y., Gao, Y., Gong, L., Guo, M., Guo, Q., Guo, Z., Hou, X., Huang, W., Huang, Y., et al.: Seedream 4.0: Toward next-generation multimodal image generation. arXiv preprint arXiv:2509.20427 (2025)
74. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. In: International Conference on Learning Representations (ICLR) (2023)
75. Sun, K., Fang, R., Duan, C., Liu, X., Liu, X.: T2i-reasonbench: Benchmarking reasoning-informed text-to-image generation. arXiv preprint arXiv:2508.17472 (2025)

76. Wallace, B., Dang, M., Rafailov, R., Zhou, L., Lou, A., Purushwalkam, S., Ermon, S., Xiong, C., Joty, S., Naik, N.: Diffusion model alignment using direct preference optimization. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
77. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
78. Wang, L., Xing, X., Cheng, Y., Zhao, Z., Li, D., Hang, T., Tao, J., Wang, Q., Li, R., Chen, C., et al.: Promptenhancer: A simple approach to enhance text-to-image models via chain-of-thought prompt rewriting. arXiv preprint arXiv:2509.04545 (2025)
79. Wang, X., Zhang, X., Luo, Z., Sun, Q., Cui, Y., Wang, J., Zhang, F., Wang, Y., Li, Z., Yu, Q., et al.: Emu3: Next-token prediction is all you need. arXiv preprint arXiv:2409.18869 (2024)
80. Wang, Z., Li, A., Li, Z., Liu, X.: Genartist: Multimodal llm as an agent for unified image generation and editing. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2024)
81. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
82. Wu, M., Wang, L., Zhao, P., Yang, F., Zhang, J., Liu, J., Zhan, Y., Han, W., Sun, H., Ji, J., et al.: Reprompt: Reasoning-augmented reprompting for text-to-image generation via reinforcement learning. arXiv preprint arXiv:2505.17540 (2025)
83. Wu, T.H., Lian, L., Gonzalez, J.E., Li, B., Darrell, T.: Self-correcting llm-controlled diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
84. Wu, X., Sun, K., Zhu, F., Zhao, R., Li, H.: Human preference score: Better aligning text-to-image models with human preference. In: ICCV (2023)
85. Xiao, G., Yin, T., Freeman, W.T., Durand, F., Han, S.: Fastcomposer: Tuning-free multi-subject image generation with localized attention. International Journal of Computer Vision (IJCV) **133**(3), 1175–1194 (2025)
86. Xie, E., Chen, J., Chen, J., Cai, H., Tang, H., Lin, Y., Zhang, Z., Li, M., Zhu, L., Lu, Y., et al.: Sana: Efficient high-resolution image synthesis with linear diffusion transformers. In: International Conference on Learning Representations (ICLR) (2025)
87. Xie, J., Mao, W., Bai, Z., Zhang, D.J., Wang, W., Lin, K.Q., Gu, Y., Chen, Z., Yang, Z., Shou, M.Z.: Show-o: One single transformer to unify multimodal understanding and generation. In: International Conference on Learning Representations (ICLR) (2025)
88. Xu, J., Liu, X., Wu, Y., Tong, Y., Li, Q., Ding, M., Tang, J., Dong, Y.: Imagereward: Learning and evaluating human preferences for text-to-image generation. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2023)
89. Xue, Z., Wu, J., Gao, Y., Kong, F., Zhu, L., Chen, M., Liu, Z., Liu, W., Guo, Q., Huang, W., et al.: Dancegrpo: Unleashing grpo on visual generation. arXiv preprint arXiv:2505.07818 (2025)
90. Yang, K., Tao, J., Lyu, J., Ge, C., Chen, J., Shen, W., Zhu, X., Li, X.: Using human feedback to fine-tune diffusion models without any reward model. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
91. Yang, Z., Wang, J., Li, L., Lin, K., Lin, C.C., Liu, Z., Wang, L.: Idea2img: Iterative self-refinement with gpt-4v for automatic image design and generation. In: European Conference on Computer Vision (ECCV) (2024)

92. Yin, T., Gharbi, M., Park, T., Zhang, R., Shechtman, E., Durand, F., Freeman, B.: Improved distribution matching distillation for fast image synthesis. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2024)
93. Yin, T., Gharbi, M., Zhang, R., Shechtman, E., Durand, F., Freeman, W.T., Park, T.: One-step diffusion with distribution matching distillation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
94. Zhang, J., Huang, Q., Liu, J., Guo, X., Huang, D.: Diffusion-4k: Ultra-high-resolution image synthesis with latent diffusion models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2025)
95. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: IEEE International Conference on Computer Vision (ICCV) (2023)
96. Zheng, K., Chen, H., Ye, H., Wang, H., Zhang, Q., Jiang, K., Su, H., Ermon, S., Zhu, J., Liu, M.Y.: Diffusionnft: Online diffusion reinforcement with forward process. In: International Conference on Learning Representations (ICLR) (2026)
97. Zhou, C., Yu, L., Babu, A., Tirumala, K., Yasunaga, M., Shamis, L., Kahn, J., Ma, X., Zettlemoyer, L., Levy, O.: Transfusion: Predict the next token and diffuse images with one multi-modal model. In: International Conference on Learning Representations (ICLR) (2025)
98. Zuo, Y., Zheng, Q., Wu, M., Jiang, X., Li, R., Wang, J., Zhang, Y., Mai, G., Wang, L.V., Zou, J., et al.: 4kagent: agentic any image to 4k super-resolution. In: Annual Conference on Neural Information Processing Systems (NeurIPS) (2025)