FineRAG: Fine-grained Retrieval-Augmented Text-to-Image Generation

Huaying Yuan[1], **Ziliang Zhao**[1], **Shuting Wang**[1], **Shitao Xiao**[2], **Minheng Ni**[3],
Zheng Liu[2*] and **Zhicheng Dou**[1*]
[1]Gaoling School of Artificial Intelligence, Renmin University of China
[2]Beijing Academy of Artificial Intelligence
[3]The Hong Kong Polytechnic University
{hyyuan, dou}@ruc.edu.cn

Abstract

Recent advancements in text-to-image generation, notably the series of Stable Diffusion methods, have enabled the production of diverse, high-quality photo-realistic images. Nevertheless, these techniques still exhibit limitations in terms of knowledge access. Retrieval-augmented image generation is a straightforward way to tackle this problem. Current studies primarily utilize coarse-grained retrievers, employing initial prompts as search queries for knowledge retrieval. This approach, however, is ineffective in accessing valuable knowledge in long-tail text-to-image generation scenarios. To alleviate this problem, we introduce **FineRAG**, a fine-grained model that systematically breaks down the retrieval-augmented image generation task into four critical stages: query decomposition, candidate selection, retrieval-augmented diffusion, and self-reflection. Experimental results on both general and long-tailed benchmarks show that our proposed method significantly reduces the noise associated with retrieval-augmented image generation and performs better in complex, open-world scenarios.

1 Introduction

Text-to-image (T2I) generation, a dynamic research area within artificial intelligence, is focused on the development of models capable of generating images in response to language prompts. The primary objective is to maintain fidelity to the prompts while minimizing the occurrence of visual concept hallucination and detail distortion. A variety of potent models have emerged in this field, including auto-regressive models (Yu et al., 2022), generative adversarial networks(Xu et al., 2018; Tao et al., 2022; Zhang et al., 2021; Zheng et al., 2021), and diffusion models(Rombach et al., 2022; Podell et al., 2023; Peebles and Xie, 2023). Among these, diffusion models have demonstrated the most promising performance (Rombach et al., 2022; Podell et al., 2023; Peebles and Xie, 2023; Bao et al., 2023). These models, by leveraging a multi-step denoising process, can generate visually appealing and accurate images corresponding to the input textual descriptions.

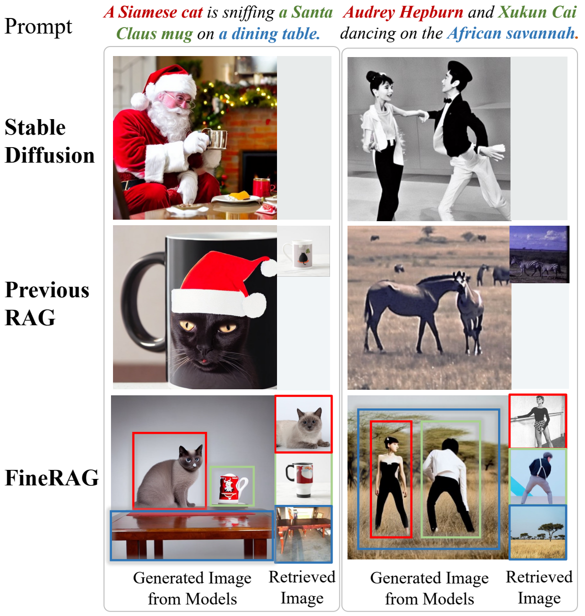

Prompt
A Siamese cat is sniffing a Santa Claus mug on a dining table.
Audrey Hepburn and Xukun Cai dancing on the African savannah.
Stable Diffusion
Previous RAG
FineRAG
Generated Image from Models
Retrieved Image
Generated Image from Models
Retrieved Image

Figure 1: Illustrative examples of FineRAG. Our model excels Stable Diffusion and previous RAG-based image generation models in accurately generating visual concepts in long-tailed complex scenarios.

Despite the remarkable progress achieved by recent diffusion models, they continue to face challenges with generating images involving less frequent entities. For instance, as illustrated in Figure 1, Stable Diffusion (Rombach et al., 2022) tends to incorrectly infer the knowledge of 'Siamese cat', which is a special breed of cat. To enhance the faithfulness of image generation, several methods based on Retrieval-Augmented Generation (RAG) have been proposed by researchers (Chen et al., 2022; Blattmann et al., 2022; Sheynin et al., 2022; Yasunaga et al., 2022). These RAG-based image generation methods operate by

[*]Corresponding author.

Proceedings of the 31st International Conference on Computational Linguistics, pages 11196–11205
January 19–24, 2025. ©2025 Association for Computational Linguistics

sourcing knowledge from an external corpus, usually containing extensive text-image pairs, as a supplement. This corpus, rich in long-tailed and contemporary information, serves as strong support for faithful image generation.

Previous RAG-based image generation models (Blattmann et al., 2022; Chen et al., 2022) have primarily used original prompts as queries for image retrieval. Such a coarse-grained retrieval approach leads to two primary challenges. **Firstly**, there may be no images in the corpus that perfectly match the prompt. For example, using the prompt 'A Siamese cat is sniffing a Santa Claus mug on a dining table.' as a query might return 'A mug with a cat painted on it.' due to the absence of exact matches to the original query in the knowledge corpus. This leads to an incomplete knowledge resource that fails to encapsulate the concepts 'A Siamese cat' and 'Santa Claus mug'. **Secondly**, even if the model identifies a set of images relevant to the prompt, not all retrieved images contribute positively to the generative process. For instance, in the case of 'Audrey Hepburn and Xukun Cai dancing on the African savannah.',[1] an image depicting a blank 'African savannah' is more useful than an image depicting 'a herd of horses on the African savannah.' Similarly, a photorealistic image of an elephant is more beneficial than a simple stick figure of an elephant for the prompt 'Two elephants are walking on the grassland.' Conclusively speaking, the RAG-based image generation task is complex and cannot be effectively addressed by merely utilizing a single prompt as a query.

To overcome these limitations, we propose a fine-grained retrieval-augmented image generation framework, **FineRAG**, which breaks down complex instructions into atomic retrievable queries, facilitating more effective knowledge sourcing and enhancing alignment between prompts and generated images. Specifically, we first decompose the original text prompts with composite knowledge into fine-grained atomic search queries, leading to more precise retrieval of useful visual information. For the image generation task, the relevance of source images does not necessarily equate to their utility. For instance, consider generating an image based on the instruction 'A Basset bleu de Gascogne and a cat are playing in front of the Bernice Pauahi Bishop Museum'. As depicted in Figure 2 at stage 2.1, both a photorealistic image of the scene and a simplistic painting of a cat may be deemed relevant. However, the latter may not be suitable for generating images that accurately align with the user's instructions. Thus, the relevance of an image should not be confused with its appropriacy for a specific image generation task. To select the most suitable images that best satisfy a user's instructions, we introduce a filtering module for further filtration. Subsequently, we use the selected images to produce a first-round image with a retrieval-augmented diffusion model. It is important to note that even though all the images are useful, accurately integrating this knowledge poses another challenge. So we introduce a reflection module, assessing whether the retrieved images can jointly result in satisfactory generation results. If not, it reasons how to improve upon the current decomposed queries and performs the RAG-based image generation for another round.

We perform extensive experiments on widely-used T2I benchmark COCO (Lin et al., 2014), where our proposed model achieves better text-image alignment compared with SOTA methods, demonstrating the general generation ability of our FineRAG model. Additionally, we construct an additional benchmark to evaluate generation ability in complex long-tailed scenarios. Experiments show that our method significantly outperforms both SOTA diffusion methods and RAG-based image generation baselines.

To summarize, our contributions are three-fold:

(1) We identify the limitations of current RAG-based image generation models, particularly the coarse-grained retrieval pipeline that leads to incomplete knowledge sourcing.

(2) To tackle these problems, we introduce FineRAG, a novel framework that deconstructs intricate instructions into elemental search queries. This framework is structured around four primary stages: query decomposition, candidate filtering, retrieval-augmented diffusion, and self-reflection.

(3) Experimental results demonstrate that our framework enhances both the prompt-image alignment and the image faithfulness.

2 Related Work

2.1 Text-to-Image Generation

Text-to-image generation models are mainly divided into three categories: auto-regressive models (Yu et al., 2022), generative adversarial networks (Xu et al., 2018; Tao et al., 2022; Zhang

[1] Xukun Cai is a famous actor in China.

et al., 2021; Zheng et al., 2021), and diffusion models (Rombach et al., 2022; Podell et al., 2023; Peebles and Xie, 2023; Saharia et al., 2022; Ramesh et al., 2022; Nichol et al., 2021). Among these, diffusion models have shown superior synthesis results, representing the current state-of-the-art. Among all the diffusion models, Stable Diffusion (Rombach et al., 2022) introduced latent diffusion models, which implement a diffusion training process within the latent space of pretrained autoencoders. Subsequently, the U-Net backbone is replaced with a transformer that operates on latent patches (Peebles and Xie, 2023; Bao et al., 2023).

While existing models primarily focus on refining image details via fine-tuning, our study seeks to employ an external knowledge corpus to bolster the coherence between textual prompts and the resultant images, particularly in scenarios characterized by complex long-tail distributions.

2.2 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) models, which integrate an external retrieval corpus to enhance the factuality, have shown significant improvements in both language (Gao et al., 2023; Zhu et al., 2023) and image generation (Chen et al., 2022; Yasunaga et al., 2022). RAG-based image generation models like KNN-Diffusion (Sheynin et al., 2022) and RDM (Blattmann et al., 2022) enable out-of-domain image generation by altering the retrieval database at inference time. Re-Imagen (Chen et al., 2022) incorporates retrieval with image-text pairs in text-to-image generation. RA-CM3 (Yasunaga et al., 2022) leverages an autoregressive architecture and concatenates all the reference images with prompts to enhance both text generation and image generation.

Despite their potential, existing retrieval-augmented image generation models predominantly employ original prompts as queries for image retrieval. This coarse-grained retrieval strategy can result in incomplete knowledge acquisition and potential overlook of the most suitable image. To address these limitations, we introduce a fine-grained retrieval-augmented image generation framework, FineRAG. This framework decomposes complex instructions into fine-grained atomic search queries, which significantly facilitate knowledge sourcing and enhance alignment between prompts and the resultant images.

3 Preliminary

Diffusion Models are generative models that simulate a diffusion process to generate data similar to a given dataset, employing a two-step process: forward and reverse process (Rombach et al., 2022).

In the forward process, original data samples $\mathbf{x}_0$ are progressively noised over T timesteps. This process follows $q(\mathbf{x}_\mathrm{t}|\mathbf{x}_\mathrm{t\text{-}1}) = \mathcal{N}(\mathbf{x}_\mathrm{t}; \sqrt{1-\beta_\mathrm{t}}\mathbf{x}_\mathrm{t\text{-}1}, \beta_\mathrm{t}\mathbf{I})$, where β_t determines the noise level at each timestep. This procedure culminates in a sequence of data samples $\{\mathbf{x}_\mathrm{t}\}_\mathrm{t=1}^{T}$. In backward process, diffusion model denoises data step-by-step, modeled as $p(\mathbf{x}_\mathrm{t\text{-}1}|\mathbf{x}_\mathrm{t}) = \mathcal{N}(\mathbf{x}_\mathrm{t\text{-}1}; \mu_\theta(\mathbf{x}_\mathrm{t}), \sigma_t^2\mathbf{I})$.

The optimization objective is to estimate $\mathbb{E}[\epsilon|x_t]$ by minimizing noise prediction loss: $\min_\theta \mathbb{E}_{t,x_0,\epsilon} \|\epsilon - \epsilon_\theta(x_\mathrm{t}, t)\|^2$, where t is uniform between 1 and T, and ϵ is the standard Gaussian noises injected to x_t. For conditional models, the condition information, c, is also incorporated:

$$\min_\theta \mathbb{E}_{t,x_0,c,\epsilon} \|\epsilon - \epsilon_\theta(x_\mathrm{t}, t, c)\|^2 \qquad (1)$$

4 Methodology

We propose a fine-grained retrieval-augmented image generation framework, FineRAG, which dissects complex instructions into fine-grained atomic search queries to facilitate accurate knowledge sourcing. The architecture of FineRAG is detailed in Figure 2. Initially, composite knowledge instructions are decomposed into atomic search queries for precise visual information retrieval. A filtering module is then employed to select the most appropriate images in line with the user's information needs. Subsequently, a retrieval-augmented diffusion model generates a preliminary image, and a reflection module is included to assess the combined potential of the retrieved images for satisfactory generation. If the output is unsatisfactory, the module optimizes the decomposed queries and initiates another round of RAG-based image generation. We will introduce the details of each component in the remaining part of this section.

4.1 Query Decomposition & Layout Design

Query decomposition is a fundamental process in the realm of image retrieval systems, which breaks down a complex or multi-faceted instruction (query) into simpler, fine-grained sub-queries. These sub-queries can then be independently processed by the retrieval system.

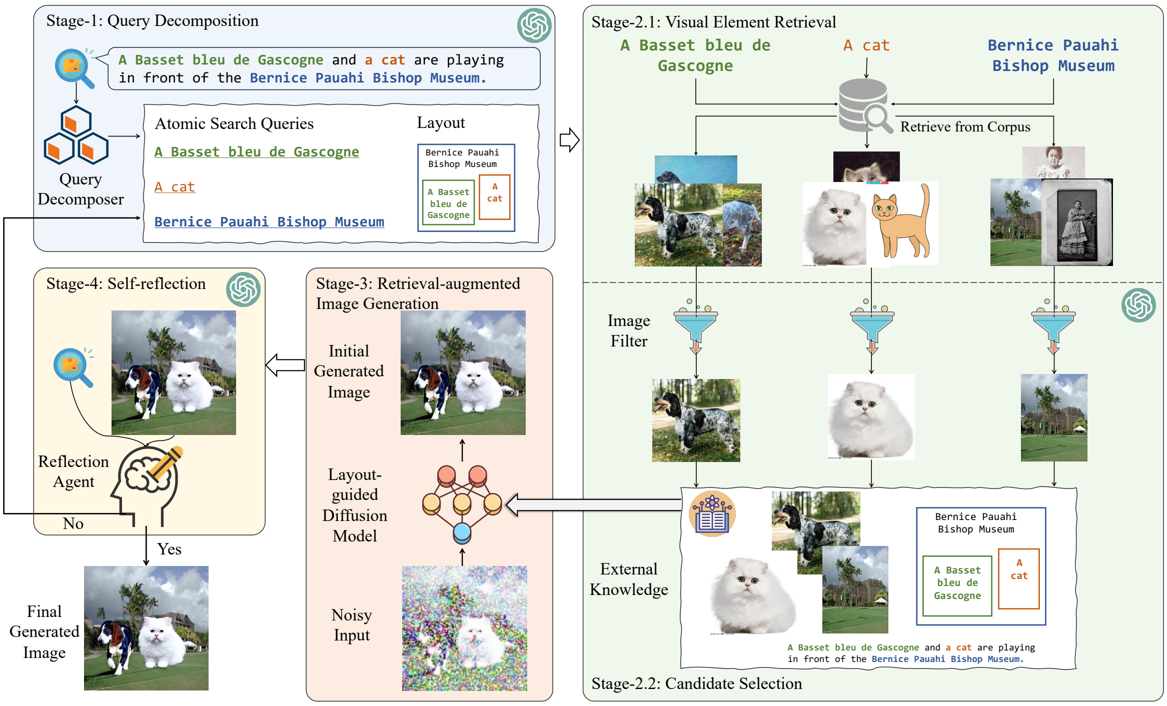

Stage-1: Query Decomposition
A Basset bleu de Gascogne and a cat are playing in front of the Bernice Pauahi Bishop Museum.
Query Decomposer
Atomic Search Queries
A Basset bleu de Gascogne
A cat
Bernice Pauahi Bishop Museum
Layout
Bernice Pauahi Bishop Museum
A Basset bleu de Gascogne
A cat
Stage-2.1: Visual Element Retrieval
A Basset bleu de Gascogne
A cat
Bernice Pauahi Bishop Museum
Retrieve from Corpus
Image Filter
External Knowledge
Bernice Pauahi Bishop Museum
A Basset bleu de Gascogne
A cat
A Basset bleu de Gascogne and a cat are playing in front of the Bernice Pauahi Bishop Museum.
Stage-2.2: Candidate Selection
Stage-3: Retrieval-augmented Image Generation
Initial Generated Image
Layout-guided Diffusion Model
Noisy Input
Stage-4: Self-reflection
Reflection Agent
No
Yes
Final Generated Image

Figure 2: The overview of our FineRAG framework. Different from existing methods that retrieve images using the original text prompt, we propose retrieving images of fine-grained entities. Given a textual instruction, FineRAG framework operates as follows: a query decomposer simplifies instructions into retrievable sub-questions and designs a layout. Relevant images are retrieved and selected by an image filter. A layout-guided diffusion model uses the selected image and layout to generate an initial image, which is then evaluated by a self-reflection module. If unsatisfactory, the process is repeated with revised sub-questions and layout.

Consider the example where the instruction is to retrieve images corresponding to the scenario “A Basset bleu de Gascogne and a cat are playing in front of the Bernice Pauahi Bishop Museum”. In this case, the query decomposition process would break down this complex instruction into atomic queries such as “A Basset bleu de Gascogne”, “A cat”, and “Bernice Pauahi Bishop Museum”. Each of these atomic queries encapsulates a specific detail and can be individually processed.

Technically, we employ LLMs as query rewriters, an approach that has demonstrated its efficacy in the field of information retrieval (Wang et al., 2023; Mao et al., 2023; Jagerman et al., 2023; Alaofi et al., 2023). Previous research has utilized LLMs to discern the object type in captions (Qu et al., 2023), for instance, identifying ’person’ in ’A long-hair woman is walking in the street’. However, within our framework, we require not just the object type but also a comprehensive phrase encapsulating the object’s details, such as ’a long-hair woman’. We provide two example inputs to the query decomposer to aid the LLM in understanding the task. In order to seamlessly incorporate the reference images of each object into the subsequent diffusion module, we also generate a layout following the previous approach (Qu et al., 2023).

By incorporating query decomposition as a preliminary step in the retrieval process, the system can more accurately uncover visual concepts inherent in instruction, thereby providing a more comprehensive knowledge base for diffusion models.

4.2 Element Retrieval & Candidate Selection

In order to minimize noise and maximize relevance, our approach incorporates a two-tiered retrieval process designed to provide useful visual knowledge for each atomic search query.

Initially, atomic search queries are used to engage a coarse retriever, which selects the top-k images for each atomic search query from external image datasets. This forms a candidate image set, denoted as I. We record the similarity score of each image, denotes as S_I.

If the average of S_I is lower than a threshold θ, we consider these images as containing noise and discard them. Otherwise, we consider them relevant and pass them to the candidate image filter. Unlike the coarse image retriever, the candidate image filter evaluates the utility of an image from a comprehensive perspective, such as style and whether it contains irrelevant items, among other factors. As depicted in Figure 2, a photorealistic image of a white cat may be more beneficial than a simplistic cat drawing given the query 'a cat'. To address this variability, we employ an image filter to sift through the candidate images. This filter is devised to select the most supportive reference image I_{i} for each atomic search query i. This mechanism considers the relevance of the scene, particularly focusing on the exclusion of irrelevant objects, and style of the image. The prompt is shown in Figure 3.

The adoption of this two-tiered retrieval process helps ensure that reference images contain as little noise as possible and contribute effectively to the image generation task.

4.3 Retrieval-augmented Diffusion Models

In this work, we utilize a layout-guided diffusion model (Li et al., 2023) as our diffusion backbone, a choice motivated by their extensive accessibility and ease of layout annotation by LLMs.

Layout-guided diffusion model is a grounded text-to-image model, which leverage the caption and grounded entities to generate image. The instruction c' can be represented as a tuple $(\mathcal{C}, \mathcal{B})$, where:

$$\text{Caption: } \mathcal{C} = [c_1, \ldots, c_{\text{L}}],$$
$$\text{Entity Bounding Boxes: } \mathcal{B} = [(b_1, l_1), \ldots, (b_{\text{N}}, l_{\text{N}})].$$

Here, L is the caption length, b_{i} is the entity embedding, l_{i} is the bounding box for the ith atomic search query. The entity embedding is calculated as follows:

$$b_{\text{i}} = \begin{cases} \text{MLP}(f_{\text{image}}(I_{\text{i}}), f_{\text{pos}}(l_{\text{i}})) & \text{if } S_{\text{i}} > \theta, \\ \text{MLP}(f_{\text{text}}(T_{\text{i}}), f_{\text{pos}}(l_{\text{i}})) & \text{otherwise}, \end{cases} \tag{2}$$

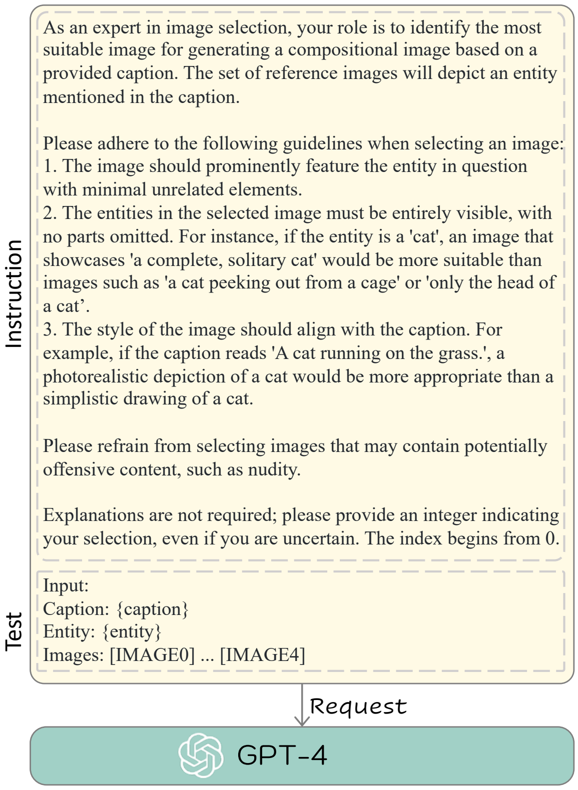
where T_{i} is the ith decomposed query, f_{image} is image encoder, f_{text} is text encoder, f_{pos} is the position embedding for bounding box, and N is the number of entities. Then, the optimization objective of the diffusion model in Equation(1) is turned to:

$$\min_{\theta} \mathbb{E}_{t,x_0,c',\epsilon} \left\| \epsilon - \epsilon_\theta(x_{\text{t}}, t, c') \right\|^2, \tag{3}$$

where θ denotes the parameters of the model, x_{t} is the image at time t, and y is the target image. The optimization aims to minimize the mean square error between the actual noise ϵ and the noise produced by the model $\epsilon_\theta(x_t, t, y)$.

4.4 Self-Reflection and Iterative Refinement

Despite the efficacy of LLMs in query decomposition and layout design, in the absence of feedback from diffusion models, these methods may yield suboptimal results in image generation tasks. To enhance the robustness of these modules, we prompt a MLLM to reflect the process of query decomposition and layout design. This reflection is based on the original caption, the results of query decomposition, layout design, and the generated image.



As an expert in image selection, your role is to identify the most suitable image for generating a compositional image based on a provided caption. The set of reference images will depict an entity mentioned in the caption.

Please adhere to the following guidelines when selecting an image:
1. The image should prominently feature the entity in question with minimal unrelated elements.
2. The entities in the selected image must be entirely visible, with no parts omitted. For instance, if the entity is a 'cat', an image that showcases 'a complete, solitary cat' would be more suitable than images such as 'a cat peeking out from a cage' or 'only the head of a cat'.
3. The style of the image should align with the caption. For example, if the caption reads 'A cat running on the grass.', a photorealistic depiction of a cat would be more appropriate than a simplistic drawing of a cat.

Please refrain from selecting images that may contain potentially offensive content, such as nudity.

Explanations are not required; please provide an integer indicating your selection, even if you are uncertain. The index begins from 0.

Input:
Caption: {caption}
Entity: {entity}
Images: [IMAGE0] ... [IMAGE4]

Figure 3: Prompt for candidate selection.

In this iterative refinement process, if the LLM assesses the generated output as satisfactory, the process concludes. Conversely, if the LLM identifies potential areas for enhancement, it detects the unsatisfactory elements and generates improved versions of both the query decomposition and layout design. This iterative self-correction mechanism has been empirically proven to significantly enhance the quality of the final image output, as evidenced by our experimental results.

5 Experiments

5.1 Experiment Settings

5.1.1 Datasets

COCO dataset. COCO dataset (Lin et al., 2014) includes 82,783 training images and 40,504 test images, spanning 80 distinct semantic classes. Each image is characterized by five textual descriptions. To have a wider comparison in different scenarios, we employ the test set restructured by the

LayoutLLM-T2I (Qu et al., 2023), which comprises five sections, numerical, spatial, semantic, mixed, and null, containing a total of 943 images.

Multi-Entity Draw Bench. We introduce the Multi-Entity Draw Bench to better evaluate T2I models' ability to generate complex, rare scenarios—something existing benchmarks inadequately assess. By prompting GPT-4, we generated 2,000 entities categorized into specific animal breeds, unique landmarks, foods, and celebrities. GPT-4 then crafted captions combining one to three entities, from which we randomly selected 200 unique captions as our test set. This approach ensures a comprehensive evaluation of models' capabilities in generating diverse and intricate images.

5.1.2 Evaluation Metrics

Our quantitative evaluation utilizes a set of metrics. For text-image alignment, we leverage the commonly-used similarity score between the generated image and the input text calculated by CLIP, denoted as **SIM**. For image quality assess, we leverage the aesthetic score[2], denoted as **AES**. For faithfulness evaluation, we borrow the faithfulness evaluator proposed in previous work (Hu et al., 2023). Following the official repo, we generate 6 VQA questions for each instruction, then we leverage GPT-4o as a strong VQA evaluator to answer these questions about generated images. The average accuracy is denoted as **TIFA**.

5.1.3 Baselines

To evaluate the effectiveness of our method, we contrast it with several popular baselines, including both text-to-image generation methods VQ Diffusion (Gu et al., 2022), Stable Diffusion (Rombach et al., 2022) and LayoutLLM (Qu et al., 2023), and retrieval-augmented image generation methods RDM (Blattmann et al., 2022) and Re-Imagen(Chen et al., 2022). All the models are listed in Table 1. All models, except for ReImage, are replicated according to their official repositories. Since Re-Imagen is not open-source, we reimplemented it ourselves to ensure a fair comparison. To maintain consistency, Re-Imagen employs the same external corpus as ours but leverages the original image captions for retrieval. This setup allows us to directly compare the impact of using large language models (LLMs) in the retrieval process. By eliminating variations due to different external corpora, we underscore the inherent effectiveness of our proposed LLM-enhanced retrieval-augmented image generation system.

5.1.4 Implementation Details

Our implementation leverages GPT-4o mini for query decomposition and layout design, with the query decomposer guided by two-shot examples from LayoutLLM-T2I's (Qu et al., 2023) policy model. For visual concept retrieval, we use CLIP index (Beaumont, 2022) from a 500 million image-text pair subset of the LAION (Schuhmann et al., 2022) dataset to retrieve the top 100 images. To enhance long-tailed entities, we use a BM25 retriever on the WIT dataset, containing 6 million text-image pairs, to retrieve the top 10 images. A cross-modal reranker, BLIP (Li et al., 2022), then narrows this pool to the top 5 images. GPT-4o mini is employed again for candidate filtering, selecting one image from the top 5. Diffusion models leverage pre-trained weights from GLIGEN.

5.2 Quantitative Analysis

To substantiate the efficacy of our FineRAG framework, we conduct a series of experiments on both COCO test set and Multi-Entity Draw Bench.

Results on Multi-Entity Draw Bench. It is observed that Re-Imagen underperforms in generating imaginary scenes that do not exist in reality due to ineffective retrieval, which is even lower than retrieval-free model: Stable Diffusion 1-4. Conversely, our model demonstrates robust performance in such scenarios, attributable to its effective query decomposition, candidate selection, and reflection pipeline.

Results on COCO dataset. The result on COCO dataset is shown as table 1 and table 3. (1) Compared to the same diffusion model without retrieval (LayoutLLM-t2i), our approach demonstrates better text-image alignment and image quality. This improvement is due to the model's ability to retrieve high-quality, fine-grained images, enabling more accurate image synthesis. (2) In comparison to RAG-based diffusion models RDM and Re-Imagen, FineRAG exhibits substantial improvements in terms of both text-image alignment (SIM) and image quality (AES) scores. Both RDM and Re-Imagen incorporate the retrieved images in an implicit way, which can potentially lead to generation inaccuracies due to the coarse matching of images. Conversely, FineRAG employs a layout-guided diffusion model as its core structure, en-

[2] https://github.com/LAION-AI/aesthetic-predictor

Models	Mixed		Numerical		Null		Semantic		Spatial		Total	
	SIM↑	AES↑	SIM↑	AES↑	SIM↑	AES↑	SIM↑	AES↑	SIM↑	AES↑	SIM↑	AES↑
w.o. **Retrieval**												
VQ-Diffusion	26.71	5.62	25.90	5.69	26.24	5.71	26.95	5.54	26.24	5.58	26.43	5.63
Stable Diffusion 1-1	27.28	6.07	26.50	5.71	26.70	5.95	27.12	5.94	26.69	5.80	26.86	5.90
Stable Diffusion 1-4	27.63	6.14	26.99	5.94	27.31	5.97	27.79	6.00	27.30	5.79	27.42	5.97
LayoutLLM-t2i	26.57	5.76	25.89	5.79	25.55	5.76	26.45	5.88	25.58	5.67	26.01	5.77
w. Retrieval												
RDM	25.91	5.12	26.00	5.25	25.65	5.17	25.87	5.04	26.00	5.01	25.88	5.11
Re-Imagen	27.48	5.90	27.57	5.89	27.31	5.92	27.43	5.94	27.45	5.87	27.44	5.90
FineRAG (Ours)	**28.40**	**6.16**	**28.38**	**6.12**	**27.70**	**6.10**	**28.23**	**6.09**	**28.67**	**6.10**	**28.27**	**6.11**

Table 1: Results on the COCO test set. 'Numerical', 'Spatial', 'Semantic', 'Mixed', and 'Null' refer to test cases with numerical descriptions, spatial relationships, semantic actions, multiple relations/descriptions, and no explicit relation keywords.

Models	SIM↑	AES↑	TIFA↑
w.o. **Retrieval**			
VQ-Diffusion	27.87	5.52	0.48
LayoutLLM-t2i	31.57	5.70	0.51
Stable Diffusion 1-1	29.64	5.61	0.50
Stable Diffusion 1-4	32.38	5.79	0.55
w. **Retrieval**			
RDM	24.76	4.65	0.44
Re-Imagen	27.88	5.63	0.47
FineRAG (Ours)	**33.58**	**6.03**	**0.67**

Table 2: Quantitative comparison of text-to-image generation models on the Multi-Entity Draw Bench.

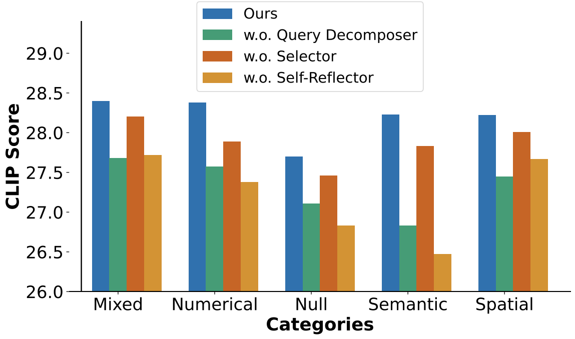


Figure 4: Ablation study on COCO test set.

Model	**FID↓**
LayoutLLM-t2i	39.3
Ours	**36.1**

Table 3: FID score on COCO test set.

abling an interpretable incorporation of external knowledge. The results confirm the effectiveness of our approach in leveraging retrieval mechanisms for enhanced image synthesis.

5.3 Ablation Study

As shown in Figure 4, we conducted ablation studies to assess the impact of each component in our method. Results indicate that removing any single component leads to a noticeable drop in performance, underscoring the importance of each part, especially the query decomposer and self-reflector. Specifically, the query decomposer simplifies complex instructions into manageable sub-queries, enhancing information retrieval accuracy. The self-reflector enables the model to iteratively refine outputs, leading to higher quality and coherence. Moreover, unlike LayoutLLM-t2i, our model incorporates retrieved images, which enhance text-image alignment and image quality. This integration results in outputs that are both visually appealing and contextually relevant.

5.4 Qualitative Analysis

To highlight the efficacy of our proposed model and its potential applications, we present examples from the Multi-Entity Draw Bench in Figure 5. RAG-based methods are compared using retrieved reference images, leading to the following conclusions: (1) Our model retrieves finer-grained, high-quality images, enabling more detailed synthesis and better text-image alignment compared to Re-Imagen. (2) While diffusion and RAG-based methods struggle with complex scenarios (e.g., 'Saussurea ussuriensis in Laut Pechora'), our model synthesizes complete, accurate images.

6 Conclusion

We propose a novel FineRAG framework that deconstructs complex instructions into atomic search queries, thus improving knowledge sourcing and

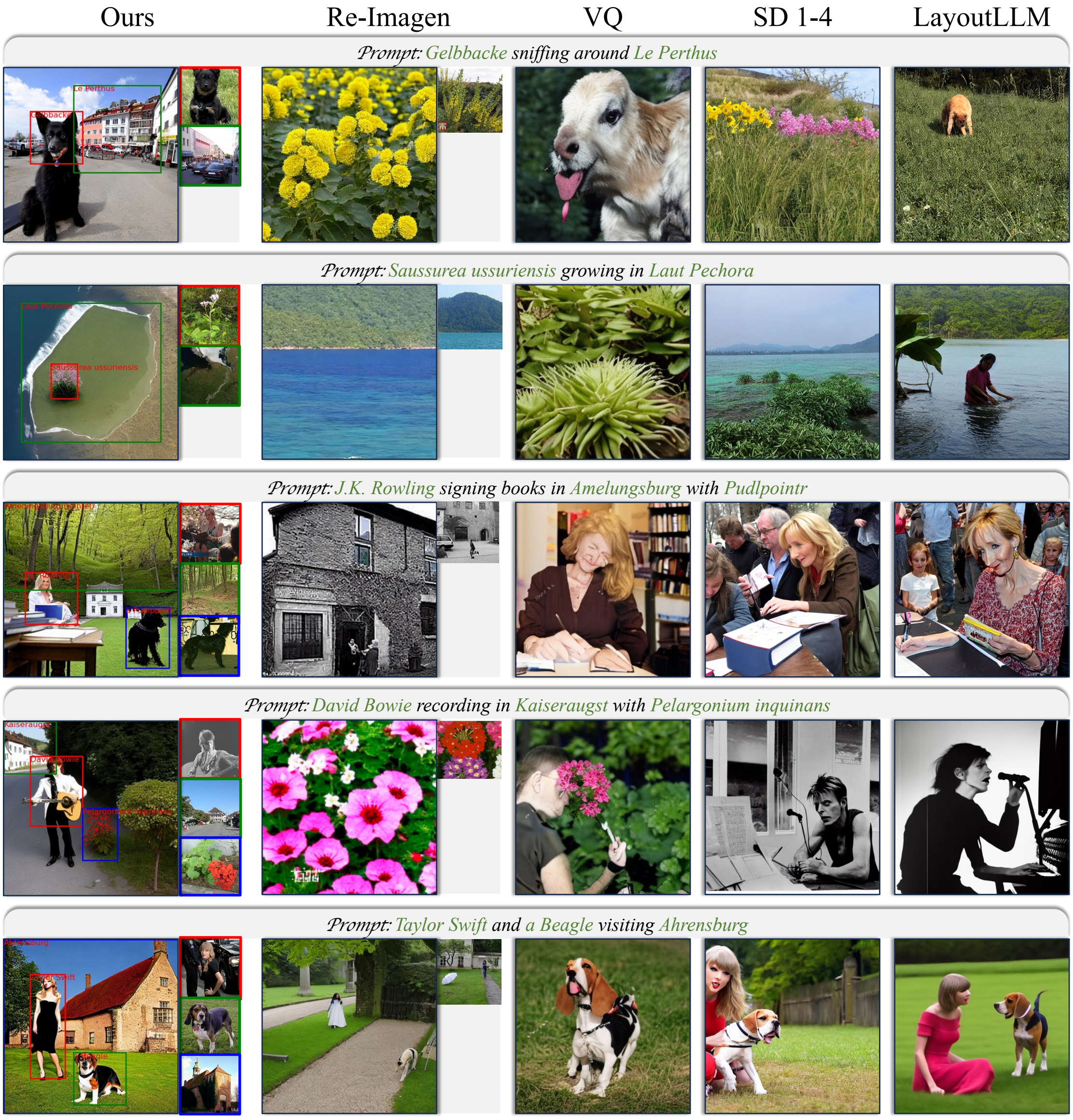

Ours
Re-Imagen
VQ
SD 1-4
LayoutLLM
Prompt: Gelbbacke sniffing around Le Perthus
Prompt: Saussurea ussuriensis growing in Laut Pechora
Prompt: J.K. Rowling signing books in Amelungsburg with Pudlpointr
Prompt: David Bowie recording in Kaiseraugst with Pelargonium inquinans
Prompt: Taylor Swift and a Beagle visiting Ahrensburg

Figure 5: Qualitative results on the COCO test set and Multi-Entity Draw Bench. The generated images from each model are presented sequentially, from left to right. 'VD' denotes VQ-Diffusion, 'SD' denotes Stable Diffusion. For better illustration, we list reference images retrieved by RAG-based methods besides the generated image.

boosting retrieval-augmented image generation. The framework handles four critical stages: information retrieval, image selection, effective integration of the retrieved knowledge into the generation process, and optimization through a self-reflective component. Experimental results show that our method significantly reduces noise and exhibits outstanding performance in complex, open-world scenarios. Our framework paves the way for future research and applications by optimizing the retrieval-augmented image generation process.

7 Limitations and Future Directions

The challenge lies in defining the capability boundaries of diffusion model. An ideal end-to-end model would have awareness of its strengths and limitations, extracting only unfamiliar information from reference images to generate accurate outputs, leading to the generation of more faithful images.

Acknowledgments

This work was supported by Beijing Natural Science Foundation No. L233008, Beijing Municipal Science and Technology Project No. Z231100010323009, National Natural Science Foundation of China No. 62272467, the fund for building world-class universities (disciplines) of Renmin University of China. The work was partially done at the Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE.

References

Marwah Alaofi, Luke Gallagher, Mark Sanderson, Falk Scholer, and Paul Thomas. 2023. Can generative llms create query variants for test collections? an exploratory study. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 1869–1873.

Fan Bao, Shen Nie, Kaiwen Xue, Yue Cao, Chongxuan Li, Hang Su, and Jun Zhu. 2023. All are worth words: A vit backbone for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22669–22679.

Romain Beaumont. 2022. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. https://github.com/rom1504/clip-retrieval.

Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. 2022. Retrieval-augmented diffusion models. *Advances in Neural Information Processing Systems*, 35:15309–15324.

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

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. 2022. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10696–10706.

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

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. 2023. Query expansion by prompting large language models. *arXiv preprint arXiv:2305.03653*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Kelong Mao, Zhicheng Dou, Fengran Mo, Jiewen Hou, Haonan Chen, and Hongjin Qian. 2023. Large language models know your contextual search intent: A prompting framework for conversational search. *arXiv preprint arXiv:2303.06573*.

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205.

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

Leigang Qu, Shengqiong Wu, Hao Fei, Liqiang Nie, and Tat-Seng Chua. 2023. Layoutllm-t2i: Eliciting layout guidance from llm for text-to-image generation. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 643–654.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3.

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

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

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. 2022. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*.

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

Liang Wang, Nan Yang, and Furu Wei. 2023. Query2doc: Query expansion with large language models. *arXiv preprint arXiv:2303.07678*.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5.

Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. 2021. Cross-modal contrastive learning for text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 833–842.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890.

Yutao Zhu, Huaying Yuan, Shuting Wang, Jiongnan Liu, Wenhan Liu, Chenlong Deng, Zhicheng Dou, and Ji-Rong Wen. 2023. Large language models for information retrieval: A survey. *arXiv preprint arXiv:2308.07107*.