

AR-RAG: Autoregressive Retrieval Augmentation for Image Generation

Jingyuan Qi^{* 1} Zhiyang Xu^{* 1} Qifan Wang² Lifu Huang³

¹Virginia Tech ²Meta ³UC Davis

jingyq1@vt.edu

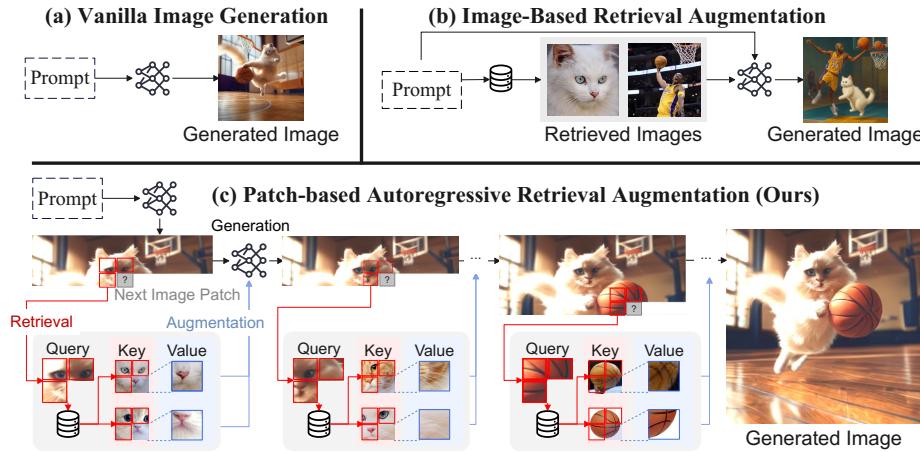


Figure 1: Comparison between Autoregressive Retrieval Augmentation (AR-RAG) for image generation in (c) and existing image generation paradigms in (a) (b). In AR-RAG, image patches in **red boxes** denote retrieval queries and keys, image patches in **blue boxes** are retrieved values, and gray boxes with the question mark are next image patches to be predicted. (Caption: A white cat is playing basketball on the court.)

Abstract

We introduce Autoregressive Retrieval Augmentation (AR-RAG), a novel paradigm that enhances image generation by autoregressively incorporating k-nearest neighbor retrievals at the patch level. Unlike prior methods that perform a single, static retrieval before generation and condition the entire generation on fixed reference images, AR-RAG performs context-aware retrievals at each generation step, using prior-generated patches as queries to retrieve and incorporate the most relevant patch-level visual references, enabling the model to respond to evolving generation needs while avoiding limitations (e.g., over-copying, stylistic bias, etc.) prevalent in existing methods. To realize AR-RAG, we propose two parallel frameworks: (1) Distribution-Augmentation in Decoding (DAiD), a training-free plug-and-use decoding strategy that directly merges the distribution of model-predicted patches with the distribution of retrieved patches, and (2) Feature-Augmentation in Decoding (FAiD), a parameter-efficient fine-tuning method that progressively smooths the features of retrieved patches via multi-scale convolution operations and leverages them to augment the image generation process. We validate the effectiveness of AR-RAG on widely adopted benchmarks,

^{*}Jingyuan Qi and Zhiyang Xu contributed equally to this work.

including Midjourney-30K, GenEval and DPG-Bench, demonstrating significant performance gains over state-of-the-art image generation models.¹

1 Introduction

Recent advancements in image generation have demonstrated remarkable capabilities in producing photorealistic images based on user prompts [31, 28, 7, 37, 10, 41, 9, 43, 45, 27, 6]. However, despite these improvements, the generated images often exhibit local distortions and inconsistencies, particularly in visual objects that possess complex structures [11], frequently interact with other objects and the surrounding scene [22, 26], or are underrepresented in the training data [8]. A promising approach to mitigating these challenges is retrieval-augmented generation (RAG), which enhances the generation process by incorporating real-world images as additional references [8, 3]. While RAG has been extensively explored in the language domain [23, 13], its application to image and multimodal generation remains largely underdeveloped. A few existing studies [3, 8, 46, 48, 49] bridge this gap by performing a single-step retrieval based on the input prompt prior to generation, conditioning the entire image generation process on fixed visual cues (Figure 1 (b)). However, as demonstrated in our pilot study (Section 5.2), such static, coarse-grained retrieval approaches [3, 8, 49] frequently introduce irrelevant or weakly aligned visual contents that persist throughout generation. Since the retrieved images are selected once, before decoding begins, and remain unchanged, these methods cannot respond to the evolving generation needs, resulting in over-copying of irrelevant details, stylistic bias, and the hallucination of unrelated visual elements. For example, as shown in Figure 1(b), a basketball player present in the retrieved references, despite being irrelevant to the input prompt, unintentionally appears in the generated image.

In this paper, we propose Autoregressive Retrieval Augmentation (AR-RAG), a novel retrieval-augmented paradigm for image generation that dynamically and autoregressively incorporates patch-level k-nearest-neighbor (k-NN) retrievals throughout the generation process (Figure 1(c)). In contrast to prior methods that rely on static, coarse-grained retrievals of entire reference images, typically using captions as retrieval queries and keys, AR-RAG performs fine-grained, step-wise retrieval at the image patch level. Specifically, as generation unfolds, AR-RAG leverages the already-generated surrounding patches as localized queries to retrieve contextually similar patches from a pre-constructed patch-level database. This database is built by encoding real-world images into latent patch features, where each entry contains a patch embedding as a value and the embeddings of its h -hop spatial neighbors as a key. During the generation of the next target patch (gray boxes in Figure 1(c)), AR-RAG retrieves the top- K most relevant patches (blue boxes) by measuring similarity between the surrounding generated context patches (red boxes) and database keys (also red boxes). These retrieved patches are then integrated into the model to inform and enhance the prediction of the next patch, enabling the model to dynamically adjust to local generation needs. By conditioning on the evolving generation context as retrieval queries, AR-RAG ensures that retrieved visual references remain relevant throughout the generation process, encouraging local semantic coherence. Moreover, the patch-level retrieval allows for precise integration of visual elements without overcommitting to entire reference images, avoiding the limitations of over-copying or irrelevant conditioning observed in static retrieval.

To realize the AR-RAG framework, we introduce two parallel implementations: (1) **Distribution-Augmentation in Decoding (DAiD)**, a training-free, plug-and-play decoding strategy that merges the model’s predicted patch distribution with that of the retrieved patches. Specifically, the top- K retrieved patches are assigned probabilities inversely proportional to their normalized ℓ_2 distances computed from the query and key patch embeddings. These probabilities are then linearly combined with the model’s native output distribution to guide the next patch prediction, enabling retrieval-aware generation without any additional training. (2) **Feature-Augmentation in Decoding (FAiD)**, a parameter-efficient fine-tuning approach that integrates retrieved patches into the generation process through learned smoothing and blending mechanisms. Specifically, when generating the next image token, FAiD operates in two stages: (1) refining the retrieved patch features by adjusting them to better fit the local context of the already generated surrounding patches, based on parameterized convolutional operations of varying kernel sizes; and (2) blending the refined features of retrieved patches with the model’s predicted feature representation for the next patch, based on compatibility

¹Code and model checkpoints can be found at <https://github.com/PLUM-Lab/AR-RAG>.

scores computed for each retrieved patch to quantify their alignment with the current generation context. To enable iterative refinement, we insert multiple FAiD modules at selected transformer layers, where the output of each FAiD module, i.e., the context-aware retrieved features blended at that layer, is forwarded as input to the next FAiD module in deeper layers. This progressive retrieval refinement mechanism allows the model to incrementally enhance its predictions as patch-level representations evolve through the network. We evaluate AR-RAG on three widely adopted benchmarks, including Midjourney-30K², Geneval [14], and DPG-Bench [18]. Experimental results demonstrate that both DAiD and FAiD significantly improve the coherence and naturalness of generated images while introducing only marginal computational overhead.

The contributions of our work can be summarized as follows:

- We propose AR-RAG, the first patch-level autoregressive retrieval augmentation framework which dynamically retrieves and integrates fine-grained visual content to enhance image generation, while avoiding limitations (e.g., over-copying, stylistic bias, etc.) prevalent in existing image-level retrieval augmentation methods.
- We introduce Distribution-Augmentation in Decoding (DAiD), a training-free, plug-and-play decoding strategy that directly integrates the distribution of retrieved patches into that predicted by the image generation models, enabling easy integration into existing architectures.
- We introduce Feature-Augmentation in Decoding (FAiD), a parameter-efficient fine-tuning framework that progressively refines and blends retrieval signals via lightweight convolutional modules, enhancing spatial coherence and visual quality across layers.
- Extensive experiments and analysis show that AR-RAG significantly improves performance of state-of-the-art image generation model across diverse metrics. In particular, Janus-Pro with FAiD achieves 6.67 FID on Midjourney-30K and 0.78 overall score on GenEval, establishing a new state of the art among autoregressive image generation models of comparable scale.

2 Preliminary

Autoregressive Image Generation Models We implement both DAiD and FAiD based on Janus-Pro [9], an autoregressive (AR) unified generation model, due to its strong performance. Janus-Pro is initialized from a transformer-based pre-trained large-language model [2], and employs a quantized autoencoder [37] to encode images into discrete image tokens. During multimodal pretraining, the model learns to predict a sequence of discrete image tokens $[v_1, v_2, \dots, v_N]$ conditioned on an input text prompt $[t_1, t_2, \dots, t_M]$. The training objective is formally defined as:

$$\arg \max_{\phi} \sum_{n=1}^N P_{\phi}(v_n | t_1, t_2, \dots, t_M, v_1, \dots, v_{n-1}) \quad (1)$$

where \mathcal{D} is the training corpus. This is the same training objective used in our FAiD method in Section 3.3. We argue that DAiD and FAiD can be extended to any image generation model that autoregressively predicts probability distributions of discrete image tokens such as LlamaGen [37], Show-o [44] and VAR [38].

Quantized Autoencoder The quantized autoencoder used in Janus-Pro consists of an encoder θ_{enc} , a decoder θ_{dec} , and a codebook \mathcal{Z} . The encoder, a convolutional neural network, downsamples and compresses raw pixel inputs into compact patch representations. During the quantization process, each patch representation is mapped to an index in the codebook by identifying its nearest neighbor vector in the codebook. In the decoding stage, these patch indices are mapped back to their corresponding vector representations via the codebook, and the decoder, another convolutional neural network, reconstructs the image from these compact representations. In our implementation, we leverage this autoencoder to build the coupled database for Janus-pro which is detailed in Section 3.1.

3 AR-RAG: Patch-based Autoregressive Retrieval Augmentation

3.1 Patch-based Retrieval Database Construction

We build a patch-based retrieval database based on several large-scale, real-world image datasets, including CC12M [5] and JourneyDB [36]. Specifically, for each image I , we encode it into N

²<https://huggingface.co/datasets/playgroundai/MJHQ-30K>

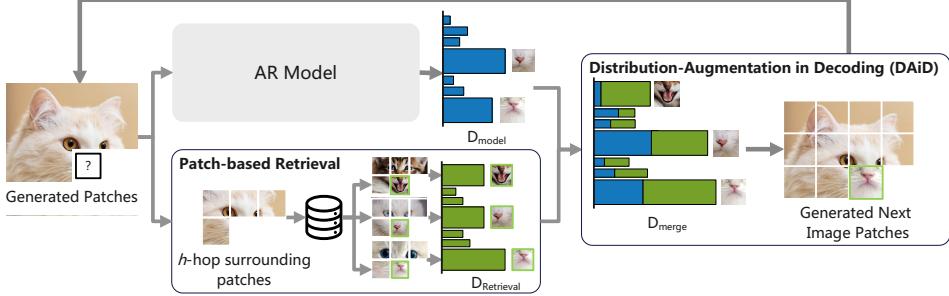


Figure 2: The decoding process in Distribution-Augmentation in Decoding (DAiD).

patches using the quantized autoencoder [37], θ_{Enc} , from Janus-Pro: $\mathbf{V} = \theta_{\text{enc}}(I) \in \mathbb{R}^{\sqrt{N} \times \sqrt{N} \times d}$, where d is the hidden dimension, and \mathbf{V}_{ij} corresponds to the latent representation of the patch at position (i, j) . We utilize each patch vector \mathbf{V}_{ij} as the value of a database entry and the representation of its h -hop surrounding patches as the key. Here, the h -hop surrounding patch representation is formed by concatenating the vectors of adjacent patches centering around (i, j) in a top-to-bottom, left-to-right order. For example, for a patch at position (i, j) , the 1-hop surrounding representation spans 8 surrounding patches $[\mathbf{V}_{(i-1)(j-1)} : \mathbf{V}_{(i-1)(j)} : \mathbf{V}_{(i-1)(j+1)} : \mathbf{V}_{(i)(j-1)} : \mathbf{V}_{(i)(j+1)} : \mathbf{V}_{(i+1)(j-1)} : \mathbf{V}_{(i+1)(j)} : \mathbf{V}_{(i+1)(j+1)}]$ where $:$ denotes the concatenation operation of image patch features. If a patch is located at the edge of the image and lacks certain surrounding patches, we substitute each missing surrounding patch with a zero vector $\mathbf{0}$.

3.2 Distribution-Augmentation in Decoding (DAiD)

Given a text prompt T , Janus-Pro autoregressively predicts a sequence of image tokens $[v_1, v_2, \dots, v_N]$ where per-token probability is defined in Equation 1. As shown in Figure 2, DAiD augments this process by incorporating probability distributions from retrieved image patches. Specifically, when Janus-Pro predicts the next image token v_{ij} , we first utilize the codebook \mathcal{Z} to convert v_{ij} 's h -hop already generated surrounding patches into patch representations. If no surrounding image tokens are available at a given position (e.g., when $i = 0$ or $j = 0$), we use the zero vector $\mathbf{0}$ as a placeholder. Once we compute the representation of v_{ij} 's h -hop surrounding patches, we leverage it as the retrieval query and retrieve the top- K most similar patch representations from the database constructed in Section 3.1 using l_2 distance. We denote the representations of the top- K retrieved patches as $[\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \dots, \hat{\mathbf{v}}_K]$ and their corresponding l_2 distances as $[s_1, s_2, \dots, s_K]$. These retrieved representations are then mapped back to discrete token indices using the codebook: $\hat{v}_k = \mathcal{Z}(\hat{\mathbf{v}}_k)$.

To augment the generation process with the retrieved image tokens $[\hat{v}_1, \hat{v}_2, \dots, \hat{v}_K]$, we create a retrieval-based distribution $D_{\text{retrieval}} \in \mathbb{R}^{|\mathcal{Z}|}$ over the entire codebook \mathcal{Z} , where $|\mathcal{Z}|$ is the codebook size. Tokens not included in the top- K retrieved set are assigned a probability of 0. For tokens within the top- K , we compute their probabilities using a softmax over their l_2 distance to the query, scaled by a retrieval temperature hyperparameter τ :

$$D_{\text{retrieval}}[v] = \begin{cases} p(\hat{v}_k) & \text{if } v = \hat{v}_k \text{ for some } m \in \{1, 2, \dots, K\} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

$$p(\hat{v}_k) = \frac{\exp(-s_k/\tau)}{\sum_{k=1}^K \exp(-s_k/\tau)}, \quad (3)$$

This creates a sparse distribution where only the top- K retrieved tokens have non-zero probabilities. Finally, we merge this retrieval distribution with the model's predicted distribution D_{model} using a weighted average:

$$D_{\text{merge}} = (1 - \lambda) \cdot D_{\text{model}} + \lambda \cdot D_{\text{retrieval}}, \quad (4)$$

where $\lambda \in [0, 1]$ is the retrieval weight hyperparameter controlling the influence of retrieved patches on the final distribution. The next token is then sampled from this merged distribution: $v_{ij} \sim D_{\text{merge}}$.

3.3 Feature-Augmentation in Decoding (FAiD)

While DAiD offers a training free approach to directly augment the probability distribution of predicted patches using retrieved ones, it suffers from noise propagation and limited flexibility in

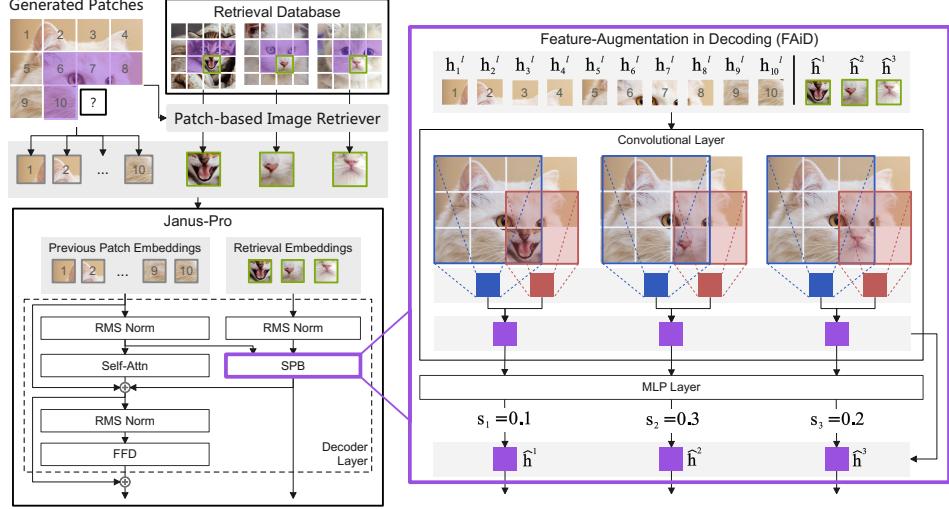


Figure 3: Overall architecture of Feature-Augmentation in Decoding (FAiD).

fully leveraging the fine-grained visual information in the retrieved patches. We thus further propose FAiD, a feature-based autoregressive augmentation strategy to enhance the image generation process. As illustrated in Figure 3, when predicting the next token v_{ij} during image generation, we employ the same retrieval process described in Section 3.2 to obtain the top- K most relevant patches and their representations $[\hat{v}_1, \hat{v}_2, \dots, \hat{v}_K]$ from our database. To effectively incorporate them into the autoregressive generation process, FAiD consists of two steps: (1) refining retrieved patches to ensure coherence with the surrounding context of v_{ij} in the generated image, and (2) adaptively blending the representation of refined patches with the hidden state of the predicted next patch based on learned compatibility scores. To enable progressive refinement of retrieved information as representations evolve through the network, we insert a FAiD module for every L/b decoder layers of the generation model, where L denotes the total number of decoder layers and b is a hyperparameter.

Multi-Scale Feature Smoothing The key of effective patch integration lies in ensuring spatial coherence between retrieved patches and the surrounding image context. To achieve this, we propose *multi-scale feature smoothing* (Algorithm 1 in Appendix A), where multi-scale convolutions are applied to retrieved patches within the generation context, so that the retrieved visual features are smoothed to preserve structural and stylistic consistency with the surrounding context of the predicted token. Specifically, at each step when predicting the next image token v_{ij} , we first construct a 2D spatial representation $\mathbf{H}^l \in \mathbb{R}^{\sqrt{N} \times \sqrt{N} \times D}$ of the current partially-generated image by arranging their hidden states $[h_1^l, h_2^l, \dots, h_{n-1}^l, h_n^l]$ from the current decoder layer l . We use $\mathbf{0}$ vectors as placeholders for positions that have not yet been generated. Then, we transform the retrieved patch representations $[\hat{v}_1, \hat{v}_2, \dots, \hat{v}_K]$ into the generation model's hidden space by mapping each patch \hat{v}_k to a discrete token index via the codebook \mathcal{Z} and embedding it through the pretrained image embedding layer Emb_{img} :

$$[\hat{h}_1, \hat{h}_2, \dots, \hat{h}_K] = \text{Emb}_{\text{img}}([\mathcal{Z}(\hat{v}_1), \mathcal{Z}(\hat{v}_2), \dots, \mathcal{Z}(\hat{v}_K)]) \quad (5)$$

For each retrieved patch \hat{h}_k , we create a copy of \mathbf{H}^l where position (i, j) (the location of v_{ij}) is replaced with \hat{h}_k . We then apply convolution operations at multiple scales (2×2 through $Q \times Q$) to capture contextual patterns at different resolutions. To maintain computational efficiency, we only perform convolution operations when the kernel covers position (i, j) , rather than processing the entire image. Each convolution kernel $\text{Conv}_{q \times q}$ produces a refined representation \hat{h}_k^q for the retrieved patch at scale q . The final refined representation for each retrieved patch is computed as a weighted sum of these multi-scale features:

$$\hat{h}_k \leftarrow \sum_{q=2}^Q \text{softmax}(\Omega)_q \cdot \hat{h}_k^q \quad (6)$$

where $\Omega = [\omega_2, \dots, \omega_Q]$ are learnable parameters that determine the importance of each scale.

Feature Augmentation After feature smoothing, some of the retrieved patch features may still not be able to fit into the surrounding neighbors and hence we need to lower their impact in the final representation. Thus, we compute a compatibility score for each of the refined patches. This is achieved by projecting each refined retrieved patch representation through a linear transformation parameterized by a weight matrix $\mathbf{W} \in \mathbb{R}^{1 \times D}$, yielding the score $s_k = \hat{\mathbf{h}}_k \mathbf{W}^T$. The final representation for the next image token v_{ij} after layer j is computed as:

$$h_{ij}^{(l+1)} = h_{ij}^l + \Delta h_{ij}^l + \sum_{k=1}^K s_k \hat{\mathbf{h}}_k \quad (7)$$

Here, h_{ij}^l is the residual, Δh_{ij}^l is the updated representation from the transformer layer l , and $\sum_{k=1}^K s_k \hat{\mathbf{h}}_k$ is the contribution of the retrieved image patches.

4 Experiment Setup

Patch-based Retrieval Database To construct our patch-level retrieval database, we randomly sample 5.7 million images from CC12M [5], 3.3 million from JourneyDB [36], and 4.6 million from DataComp [12], while ensuring that any samples included in the testing set are excluded to prevent data leakage. Each image is encoded into a sequence of patch-level representations and image tokens using the same image tokenizer employed in the Janus-Pro model. For efficient similarity search, we implement our retriever using the FAISS library [21].

Training Setup We adopt Janus-Pro-1B [9] and Show-o [44] as our backbone models and fine-tune them on a dataset of 50,000 image-caption pairs sampled from CC12M [5] and Midjourney-v6³. We empirically determine the optimal hyperparameters for DAID and FAID, and the complete hyperparameter optimization experiment results can be found in Appendix C.2. Further details regarding the training dataset construction and implementation can be found in Appendix B.3.

Baselines To evaluate the effectiveness of our proposed methods, we adopt several state-of-the-art image generation approaches as baselines, including non-retrieval models such as LlamaGen [37], LDM [32], Stable Diffusion (SDv1.5 and SDv3) [31, 10], PixArt-alpha [7], DALL-E 2 [30], Show-o[44], and Janus-Pro [9], and image-based retrieval augmentation methods, including RDM [3], RA-CM3 [46], and ImageRAG[33]. Since pretrained models of RA-CM3 are not publicly available, we try our best to replicate their method based on Janus-Pro to ensure a fair comparison. More details of training and implementation of RA-CM3 can be found in Appendix B.1.

Evaluation Benchmarks and Metrics To comprehensively evaluate our proposed methods, we employ three benchmarks: (1) GenEval [14], which assesses models’ ability to generate images with specific attributes and relationships described in text prompts; (2) DPG-Bench [18], which evaluates performance on detailed prompts with complex requirements; and (3) Midjourney-30k [40], where we employ three complementary metrics: FID [17] for measuring statistical similarity between generated and real image distributions, CMMI [20] for assessing alignment with human perception using CLIP embeddings, and FWD [39] for evaluating spatial and frequency coherence through wavelet packet coefficients. For all three metrics, lower scores indicate higher quality generated images. Detailed descriptions of these benchmarks and metrics can be found in Appendix B.4.

5 Results and Discussion

5.1 Text-to-Image Generation Results

Tables 1, 2, and 3 present performance comparisons across multiple benchmarks, where our AR-RAG methods consistently outperform existing approaches. Notably, previous retrieval-augmented approaches such as RDM and ImageRAG perform worse than their non-retrieval counterparts (LDM and SDXL, respectively) on both GenEval and DPG-Bench. We provide detailed analysis for existing image-level retrieval methods and highlight the unique advantages of our AR-RAG frameworks in the following discussion and Section 5.2. Appendix C.1 provides a benchmark analysis to demonstrate the effectiveness of patch-level retrieval in our AR-RAG methods.

³<https://huggingface.co/datasets/brivang1/midjourney-v6-llava>

Method	Params	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall ↑
<i>Non Retrieval-Augmented Model</i>								
PixArt- α	0.6B	0.98	0.50	0.44	0.80	0.08	0.07	0.48
LlamaGen	0.8B	0.71	0.34	0.21	0.58	0.07	0.04	0.32
SDv1.5	0.9B	0.97	0.38	0.35	0.76	0.04	0.06	0.43
SDv2.1	0.9B	0.98	0.51	0.44	0.85	0.07	0.17	0.50
Janus-Pro	1.0B	0.98	0.77	0.52	0.84	0.61	0.55	0.71
Show-o	1.3B	0.98	0.80	0.66	0.84	0.31	0.50	0.68
LDM	1.4B	0.92	0.29	0.23	0.7	0.02	0.05	0.37
SD3 (d=24)	2.0B	0.98	0.74	0.63	0.67	0.34	0.36	0.62
SDXL	2.6B	0.98	0.74	0.39	0.85	0.15	0.23	0.55
DALL-E 2	6.5B	0.94	0.66	0.49	0.77	0.10	0.19	0.52
DALL-E 3	-	0.96	0.87	0.47	0.83	0.43	0.45	0.67
Transfusion	7.3B	-	-	-	-	-	-	0.63
Chameleon	34B	-	-	-	-	-	-	0.39
<i>Retrieval-Augmented Model</i>								
RDM	1.4B	0.91	0.21	0.28	0.71	0.02	0.04	0.36
ImageRAG	3.5B	0.93	0.06	0.03	0.37	0.01	0.03	0.24
Janus-Pro	1.0B	0.98	0.78	0.41	0.84	0.42	0.49	0.65 (-0.06)
+ RA-CM3	1.0B	0.98	0.82	0.54	0.87	0.63	0.49	0.72 (+0.01)
+ DAID (ours)	1.0B	1.00	0.88	0.50	0.86	0.70	0.73	0.78 (+0.07)
+ FAID (ours)	1.2B							

Table 1: Evaluation of text-to-image generation ability on GenEval benchmark. Note our methods are based on Janus-Pro highlighted in gray.

Method	Params	Global	Entity	Attribute	Relation	Other	Overall ↑
<i>Non Retrieval-Augmented Model</i>							
PixArt- α	0.6B	74.97	97.32	78.60	82.57	76.96	71.11
SDv1.5	0.9B	74.63	74.23	75.39	73.49	67.81	63.18
Janus-Pro	1.0B	81.76	84.53	84.34	92.22	75.20	77.26
Lumina-Next	2.0B	82.82	88.65	86.44	80.53	81.82	74.63
SDXL	3.5B	83.27	82.43	80.91	86.76	80.41	74.65
<i>Retrieval-Augmented Model</i>							
RDM	1.4B	62.36	40.46	60.20	69.16	24.68	26.51
ImageRAG	3.5B	61.35	32.77	53.87	60.38	18.42	19.82
Janus-Pro	1B	81.76	81.03	83.32	90.60	70.80	73.76 (-3.50)
+ RA-CM3	1.0B	83.58	84.46	84.76	91.49	76.40	77.88 (+0.62)
+ DAID (ours)	1.0B	82.67	85.80	85.38	92.30	76.80	79.36 (+2.10)
+ FAID (ours)	1.2B						

Table 2: Evaluation of text-to-image generation ability on DPG-Bench. Note our methods are based on Janus-Pro highlighted in gray.

On GenEval, our methods show significant improvements in categories such as “Two Obj.” and “Position,” which demand accurate multi-object generation and spatial arrangement. These gains are largely due to the local and dynamic nature of our autoregressive patch-level retrieval. Consider the prompt “*a green couch and an orange umbrella*”, a combination that rarely co-occurs in real-world images. Static full-image retrieval methods may retrieve references containing only one of the objects. Taking these references as a global visual prior throughout the generation can lead the model to overfit to irrelevant layouts or dominant visual structures in the retrieved examples. On DPG-Bench, which features dense and highly detailed prompts, the performance gap between our method and prior retrieval-augmented approaches becomes even more substantial. Similar as GenEval, existing image-level retrieval augmentation methods struggle to retrieve meaningful references when the number of distinct entities and attributes in a prompt increases. In contrast, our autoregressive augmentation framework overcomes this limitation by dynamically retrieving patch-level visual features based on the evolving image context rather than the original prompt, enabling more targeted and effective augmentation.

On Midjourney-30K, our proposed methods consistently outperform both Janus-Pro and Show-o baselines across all three evaluation metrics. Notably, despite operating locally at the patch level, our approach leads to a significant reduction in FID

Table 3: Evaluation of text-to-image generation ability on the Midjourney-30K benchmark.

scores, indicating improved global visual quality and closer alignment with the distribution of real images. This suggests that context-aware, auto-regressive retrieval and refinement can propagate to enhance holistic image fidelity. Furthermore, the improvements in CMMD and FWD metrics confirm our method’s effectiveness in reducing visual distortions and enhancing coherence. These results also demonstrate that AR-RAG delivers robust and architecture-agnostic improvements, validating its broad applicability across different image generation backbones.

5.2 Qualitative Analysis

Figure 4 illustrates these quantitative improvements with representative examples from DPG-Bench (left three columns) and GenEval (right two columns). These examples demonstrate how autoregressive retrieval augmentation improves the vanilla image generation models. The vanilla model struggles with *object interactions* (e.g., column 3, where shoes merge with a coffee machine in the background), *complex structures* (e.g., columns 2 and 5, where camels and sheep have anatomically incorrect numbers of organs), and *implausible configurations* (e.g., column 4, where a chair exhibits an impossible design). Both DAiD and FAiD substantially reduce such local distortions, with FAiD yielding the highest visual quality. These results confirm that autoregressive retrieval effectively maintains object consistency and structural integrity throughout the generation process, particularly for complex objects and multi-object scenes.

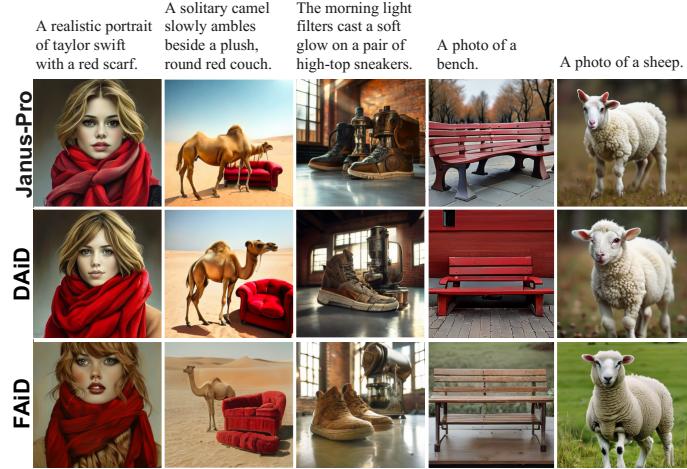


Figure 4: Qualitative results of DAiD, FAiD and baselines. Both DAiD and FAiD substantially reduce such local distortions, with FAiD yielding the highest visual quality. These results confirm that autoregressive retrieval effectively maintains object consistency and structural integrity throughout the generation process, particularly for complex objects and multi-object scenes.

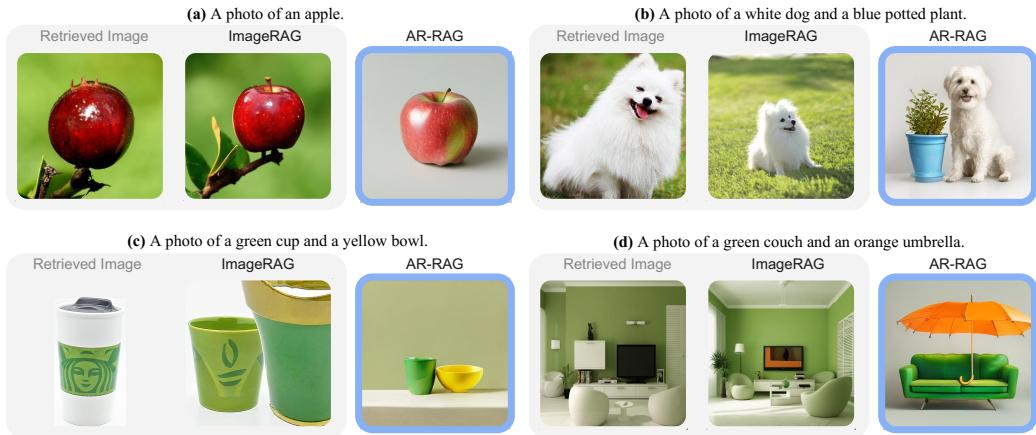


Figure 5: Images generated by ImageRAG [33] and our AR-RAG. ImageRAG excessively copies retrieved images and does not follow user prompts.

Figure 5 presents a comparative analysis of conventional image-level and our autoregressive patch-level retrieval augmentation methods. By comprehensively examining images produced by ImageRAG alongside their corresponding retrieved reference images, we identify two critical challenges inherent in image-level retrieval augmentation approaches. First, these methods tend to overcopy irrelevant visual elements from retrieved reference images into the generation outputs. As illustrated in Figure 5 (a), when generating an image of an apple, image-level retrieval approaches retrieve a reference image showing an apple on a tree branch and subsequently incorporate both the apple and the surrounding branches, despite the prompt making no mention of them. Similarly, for the prompt “*a green cup and a yellow bowl*” in Figure 5 (b), the image-level retrieval augmentation approach retrieves a green Starbucks cup and reproduces the pattern on the cup in the generated image,

despite this element not being part of the original instruction. This overcopying behavior directly compromises the instruction-following capability of generative models. Figure 5 (c) demonstrates that when prompted to generate “*A photo of a white dog and a blue potted plant*,” image-level retrieval methods produce an image containing only the white dog, omitting the blue potted plant entirely. Similarly, for “*a photo of a green couch and an orange umbrella*” in Figure 5 (d), the generated image fails to include the umbrella. This degradation in instruction following occurs because image-level retrieval biases the generation process toward the compositional structure of retrieved reference images, which may not align with the multi-object relationships specified in the prompt. In contrast, by autoregressively retrieving and integrating visual information at the fine-grained patch level rather than the image level, AR-RAG enables selective incorporation of relevant visual elements while maintaining independence from irrelevant contextual features present in the reference images.

5.3 Inference Time Cost

Table 4 shows the inference time comparisons across different models when generating 100 images using both a single L40 GPU. The DAiD method introduces only a minimal increase in inference time compared to the base Janus-Pro-1B model, with an average overhead of just 0.22%, demonstrating that DAiD maintains high computational efficiency. FAiD shows a more noticeable overhead of 36.03% on a single GPU due to its autoregressive retrieval and feature blending operations. However, this increase remains reasonable given the substantial performance gains in generation quality. Overall, both DAiD and FAiD do not significantly compromise the inference efficiency of Janus-Pro, making them practical for real-world applications.

6 Related Work

Retrieval-augmented generation (RAG) has emerged as a powerful paradigm that enhances generative models by incorporating external knowledge during decoding [23, 13, 16, 47, 46, 15, 24, 25, 42]. Originally developed for natural language processing, RAG enables models to retrieve relevant documents to supplement parametric knowledge during response generation [4], and has been widely adopted in many downstream tasks, such as knowledge-intensive tasks [23], document fusion [19], model pretraining [16], dialogue generation [35, 1], and so on.

Beyond the text domain, prior research has explored enhancing image generation by incorporating external visual references. Early approaches [8, 3] condition the diffusion process on retrieved images, typically encoded via CLIP or VAE encoders, to guide generation toward higher visual fidelity. KNN-Diffusion [34] extends this idea by leveraging k -nearest neighbor images to improve zero-shot generalization to novel domains. Building on this retrieval-augmented framework, more recent methods [49, 33] introduce adaptive retrieval pipelines that iteratively refine retrieved images based on feedback from multimodal large language models (MLLMs) analyzing the generated outputs. These methods enable context-aware and prompt-sensitive guidance during generation. Another line of work [46] encodes multimodal retrievals into discrete visual and text tokens, and uses them directly as contextual input to augment the generation process of a multimodal large language model. All of these works differ from our method by that our method works on patch-level, enabling more fine grain retrievals and can dynamically adjust retrievals based on evolving generation states.

7 Conclusion

In this work, we propose Autoregressive Retrieval Augmentation (AR-RAG), a novel retrieval paradigm that enhances image synthesis by leveraging k -nearest neighbor retrievals at the patch level. Unlike traditional image-level retrieval approaches, AR-RAG enables fine-grained visual element integration while maintaining compositional flexibility. We introduce two parallel frameworks: (1) Distribution-Augmentation in Decoding (DAiD), a training-free approach that integrates retrieved patch distributions directly into generation, and (2) Feature-Augmentation in Decoding (FAiD), which employs parameter-efficient fine-tuning with multi-scale feature smoothing and compatibility-based feature augmentation. Extensive experiments across GenEval, DPG-Bench, and Midjourney-30K demonstrate that AR-RAG significantly outperforms both conventional and retrieval-augmented baselines, particularly in handling complex prompts with multiple objects and specific spatial rela-

Model	Single GPU (L40)	
	Total (s)	Average (s)
ImageRAG	879.64	8.80
Janus-Pro	457.74	4.58
+ DAiD	459.34	4.59 (+0.22%)
+ FAiD	623.01	6.23 (+36.03%)

Table 4: Inference time for generating 100 images on a single L40 card.

tionships. Our methods substantially reduce local distortions in generated images, improving object consistency and structural integrity.

References

- [1] Trevor Ashby, Adithya Kulkarni, Jingyuan Qi, Minqian Liu, Eunah Cho, Vaibhav Kumar, and Lifu Huang. Towards effective long conversation generation with dynamic topic tracking and recommendation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 540–556, 2024.
- [2] Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, et al. Deepseek llm: Scaling open-source language models with longtermism. *arXiv preprint arXiv:2401.02954*, 2024.
- [3] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR, 17–23 Jul 2022.
- [5] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *CoRR*, abs/2102.08981, 2021.
- [6] Juhai Chen, Zhiyang Xu, Xichen Pan, Yushi Hu, Can Qin, Tom Goldstein, Lifu Huang, Tianyu Zhou, Saining Xie, Silvio Savarese, Le Xue, Caiming Xiong, and Ran Xu. Blip3-o: A family of fully open unified multimodal models-architecture, training and dataset, 2025.
- [7] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James T. Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis. *CoRR*, abs/2310.00426, 2023.
- [8] Wenhui Chen, Hexiang Hu, Chitwan Saharia, and William W. Cohen. Re-imagen: Retrieval-augmented text-to-image generator. In *The Eleventh International Conference on Learning Representations*, 2023.
- [9] Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *CoRR*, abs/2501.17811, 2025.
- [10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024.
- [11] Wan-Cyuan Fan, Yen-Chun Chen, Dongdong Chen, Yu Cheng, Lu Yuan, and Yu-Chiang Frank Wang. Frido: Feature pyramid diffusion for complex scene image synthesis. *CoRR*, abs/2208.13753, 2022.
- [12] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga

- Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023.
- [13] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997, 2023.
 - [14] Dhruba Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *CoRR*, abs/2310.11513, 2023.
 - [15] Liangke Gui, Borui Wang, Qiuyuan Huang, Alex Hauptmann, Yonatan Bisk, and Jianfeng Gao. Kat: A knowledge augmented transformer for vision-and-language. *arXiv preprint arXiv:2112.08614*, 2021.
 - [16] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR, 2020.
 - [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS’17, page 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc.
 - [18] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment, 2024.
 - [19] Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online, April 2021. Association for Computational Linguistics.
 - [20] Sadeep Jayasumana, Sri Kumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *CVPR*, pages 9307–9315, 2024.
 - [21] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.
 - [22] Xuan Ju, Ailing Zeng, Chenchen Zhao, Jianan Wang, Lei Zhang, and Qiang Xu. Humansd: A native skeleton-guided diffusion model for human image generation. *CoRR*, abs/2304.04269, 2023.
 - [23] Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc’Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.
 - [24] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36:22820–22840, 2023.
 - [25] Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. Preflmr: Scaling up fine-grained late-interaction multi-modal retrievers. *arXiv preprint arXiv:2402.08327*, 2024.
 - [26] Wenquan Lu, Yufei Xu, Jing Zhang, Chaoyue Wang, and Dacheng Tao. Handrefiner: Refining malformed hands in generated images by diffusion-based conditional inpainting. In Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian,

Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu, editors, *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, pages 7085–7093. ACM, 2024.

- [27] Xichen Pan, Satya Narayan Shukla, Aashu Singh, Zhuokai Zhao, Shlok Kumar Mishra, Jialiang Wang, Zhiyang Xu, Juhai Chen, Kunpeng Li, Felix Juefei-Xu, Ji Hou, and Saining Xie. Transfer between modalities with metaqueries. *CoRR*, abs/2504.06256, 2025.
- [28] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
- [29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. pages 8748–8763, 2021.
- [30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.
- [31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021.
- [32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.
- [33] Rotem Shalev-Arkushin, Rinon Gal, Amit H. Bermano, and Ohad Fried. Imagerag: Dynamic image retrieval for reference-guided image generation, 2025.
- [34] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. kNN-diffusion: Image generation via large-scale retrieval. In *The Eleventh International Conference on Learning Representations*, 2023.
- [35] Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. Retrieval augmentation reduces hallucination in conversation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
- [36] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang, Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li. Journeydb: A benchmark for generative image understanding. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.
- [37] Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *CoRR*, abs/2406.06525, 2024.
- [38] Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 84839–84865. Curran Associates, Inc., 2024.
- [39] Lokesh Veeramacheneni, Moritz Wolter, Hilde Kuehne, and Juergen Gall. Fréchet wavelet distance: A domain-agnostic metric for image generation. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [40] Vivym. Midjourney prompts dataset. <https://huggingface.co/datasets/vivym/midjourney-prompts>, 2023. Accessed: 2024-04-11.

- [41] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *CoRR*, abs/2409.18869, 2024.
- [42] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*, 2023.
- [43] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: efficient high-resolution image synthesis with linear diffusion transformers. *CoRR*, abs/2410.10629, 2024.
- [44] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. *CoRR*, abs/2408.12528, 2024.
- [45] Zhiyang Xu, Minqian Liu, Ying Shen, Joy Rimchala, Jiaxin Zhang, Qifan Wang, Yu Cheng, and Lifu Huang. Modality-specialized synergizers for interleaved vision-language generalists. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.
- [46] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. Retrieval-augmented multimodal language modeling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39755–39769. PMLR, 2023.
- [47] Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. *CoRR*, abs/2310.01558, 2023.
- [48] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning. *CoRR*, abs/2309.02591, 2023.
- [49] Huaying Yuan, Ziliang Zhao, Shuteng Wang, Shitao Xiao, Minheng Ni, Zheng Liu, and Zhicheng Dou. FineRAG: Fine-grained retrieval-augmented text-to-image generation. In Owen Ramshaw, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 11196–11205, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics.

A Multi-Scale Feature Smoothing Algorithm

Algorithm A illustrates the multi-scale feature smoothing, which is the core computational procedure for refining retrieved patch representations within their generation context. This algorithm ensures that retrieved visual elements are spatially and stylistically coherent with the surrounding image content through systematic multi-scale convolution operations.

The algorithm processes each retrieved patch representation $\hat{\mathbf{h}}_i$ independently, applying convolution operations at multiple scales ranging from 2×2 to $Q \times Q$ kernels. For each scale q , the algorithm initializes a temporary feature tensor $\mathbf{M} \in \mathbb{R}^{Q \times Q \times D}$ and an accumulation vector $\hat{\mathbf{h}}_q \in \mathbb{R}^D$. The nested loops over indices m and n systematically extract local patch features from different spatial windows around the target position (i, j) . Each extraction operation $\mathbf{H}_{\text{loc}}^l \leftarrow \mathbf{H}^l[i - m : i + q - m, j - n : j + q - n]$ captures a local neighborhood of size $q \times q$ centered at varying offsets from the target position.

Algorithm 1: Multi-Scale Feature Smoothing

Input: Image Representations $\mathbf{H}^l \in \mathbb{R}^{\sqrt{N} \times \sqrt{N} \times D}$,

Retrieved Patch Representations

$[\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \dots, \hat{\mathbf{h}}_K]$, Next Patch Index (i, j)

Output: Updated hidden states $[\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, \dots, \hat{\mathbf{h}}_K]$

```

1   foreach  $\hat{\mathbf{h}}_i \in [\hat{\mathbf{h}}_1, \dots, \hat{\mathbf{h}}_K]$  do
2       for  $q = 2$  to  $Q$  do
3           Initialize tensor:  $\mathbf{M} \leftarrow \mathbf{0} \in \mathbb{R}^{Q \times Q \times D}$ ;
4           Initialize tensor:  $\hat{\mathbf{h}}_q \leftarrow \mathbf{0} \in \mathbb{R}^D$ ;
5           for  $m = q$  down to 1 do
6               for  $n = q$  down to 1 do
7                    $\mathbf{H}_{\text{loc}}^l \leftarrow \mathbf{H}^l[i - m : i + q - m, j - n : j + q - n]$ ;
8                    $\mathbf{M}_{mn} \leftarrow \text{Conv}_{q \times q}^1(\mathbf{H}_{\text{loc}}^l)$ ;
9                    $\hat{\mathbf{h}}_q += \text{Conv}_{q \times q}^2(\mathbf{M})$ ;
10                   $\hat{\mathbf{h}}_i \leftarrow \frac{\hat{\mathbf{h}}_q}{Q - 1}$ ;

```

The extracted local features undergo two-stage convolution processing. The first convolution operation $\text{Conv}_{q \times q}^1$ transforms the local patch features into an intermediate representation stored in \mathbf{M}_{mn} , effectively capturing contextual relationships within each local window. The second convolution operation $\text{Conv}_{q \times q}^2$ processes the accumulated intermediate features to produce scale-specific refined representations. This two-stage design enables the algorithm to first capture local contextual patterns and then integrate them into a coherent scale-specific feature representation.

After processing all scales for a given retrieved patch, the algorithm computes the final refined representation by averaging the scale-specific features. The normalization factor $(Q - 1)$ accounts for the number of scales processed, ensuring consistent feature magnitudes across different retrieved patches. This averaging operation effectively combines multi-scale contextual information into a single refined representation that preserves both fine-grained details from smaller kernel sizes and broader contextual patterns from larger kernel sizes. The resulting refined patch representations maintain spatial coherence with the surrounding generation context while preserving the essential visual characteristics of the retrieved content.

B Experiment Setup

B.1 RA-CM3 Implementation Details

Since the pretrained RA-CM3 model is not publicly available, we implement our own version following the methodology described in the original paper to serve as a representative baseline for image-level retrieval-augmented generation. Our implementation uses Janus-Pro as the backbone model to ensure fair comparison with our proposed methods, as both approaches operate on the same foundation architecture.

We construct an image-level retrieval database using the same CC12M [5] and JourneyDB [36] datasets employed for our patch-level retrieval database to maintain consistency in the underlying data distribution. All images in the database are encoded into 512 dimensional vector representations using a pretrained CLIP [29] model. For each training instance in our 50,000 sample training set, we retrieve the most relevant reference image by encoding the corresponding text prompt with the same CLIP model, extracting the [CLS] token as the text representation, and computing cosine similarity scores between the text representation and all image representations in the database. The image

with the highest similarity score is selected as the retrieved reference. Each retrieved image is then processed through the quantized autoencoder from Janus-Pro to obtain image tokens $[v_1, \dots, v_N] = \mathcal{Z}(\theta_{Enc}(I))$, which are subsequently encoded into 2048 dimensional vector representations in the language model’s latent space using the image embedding and aligning layers in Janus-Pro. These retrieved image representations are concatenated with the text embeddings of the input prompts to form the augmented input for training the retrieval-enhanced model, which is the same training strategy used in RA-CM3.

During inference, given a text prompt for image generation, we follow the same retrieval process used in training. The input prompt is encoded using the CLIP text encoder, and we compute cosine similarity with all images in the database to identify the most relevant reference image. The retrieved image is processed through the same pipeline to obtain its representation in the language model’s latent space. This representation is then prepended to the text prompt embedding to provide the model with both textual and visual context for generation. The augmented input is fed into the fine-tuned Janus-Pro model to generate the output image following the standard autoregressive generation procedure.

B.2 Show-o Implementation Details

Our patch-based autoregressive retrieval augmentation methods can be theoretically adapted to any model that generates images through discrete tokens. To demonstrate this generalizability, we implement both DAID and FAID on the Show-o [44] model, which generates images through a masked token decoding process rather than strict left-to-right autoregression. Show-o decodes multiple image tokens simultaneously at each time step by converting masked tokens to specific image tokens based on a learned probability matrix. This fundamental difference in generation strategy necessitates several architectural adaptations to effectively incorporate our patch-based retrieval mechanisms while maintaining the model’s inherent generation capabilities.

DAID on Show-o The implementation of DAID on Show-o requires three key modifications to accommodate its non-autoregressive generation strategy. First, instead of constructing retrieval queries from upper-left neighboring patches as in autoregressive models, we utilize all eight surrounding patches to form the h -hop neighborhood representation for each target token position (i, j) . This comprehensive neighborhood encoding is computed as $[\mathbf{V}_{(i-1)(j-1)} : \mathbf{V}_{(i-1)(j)} : \mathbf{V}_{(i-1)(j+1)} : \mathbf{V}_{(i)(j-1)} : \mathbf{V}_{(i)(j+1)} : \mathbf{V}_{(i+1)(j-1)} : \mathbf{V}_{(i+1)(j)} : \mathbf{V}_{(i+1)(j+1)}]$, where missing positions are filled with zero vectors $\mathbf{0}$. Second, to mitigate retrieval noise arising from sparse neighborhood information in early time steps, we apply patch-level retrieval only during the final half of Show-o’s decoding process when sufficient contextual information is available. Third, since Show-o simultaneously predicts tokens for all patch positions at each time step rather than sequentially, we perform retrieval for all patch positions concurrently. At each qualifying time step t , for every patch position (i, j) in the partially generated image, we extract the eight-neighborhood representation as the retrieval query and obtain the top- K most similar patches $[\hat{\mathbf{v}}_1^{(i,j)}, \hat{\mathbf{v}}_2^{(i,j)}, \dots, \hat{\mathbf{v}}_K^{(i,j)}]$ from our database. We then construct position-specific retrieval distributions $D_{\text{retrieval}}^{(i,j)} \in \mathbb{R}^{|\mathcal{Z}|}$ using the same softmax formulation over retrieval distances as described in the main paper. These retrieval distributions are merged with Show-o’s predicted distributions for each patch position using the weighted average $D_{\text{merge}}^{(i,j)} = (1 - \lambda) \cdot D_{\text{model}}^{(i,j)} + \lambda \cdot D_{\text{retrieval}}^{(i,j)}$, where λ controls the retrieval influence across all positions.

FAID on Show-o The adaptation of FAID to Show-o involves both training and inference modifications to accommodate the model’s masked token generation process. During training, we prepare the training dataset by applying Show-o’s noise injection process to generate intermediate noisy representations at each time step, which serve as ground truth targets for the denoising process. For each training instance, we save these intermediate representations and apply patch-level retrieval to obtain relevant patches for all time steps. The training objective remains consistent with the standard Show-o formulation, but with augmented input representations that incorporate retrieved patch information. We insert FAID modules into every L/b decoder layers of Show-o’s Φ model, where each module processes all patch positions simultaneously rather than focusing on a single next token. At each qualifying time step and for each FAID-equipped layer l , we construct the 2D spatial representation $\mathbf{H}^l \in \mathbb{R}^{\sqrt{N} \times \sqrt{N} \times D}$ from the current hidden states and perform multi-scale feature smoothing for all patch positions. For each position (i, j) and its corresponding retrieved patches

$[\hat{\mathbf{h}}_1^{(i,j)}, \hat{\mathbf{h}}_2^{(i,j)}, \dots, \hat{\mathbf{h}}_K^{(i,j)}]$, we apply the convolution operations $\{\text{Conv}_{2\times 2}, \text{Conv}_{3\times 3}, \dots, \text{Conv}_{Q\times Q}\}$ to capture contextual patterns at multiple scales. The refined representations are computed as $\hat{\mathbf{h}}_k^{(i,j)} \leftarrow \sum_{q=2}^Q \text{softmax}(\Omega)_q \cdot \hat{\mathbf{h}}_{k,q}^{(i,j)}$, where $\hat{\mathbf{h}}_{k,q}^{(i,j)}$ represents the output of the $q \times q$ convolution for patch k at position (i, j) . The final augmented representation for each position is calculated as $h_{ij}^{(l+1)} = h_{ij}^l + \Delta h_{ij}^l + \sum_{k=1}^K s_k^{(i,j)} \hat{\mathbf{h}}_k^{(i,j)}$, where Δh_{ij}^l represents the standard transformer layer updates including self-attention and feed-forward components, and $s_k^{(i,j)}$ are position-specific compatibility scores computed through learned linear projections. During inference, we follow the same procedure but apply retrieval and feature blending only during the final half of the generation time steps to ensure sufficient contextual information is available for effective patch integration.

B.3 Training Setup

Training Datasets For model training, we utilize two large-scale image-caption datasets: CC12M [5] and Midjourney-v6⁴. From the training sets of these datasets, we randomly sample a total of 50,000 image-caption pairs (25,000 from each dataset) to fine-tune our model. Each image is encoded into 576 patch features and corresponding image tokens with the same image tokenizer [37] employed in the Janus-Pro model. For each image patch, we further retrieve the top- K image tokens from our retrieval database that exhibit similar neighborhood relationships. Consequently, each training instance comprises: (1) a textual image caption that serves as the conditioning input, (2) a sequence of 576 image tokens representing the ground-truth image, where each token is paired with K relevant image tokens retrieved from the database based on similar contextual features.

Training Details For the implementation of our FAiD approach, we fine-tune two pre-trained text-to-image generation models using the training dataset of 50K text-image pairs that we constructed. We select Janus-Pro-1B [9] and Show-o [44] as our base models. The fine-tuning process is conducted on 4 NVIDIA A100 (80GB) GPUs with a global batch size of 256 for a single epoch. We utilize the AdamW optimizer without weight decay, incorporating a 10% linear warm-up schedule followed by a constant learning rate of 2e-4.

B.4 Evaluation Benchmarks and Metrics

To comprehensively evaluate our proposed methods, we adopt multiple widely used benchmarks that assess different aspects of image generation quality:

- **GenEval** [14] is a benchmark designed to evaluate models’ ability to understand and generate images based on specific attributes and relationships described in text prompts. It comprises multiple categories such as single object generation, two-object composition, counting, colors, positioning, color attribution and so on. Performance is measured as the percentage of generated images that correctly align with the text descriptions.
- **DPG-Bench** [18] (Detailed Prompt Generation Benchmark) evaluates how well image generation models handle detailed prompts with complex requirements, covering categories such as global image quality, entity generation, attribute accuracy, relationship modeling, and other complex generation tasks. Scores are reported as percentages.
- For the **Midjourney-30k benchmark** [40], we employ three complementary metrics to evaluate the quality of generated images, including (1) Fréchet Inception Distance (FID) [17], which measures the statistical similarity between the distributions of generated and real images in the feature space of a pre-trained Inception network; (2) CLIP-MMD (CMMMD) [20], which measures the distance between real and generated images using CLIP embeddings and the Maximum Mean Discrepancy, and is specifically designed to better align with human perception of image quality and addresses several limitations of FID, including poor sample efficiency and incorrect normality assumptions; and (3) Fréchet Wavelet Distance (FWD) [39], which measures the distance between real and generated images in the wavelet packet coefficient space. FWD captures both spatial and frequency information without relying on pre-trained networks, making it domain-agnostic and robust to domain shifts across various image types. *For all three metrics, lower scores indicate higher-quality image generation, with both CMMMD and FWD particularly effective in capturing distortions in generated images in ways that better correlate with human judgements.*

⁴<https://huggingface.co/datasets/brivangl/midjourney-v6-llava>

C Experiment Results and Discussion

C.1 Accuracy of Patch-based Autoregressive Retrieval

To assess the effectiveness of our patch-level autoregressive retrieval mechanism, we conduct a comparative analysis between the top- K retrieved image tokens and the ground-truth tokens to be generated. Specifically, we randomly sampled 1,000 instances from our training set, each comprising 576 image tokens and $576 \times k$ retrieved tokens. To demonstrate the accuracy of the retrieved image tokens, for each ground-truth image token, we also randomly sample a vocabulary code as non-relevant tokens. Using the shared codebook, we transform all image tokens into vector representations and compute the l_2 distances between each ground-truth image token and its top- K retrieved counterparts. Similarly, we also compute the mean of the l_2 distance between each ground-truth token and the randomly sampled tokens. As shown in Figure 6, the l_2 distance between retrieved tokens and ground-truth image tokens is significantly smaller than the distance between randomly sampled tokens and ground-truth tokens. As k increases, the distance between the k -th retrieved token and the ground-truth token also increases, demonstrating the effectiveness of the retrieval approach and our assumption that image patches with similar neighbors usually exhibit inherent similarities.

C.2 Hyperparameter Optimization

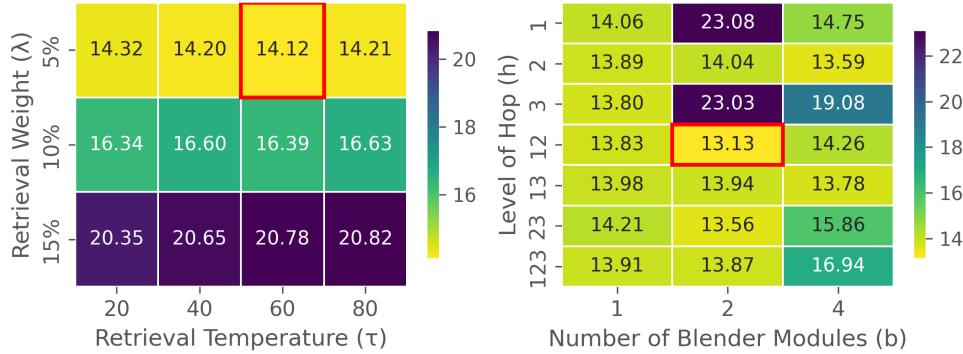


Figure 7: Hyperparameter optimization results for DAiD and FAiD on FID scores. Left: FID scores for DAiD across different combinations of retrieval temperature τ and merging weight λ . Right: FID scores for FAiD across varying levels of hop h and numbers of blender modules b . All experiments conducted on the Midjourney-10K benchmark, with optimal configurations highlighted by red borders.

Both DAiD and FAiD require careful optimization of distinct sets of hyperparameters. For DAiD, we optimize the retrieval temperature τ and merging weight λ , which control the retrieval-based probability distribution sharpness and the balance between retrieval and model predictions, respectively. For FAiD, we focus on the level of hop (h) and number of blender modules (b), determining the spatial context incorporated during retrieval and extent of feature blending. To identify optimal

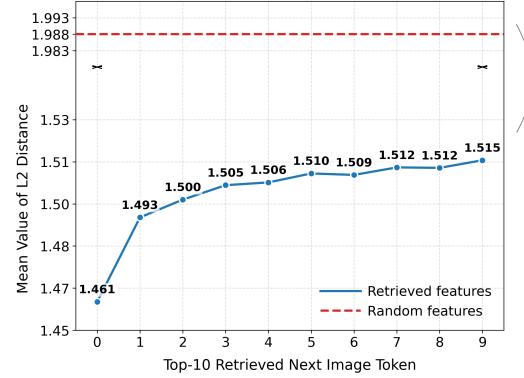


Figure 6: l_2 distance between ground-truth tokens and top-10 retrieved tokens (blue line) compared to randomly sampled tokens (red dashed line). The curved arrow indicates a broken y-axis that accommodates the large gap between the retrieved token and the random token baseline.

configurations, we conducted a systematic ablation study on the Midjourney-10K benchmark using Fréchet Inception Distance (FID) as the performance metric.

Figure 7 presents the FID scores for DAiD across different combinations of λ and τ , and for FAiD across varying levels of (h) and (b) , where composite hop levels such as “12” indicate combined use of multiple hop distances. Analysis of the DAiD results reveals that performance degrades as λ increases, suggesting that modest integration of retrieval information enhances performance while excessive reliance impairs generative flexibility. The retrieval temperature τ demonstrates less pronounced effects, though a moderate value of 0.6 provides marginal benefits. For FAiD, configurations incorporating multiple hop levels generally outperform single hop levels, with the “12” configuration yielding optimal results. Regarding blender modules, an intermediate value consistently delivers the best performance, implying that moderate feature blending optimizes the incorporation of retrieved patches while avoiding both under-utilization and over-smoothing. Based on this analysis, we selected $\lambda = 0.05$ and $\tau = 0.6$ for DAiD, and hop levels “12” with 2 blender modules for FAiD, achieving FID scores of 14.12 and 13.13, respectively. These configurations effectively harness retrieval information while preserving the generative strengths of the underlying Janus-Pro model, as demonstrated by their superior performance on the benchmark.

D Limitations

While our AR-RAG framework demonstrates strong performance across multiple benchmarks, several limitations should be acknowledged. First, our approach relies on discrete image tokenization and targets discrete token-based models, so it may not be directly applied to continuous diffusion models operating in latent spaces. Second, due to computational resource limitations, our retrieval database remains smaller than billion-scale databases. This limitation may introduce visual pattern biases, as the database may not fully capture the diversity of real-world visual patterns, potentially affecting the generation of underrepresented visual elements. Third, our implementation focuses exclusively on 2D image generation. While the underlying patch-based retrieval concept could theoretically extend to other structured generation tasks such as 3D point cloud generation, we have not explored these applications.

E Broader impacts

We propose a novel retrieval-augmented approach to enhance existing image generation models. Our method is both highly efficient and readily adaptable to a wide range of applications, making it valuable for both academic research and industrial deployment. However, as our approach builds upon existing generative models, it may inherit their biases and could potentially produce inappropriate outputs in the absence of additional safety mechanisms. Furthermore, the large-scale retrieval database may contain unsafe or undesirable content, which can be reflected in the retrieved image patches. To ensure safe deployment in real-world scenarios, additional safeguards and filtering measures are necessary to mitigate these risks.