The Future of RAG-Augmented Image Generation

Originally published at https://www.unite.ai/the-future-of-rag-augmented-image-generation/

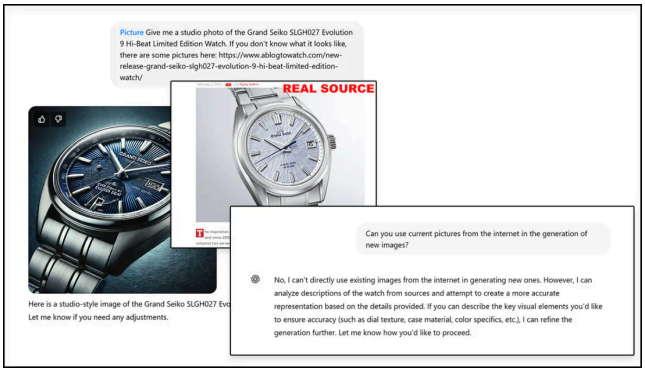
Generative diffusion models like Stable Diffusion, Flux, and video models such as Hunyuan rely on knowledge acquired during a single, resource-intensive training session using a fixed dataset. Any concepts introduced after this training – referred to as the knowledge cut-off – are absent from the model unless supplemented through fine-tuning or external adaptation techniques like Low Rank Adaptation (LoRA).

It would therefore be ideal if a generative system that outputs images or videos could *reach out to online sources* and bring them into the generation process as needed. In this way, for instance, a diffusion model that knows nothing about the very latest Apple or Tesla release could still produce images containing these new products.

In regard to language models, most of us are familiar with systems such as Perplexity, Notebook LM and ChatGPT-4o, that can incorporate novel external information in a Retrieval Augmented Generation (RAG) model.

RAG processes make ChatGPT 4o's responses more relevant. Source: https://chatgpt.com/

However, this is an uncommon facility when it comes to generating images, and ChatGPT will confess its own limitations in this regard:



ChatGPT 4o has made a good guess about the visualization of a brand new watch release, based on the general line and on descriptions it has interpreted; but it cannot 'absorb' and integrate new images into a DALL-E-based generation.

Incorporating externally retrieved data into a generated image is challenging because the incoming image must first be broken down into tokens and embeddings, which are then mapped to the model's nearest trained domain knowledge of the subject.

While this process works effectively for post-training tools like ControlNet, such manipulations remain largely superficial, essentially funneling the retrieved image through a rendering pipeline, but without deeply integrating it into the model's internal representation.

As a result, the model lacks the ability to generate novel perspectives in the way that neural rendering systems like NeRF can, which construct scenes with true spatial and structural understanding.

Mature Logic

A similar limitation applies to RAG-based queries in Large Language Models (LLMs), such as Perplexity. When a model of this type processes externally retrieved data, it functions much like an adult drawing on a lifetime of knowledge to infer probabilities about a topic.

However, just as a person cannot retroactively integrate new information into the cognitive framework that shaped their fundamental worldview – when their biases and preconceptions were still forming – an LLM cannot seamlessly merge new knowledge into its pre-trained structure.

Instead, it can only 'impact' or juxtapose the new data against its existing internalized knowledge, using learned principles to analyze and conjecture rather than to synthesize at the foundational level.

This short-fall in equivalency between *juxtaposed* and *internalized* generation is likely to be more evident in a generated image than in a language-based generation: the deeper network connections and increased creativity of 'native' (rather than RAG-based) generation has been established in various studies.

Hidden Risks of RAG-Capable Image Generation

Even if it were technically feasible to seamlessly integrate retrieved internet images into newly synthesized ones in a RAG-style manner, safety-related limitations would present an additional challenge.

Many datasets used for training generative models have been curated to minimize the presence of explicit, racist, or violent content, among other sensitive categories. However, this process is imperfect, and residual associations can persist. To mitigate this, systems like DALL·E and Adobe Firefly rely on secondary filtering mechanisms that screen both input prompts and generated outputs for prohibited content.

As a result, a simple NSFW filter – one that primarily blocks overtly explicit content – would be insufficient for evaluating the acceptability of retrieved RAG-based data. Such content could still be offensive or harmful in ways that fall outside the model's predefined moderation parameters, potentially introducing material that the AI lacks the contextual awareness to properly assess.

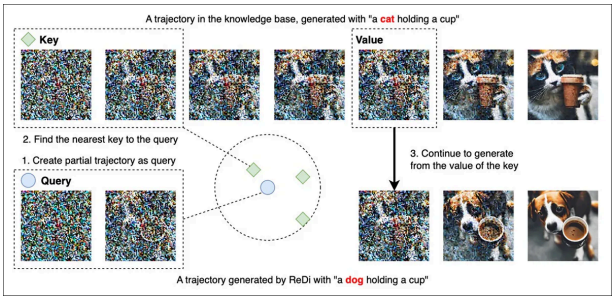
Discovery of a recent vulnerability in the CCP-produced DeepSeek, designed to suppress discussions of banned political content, has highlighted how alternative input pathways can be exploited to bypass a model's ethical safeguards; arguably, this applies also to arbitrary novel data retrieved from the internet, when it is intended to be incorporated into a new image generation.

RAG for Image Generation

Despite these challenges and thorny political aspects, a number of projects have emerged that attempt to use RAG-based methods to incorporate novel data into visual generations.

ReDi

The 2023 Retrieval-based Diffusion (ReDi) project is a learning-free framework that speeds up diffusion model inference by retrieving similar *trajectories* from a precomputed knowledge base.



Values from a dataset can be 'borrowed' for a new generation in ReDi. Source: https://arxiv.org/pdf/2302.02285

In the context of diffusion models, a trajectory is the step-by-step path that the model takes to generate an image from pure noise. Normally, this process happens gradually over many steps, with each step refining the image a little more.

ReDi speeds this up by skipping a bunch of those steps. Instead of calculating every single step, it retrieves a similar past trajectory from a database and jumps ahead to a later point in the process. This reduces the number of calculations needed, making diffusion-based image generation much faster, while still keeping the quality high.

ReDi does not modify the diffusion model's weights, but instead uses the knowledge base to skip intermediate steps, thereby reducing the number of function estimations needed for sampling.

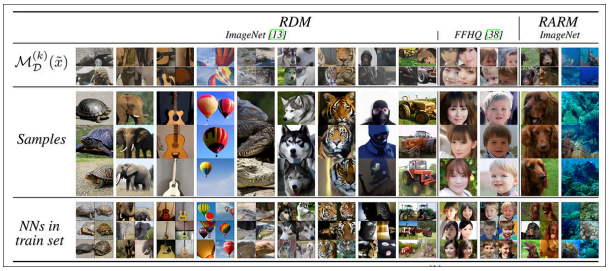
Of course, this is not the same as incorporating specific images at will into a generation request; but it does relate to similar types of generation.

Released in 2022, the year that latent diffusion models captured the public imagination, ReDi appears to be among the earliest diffusion-based approach to lean on a RAG methodology.

Though it should be mentioned that in 2021 Facebook Research released Instance-Conditioned GAN, which sought to condition GAN images on novel image inputs, this kind of *projection* into the latent space is extremely common in the literature, both for GANs and diffusion models; the challenge is to make such a process training-free and functional in real-time, as LLM-focused RAG methods are.

RDM

Another early foray into RAG-augmented image generation is Retrieval-Augmented Diffusion Models (RDM), which introduces a semi-parametric approach to generative image synthesis. Whereas traditional diffusion models store all learned visual knowledge within their neural network parameters, RDM relies on an external image database:



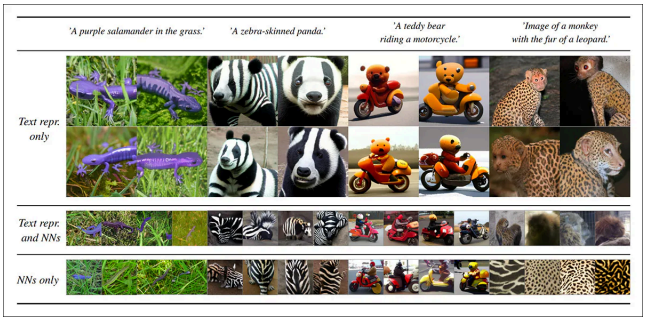
Retrieved nearest neighbors in an illustrative pseudo-query in RDM.*

During training the model retrieves *nearest neighbors* (visually or semantically similar images) from the external database, to guide the generation process. This allows the model to condition its outputs on real-world visual instances.

The retrieval process is powered by CLIP embeddings, designed to force the retrieved images to share meaningful similarities with the query, and also to provide novel information to improve generation.

This reduces reliance on parameters, facilitating smaller models that achieve competitive results without the need for extensive training datasets.

The RDM approach supports *post-hoc* modifications: researchers can swap out the database at inference time, allowing for zero-shot adaptation to new styles, domains, or even entirely different tasks such as stylization or class-conditional synthesis.



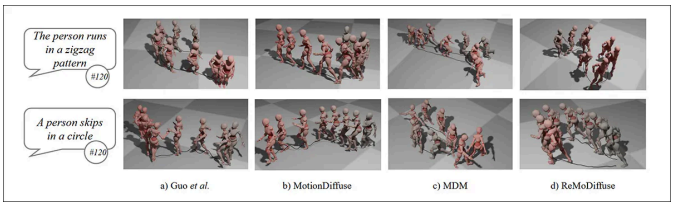
In the lower rows, we see the nearest neighbors drawn into the diffusion process in RDM.*

A key advantage of RDM is its ability to improve image generation without retraining the model. By simply altering the retrieval database, the model can generalize to new concepts it was never explicitly trained on. This is particularly useful for applications where domain shifts occur, such as generating medical imagery based on evolving datasets, or adapting text-to-image models for creative applications.

Negatively, retrieval-based methods of this kind depend on the quality and relevance of the external database, which makes data curation an important factor in achieving high-quality generations; and this approach remains far from an image synthesis equivalent of the kind of RAG-based interactions typical in commercial LLMs.

ReMoDiffuse

ReMoDiffuse is a retrieval-augmented motion diffusion model designed for 3D human motion generation. Unlike traditional motion generation models that rely purely on learned representations, ReMoDiffuse retrieves relevant motion samples from a large motion dataset and integrates them into the denoising process, in a schema similar to RDM (see above).

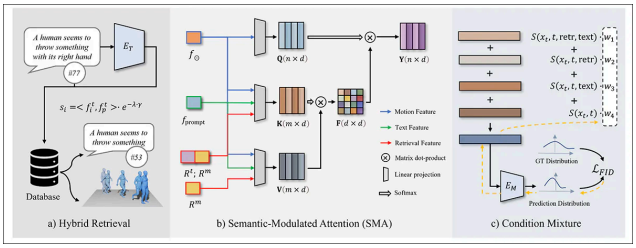


Comparison of RAG-augmented ReMoDiffuse (right-most) to prior methods. Source: https://arxiv.org/pdf/2304.01116

This allows the model to generate motion sequences designed to be more natural and diverse, as well as semantically faithful to the user's text prompts.

ReMoDiffuse uses an innovative *hybrid retrieval mechanism*, which selects motion sequences based on both semantic and kinematic similarities, with the intention of ensuring that the retrieved motions are not just thematically relevant but also physically plausible when integrated into the new generation.

The model then refines these retrieved samples using a *Semantics-Modulated Transformer*, which selectively incorporates knowledge from the retrieved motions while maintaining the characteristic qualities of the generated sequence:



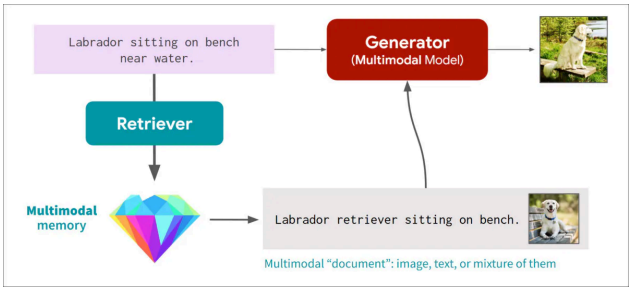
Schema for ReMoDiffuse's pipeline.

The project's *Condition Mixture* technique enhances the model's ability to generalize across different prompts and retrieval conditions, balancing retrieved motion samples with text prompts during generation, and adjusting how much weight each source gets at each step.

This can help prevent unrealistic or repetitive outputs, even for rare prompts. It also addresses the scale sensitivity issue that often arises in the classifier-free guidance techniques commonly used in diffusion models.

RA-CM3

Stanford's 2023 paper Retrieval-Augmented Multimodal Language Modeling (RA-CM3) allows the system to access real-world information at inference time:



Stanford's Retrieval-Augmented Multimodal Language Modeling (RA-CM3) model uses internet-retrieved images to augment the generation process, but remains a prototype without public access. Source: https://cs.stanford.edu/~myasu/files/RACM3_slides.pdf

RA-CM3 integrates retrieved text and images into the generation pipeline, enhancing both text-to-image and image-to-text synthesis. Using CLIP for retrieval and a Transformer as the generator, the model refers to pertinent multimodal documents before composing an output.

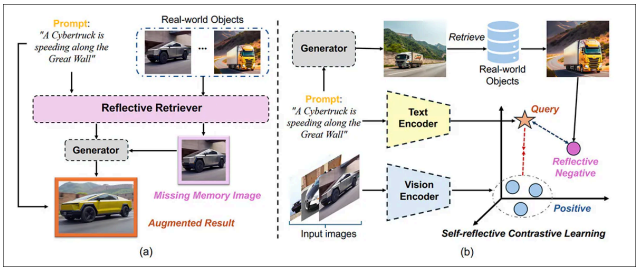
Benchmarks on MS-COCO show notable improvements over DALL-E and similar systems, achieving a 12-point Fréchet Inception Distance (FID) reduction, with far lower computational cost.

However, as with other retrieval-augmented approaches, RA-CM3 does not seamlessly internalize its retrieved knowledge. Rather, it superimposes new data against its pre-trained network, much like an LLM augmenting responses with search results. While this method can improve factual accuracy, it does not replace the need for training updates in domains where deep synthesis is required.

Furthermore, a practical implementation of this system does not appear to have been released, even to an API-based platform.

RealRAG

A new release from China, and the one that has prompted this look at RAG-augmented generative image systems, is called *Retrieval-Augmented Realistic Image Generation* (RealRAG).



External images drawn into RealRAG (lower middle). Source: https://arxiv.o7rg/pdf/2502.00848

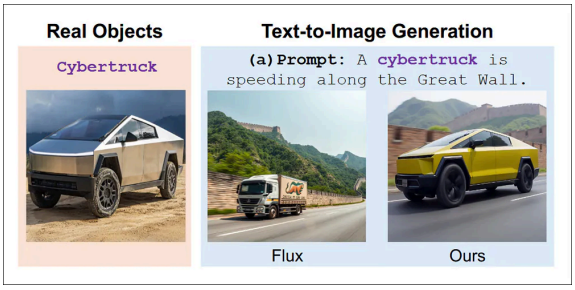
RealRAG retrieves actual images of relevant objects from a database curated from publicly available datasets such as ImageNet, Stanford Cars, Stanford Dogs, and Oxford Flowers. It then integrates the retrieved images into the generation process, addressing knowledge gaps in the model.

A key component of RealRAG is *self-reflective contrastive learning*, which trains a retrieval model to find informative reference images, rather than just selecting *visually similar* ones.

The authors state:

> *'Our key insight is to train a retriever that retrieves images staying off the generation space of the generator, yet closing to the representation of text prompts.*
>
> *'To this [end], we first generate images from the given text prompts and then utilize the generated images as queries to retrieve the most relevant images in the real-object-based database. These most relevant images are utilized as reflective negatives.'*

This approach ensures that the retrieved images contribute *missing knowledge* to the generation process, rather than reinforcing existing biases in the model.



Left-most, the retrieved reference image; center, without RAG; rightmost, with the use of the retrieved image.

However, the reliance on retrieval quality and database coverage means that its effectiveness can vary depending on the availability of high-quality references. If a relevant image doesn't exist in the dataset, the model may still struggle with unfamiliar concepts.

RealRAG is a very modular architecture, offering compatibility with multiple other generative architectures, including U-Net-based, DiT-based, and autoregressive models.

In general the retrieving and processing of external images adds computational overhead, and the system's performance depends on how well the retrieval mechanism generalizes across different tasks and datasets.

Conclusion

This is a representative rather than exhaustive overview of image-retrieving multimodal generative systems. Some systems of this type use retrieval solely to improve vision understanding or dataset curation, among other diverse motives, rather than seeking to generate images. One example is Internet Explorer.

Many of the other RAG-integrated projects in the literature remain unreleased. Prototypes, with only published research, include Re-Imagen, which – despite its provenance from Google – can only access images from a local custom database.

Also, In November 2024, Baidu announced *Image-Based Retrieval-Augmented Generation* (iRAG), a new platform that uses retrieved images 'from a database'. Though iRAG is reportedly available on the Ernie platform, there seem to be no further details about this retrieval process, which looks to rely on a *local* database (i.e., local to the service and not directly accessible to the user).

Further, the 2024 paper *Unified Text-to-Image Generation and Retrieval* offers yet another RAG-based method of using external images to augment results at generation time – again, from a local database rather than from *ad hoc* internet sources.

Excitement around RAG-based augmentation in image generation is likely to focus on systems that can incorporate internet-sourced or user-uploaded images directly into the generative process, and which allow users to participate in the choices or sources of images.

However, this is a significant challenge for at least two reasons; firstly, because the effectiveness of such systems usually depends on deeply integrated relationships formed during a resource-intensive training process; and secondly, because concerns over safety, legality, and copyright restrictions, as noted earlier, make this an unlikely feature for an API-driven web service, and for commercial deployment in general.

* Source: https://proceedings.neurips.cc/paper_files/paper/2022/file/62868cc2fc1eb5cdf321d05b4b88510c-Paper-Conference.pdf

First published Tuesday, February 4, 2025

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More