Retrieval-Augmented Multimodal Language Modeling

Michihiro Yasunaga,[1][*] **Armen Aghajanyan,**[2] **Weijia Shi,**[3] **Rich James,**[2] **Jure Leskovec,**[1] **Percy Liang**[1]
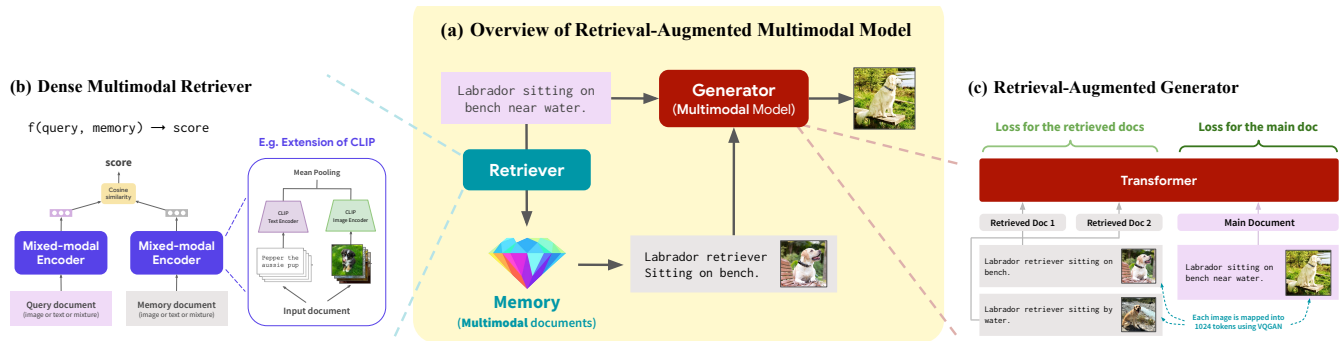**Mike Lewis,**[2] **Luke Zettlemoyer,**[3][2] **Wen-tau Yih**[2]

Figure 1. **Our approach, retrieval-augmented multimodal modeling**. **(a)** Overview: given an input multimodal document, we use a **retriever** to retrieve relevant multimodal documents from an external memory, and use the **generator** to refer to the retrieved documents and make predictions for the input (e.g., generate the continuation). **(b)** The multimodal retriever is a dense retriever with a mixed-modal encoder that can encode mixture of text and images (§3.2). **(c)** The retrieval-augmented generator uses the CM3 Transformer architecture, and we prepend the retrieved documents to the main input document that we feed into the model (§3.3).

Abstract

Recent multimodal models such as DALL-E and CM3 have achieved remarkable progress in text-to-image and image-to-text generation. However, these models store all their knowledge (e.g., the appearance of the Eiffel Tower) in the model parameters, requiring increasingly larger models and training data to capture more knowledge. To integrate knowledge in a more scalable and modular way, we propose a retrieval-augmented multimodal model, which enables a base multimodal model (*generator*) to refer to relevant text and images fetched by a *retriever* from external memory (e.g., documents on the web). Specifically, for the retriever, we use a pretrained CLIP, and for the generator, we train a CM3 Transformer on the LAION dataset. Our resulting model, named Retrieval-Augmented CM3 (RA-CM3), is the first multimodal model that can retrieve and generate both text and images. We show that RA-CM3 significantly outperforms baseline multimodal mod-
els such as DALL-E and CM3 on both image and caption generation tasks (12 FID and 17 CIDEr improvements on MS-COCO), while requiring much less compute for training ($< 30\%$ of DALL-E). Moreover, we show that RA-CM3 exhibits novel capabilities, such as faithful image generation and multimodal in-context learning (e.g., image generation from demonstrations).

1. Introduction

Recent multimodal models have achieved remarkable progress in image and text generation. DALL-E (Ramesh et al., 2021) and Parti (Yu et al., 2022) perform image generation from text, Flamingo (Alayrac et al., 2022) performs text generation from images, and CM3 (Aghajanyan et al., 2022) offers a unified Transformer model that generates both text and images. Typically, these models store all their knowledge (e.g., the appearance of the Eiffel Tower) implicitly in the parameters of the underlying neural network, requiring a lot of parameters (e.g., 10–80B) and training data (e.g., 1–10B images) to cover all the knowledge. This motivates the development of multimodal models that can refer to an external memory of knowledge (e.g., web data) for increased knowledge capacity. Access to external memory is useful to accommodate the growth and update of knowledge through time, and is especially helpful for tasks that involve entity knowledge, such as generating images

[1]Stanford University [2]Meta AI [3]University of Washington. [*]Majority of research was conducted as an intern at Meta AI. Correspondence to: Michihiro Yasunaga <myasu@cs.stanford.edu>.

Approach	Model type	Image generation	Text generation	Retrieval	In-context learning
DALL-E, Parti (Ramesh et al.; Yu et al.)	Autoregressive	✔			
DALL-E 2, Imagen (Ramesh et al.; Saharia et al.)	Diffusion	✔			
Re-Imagen (Chen et al.)	Diffusion	✔		✔	
Flamingo, MetaLM (Alayrac et al.; Hao et al.)	Autoregressive		✔		✔
MuRAG (Chen et al.)	Autoregressive		✔[†]	✔	
CM3 (Aghajanyan et al.)	Autoregressive	✔	✔		
RA-CM3 (Ours)	Autoregressive	✔	✔	✔	✔

Table 1. **Comparison with other multimodal models**. Our RA-CM3 is the first retrieval-augmented model that can perform both image and text generation. RA-CM3 also exhibits strong in-context learning abilities thanks to the proposed retrieval-augmented training (§3.3). [†]Focus on question answering.

for entity-rich captions like "George Washington standing in front of the Eiffel Tower". Reference to external memory may also offer benefits such as explainable and faithful predictions (Metzler et al., 2021).

Recently, retrieval-augmented language models have shown promise in natural language processing (NLP) (Karpukhin et al., 2020; Guu et al., 2020; Lewis et al., 2020b; Borgeaud et al., 2022). Given input text, such a model uses a *retriever* that retrieves relevant documents from an external memory, and uses a *generator* to generate predictions given the retrieved documents. However, these retrieval-augmented methods are studied originally for text, and extending them to the multimodal setting remains an open problem with challenges. Specifically, we need to design a retriever and a generator that handle multimodal documents, consisting of *both* images and text. Several concurrent works study retrieval for multimodal data (Chen et al., 2022a;b), but their generators are each limited to a single modality, either text generation or image generation (Table 1).

In this work, we address the above challenge and present the first retrieval-augmented multimodal model that can retrieve and generate *both* text and images. As in Figure 1, our input data and external memory comprise a set of *multimodal documents*, each of which is an arbitrary sequence of text/images (e.g., text, image, or their combinations like caption-image pair). First, to obtain a multimodal retriever, we use the Dense Retrieval method (Karpukhin et al., 2020) with a mixed-modal encoder that can encode combinations of text and images (e.g., pretrained CLIP; Radford et al. 2021). Given this retriever, we design a technique to retrieve diverse and informative documents for the input document. Second, we design the retrieval-augmented generator based on the CM3 architecture (Aghajanyan et al., 2022), which is a Transformer sequence model capable of both text and image generation. Concretely, we prepend the retrieved documents as in-context examples to the main input document, and train the generator by optimizing token prediction loss jointly for the main document and retrieved documents.

We train our retrieval-augmented CM3 (*RA-CM3*), using 150M text-image pairs from the LAION dataset (Schuhmann et al., 2021). RA-CM3 achieves strong performance on MS-COCO image and caption generation, significantly outperforming the baseline CM3 with no retrieval (12 FID and 17 CIDEr improvements). It also outperforms existing models such as DALL-E and Flamingo, despite using fewer parameters ($< 30\%$) and compute for training ($< 30\%$).

We further demonstrate novel capabilities of RA-CM3 (§5). First, it can perform faithful generation for tasks that require entity knowledge, for which existing models struggle (Figure 3, 4). Second, RA-CM3 exhibits a multimodal in-context learning ability: it can perform controlled image generation by prompting with demonstration examples in context (Figure 7), and it can also perform few-shot image classification. RA-CM3 is the first model that can perform in-context learning for both text and image generation (Table 1).

More broadly, our work offers a general and modular retrieval augmentation framework for multimodal models, and opens up various research avenues, such as further advancement of multimodal retrievers and generators.

2. Related work

We discuss related works in detail in §B.

3. Approach

We present a retrieval-augmented multimodal model that can retrieve and generate both text and images. As illustrated in Figure 1, given an input multimodal document (i.e., arbitrary sequence of text/images), we use a **retriever** that retrieves relevant multimodal documents from an external memory, and uses a **generator** to refer to the retrieved documents and make predictions for the input document (i.e., generate the continuation). We design the multimodal retriever as a dense retriever with a mixed-modal encoder that can encode combinations of text and images (e.g., pretrained CLIP; §3.2). We build the retrieval-augmented generator using the CM3 Transformer architecture, and we prepend the retrieved documents to the main input document that we feed into the generator (§3.3). We describe how we train this model and use it for text-to-image or image-to-text generation in §3.4.

Notably, our resulting model, *Retrieval-Augmented CM3 (RA-CM3)*, is the first multimodal model that can retrieve and generate *combinations* of text and images, which is the most general capability among existing multimodal models (Table 1). Moreover, while we build on existing techniques such as CLIP and CM3, we are the first to establish a method to unify them into a performant retrieval-augmented model through extensive analyses of design choices (§C.3).

3.1. Preliminaries

Retrieval augmented language model. The framework consists of a retrieval module R and a generator module G (e.g., language model). The retrieval module R takes an input sequence x and an external memory of documents $\mathcal{M}$, and returns a list of documents $M \subseteq \mathcal{M}$. The generator G then takes the input sequence x and the retrieved documents $M = (m_1, ..., m_K)$, and returns the target y, where y is the continuation of x in a typical language modeling task.

Causal masked multimodal model (CM3). CM3 (Aghajanyan et al., 2022) is a Transformer decoder (Vaswani et al., 2017) model for multimodal documents. A *multimodal document* is defined as an arbitrary sequence of text/images (e.g., text, image, or their combinations like caption-image pair). In particular, CM3 formats each multimodal document as an HTML sequence, such as "<img alt=[text] src=[image]>", where [text] is a sequence of text tokens, and [image] is a sequence of image tokens obtained by an image tokenizer such as VQGAN (Esser et al., 2021), which maps a raw image into 1024 tokens.

At training time, CM3 either takes the original sequence as input (e.g., $x_{\text{input}} =$ "Photo of a cat: [image]") or converts it into an infilling instance by masking some spans and moving them to the end (e.g., $x_{\text{input}} =$ "Photo of <mask>: [image] <infill> a cat"), and then optimizes the standard next token prediction loss for the input, $-\log p(x_{\text{input}})$. This provides a flexible model that learns to perform infilling besides standard autoregressive generation. In particular, the model can perform both image and text generation: for caption-to-image, CM3 generates a continuation from the prompt "Photo of a cat:". For image-to-caption, CM3 generates from the prompt "Photo of <mask>: [image] <infill>".

Our setup. We aim to generalize the retrieval-augmented language model framework to the multimodal setting. Our input x + target y will be a multimodal document, and our memory $\mathcal{M}$ will be a set of multimodal documents. We design the retrieval module R for multimodal data (§3.2), and design the multimodal generator G based on CM3 (§3.3).

3.2. Multimodal retrieval

Dense retriever. A retriever r takes a query q (e.g., the input sequence x) and a candidate document m from the memory $\mathcal{M}$, and returns a relevance score $r(q, m)$. We follow the Dense Retrieval method (Karpukhin et al., 2020), in which the retriever r is a bi-encoder architecture,

$$r(q, m) = E_Q(q)^\top E_M(m) \tag{1}$$

where the query encoder E_Q and memory encoder E_M produce dense vectors for the query and memory document, respectively (Figure 1b). As our input and memory are multimodal documents, we let E_Q and E_M be **mixed-modal encoders** that encode a combination of text and images. While any mixed-modal encoders could be used in our framework, we find that a simple extension of CLIP (Ramesh et al., 2021) works well empirically, so we adopt it in our final system. Concretely, as shown in Figure 1b (right), given a multimodal document, we split it into a text part and an image part, encode the two parts separately using off-the-shelf frozen CLIP text and image encoders, and then average the two, with the L2 norm scaled to 1, as the vector representation of the document. We use this encoding method for both E_Q and E_M. Intrinsic evaluation of this CLIP-based retriever can be found in §C.1.

Given this retriever r, we perform Maximum Inner Product Search (MIPS; §4.1) over the memory to obtain a list of candidate documents sorted by the relevance score. We then sample the final K retrieved documents from this list.

Retrieval strategy. We discuss three key factors in obtaining/sampling informative retrieved documents for the generator in practice.

Relevance: The retrieved documents need to be relevant to the input sequence; otherwise, the retrieved documents do not provide useful information for modeling the main input sequence (see §C.3 for the ablation study). The dense retriever score based on CLIP captures this relevance factor.

Modality: While existing works on retrieval (Chen et al., 2022b) typically retrieve either an image or text only for the generator, we find that retrieving a multimodal document that consists of both images and text leads to better generator performance (see §C.3). Our intuition is that a multimodal document can be more informative because the text and image within it can contextualize each other. Hence, in our final system, we retrieve the raw multimodal documents that keep both images and text for the generator.

Diversity: We find that ensuring diversity in retrieved documents is important. First, simply sampling or taking the top K from the document list based on the relevance score can result in duplicate or highly similar images or text, leading to poor generator performance. This is especially important in the multimodal setting because even when two multimodal documents are not duplicates by themselves, the images or text contained in them can be duplicates, hurting the generator performance. To avoid redundancy, when we take

documents from the top of the list, we skip a candidate if it is too similar (e.g., relevance score > 0.9) to the query or to the documents we already retrieved. Second, to further encourage diversity, we also propose Query Dropout, which drops some tokens of the query used in retrieval (e.g., 20% of tokens). This technique serves as regularization for training, and leads to further improvement in generator performance. Hence, our final system uses these two techniques (Avoid Redundancy + Query Dropout) for training, and uses Avoid Redundancy for inference. See §C.3 for detailed analysis.

3.3. Multimodal generator

We use CM3 as the base of our multimodal generator G. To incorporate the retrieved documents $M = (m_1, ..., m_K)$ into the generator, we prepend them to the main input sequence x, and feed the resulting sequence $(m_1, ..., m_K, x)$ to the Transformer (Figure 1c). In other words, the retrieved documents are in-context examples for the main input.

To train the generator, we optimize the following loss:

$$L = L_{\text{main}} + \alpha L_{\text{retr}} \qquad (2)$$
$$= -\log p(x \,|\, m_1, ..., m_K) - \alpha \log p(m_1, ..., m_K) \quad (3)$$

where L_{main} and L_{retr} are the CM3 token prediction loss for the main input sequence x and for the retrieved documents $(m_1, ..., m_K)$, respectively. Here we propose optimizing the two loss terms jointly, with $\alpha \geq 0$. Existing retrieval-augmented models (e.g., Lewis et al. 2020b) typically only optimize the loss for the main sequence, L_{main} (i.e., $\alpha = 0$). However, as the Transformer computes logits for tokens in the retrieved documents when it computes logits for tokens in the main sequence, we can easily include the loss for the retrieved documents, L_{retr}. Thus, $\alpha > 0$ offers an effect analogous to increasing the batch size (the number of tokens involved in optimization) without much extra compute, and boosts training efficiency. This technique is especially useful in the multimodal modeling setting, because each image takes many tokens (e.g., 1024 tokens), and $\alpha = 0$ would throw away computation used for the image tokens in retrieved documents. In practice, we find $\alpha = 0.1$ works well. See §C.3 for detailed analysis.

3.4. Training and inference

Training. Given a full input document x, we use either its text part or its image part as the query q for retrieving documents (§3.2). We then optimize the generator token prediction loss over the whole concatenated sequence (Equation 2) by standard teacher forcing. We only use the text or image part as the query because (1) retrieving documents based on the full input document could make the generator's token prediction task too easy during training, and (2) this training setting is close to the typical inference scenarios of

text-to-image and image-to-text generation.

Since our off-the-shelf CLIP-based retriever already performs well, we fix the retriever and only train the generator in this work. An interesting future research direction would be the exploration of co-training or fine-tuning the retriever.

Inference. Our method takes an input sequence (prompt) x, uses x as the query for retrieval, and then lets the generator take the retrieved documents as part of the input to decode the continuation of x. For instance, for text-to-image generation, prompt x takes the source caption, and the continuation will be the target image. For image-to-text, prompt x takes the source image, and the continuation will be the target caption. Thus, the retriever only uses the prompt as a query and never sees the ground-truth continuation y to be evaluated, ensuring no information leakage.

4. Experiments

To experiment with our proposed approach, we train models using the LAION multimodal dataset (§4.1), and evaluate on the MS-COCO image and caption generation tasks (§4.2). We show that our retrieval-augmented model (RA-CM3) significantly improves both image and text generation performance (§4.3). We then analyze the scaling laws and key design choices of our model (§C.2, C.3). Finally, §5 presents qualitative results and capabilities of our model, such as knowledge intensive generation and in-context learning.

4.1. Training setup

Data. To train our model, we use LAION (Schuhmann et al., 2021), an open-sourced dataset that consists of text-image pairs collected from the web. Following the preprocessing step of Stable Diffusion (Rombach et al., 2022), we cleaned a subset of LAION[1] and obtained 150M text-image pairs in total. Following CM3, we format each text-image pair as an HTML document, "`<img alt=[text] src=[image]>`", where `[image]` is a sequence of 1024 image tokens obtained by tokenizing the raw image using VQGAN (Esser et al., 2021; Gafni et al., 2022). These 150M documents are used as our model's final training data.

We also use the same 150M documents for our external memory $\mathcal{M}$.

Implementation. In our retrieval module R, we use the off-the-shelf CLIP model (`ViT-L/14`) (Radford et al., 2021) for both the query and memory encoders E_Q and E_M. We use FAISS (Johnson et al., 2019) to index the external memory $\mathcal{M}$ (Flat Index) and perform MIPS-based retrieval.

---

[1] We filter out images with watermark probability above 0.5, unsafe probability above 0.5, or resolution below 256×256.

Approach	Model type	MS-COCO FID (↓)	
		Not finetuned	Finetuned
Retrieval Baseline	-	17.97	-
KNN-Diffusion (Ashual et al., 2022)	Diffusion	16.66	-
Stable Diffusion (Rombach et al., 2022)	Diffusion	12.63	-
GLIDE (Nichol et al., 2021)	Diffusion	12.24	-
DALL-E 2 (Ramesh et al., 2022)	Diffusion	10.39	-
Imagen (Saharia et al., 2022)	Diffusion	7.27	-
Re-Imagen (Chen et al., 2022b)	Diffusion	6.88	5.25
DALL-E (12B) (Ramesh et al., 2021)	Autoregressive	∼28	-
CogView (4B) (Ding et al., 2021)	Autoregressive	27.1	-
CogView2 (6B) (Ding et al., 2022)	Autoregressive	24.0	17.7
Parti (20B) (Yu et al., 2022)	Autoregressive	7.23	3.22
Vanilla CM3	Autoregressive	29.5	-
RA-CM3 (2.7B) (Ours)	Autoregressive	**15.7**	-

Table 2. **Caption-to-image generation performance on MS-COCO**. Our retrieval-augmented CM3 significantly outperforms the baseline CM3 with no retrieval, as well as other models such as DALL-E (12B parameters). Moreover, our model achieves strong performance with much less training compute than existing models; see Figure 2 for details.
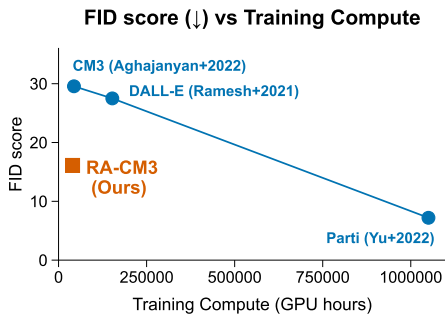

FID score (↓) vs Training Compute

Figure 2. **Image generation quality vs training compute** for our RA-CM3 model and baseline models. x-axis is the amount of training compute used in terms of A100 GPU hours. y-axis is the MS-COCO FID score (the lower, the better). Our retrieval-augmented method achieves significantly better training efficiency than existing works under a similar autoregressive Transformer paradigm (e.g., CM3, DALL-E, Parti).

For our generator G, we use a Transformer (Vaswani et al., 2017) of 2.7B parameters. The sequence length is 4096, which can take up to 3 documents. For each input document x, we retrieve $K \sim \text{Uniform}(\{0, 1, 2\})$ documents and prepend them to x. At inference time, we may also retrieve and add $K > 2$ documents via ensemble (see §5.4).

The model is trained from scratch for five days on 256 A100 GPUs. Our implementation is in PyTorch (Paszke et al., 2019) using Metaseq (Zhang et al., 2022). We use model parallelism over 4 GPUs and a batch size of 16 sequences per GPU. The optimization uses a linear learning rate decay with 1500 warmup steps, a peak learning rate of 1e-4, a gradient clipping of 1.0, and the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ (Kingma & Ba, 2015).

Approach	CIDEr (↑)
Retrieval Baseline	84.1
DALL-E[Small] (Wang, 2021)	20.2
ruDALL-E-XL (Forever, 2021)	38.7
minDALL-E (Kim et al., 2021)	48.0
X-LXMERT (Cho et al., 2020)	55.8
Parti (Yu et al., 2022)	83.9
Flamingo (3B; 4-shot) (Alayrac et al., 2022)	85
Flamingo (80B; 4-shot) (Alayrac et al., 2022)	103
Vanilla CM3	71.9
RA-CM3 (2.7B) (Ours)	**89.1**

Table 3. **Image-to-caption generation performance on MS-COCO** (with no finetuning). Our retrieval-augmented CM3 significantly outperforms the baseline CM3 with no retrieval. Moreover, our model outperforms other strong models such as Parti (20B parameters) and Flamingo (3B; 4-shot), despite using just ∼3B parameters and 2-shot in-context examples.

Baseline. For our baseline, we train a vanilla CM3 with no retrieval augmentation, using the same model architecture, training data, and amount of compute, for fair comparison. Since RA-CM3's external memory consists of the same training data, the total information accessible to RA-CM3 and vanilla CM3 are controlled to be the same.

4.2. Evaluation setup

For the main evaluation, we use the standard benchmark, MS-COCO (Lin et al., 2014), to evaluate both text-to-image and image-to-text generation. We evaluate our trained model with no further finetuning.

For text-to-image, following prior works (Ramesh et al., 2021; Nichol et al., 2021), we generate images for the MS-COCO validation set captions and measure the FID score (Heusel et al., 2017) against ground-truth images. To generate an image for each caption, we sample 10 images from the model and then take the top image based on the CLIP score (Radford et al., 2021) with respect to the input caption, as done in Aghajanyan et al. (2022).

For image-to-text, following prior works (Alayrac et al., 2022), we generate captions for the MS-COCO validation set images and measure the CIDEr score (Vedantam et al., 2015) against ground-truth captions. To generate an caption for each image, we sample 32 captions from the model and take the top caption based on perplexity (Fried et al., 2022).

4.3. Main results

Caption-to-image generation. Table 2 shows the caption-to-image generation performance on MS-COCO. The metric is FID score, where lower is the better. Our retrieval-augmented CM3 achieves an FID score of 16 without finetuning, significantly outperforming the baseline CM3 with no retrieval (FID 29) and other models such as DALL-E

Figure 3. **Text-to-image generation involving world knowledge.** Our retrieval-augmented model (RA-CM3) can generate correct images from entity-rich captions thanks to the access to retrieved images in the context. For example, RA-CM3's outputs faithfully capture the visual characteristics of various entities (e.g., the shape and painting of Ming Dynasty vase, the amount of Callanish standing stones). On the other hand, baseline models without retrieval capabilities (vanilla CM3, Stable Diffusion) tend to struggle, especially when the caption involves rare entities (e.g., "Ming Dynasty vase", "Oriental Pearl tower", "Dragon and Tiger Pagodas").

(FID 28), which is 3x bigger than our model. This suggests that retrieval augmentation provides significant help in generating higher-quality images.

To also factor in training efficiency, Figure 2 visualizes the image generation performance (y-axis: FID score) vs the amount of compute used in model training (x-axis: normalized A100 GPU hours) for our RA-CM3 model and baseline models. We find that existing models in the autoregressive Transformer paradigm follow a negatively sloped line in this chart (the blue dots and line in Figure 2). RA-CM3 is located significantly below this line, i.e., obtaining a better FID with less training compute. This suggests that the proposed retrieval-augmented method achieves significantly better training efficiency than existing works.

Our intuition is that retrieval augmentation allows the model to focus on learning how to use the retrieved documents in the context rather than fitting all the documents into the parameters of the model, speeding up the training process.

Image-to-caption generation. Table 3 shows the image-to-caption generation performance on MS-COCO, with no finetuning. The metric is the CIDEr score, where the higher is the better. Our retrieval-augmented CM3 achieves a CIDEr score of 89, significantly outperforming the baseline CM3 with no retrieval (CIDEr 72). Moreover, RA-CM3 outperforms other strong models such as Parti (20B parameters) and Flamingo (3B; 4-shot), despite using just ∼3B

parameters and 2-shot in-context examples.

These results confirm that our model can perform both image and text generation well, offering the first unified retrieval-augmented multimodal model (Table 1).

4.4. Analysis

We analyze the scaling laws of RA-CM3 in §C.2 and the key design choices of RA-CM3 in §C.3.

5. Qualitative results

We show novel qualitative capabilities of our RA-CM3, such as knowledge-intensive multimodal generation (§5.1) and multimodal in-context learning (§5.2, 5.3, 5.4). While GPT-3 (Brown et al., 2020) and Flamingo (Alayrac et al., 2022) showed in-context learning for text-to-text or image-to-text generation, we show that RA-CM3 can do in-context learning for both text (§5.4) and image (§5.2, 5.3) generation.

5.1. Knowledge-intensive multimodal generation

Because of the retrieval capability, RA-CM3 is especially good at tasks that require world knowledge or composition of knowledge (knowledge-intensive generation). Figure 3, 4 show example outputs from RA-CM3. For each caption, the output images were obtained by sampling 256 images
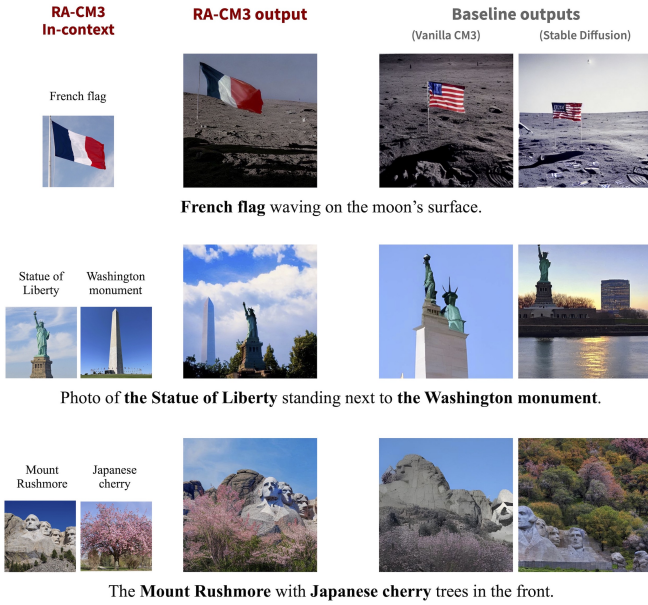
French flag waving on the moon's surface.

Photo of **the Statue of Liberty** standing next to **the Washington monument**.

The **Mount Rushmore** with **Japanese cherry** trees in the front.

Figure 4. **Text-to-image generation involving rare *composition* of knowledge.** Our retrieval-augmented model (RA-CM3) can generate faithful images from captions that contain a rare or unseen composition of entities (e.g., "French flag" + "moon", "Mount Rushmore" + "Japanese cherry"). On the other hand, baseline models without retrieval capabilities (vanilla CM3, Stable Diffusion) tend to struggle on these examples, e.g., generate a US flag instead of a French flag on the moon.
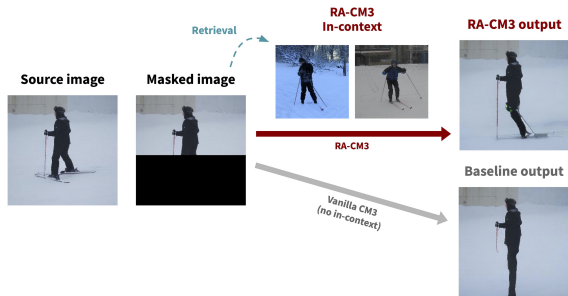

Figure 5. **Our model can perform better image infilling.** Infilling an image requires world knowledge, e.g., to recover the masked patches of the above image, the model needs to know about skiing. While the vanilla CM3 (no retrieval) tends to simply infill legs, our RA-CM3 (with retrieval) successfully recovers both legs and skis.

Figure 6. **Our model can perform image editing.** Instead of using retrieved examples in our RA-CM3's context (Figure 5), we can also intervene and manually specify the in-context examples to control image infilling. For instance, we can place an image of a person wearing a red jacket in the context to edit the black jacket in the original image to be red (Figure top).

Photo of a house taken on an autumn day.

Painting of red roses.

Figure 7. **Controllable image generation.** Our RA-CM3 model can control the style of caption-to-image generation by prepending demonstration examples in the generator's context. For instance, when generating an image of "a house taken on an autumn day" (Figure top), we can specify a concrete style by providing demonstration images (e.g., image of a triangular wooden house and image of orange autumn leaves background). Consequently, RA-CM3 generates an image that follows the visual characteristics of these in-context images.

from the model and then re-ranking them using the CLIP score with respect to the input caption. We then apply an off-the-shelf super-resolution tool (Rombach et al., 2022).

World knowledge. Figure 3 shows model outputs for caption-to-image generation that involves world knowledge (e.g., specific entities). We find that our RA-CM3 model can generate correct images from entity-rich captions thanks to the access to retrieved images in the context. For example, RA-CM3's outputs faithfully capture the visual characteristics of various entities (e.g., the shape and painting of Ming
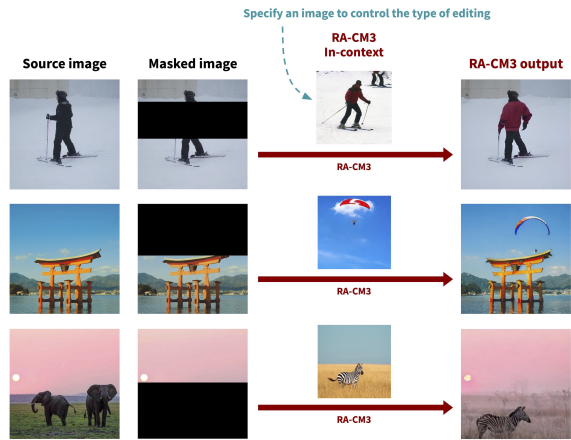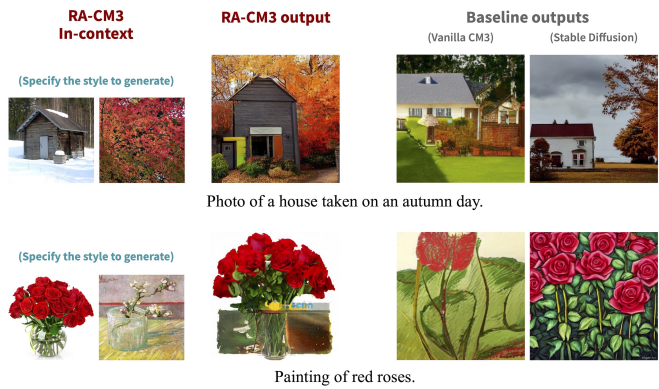
Dynasty vase, the amount of Callanish standing stones). On the other hand, baseline models without retrieval capabilities (vanilla CM3, Stable Diffusion) tend to struggle, especially when the caption involves rare entities (e.g., "Ming Dynasty vase", "Oriental Pearl tower", "Dragon and Tiger Pagodas").

Composition of knowledge. Figure 4 shows model outputs for caption-to-image generation that involves rare *composition* of knowledge. We find that our retrieval-augmented model can generate faithful images from captions that contain a rare or unseen composition of entities (e.g., "French
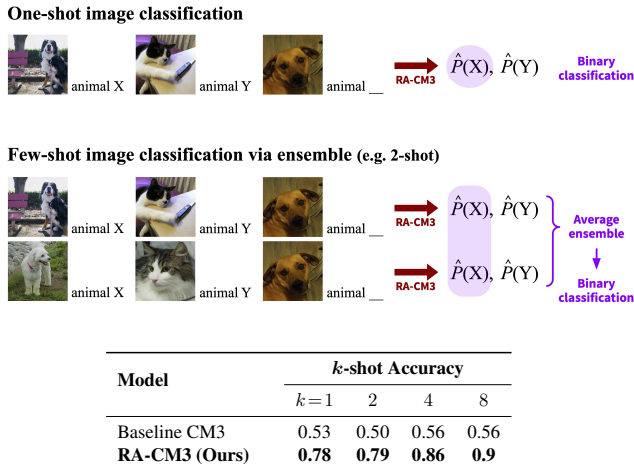
One-shot image classification

Few-shot image classification via ensemble (e.g. 2-shot)

Model	k-shot Accuracy			
	$k=1$	2	4	8
Baseline CM3	0.53	0.50	0.56	0.56
RA-CM3 (Ours)	**0.78**	**0.79**	**0.86**	**0.9**

Figure 8. **Our model performs one/few-shot image classification via in-context learning**. To assess the in-context learning ability, we consider a binary image classification task with non-semantic labels (e.g., "animal X" and "animal Y" instead of "dog" and "cat"). For one-shot classification (Figure top), we feed into the model one pair of demonstration examples, followed by a test example ([test image], "animal _"), for which we predict the probability of "X" and "Y". For k-shot classification (Figure middle), we repeat the above procedure k times, each using a different pair of demonstration examples, and take the average ensemble of the predicted probability ("X" and "Y") across the k passes.

The table (Figure bottom) shows the results of k-shot classification accuracy, with $k = 1, 2, 4, 8$. Across all k's, our RA-CM3 improves on the baseline CM3 by large margins. Increasing k consistently improves accuracy for the k values above.

flag" + "moon", "Mount Rushmore" + "Japanese cherry"). On the other hand, baseline models without retrieval capabilities (vanilla CM3, Stable Diffusion) tend to struggle on these examples, e.g., generate a US flag instead of a French flag on the moon (Figure 4 top). This is likely because the US flag was the most common flag that co-occurred with the moon in the training data.

5.2. Image infilling and editing

Because our model builds on CM3, it can also perform *infilling*.[2] Figure 5 shows that our RA-CM3 can perform improved image infilling because of the retrieval capability. Infilling an image requires world knowledge, e.g., to recover the masked patches of the image in Figure 5, the model needs to know about skiing. While the vanilla CM3 (no retrieval) tends to simply infill legs, RA-CM3 (with retrieval) successfully recovers both legs and skis.

---

[2]To perform image infilling, the model is given "[unmasked part of image] <mask> [unmasked part of image]" as the prompt and generates the <mask> part as the completion. The final output image is constructed by plugging the generated output to the <mask> part of the input.

Moreover, instead of using retrieved examples in the RA-CM3 context, we can also intervene and manually specify the in-context examples to control image infilling. Figure 6 shows examples. For instance, we can place an image of a person wearing a red jacket in the context to edit the black jacket in the original image to be red (Figure 6 top).

5.3. Controlled image generation

Controlled generation—controlling the behavior of models in generation (e.g., style of outputs)—is a key problem in generative models (Keskar et al., 2019; Li et al., 2019).

Our RA-CM3 can control the style of caption-to-image generation by prepending demonstration examples in the generator's context (Figure 7). For instance, when generating an image for "Photo of a house taken on an autumn day" (Figure 7 top), we can specify a concrete style by providing demonstration images (e.g., an image of a triangular wooden house and an image of orange autumn leaves background). Consequently, RA-CM3 can generate an image that actually follows the visual characteristics of these in-context images. This is a very useful capability because we can control generation not only via text (captions) but also via image demonstrations—especially helpful when some visual characteristics we want to specify might be difficult to express in text.

Moreover, the finding that RA-CM3 can use in-context examples for controlled generation suggests that it has acquired a form of **multimodal in-context learning** ability. Our intuition is that because the RA-CM3 generator has seen relevant multimodal documents prepended to the main document in context during retrieval-augmented training, it has learned how to use in-context examples effectively.

5.4. One-shot and few-shot image classification

So far we have seen RA-CM3's in-context learning behavior for image generation (§5.3). Here we study its in-context learning ability for image-to-text generation, through one-shot and few-shot image classification.

Figure 8 illustrates the experiment. To assess the true in-context learning ability that factors out prior knowledge of the model, we consider a binary image classification task with non-semantic labels (e.g., "animal X" and "animal Y" instead of "dog" and "cat"). Specifically, we use ImageNet (Deng et al., 2009) to construct such evaluation sets where each class (e.g., animal X or Y) contains the same number of test images (e.g., 100 images). For one-shot classification (Figure 8 top), we feed into the model one pair of demonstration examples ([image X], "animal X", [image Y], "animal Y"), followed by a test example ([test image], "animal _"), for which we predict the probability of "X" and "Y". For k-shot classification (Figure 8 middle), we repeat the above

procedure k times, each using a different pair of demonstration examples, and take the average ensemble of the predicted probability ("X" and "Y") across the k passes.[3]

The table in Figure 8 bottom shows the results of k-shot (binary) classification accuracy, with $k = 1, 2, 4, 8$. Across all k's, our RA-CM3 obtains significantly improved accuracy over the baseline CM3, which were not trained with retrieved documents in context. In particular, RA-CM3 already performs reasonably well at one-shot (0.78 accuracy at $k = 1$). This result suggests that RA-CM3 has acquired a strong in-context learning ability, especially given that we use non-semantic labels for image classes in this evaluation. Moreover, we find that increasing k consistently improves accuracy for the k values above (0.90 accuracy at $k = 8$). This observation suggests that ensemble is an effective method to increase the number of in-context examples to provide for the model.

6. Conclusion

We presented a retrieval-augmented multimodal model that can retrieve and refer to an external memory for generating images and text. Specifically, we implemented a multimodal retriever using the pretrained CLIP and designed a retrieval-augmented generator using the CM3 architecture. Our resulting model, named RA-CM3, outperforms existing multimodal models on both image and caption generation tasks, while requiring much less training compute. Moreover, RA-CM3 exhibits novel capabilities such as knowledge-intensive image generation and multimodal in-context learning.

This work aims to offer a general and modular retrieval augmentation framework for multimodal models. We believe this opens up various exciting research avenues, such as improving the multimodal retriever and generator, extending modalities beyond image and text, and further investigating multimodal prompting and in-context learning.

Acknowledgements

---

[3]An alternative way to use k-shot examples could be to prepend all the k pairs of demonstrations directly in RA-CM3's context, but this would take a significant sequence length in Transformer and might not be easy to scale. We find that the ensemble-based method performs well empirically, and comes with benefits such as being more scalable (parallel runs of shorter-length passes) and principled (agnostic to the order of the k examples).

References

Agarwal, O., Ge, H., Shakeri, S., and Al-Rfou, R. Knowledge graph based synthetic corpus generation for knowledge-enhanced language model pre-training. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

Aghajanyan, A., Huang, B., Ross, C., Karpukhin, V., Xu, H., Goyal, N., Okhonko, D., Joshi, M., Ghosh, G., Lewis, M., and Zettlemoyer, L. CM3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

Ashual, O., Sheynin, S., Polyak, A., Singer, U., Gafni, O., Nachmani, E., and Taigman, Y. Knn-diffusion: Image generation via large-scale retrieval. *arXiv preprint arXiv:2204.02849*, 2022.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021a.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Ethical considerations of generative AI. *AI for Content Creation Workshop, CVPR*, 2021b.

Blattmann, A., Rombach, R., Oktay, K., and Ommer, B. Retrieval-augmented diffusion models. *arXiv preprint arXiv:2204.11824*, 2022.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Van Den Driessche, G. B., Lespiau, J.-B., Damoc, B., Clark, A., et al. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning (ICML)*, 2022.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Chen, W., Hu, H., Chen, X., Verga, P., and Cohen, W. W. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2022a.

Chen, W., Hu, H., Saharia, C., and Cohen, W. W. Reimagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022b.

Cho, J., Lu, J., Schwenk, D., Hajishirzi, H., and Kembhavi, A. X-lxmert: Paint, caption and answer questions with multi-modal transformers. *arXiv preprint arXiv:2009.11278*, 2020.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Ding, M., Yang, Z., Hong, W., Zheng, W., Zhou, C., Yin, D., Lin, J., Zou, X., Shao, Z., Yang, H., et al. Cogview: Mastering text-to-image generation via transformers. In *Neural Information Processing Systems (NeurIPS)*, 2021.

Ding, M., Zheng, W., Hong, W., and Tang, J. Cogview2: Faster and better text-to-image generation via hierarchical transformers. *arXiv preprint arXiv:2204.14217*, 2022.

Esser, P., Rombach, R., and Ommer, B. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Forever, A. rudall-e. https://github.com/ai-forever/ru-dalle, 2021.

Fried, D., Aghajanyan, A., Lin, J., Wang, S., Wallace, E., Shi, F., Zhong, R., Yih, W.-t., Zettlemoyer, L., and Lewis, M. Incoder: A generative model for code infilling and synthesis. *arXiv preprint arXiv:2204.05999*, 2022.

Gafni, O., Polyak, A., Ashual, O., Sheynin, S., Parikh, D., and Taigman, Y. Make-a-scene: Scene-based text-to-image generation with human priors. *arXiv preprint arXiv:2203.13131*, 2022.

Gur, S., Neverova, N., Stauffer, C., Lim, S.-N., Kiela, D., and Reiter, A. Cross-modal retrieval augmentation for multi-modal classification. *arXiv preprint arXiv:2104.08108*, 2021.

Guu, K., Lee, K., Tung, Z., Pasupat, P., and Chang, M.-W. Realm: Retrieval-augmented language model pre-training. In *International Conference on Machine Learning (ICML)*, 2020.

Hao, Y., Song, H., Dong, L., Huang, S., Chi, Z., Wang, W., Ma, S., and Wei, F. Language models are general-purpose interfaces. *arXiv preprint arXiv:2206.06336*, 2022.

Hashimoto, T. B., Guu, K., Oren, Y., and Liang, P. S. A retrieve-and-edit framework for predicting structured outputs. In *Neural Information Processing Systems (NeurIPS)*, 2018.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Neural Information Processing Systems (NeurIPS)*, 2017.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 2019.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense passage retrieval for open-domain question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

Keskar, N. S., McCann, B., Varshney, L. R., Xiong, C., and Socher, R. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*, 2019.

Khandelwal, U., Levy, O., Jurafsky, D., Zettlemoyer, L., and Lewis, M. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations (ICLR)*, 2019.

Kim, S., Cho, S., Kim, C., Lee, D., and Baek, W. mindall-e on conceptual captions. https://github.com/kakaobrain/minDALL-E, 2021.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

Lewis, M., Ghazvininejad, M., Ghosh, G., Aghajanyan, A., Wang, S., and Zettlemoyer, L. Pre-training via paraphrasing. In *Neural Information Processing Systems (NeurIPS)*, 2020a.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020b.

Li, B., Qi, X., Lukasiewicz, T., and Torr, P. Controllable text-to-image generation. *Neural Information Processing Systems (NeurIPS)*, 2019.

Li, Y., Pan, Y., Yao, T., and Mei, T. Comprehending and ordering semantics for image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014.

Metzler, D., Tay, Y., Bahri, D., and Najork, M. Rethinking search: making domain experts out of dilettantes. In *ACM SIGIR Forum*, 2021.

Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., and Chen, M. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. Pytorch: An imperative style, high-performance deep learning library. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, 2021.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Ramos, R., Martins, B., Elliott, D., and Kementchedjhieva, Y. Smallcap: Lightweight image captioning prompted with retrieval augmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Ghasemipour, S. K. S., Ayan, B. K., Mahdavi, S. S., Lopes, R. G., et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022.

Sarto, S., Cornia, M., Baraldi, L., and Cucchiara, R. Retrieval-augmented transformer for image captioning. In *Proceedings of the 19th International Conference on Content-based Multimedia Indexing*, 2022.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., Zettlemoyer, L., and Yih, W.-t. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in neural information processing systems (NeurIPS)*, 2017.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Wang, P. Dall-e in pytorch. https://github.com/lucidrains/DALLE-pytorch, 2021.

Wang, Y., Yasunaga, M., Ren, H., Wada, S., and Leskovec, J. Vqa-gnn: Reasoning with multimodal semantic graph for visual question answering. *arXiv preprint arXiv:2205.11501*, 2022a.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations (ICLR)*, 2022b.

Xie, T., Wu, C. H., Shi, P., Zhong, R., Scholak, T., Yasunaga, M., Wu, C.-S., Zhong, M., Yin, P., Wang, S. I., Zhong, V., Wang, B., Li, C., Boyle, C., Ni, A., Yao, Z., Radev, D., Xiong, C., Kong, L., Zhang, R., Smith, N. A., Zettlemoyer, L., and Yu, T. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2022.

Yasunaga, M., Ren, H., Bosselut, A., Liang, P., and Leskovec, J. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021.

Yasunaga, M., Bosselut, A., Ren, H., Zhang, X., Manning, C. D., Liang, P., and Leskovec, J. Deep bidirectional language-knowledge graph pretraining. In *Neural Information Processing Systems (NeurIPS)*, 2022a.

Yasunaga, M., Leskovec, J., and Liang, P. LinkBERT: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)*, 2022b.

Yu, J., Xu, Y., Koh, J. Y., Luong, T., Baid, G., Wang, Z., Vasudevan, V., Ku, A., Yang, Y., Ayan, B. K., et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. Ernie: Enhanced language representation with informative entities. In *Association for Computational Linguistics (ACL)*, 2019.

A. Ethics and societal impact

Multimodal generative models, including our RA-CM3 and other models like DALL-E, Parti and Stable Diffusion, are typically trained on large, noisy image-text datasets collected from the web. These datasets may contain biases about race, gender and other demographic attributes, and unsafe content such as pornography and violence (Birhane et al., 2021a). We performed extensive data filtering to remove problematic content in training data following existing works (§4.1), but it may still be possible that RA-CM3 outputs problematic text or images. RA-CM3 is a **research prototype**, and we do not encourage using it in high-risk or sensitive domains or for generating images of people. For a more general, detailed discussion on the ethical considerations of multimodal generative models, we refer readers to e.g., Birhane et al. (2021b).

We also highlight the potential societal benefits of our retrieval-augmented multimodal model. First, as our model requires much less compute for training than existing models (§4.3), it can provide **energy savings**. Second, retrieval helps capture long-tail knowledge (e.g., rare entities or minority groups), which can contribute to more **fair** multimodal models. Third, retrieval naturally provides the provenance of knowledge, offering better interpretability and **explainability** about model predictions. Retrieval augmentation also helps make image/text generation **faithful** to the retrieved evidence documents (§5.1), potentially helping reduce unintentionally fake or hallucinated outputs.

B. Related work

Vision-language multimodal models. Various models have been developed for text-to-image generation. Typically, these models are autoregressive Transformer-based, e.g., DALL-E (Ramesh et al., 2021) and Parti (Yu et al., 2022), or diffusion-based, e.g., Imagen (Saharia et al., 2022), DALL-E 2 (Ramesh et al., 2022) and Stable Diffusion (Rombach et al., 2022). Meanwhile, several works also study image-to-text generation (Cho et al., 2020; Wang et al., 2022b). In particular, Flamingo (Alayrac et al., 2022) is a Transformer-based image-to-text generation model, with in-context learning ability. Recently, CM3 (Aghajanyan et al., 2022) provides a unified model that uses a Transformer to perform both text and image generation. To make use of this generality, we will build on CM3 to design our model.

While the above models have achieved strong performance on image and text generation, they store all their knowledge inside the model, which tends to require a lot of parameters (e.g., 10B) and training data (e.g., 1B images). To address this limitation, we augment them with an ability to refer to relevant examples from an external memory when generating images/text. With this augmentation, our model

outperforms the existing models by using less training data (150M images), compute and parameters ($<30\%$) (§4).

Retrieval-augmented language models. Retrieval augmentation has shown promise in NLP (Lewis et al., 2020b). To incorporate knowledge into a language model (LM), this line of work retrieves documents relevant to input text from an external memory, and lets the LM (generator) use the retrieved documents to make more informed predictions. The external memory used is typically a collection of text passages (Hashimoto et al., 2018; Khandelwal et al., 2019; Karpukhin et al., 2020; Guu et al., 2020; Lewis et al., 2020b;a; Yasunaga et al., 2022b; Borgeaud et al., 2022; Shi et al., 2023) or a structured knowledge base (Zhang et al., 2019; Agarwal et al., 2021; Xie et al., 2022; Yasunaga et al., 2021; 2022a). Here we generalize the scope of the retrieval-augmented LM framework and consider *multimodal* documents for both our input data and external memory, which can be arbitrary sequences of text and images.

Retrieval in multimodal models. Besides the retrieval augmentation for language models, recent works also study retrieval for computer vision models (Ashual et al., 2022; Blattmann et al., 2022; Gur et al., 2021; Sarto et al., 2022; Li et al., 2022; Ramos et al., 2023; Wang et al., 2022a). More recently, some concurrent works study retrieval in multimodal models: Re-Imagen (Chen et al., 2022b) performs diffusion-based caption-to-image generation using retrieved images; MuRAG (Chen et al., 2022a) performs natural language question answering using retrieved images. While these works focus on generating a single modality, either text or image, we develop a general and unified model that can retrieve, encode, and generate *both* images and text (Table 1). Moreover, our retrieval-augmented training allows the generator model to acquire novel in-context learning ability such as controlled image generation (§5).

C. Additional results

C.1. Intrinsic evaluation of CLIP-based retriever

Method	Recall ($\uparrow$)		
	@1	@3	@5
CLIP text-to-image retrieval	48	65	78
CLIP text-to-mixture retrieval	61	77	85
CLIP image-to-text retrieval	56	75	84
CLIP image-to-mixture retrieval	78	84	87

Table 4. **Multimodal retrieval performance on MS-COCO.** We use the frozen pretrained CLIP. "text-to-image retrieval" and "image-to-text retrieval" use the CLIP text/image encoder as it is. "text-to-mixture retrieval" and "image-to-mixture retrieval" use our mixed-modal encoder based on CLIP (§3.2). In all these cases, the CLIP-based retrieval method performs reasonably well.
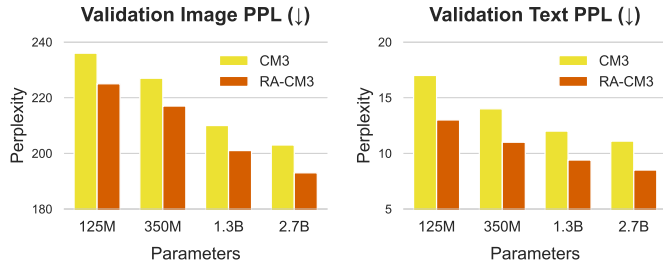

Figure 9. **Perplexity-based scaling laws for our RA-CM3 model.** We train RA-CM3 and vanilla CM3 of various parameter counts using the same amount of compute, and evaluate perplexity on the held-out validation set of MS-COCO. RA-CM3 provides consistent improvements over vanilla CM3 across different scales.

Method design	Choice	Image ppl ($\downarrow$)	Text ppl ($\downarrow$)
Retrieval relevance (§3.2)	Random at train & infer	246	23
	Retrieve at train, random at infer	246	24
	Random at train, retrieve at infer	243	18
	Retrieve at train & infer (**final**)	**227**	**13**
Retrieval modality (§3.2)	Only image or only text	234	15
	Multimodal document (**final**)	**227**	**13**
Retrieval diversity (§3.2)	Simply take top K	244	17
	Avoid redundancy	235	15
	Avoid redundancy + Query dropout (**final**)	**227**	**13**
Generator training (Equation 2)	$\alpha = 0$	239	17
	$\alpha = 1$	240	17
	$\alpha = 0.3$	231	14
	$\alpha = 0.1$ (**final**)	**227**	**13**

Table 5. **Analysis of our method's design choices.** As the metric, we use the perplexity of image/text generation on the MS-COCO validation set. We find that key methods to achieve the best performance are: ensure *relevance* in retrieved documents (table top); retrieve *multimodal* documents instead of only images or text (table second from top); encourage *diversity* in retrieved documents during training (table second from bottom); and train the token prediction loss for *both* the main input document and the retrieved documents, in particular, with a weight of $\alpha = 0.1$ (table bottom). Note that images naturally have higher perplexity than text, as also observed in prior works (e.g., Aghajanyan et al. 2022).

C.2. Scaling laws of RA-CM3

To study the scaling laws of retrieval augmentation for multimodal models, we train the retrieval-augmented CM3 and vanilla CM3 of various sizes (125M, 350M, 1.3B, 2.7B parameters) using the same amount of compute (two days on 256 GPUs), and then evaluate the models' perplexity on the MS-COCO validation set (Figure 9). We observe that RA-CM3 provides consistent improvements over vanilla CM3 across different scales. We do not observe any diminishing returns in the 125M–2.7B range that we studied. This suggests that retrieval augmentation is also promising at a larger scale.

C.3. Analysis of RA-CM3 designs

We analyze key design choices of the retrieval-augmented multimodal model, such as the strategies of retrieval (§3.2) and generator training (§3.3). Here, our experiments used a 2.7B parameter-model, trained for a day on 256 GPUs.

Retrieval relevance (Table 5 top). A main contribution of our work is retrieval-augmented training of the generator (§3.3). While our final RA-CM3 prepends documents retrieved by our CLIP-based retriever at both train and inference time ("Retrieve at train & infer" row in the table), one natural baseline is to train the model using random documents without retrieval (i.e., vanilla CM3) but use retrieved documents at inference time ("Random at train, retrieve at infer"). This baseline leads to a significant performance drop, suggesting that having relevant documents in context is crucial for model training. We also study other baselines, such as using retrieved documents at train time but random documents at inference time, or using random documents at both train and inference times. Both result in significant performance drops. These results confirm that relevance is a crucial factor in retrieval at both train and inference times.

Retrieval modality (Table 5 second from top). While existing works on retrieval (Chen et al., 2022b) typically retrieve either an image or text only for the generator, our retriever based on a mixed-modal encoder (§3.2), can retrieve a multimodal document that consists of both images and text. We find that retrieving multimodal documents performs better than retrieving only images or text. Our intuition is that a multimodal document can be more informative because the text and image within it can contextualize each other.

Retrieval diversity (Table 5 second from bottom). As discussed in §3.2, encouraging diversity in retrieved documents is important. Simply taking the top K (e.g., 2) from the list of candidate documents sorted by the retriever scores leads to poor performance—in fact, slightly worse than the baseline with no retrieval augmentation. Our first technique which avoids redundancy in retrieved documents leads to significant performance improvement. We also find that the second technique, Query Dropout, which encourages more diversity in retrieval during training leads to a further boost in evaluation performance.

Generator training (Table 5 bottom). A key design of our generator is that we optimize token prediction loss jointly for the main input document and the retrieved documents, with a weighting α (§3.3; Equation 2). Existing retrieval-augmented models typically optimize loss for the main document only ($\alpha = 0$), but we find that joint optimization ($\alpha > 0$) facilitates training and improves performance. We find that $\alpha = 0.1$ works well. Setting α to be too large (e.g., $\alpha = 1$) hurts training because this would place too much weight on modeling retrieved documents instead of the main

document.

D. Additional discussions

D.1. Fair comparison of the retrieval-augmented model and non-retrieval-augmented model

Experiments: Our baseline is the vanilla CM3 with no retrieval augmentation. To make a fair comparison between the retrieval-augmented model (RA-CM3) and the baseline, RA-CM3 is trained using the same generator architecture, same training data, and same amount of compute as the vanilla CM3 (§4.1). RA-CM3's external memory used for retrieval also consists of the same training data. Thus, we ensured that no additional data or training compute is used for the retrieval-augmented model compared to the non-retrieval-augmented models. Under this controlled experiment, RA-CM3 substantially outperforms the vanilla CM3 in image and text generation (Table 2, 3). Figure 9 further indicates that RA-CM3 with fewer parameters (1.3B) can also outperform the vanilla CM3 with more parameters (2.7B). We also include the "Retrieval Baseline" in Table 2 and 3, which simply returns the retrieved images or text as model outputs. RA-CM3 outperforms this retrieval baseline.

Usage: Both the retrieval-augmented model and non-retrieval-augmented models take the same input from users, e.g., a source caption for image generation or a source image for caption generation. In the retrieval-augmented model, the retriever will automatically take this input prompt, fetch relevant images/text, and add them to the context of the generator, so no additional input is needed from the user. Of course, the user may also intervene and self-specify in-context examples for the generator, so the retrieval-augmented model provides more flexibility and controllability for users (§5.3). The retrieval step can be performed efficiently using FAISS (§4.1) in less than a second.

Thus, while the retrieval-augmented model operates within the same (fair) task definition as non-retrieval-augmented models, it opens up various benefits and new capabilities, such as improved explainability, faithfulness, and controllability in generation (§A).

D.2. Taking an existing model (e.g. vanilla CM3) and finetune it with retrieval-augmentation, instead of training the retrieval-augmented model (RA-CM3) from scratch

In practice, finetuning from the vanilla model can save compute for training RA-CM3 and is useful. The reason we trained RA-CM3 and vanilla CM3 both from scratch in our main experiments was to make a fair comparison between them by training with the same amount of compute, and to systematically study the effect of retrieval augmentation.

D.3. How the number of retrieved documents used for the generator (K) was set

We set K to be up to 2 (§4.1), primarily in consideration of the Transformer sequence length. Using the recent image tokenizer (e.g., Esser et al. 2021), each image is mapped to 1K tokens. Hence, the concatenation of $K=2$ retrieved documents and the main input document takes 3–4K tokens in total in Transformer. Increasing the Transformer sequence length beyond 4K incurs a significant burden in computation and GPU memory, so we decided on $K=2$. $K=2$ also worked reasonably well in practice (§4.3).

During inference, we may do ensemble (see §5.4) and take more than two retrieved documents in total into account. We conducted an analysis of varied $K = 1, 2, 4, 8$ in the MS-COCO caption-to-image generation evaluation (Table 6). We find that $K=2$ worked the best in this experiment. Our intuition is that most of the MS-COCO captions involve 1–2 objects, so retrieving the top two multimodal documents may be sufficient for generating corresponding images. It is an interesting future research to investigate larger K's on image generation tasks that involve more entities/objects.

Model	Image Perplexity ($\downarrow$)			
	$K=1$	2	4	8
RA-CM3	228	227	228	232

Table 6. **MS-COCO caption-to-image generation performance** when the number of retrieved multimodal documents (K) is varied.