MORE: Multi-mOdal REtrieval Augmented Generative Commonsense Reasoning

Anonymous ACL submission

Abstract

Since commonsense information has been recorded significantly less frequently than its existence, language models pre-trained by text generation have difficulty to learn sufficient commonsense knowledge. Several studies have leveraged text retrieval to augment the models' commonsense ability. Unlike text, images capture commonsense information inherently but little effort has been paid to effectively utilize them. In this work, we propose a novel **M**ulti-m**O**dal **RE**trieval (MORE) augmentation framework, to leverage both text and images to enhance the commonsense ability of language models. Extensive experiments on the Common-Gen task have demonstrated the efficacy of MORE based on the pre-trained models of both single and multiple modalities.

1 Introduction

Language Models (LMs) have gained increasing prominence in artificial intelligence, especially Large Language Models (LLMs) such as LLaMA (Touvron et al., 2023), GPT-3.5 (OpenAI, 2022), and GPT-4 (Achiam et al., 2023) that have achieved compelling performance across various tasks. However, even LLMs still lack robust commonsense capabilities and can sometimes generate sentences that violate commonsense knowledge. Figure 1 illustrates an instance of composing a sentence given several words, where both GPT-3.5 and GPT-4 consider that music can decorate the tree, which makes nonsense.

Due to the well-recognized reporting bias (Gordon and Durme, 2013), i.e., the recording of commonsense information is significantly less than its existence in reality (Grice, 1975; Havasi et al., 2007), it is inherently difficult for LMs to learn enough commonsense knowledge from modeling text generation. To enhance their commonsense ability, there have been a few attempts to retrieve external commonsense text information (He et al., 2022; Li et al., 2021; Liu et al., 2022) to augment the LM generation, which have been shown to be effective on commonsense reasoning tasks.

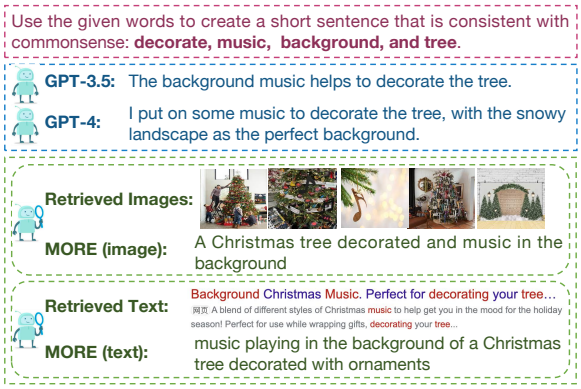


Figure 1: Sentences made by GPT3.5, GPT-4, and MORE given some concept words.

In contrast to text, commonsense knowledge is naturally recorded in the visual data. Additionally, text is used primarily for communication and may include subjective statements while images often record the physical world more objectively. Thus, images can be supplementary to text for LMs to enhance commonsense abilities. This can also be confirmed by the fact that humans acquire knowledge from both textual and visual data (Gambrell and Bales, 1986; Bloom, 2000; Joffe et al., 2007). Aware of this, instead of retrieving text snippets to assist the models in conducting commonsense tasks (He et al., 2022; Yu et al., 2022; Li et al., 2021; Liu et al., 2022), we propose a **M**ulti-m**O**dal **RE**trieval (MORE) augmentation framework to incorporate both text and images. For LLMs pre-trained with multi-modal data (e.g., BLIP2 (Li et al., 2023)), multi-modal retrieval augmentation can also be beneficial since it explicitly provides the text snippets and images carrying related commonsense information to the current sample. To effectively incorporate the multi-modal information into LMs, there are two major challenges:

1) How can we enable LMs to effectively extract useful knowledge from multi-modal retrieved results? This is even more challenging for text-based LMs because of the modality differences. To address this challenge, we propose a plug-and-play integrator that adopts a cross-attention mechanism to weigh each of the multi-modal results based on the query input and extract beneficial information. For text-based LMs, we employ a multi-modal encoder (e.g., the Qformer of BLIP2) to ingest results of images and text. In this case, the integrator also acts as a bridge that transforms the encoded semantic space of the retrieved results into the representation space used by the LMs.

2) Since the retrieval quality could vary considerably, how can we ensure the LMs do not ignore the retrieved results and also not trust them blindly? On the one hand, to prevent LMs from disregarding the entire retrieved results due to the noise they may contain, we introduce a training mechanism in MORE, i.e., query dropout, that masks the query input to the LMs at a certain ratio to urge the LMs to leverage the retrieved results for generation. On the other hand, to avoid too much dependence on the results that could be noisy, when queries are dropped out, we randomly replace the results with irrelevant ones and guide the LMs to output empty in such cases, so that the LMs can learn that it is not necessary to use retrieval all the time.

We evaluate our approach on a generative commonsense reasoning task, i.e., CommonGen (Lin et al., 2020). This task requires models to generate reasonable sentences using given concepts. Experimental results show that MORE can significantly boost the performance on CommonGen by incorporating multi-modal retrieved results for the LMs pre-trained with data of single or multiple modalities. MORE also significantly outperforms representative retrieval augmentation baselines and LLMs like GPT-3.5 and GPT-4, demonstrating the effectiveness of its architecture and training mechanism.

We summarize our contributions as follows: (1) We propose a novel multi-modal retrieval augmented language modeling framework for enhancing text generation of LMs. (2) Evaluations on the generative commonsense reasoning task, i.e., CommonGen, demonstrate the effectiveness of MORE on single/multi-modal LMs. (3)We conduct comprehensive analyses to verify the effectiveness of MORE under various settings and illustrate its advantages compared to LLMs like GPT-3.5 and GPT-4 through case studies.

2 Related Work

2.1 Retrieval Augmented Generation

The effectiveness of introducing additional contexts in the generation task has been demonstrated. Specifically, utilizing the input as a query, a retriever initially retrieves a set of documents from a corpus. Then a LM integrates these retrieved documents as supplementary information to generate a final prediction. For instance, Atlas (Izacard et al., 2022) finetunes a LM jointly with the retriever with very few training examples. RETRO (Borgeaud et al., 2022) modifies the decoder-only architecture to incorporate retrieved texts and pretrains the LM from scratch. Both methods necessitate updating model parameters through gradient descent, a process not applicable to Large Language Models (LLMs).

Given that the cost of fine-tuning LMs may not always be acceptable, recent research has explored retrieval augmentation for frozen LMs. Mallen et al. (2023); Si et al. (2023); Ram et al. (2023) demonstrate that directly prepending the documents returned by a frozen retriever to the input can improve LMs performance on open-domain QA. To support a large number of documents, FiD (Izacard and Grave, 2021) processes each input passage in parallel in the encoder. RePlug (Shi et al., 2023) further finetunes the retriever based on feedback from the frozen LM to get more helpful retrieved results. On these bases, compressing the retrieved results at the sentence level (Xu et al., 2023) or token level (Liu et al., 2023; Berchansky et al., 2023) can boost performance by filtering irrelevant information retrieved and improve computing efficiency.

2.2 Image Enhanced Text Generation

VisCTG (Feng et al., 2022) enhances the commonsense ability of LMs by retrieving images and using image captions as input augmentation. In addition to explicitly retrieving images, VAWI (Guo et al., 2022) leverages information from vision-language models, i.e. CLIP (Radford et al., 2021), to aid natural language understanding. I&V (Wang et al., 2022) train an imagination model to generate a scene graph given an input under the supervision of images and then train LMs to generate sentences based on both input and scene graph. The above methods either do not directly use images as non-

verbal data or require fine-tuning the whole pre-trained LMs to adapt to visual input. Drawing on the importance of imagination to human writing, iNLG (Zhu et al., 2022) and LIVE (Tang et al., 2023) use images generated by an image-generative model based on text inputs as supplementary information and train the LM to generate under visual guidance. However, the generated images may not necessarily carry commonsense information, such as cartoon images.

3 Generative Commonsense Reasoning

We focus on the task of CommonGen (Lin et al., 2020) to investigate and enhance the common sense reasoning capabilities of LMs.

3.1 Preliminaries

Problem Statement: The generative commonsense reasoning task in CommonGen asks the LM to make a sentence y that contains all the concept words in the given set $C = \{c_1, ..., c_K\}$, where c_i denotes the i-th concept and y should describe a common scenario in our daily life.

Training Objectives: It is usually modeled as a sequence generation task and is optimized by minimizing the cross-entropy loss between the predicted token distribution and the reference distribution: $L = -\sum_{t=1}^{|y|} \log P(y_t|C, y_{<t})$. In this work, to ensure parameter efficiency and applicability to LLMs, we use prompt tuning (Lester et al., 2021; Liu et al., 2021) instead of fine-tuning LMs. We only tune a task prompt, which is prepended to the input word embeddings in the first layer. When retrieval augmentation is enabled, a set of retrieved items D, which is retrieved with the concepts as query words, is also used as input and the new loss function becomes:

$$L = -\sum_{t=1}^{|y|} \log P(y_t|C, D, y_{<t}). \quad (1)$$

3.2 Multi-Modal Retrieval Augmention

As shown in Figure 2, the **M**ulti-m**O**dal **RE**trieval (MORE) augmented framework for text generation has four core components: retrieving relevant images and texts based on the concept words (§ 3.2.1), encoding the retrieved results with an encoder (§ 3.2.2), extracting useful information to yield a retrieval augmented prompt with an integrator (§ 3.2.3), and generating sentences based on the retrieval augmented prompt, task prompt, and concept embeddings with the frozen LM backbone (§ 3.2.4).

3.2.1 Retrieval Results for Augmentation

Previous work (He et al., 2022; Li et al., 2021; Liu et al., 2022) that incorporates retrieval augmentation on this task consider the image/video captions (Krishna et al., 2017; Williams et al., 2017; Wang et al., 2019; Bowman et al., 2015; Lin et al., 2014) that CommonGen is built on as the retrieval candidates, which is obviously impractical. In such a setting, we find that the captions retrieved by BM25 (Robertson et al., 2009) can achieve comparable performance with the state-of-the-art (SOTA) methods that train retrievers for augmenting LLMs (shown in Appendix A), making the investigation less meaningful.

To accommodate the task in real-world scenarios, in this paper, we retrieve the image and text results from a general Web search engine, i.e., Bing, for retrieval augmentation. We employ Bing as a reasonable off-the-shelf retriever since our focus is on how to incorporate the supporting items rather than training a powerful retriever. Formally speaking, given a concept set C, we retrieved M images and N text snippets by Bing using words in C as the query, comprising the set of items $D = \{d_1^v, ...d_M^v, d_1^t, ..., d_N^t\}$.

Specifically, after we preprocessed the retrieved results by removing duplicate images and noisy text, we collected a total of 500,100 images and 787,970 text snippets. On average, each concept set has 14 images and 23 passages, which means that M and N can be at most 14 and 23 respectively. See the Appendix B for more details.

3.2.2 Multi-Modal Encoder

Then we use an encoder to get the initial representation e_i^{ra} for each retrieved image or passage d_i:

$$e_i^{ra} = \text{Encode}(d_i). \quad (2)$$

This results in a sequence of representations with d_{enc} dimension. To align the encodings of text and images in the same semantic space, we adopt a multi-modal encoder - the frozen Q-Former of the pre-trained BLIP2 (Li et al., 2023). Unlike the other commonly used multi-modal encoder - CLIP (Radford et al., 2021), that encodes the input to a single final embedding, the Q-Former of BLIP2 outputs a sequence of embeddings and thus can retain more information.

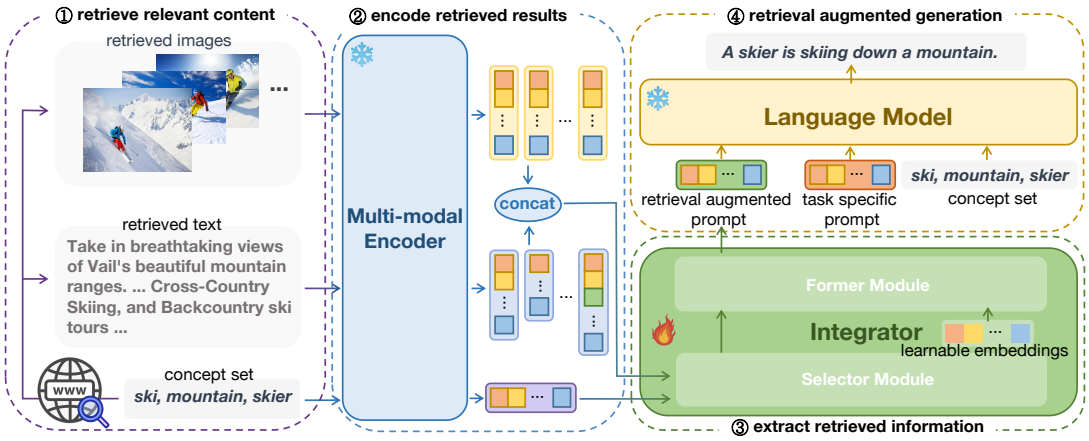


Figure 2: The process of our framework generating the sentence given input concepts based on multi-modal retrieval augmentation.

3.2.3 Retrieved Information Integrator

The integrator extracts useful information from the representations of multiple retrieved results and condenses it into a retrieval augmented prompt. The Integrator has a Selector module and a Former module.

Selector: This module extracts useful information from the retrieval representations based on the input concept and outputs a variable-length retrieval augmentation representation. It receives the concatenation of multiple initial representations $e^{ra} = [e^{ra}_1; ...; e^{ra}_{M+N}] \in \mathbb{R}^{(M+N)\times d_{enc}}$ and the embeddings of the concept words $e^c \in \mathbb{R}^{l_c \times d_{enc}}$ as input, in which l_c is the number of tokens in the concept set. The Selector is composed of two stacks of identical layers. Each layer consists of a self-attention network, which is used for interaction between retrieved content, a cross-attention network, which is used for interaction between retrieved content and concepts, and a fully connected feed-forward network:

$$\begin{aligned} h_i^{self} &= \text{Attn}(h_{i-1}W_i^Q; h_{i-1}W_i^K; h_{i-1}W_i^V) \\ h_i^{cross} &= \text{Attn}(h_i^{self}M_i^Q; E^{ra}M_i^K; E^{ra}M_i^V) \\ h_i &= h_i^{cross}F_i. \end{aligned} \tag{3}$$

$\text{Attn}(Q, K, V)$ is the multi-head attention layer as in Transformer (Vaswani et al., 2017). $W \in \mathbb{R}^{d_{enc}\times d_{int}}$, $M \in \mathbb{R}^{d_{enc}\times d_{int}}$,and $F \in \mathbb{R}^{d_{int}\times d_{enc}}$ are projection matrices, in which d_{int} is the dimension of the hidden states of Integrator, and h_0 is set to e^c. The output of the Selector module is a variable-length retrieval augmentation representation $h_2 \in \mathbb{R}^{l_c \times d_{int}}$.

Former: This module converts the representation produced by the Selector into fixed-length and projects it into the input embedding space of the LM. This results in the final retrieval augmentation prompt p^{ra}. The Former comprises a cross-attention network and a fully connected feed-forward network:

$$\begin{aligned} p^{ra'} &= \text{Attn}(qM^{Q'}; h_2M^{K'}; h_2M^{V'}) \\ p^{ra} &= p^{ra'}O, \end{aligned} \tag{4}$$

in which $q \in \mathbb{R}^{l_q \times d_{int}}$ is a learnable embeddings with fixed-length l_q. $O \in \mathbb{R}^{d_{int}\times d_{lm}}$ is the projection matrix used for spatial mapping and d_{lm} is the dimension of the input embedding of the LM.

3.2.4 Soft Prompt Based Text Generation

To ensure training efficiency especially based on LMs, we freeze the parameters of the LMs and adopt the Prompt-tuning (Lester et al., 2021) technique, which incorporates the fixed-length embeddings produced by the Integrator as a plug-and-play soft prompt.

During text generation, the LM receives the concatenation of the task-specific prompt p^{task} and the concept set C as input, and generates sentence y as output, denoted as $y = \text{LM}([p^{task}; C])$. When using retrieval augmentation, besides the task-specific prompt, we also prepend the retrieval-augmented (RA) prompt to the input. Consequently, the input to the model becomes $[p^{ra}; p^{task}; C]$.

3.2.5 Training Strategy

Query Concept Dropout: The retrieval quality can vary considerably, so the model may simply ignore the retrieval input instead of learning to extract useful information. To enhance the uti-

lization of retrieval augmented inputs, we propose a query dropout training strategy. Specifically, we randomly mask the query concept C input to the LMs with probability p in the initial T training steps, and let the model generate sentences only based on retrieved results. Please note that query dropout is only applied to the input of LMs, and C is always input to the Integrator to guide the model in extracting beneficial information from the retrieved results. The probability p decreases as the number of training steps increases: $p = 0.5 \times (1 - \sin(\pi(\min(\frac{t}{T}, 1) - 0.5)))$.

Noisy RA Input: It is also important to ensure that the model can learn to ignore noise rather than blindly trust the retrieved results. Therefore, we artificially introduce noise during query dropout by replacing the retrieval input with irrelevant results from other samples and correspondingly changing the target output with an 'EOS' token with probability $\hat{p}$.

4 Experiments Settings

4.1 Dataset

We validate our method on the CommonGen dataset[1] (Lin et al., 2020). It is designed for generative commonsense reasoning tasks involving the composition of discrete concepts into sentences depicting everyday scenarios. The dataset comprises 32,651, 993, and 1,497 unique concept sets for training, development, and testing, respectively. Each concept set has multiple associated gold target sentences, yielding 67,389, 4,018, and 6,042 sentences for reference in total. When retrieval augmentation is enabled, we used the retrieved results from Bing as introduced in Section 3.2.1. We will release the crawled images and text to encourage future research in multi-modal retrieval augmentation under a practical setting for CommonGen.

4.2 Methods for Comparisons

Text-based/Multi-modal LMs: For text-based LMs, we employ **T5$_{\text{BASE}}$** as well as **T5$_{\text{LARGE}}$** (Raffel et al., 2019) to represent small pre-trained LMs, and **OPT**$_{2.7b}$ (Zhang et al., 2022) to represent the LLMs. We also query the close source model **gpt-3.5-turbo** (OpenAI, 2023a) through API with the prompt "Use the given words to make a short sentence that is consistent with commonsense. Words: {...}". For Multi-modal LMs (MLMs) we compare with **BLIP2-OPT**$_{2.7b}$ (Li et al., 2023), an open source model, and **gpt-4-vision-preview** (OpenAI, 2023b), a close source model. Since MLMs can accept images and text as input, we directly input the retrieved items into MLMs. All the above open source models are based on huggingface[2] and are under Apache License 2.0.

Retrieval Augmented Generation Baselines: We consider two types of textual retrieval augmented models. One is **Prepend** (Mallen et al., 2023; Si et al., 2023; Ram et al., 2023), which prepends the top-k text results to the concepts as input. The other one is **FiD** (Izacard and Grave, 2021), which concatenates each retrieved passage with the concept words separately to encode in parallel for better handling of long text. For the visual retrieval augmented model, we compare with **VisCTG** (Feng et al., 2022), which generates a caption for each image with an image captioning model (Luo et al., 2018) and pretends the captions to the input for augmentation. All the above models use T5$_{\text{BASE}}$ as the backbone and are tuned with prompt-tuning.

MORE with Various Backbones: We test MORE with various backbones to explore whether it can be effectively used in different model architectures. Specifically, **T5$_{\text{BASE}}$** and **T5$_{\text{LARGE}}$** represent small LMs and are encoder-decoder architecture. **BLIP2-OPT**$_{2.7b}$ represent MLMs. It should be noted that BLIP2-OPT$_{2.7b}$ is equivalent to OPT$_{2.7b}$ when not receiving image input. Therefore it can also be regarded as a variant based on **OPT**$_{2.7b}$, which represents LLMs and is decoder-only architecture.

4.3 Evaluation Metrics

To assess the generation performance, we use standard metrics: BLEU (Papineni et al., 2002) quantifying the overlap between predictions and references based on n-gram precision and ROUGE (Lin, 2004) measuring the n-gram recall. METEOR (Banerjee and Lavie, 2005) is an improved version of BLEU and considers both exact word matches and semantic similarities. CIDEr (Vedantam et al., 2014) focuses on capturing sentence semantic similarity. SPICE (Anderson et al., 2016) quantifies the semantic propositional content of generations by leveraging scene graphs and this score is most consistent with human evaluation. We use the entire test dataset to obtain the main results and randomly sample 500 data in the test

[1] https://inklab.usc.edu/CommonGen/. Under MIT license.

[2] https://github.com/huggingface

Table 1: Test results on CommonGen(V1.0). The best results are bolded, and '*' indicates that the results are significantly improved ($p < 0.05$) compared to the best baseline model (be underlined) under the significance test. In the last block, we also mark results with † that are significantly improved compared with the sub-optimal baseline.

Model	Bleu_4	METEOR	ROUGE_L	CIDEr	SPICE
gpt-3.5-turbo (0 shot)	21.44	28.93	48.80	13.03	26.69
gpt-3.5-turbo (3 shot)	28.91	31.14	53.25	15.92	28.89
T5_{BASE}	28.93	29.48	54.25	15.36	30.95
$\text{Prepend}_{\text{BASE}}$	26.68	28.43	53.29	14.39	29.95
FiD_{BASE}	28.32	28.72	54.07	14.78	30.25
$\text{VisCTG}_{\text{BASE}}$	27.67	28.82	53.71	14.77	30.24
$\text{MORE-T5}_{\text{BASE}}$ (text)	29.87*	30.15*	**55.20***	15.70*	31.57*
$\text{MORE-T5}_{\text{BASE}}$ (image)	29.98*	30.21*	55.07*	15.92*	31.63*
$\text{MORE-T5}_{\text{BASE}}$ (multi-modal)	**30.27***	30.28*	55.18*	**16.02***	**31.94***
T5_{LARGE}	31.16	30.48	55.68	16.14	31.62
$\text{MORE-T5}_{\text{LARGE}}$ (text)	32.03*	31.05*	56.14	16.37	32.00
$\text{MORE-T5}_{\text{LARGE}}$ (image)	**32.37***	**31.31***	56.60*	**16.67***	32.31*
$\text{MORE-T5}_{\text{LARGE}}$ (multi-modal)	32.29*	30.90*	**56.62***	16.63*	**32.34***
$\text{OPT}_{\text{2.7b}}$	31.53	31.43	55.95	16.76	32.24
$\text{BLIP2}_{\text{opt-2.7b}}$ (multi-modal)	31.92	31.70	56.22	16.73	32.44
$\text{MORE-OPT}_{\text{2.7b}}$ (multi-modal)	**32.78*†**	**32.15†**	**57.07*†**	**17.03*†**	**32.94*†**

Table 2: Test results compared with closed source LLMs on 500 randomly sampled data. '*' and '†' indicate the results are significantly improved compared to GPT-4 and GPT-3.5 (3 shot), respectively.

Model	Bleu_4	CIDEr	SPICE
GPT-4	28.53	16.41	29.05
GPT-4 (1i1t)	19.86	12.43	26.2
GPT-3.5 (0 shot)	22.97	13.93	27.25
GPT-3.5 (3 shot)	28.35	16.14	29.13
$\text{MORE}_{\text{OPT-2.7b}}$ (1i1t)	31.81*†	17.08*†	31.81*†
$\text{MORE}_{\text{OPT-2.7b}}$ (3i3t)	**32.53***†	**17.30***†	**32.81***†

Table 3: Ablation study results based on $\text{MORE}_{\text{BASE}}$.

Model	Bleu_4	CIDEr	SPICE
T5	28.93	15.36	30.95
MORE	30.27	16.02	31.94
w/o concept-input	29.45	15.33	31.18
w/o query-dropout	29.81	15.52	31.14
w/o noisy-RA	29.54	15.42	30.88

set to compare with GPT-4 for the sake of limit budget.

4.4 Implementation Details

The same set of hyper-parameters is used for all the models. We use the AdamW (Loshchilov and Hutter, 2017) optimizer with $\beta 1 = 0.9$, $\beta 2 = 0.999$ and the weight decay is 0.05. The batch size is selected from {64, 128}. Models were trained with at most 20,000 steps with a 1% warm-up period. For retrieval augmentation, we train the model with an additional $T = 2000$ steps with query dropout, and the noisy RA input ratio $\hat{p}$ is set to 0.3. The learning rates of the task prompt and the retrieval augmented prompt are selected from $\{1e-4, 5e-4, 1e-3\}$ and $\{1e-5, 3e-5\}$ respectively. During decoding, we use beam search with size 5. We train the models under each setting with 3 random seeds and choose the best ones according to the performance on the development set for testing. The prompt length of task and retrieval augmentation are both set to 32.

5 Results and Analysis

5.1 Overall Results

As shown in Table 1, after incorporating text and images to LMs, our method can boost the generation performance significantly based on various backbones. Comparing images and text, we find that images are better in improving commonsense ability, and incorporating both of them yields even

Table 4: Analyze the influence of additional parameters on the results. All models are based on $T5_{BASE}$. 'n pl' means using a prompt with n length. 'rand' means training models with irrelevant random search results.

Model	$Bleu_4$	CIDEr	SPICE
T5 (32 pl)	28.93	15.36	30.95
T5 (64 pl)	28.74	14.84	30.9
MORE (rand-RA)	29.33	15.54	30.73
MORE	30.27	16.02	31.94

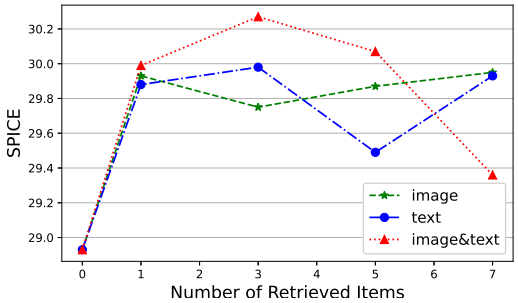


Figure 3: SPICE with different retrieval content numbers.

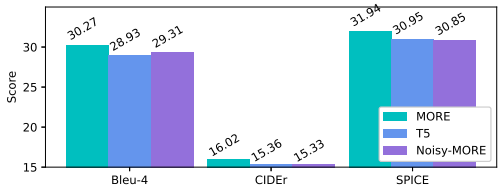


Figure 4: Test result of the baseline model, MORE augmented with relevant content, and MORE augmented with irrelevant content.

better performance.

As shown in Table 2, Although based on a smaller model, MORE can achieve better results than GPT-3.5 and GPT-4. This fully illustrates the effectiveness of our method. Considering that GPT-4 is a multi-modal language model, we also test its performance when retrieval augmented items are provided. However, GPT-4 cannot effectively utilize the retrieved inputs, leading to deteriorated performance. Retrieval augmentation methods for the GPT-4 model are worth exploring in the future.

In terms of incorporating retrieved results, MORE is better than Prepend and FiD, which are textual augmented models. Although previous work found that using captions can have better results, this also risks leaking the answer. The method of directly inputting retrieved results becomes invalid after the retrieval content changes. MORE is also better than VisCTG, which is a visual augmented model. As shown in Appendix C, the generated captions are not always accurate and may lack the required information, so pre-converting images to captions is not a proper approach to leverage images.

5.2 Ablation Study

We examine the effectiveness of various components within MORE by creating several variants, selectively removing or substituting each component, with results detailed in Table 3.

First, we replace the concepts that are input into the integrator with a randomly initialized learnable token sequence (w/o concept-input). The drop in performance highlights the importance of using concept words for references when extracting beneficial information from the retrieved results. Second, we remove the query dropout strategy (w/o query-dropout). The performance drop demonstrates its importance in effectively leveraging the retrieved results. Finally, we further exclude the noisy retrieval augmented input (w/o noisy-RA). Performance degradation indicates that blindly trust in retrieved input can harm model performance. It is necessary to explicitly instruct the model to learn to ignore the irrelevant augmentation results.

Please note that when the query dropout strategy is removed, it means that noisy retrieval augmented input is also not contained. However, the results of 'w/o query-dropout' is better than the results of 'w/o noisy-RA'. This emphasizes the disadvantages of blindly trusting retrieval items, and it is necessary for the model to learn to distinguish irrelevant retrieval input.

5.3 Analysis

Are the improvements of MORE attributed to additional parameters? Considering that our framework introduces more parameters, we investigate whether the performance improvements are attributed to these additional parameters, which arise from two aspects: 1) The retrieval augmented prompt results in an extended input length. To investigate, we adjust the task prompt length from 32 to 64, aligning with the total input length of MORE. 2) The integrator introduces more learnable parameters. To assess whether this would affect performance, we replace the retrieval inputs of each sample with irrelevant retrieved results during

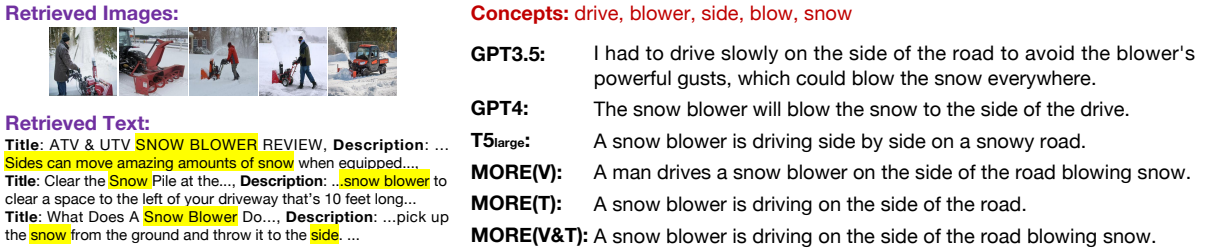


Figure 5: Generated sentences that benefit from retrieval augmentation

training, denoted as rand-RA. This maintains consistency in the learnable parameters with MORE. The experimental results are recorded in Table 4. Neither of them shows significant improvements over the backbone T5-base, showing that the benefit of MORE does not come from the extra parameters.

Will utilizing more retrieved results enhance the model performance? In retrieval-augmented methodologies, a crucial factor influencing the final results is the number of retrieval items. We integrate varying numbers of retrieval content for both single-modal and multi-modal settings. The SPICE results are illustrated in Figure 3, with additional metric results available in Appendix D. No matter which modality is used, the model performance first increases and then decreases as the number of retrieval inputs increases. This is because too few retrieval inputs may lead to insufficient coverage of required information, while an excess of inputs may introduce redundancy and noise.

Is MORE robust to noise in retrieval augmentation? The retrieved results may not always be of high quality and occasionally may be even irrelevant to the query. Therefore it is also important for retrieval augmented models to be robust to the noisy retrieval outcome. We test the model's robustness in the face of poorly retrieved results by feeding it with only irrelevant retrieval content during testing (denoted as Noisy-MORE). As illustrated in Figure 4, Noisy-MORE performs similarly to its backbone T5 when using irrelevant results for augmentation. This indicates that MORE is robust to the noise in the retrieved items by not blindly trusting the augmentation input.

5.4 Case Study

We conduct case studies to qualitatively analyze how MORE enhances text generation through retrieval augmentation. As shown in Figure 5, small LMs like $T5_{LARGE}$ make a nonsensical sentence that 'only one blower can not drive side by side'. The retrieved images show a 'blower located on the side of the road' scene, and the retrieved text describes 'throw the snow to the side of the road', thereby helping the model clarify the usage of 'side' and correct generation errors. More cases can be found in Appendix E.

As for LLMs, we find that they sometimes make nonsensical sentences, as shown in Figure 1. This shows that even if a huge amount of parameters and training data are used, the LLMs are still not able to grasp commonsense knowledge completely. Therefore, it is also necessary for LLMs to use retrieval augmentation to provide reference. The other characteristic is that the sentences made by GPT-3.5 and GPT-4 are usually long. To connect the given concepts and output reasonable sentences, they may need more words or information. This also reflects the lack of commonsense knowledge that humans are well aware of.

6 Conclusion and Future Work

To sum up, we introduce MORE, a multi-modal retrieval augmentation framework. Our approach is capable of extracting useful information and disregarding irrelevant noise from visual and textual results of variable quality, thereby assisting language models in generating reasonable sentences. Extensive experiments on the CommonGen task demonstrated the effectiveness of our method. This novel approach may offer a new perspective for retrieval-augmented language models.

We focus on the generation task in this work and the application of multi-modal retrieval augmentation on more tasks is worth exploring in the future. Besides, the current method concentrates on 'how to incorporate multimodal retrieved items' and does not involve optimization of the retrieving step, which is left for future work.

7 Limitations

Our method uses soft-prompt, making it unsuitable for LMs accessible solely through the API as it cannot convey input in natural language form. In addition, to avoid changing the internal structure of the LMs, we adopted the p-tuning in this work. Using more advanced methods such as LoRA (Hu et al., 2021) to achieve better results can be considered in future work.

The retrieved results are from public data on the Internet and we did not collect any privately identifiable information in our study. However, it may be inevitable to crawl some public photos and other data, which may, which may still include some personal information, such as faces. We followed Bing's authorization requirements for the use of data and did not modify or use it commercially. We call on anyone using our framework to follow the licensing requirements and not misuse the technology.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. *ArXiv*, abs/1607.08822.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Moshe Berchansky, Peter Izsak, Avi Caciularu, Ido Dagan, and Moshe Wasserblat. 2023. Optimizing retrieval-augmented reader models via token elimination. *arXiv preprint arXiv:2310.13682*.

Paul Bloom. 2000. How children learn the meanings of words.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Conference on Empirical Methods in Natural Language Processing*.

Steven Y Feng, Kevin Lu, Zhuofu Tao, Malihe Alikhani, Teruko Mitamura, Eduard Hovy, and Varun Gangal. 2022. Retrieve, caption, generate: Visual grounding for enhancing commonsense in text generation models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10618–10626.

Linda B. Gambrell and Ruby J. Bales. 1986. Mental imagery and the comprehension-monitoring performance of fourth- and fifth-grade poor readers. *Reading Research Quarterly*, 21:454.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Conference on Automated Knowledge Base Construction*.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Hangyu Guo, Kun Zhou, Wayne Xin Zhao, Qinyu Zhang, and Ji rong Wen. 2022. Visually-augmented pretrained language models for nlp tasks without images. *ArXiv*, abs/2212.07937.

Catherine Havasi, Robert Speer, and Jason Alonso. 2007. Conceptnet 3: a flexible, multilingual semantic network for common sense knowledge. In *Recent advances in natural language processing*, pages 27–29. John Benjamins Philadelphia, PA.

Xingwei He, Yeyun Gong, A-Long Jin, Weizhen Qi, Hang Zhang, Jian Jiao, Bartuer Zhou, Biao Cheng, Sm Yiu, and Nan Duan. 2022. Metric-guided distillation: Distilling knowledge from the metric to ranker and retriever for generative commonsense reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 839–852.

J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685.

Gautier Izacard and Édouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*.

Victoria L Joffe, Kate Cain, and Nataša Marić. 2007. Comprehension problems in children with specific language impairment: Does mental imagery training help? *International Journal of Language & Communication Disorders*, 42(6):648–664.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Conference on Empirical Methods in Natural Language Processing*.

Haonan Li, Yeyun Gong, Jian Jiao, Ruofei Zhang, Timothy Baldwin, and Nan Duan. 2021. Kfcnet: Knowledge filtering and contrastive learning for generative commonsense reasoning. In *Conference on Empirical Methods in Natural Language Processing*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Bill Yuchen Lin, Wangchunshu Zhou, Ming Shen, Pei Zhou, Chandra Bhagavatula, Yejin Choi, and Xiang Ren. 2020. Commongen: A constrained text generation challenge for generative commonsense reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*.

Junyi Liu, Liangzhi Li, Tong Xiang, Bowen Wang, and Yiming Qian. 2023. Tcra-llm: Token compression retrieval augmented large language model for inference cost reduction. *arXiv preprint arXiv:2310.15556*.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021. Gpt understands, too. *ArXiv*, abs/2103.10385.

Xin Liu, Dayiheng Liu, Baosong Yang, Haibo Zhang, Junwei Ding, Wenqing Yao, Weihua Luo, Haiying Zhang, and Jinsong Su. 2022. Kgr4: Retrieval, retrospect, refine and rethink for commonsense generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11029–11037.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Ruotian Luo, Brian L. Price, Scott D. Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6964–6974.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

OpenAI. 2023a. Gpt-3.5 turbo fine-tuning and api updates.

OpenAI. 2023b. Gpt-4.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Annual Meeting of the Association for Computational Linguistics*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Lee Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. In *The Eleventh International Conference on Learning Representations*.

Tianyi Tang, Yushuo Chen, Yifan Du, Junyi Li, Wayne Xin Zhao, and Ji rong Wen. 2023. Learning to imagine: Visually-augmented natural language generation. *ArXiv*, abs/2305.16944.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard

Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

PeiFeng Wang, Jonathan Zamora, Junfeng Liu, Filip Ilievski, Muhao Chen, and Xiang Ren. 2022. Contextualized scene imagination for generative commonsense reasoning. In *International Conference on Learning Representations*.

Xin Eric Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4580–4590.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. In *North American Chapter of the Association for Computational Linguistics*.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. 2023. Recomp: Improving retrieval-augmented lms with compression and selective augmentation. *arXiv preprint arXiv:2310.04408*.

Wenhao Yu, Chenguang Zhu, Zhihan Zhang, Shuohang Wang, Zhuosheng Zhang, Yuwei Fang, and Meng Jiang. 2022. Retrieval augmentation for commonsense reasoning: A unified approach. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4364–4377.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pre-trained transformer language models. *ArXiv*, abs/2205.01068.

Wanrong Zhu, An Yan, Yujie Lu, Wenda Xu, Xin Eric Wang, Miguel Eckstein, and William Yang Wang. 2022. Visualize before you write: Imagination-guided open-ended text generation. *arXiv preprint arXiv:2210.03765*.

Table 5: Test results on CommonGen(V1.1) by directly using the captions retrieved through BM25 as output and other existing methods.

Model	$Bleu_4$	CIDEr	SPICE
BM25	44.07	18.64	33.47
$DKMR^2$	44.33	19.54	34.59
KFCNet	43.62	18.85	33.91
KGR^4	42.82	18.42	33.56

A Test Result Using Captions

Since the Commongen dataset itself relies on caption data during the construction process, and most existing methods use the retrieved caption as a reference for generation, such as DKMR2 (He et al., 2022), KFCNet (Li et al., 2021), and KGR4 (Liu et al., 2022). We suspect and test whether the retrieved caption itself reveals the correct answer to some extent. Specifically, we use concepts as query, and then simply use BM25 as the retrieval method to retrieve captions from image captions and video captions. The retrieved caption will be directly used as the prediction result without modification and the results compared with other methods are shown in Table 5. It can be seen that even without a tunable retriever and any modification to the caption, good results can be achieved.

B Retrieved Inputs Crawling and Preprocessing

For the image, we use the template 'a photo of {...}' and crawl the first 20 image results returned by the search engine. We further removed duplicate images based on the dHash algorithm. For text, we crawl the results from the first two pages. Considering that the full document associated with each result may be very long, the search engine has provided a concise text summary of the webpage aligning with the search keywords, we only keep the title and description in the snapshot. We also removed the URL and non-English parts.

C Examples of Generated Captions from VisCTG

We use an example to illustrate why it is better to directly use the raw image than to convert the image into a caption. There are two main reasons: 1) the generated caption may be inaccurate. As shown in Figure 6, due to the error of the model and the bias in training data, when 'dog' appears, the caption model always generates sentences containing 'frisbee' or 'ball', even though these objects do not appear in the image. Inaccurate captions will further mislead follow-up text generation. 2) the pre-generated captions may lack the required information. In the example, the T5 model incorrectly generates 'the back of the grass', and the information needed is 'dog rolls on their back'. Although the images contain relevant information, it is not included in the captions, so that the original generation cannot be corrected.

Figure 6: An example of the generation of VisCTG, the generated captions as well as corresponding images. We also show the generation of MORE and the retrieval content ii use. Since the captions used by VisCTG are ranked by their coverage of the concept words in descending order, the order of images of VisCTG and MORE may be different.

D Results of Other Metrics

Combining various metrics shown in Figure 7, it can be seen that using three retrieval contents is a better choice. The conclusion that using too many or too few retrieved results will not lead to optimal results has not changed.

E Cases of Generation

We show two additional examples in Figure 8 to help intuitively understand how multi-modal retrieval augmentation helps the model generate more reasonable sentences.

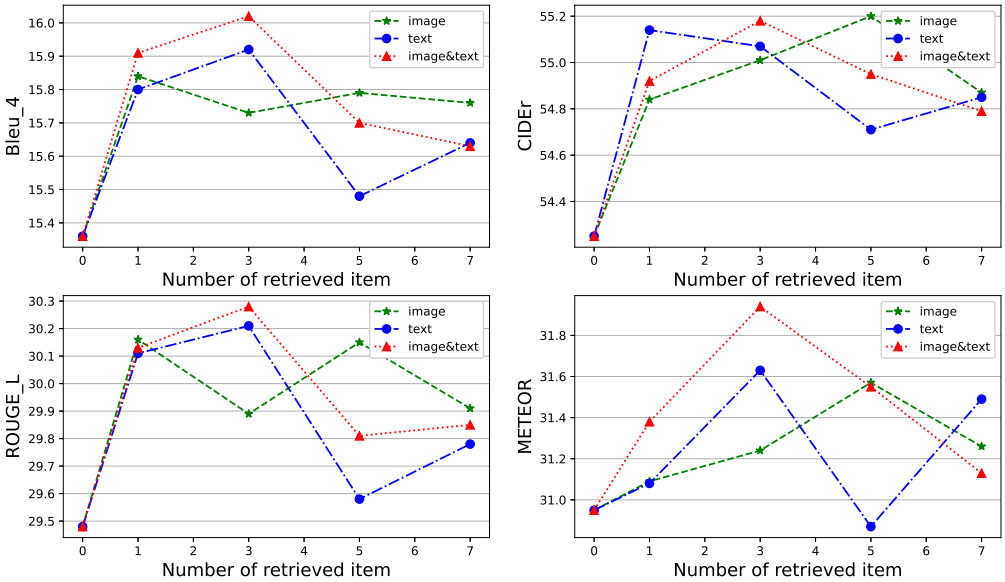
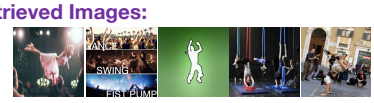


Figure 7: Scores with different retrieval content numbers.

Retrieved Images:



Retrieved Text:

Title: Fist Pump Dance GIFs, **Description**: ... add popular Fist Pump Dance animated GIFs to your conversations...,
Title: The Jersey Floor Workout, **Description**: ... extend your dominant arm into the air, form a tight fist, and pump it from a...
Title: Donald Trump's Fist Pump Body Language Is Pretty, **Description**: Trump's Fist Pump At A 9/11 Memorial Revealed...

Concepts: pump, air, dance, room, fist

GPT3.5:	As the music filled the room, she used her fist to pump the air with excitement, breaking into a spontaneous dance.
GPT4:	I watched as people filled the room, dancing with fists pumping in the air.
T5$_{large}$:	A man pumps air into a room and dances with his fists.
MORE(V):	A man pumps the air with his fist as he dances in a room.
MORE(T):	A man pumps the air with his fist as he dances in the room.
MORE(V&T):	A woman dances in a room with her fists pumping the air.

Retrieved Images:

Retrieved Text:

Title1: Amazon.com: Bagpiper Uniform | Scottish Kilt, Description1: Choose from our extensive collection of over...,
Title2: Bagpiper Outfit | Custom Bagpiper Outfit, Description2: ...a kilt outfit that will make you look like a true Scottish gentleman?...
Title3: Bagpipe Band Uniforms..., Description3: ... genuine Scottish kilt and accessories from Claymore Imports...

Concepts: bagpipe, dress, front, kilt, stand

GPT3.5:	The man in a kilt stood in front, playing the bagpipe while the others in dresses watched.
GPT4:	The bagpipe player stood in front of the crowd in his kilt and dress uniform.
T5$_{large}$:	A woman in a kilt stands in front of a bagpipe
MORE(V):	A man in a kilt stands in front of a bagpipe.
MORE(T):	A man in a kilt stands in front of a bagpipe.
MORE(V&T):	A man in a kilt stands in front of a bagpipe.

Figure 8: Generated sentences with/without retrieval augmentation