# Retrieving Multimodal Information for Augmented Generation: A Survey

**Ruochen Zhao**[1] **Hailin Chen**[1] **Weishi Wang**[1] **Fangkai Jiao**[1]
**Xuan Long Do**[1] **Chengwei Qin**[1] **Bosheng Ding**[1] **Xiaobao Guo**[1]
**Minzhi Li** [2] **Xingxuan Li** [1] **Shafiq Joty**[1,3]*
[1]Nanyang Technological University, Singapore
[2]National University of Singapore, Singapore
[3]Salesforce Research
{ruochen002, hailin001, xuanlong001, weishi001, fangkai002, chengwei003, bosheng001}@e.ntu.edu.sg
{xiaobao001, xingxuan001}@e.ntu.edu.sg, li.minzhi@u.nus.edu, srjoty@ntu.edu.sg



## Abstract

In this survey, we review methods that retrieve multimodal knowledge to assist and augment generative models. This group of works focuses on retrieving grounding contexts from external sources, including images, codes, tables, graphs, and audio. As multimodal learning and generative AI have become more and more impactful, such retrieval augmentation offers a promising solution to important concerns such as factuality, reasoning, interpretability, and robustness. We provide an in-depth review of retrieval-augmented generation in different modalities and discuss potential future directions. As this is an emerging field, we continue to add new papers and methods.

## 1 Introduction

Generative Artificial Intelligence (GAI) has demonstrated impressive performances in tasks such as text generation (Ouyang et al., 2022; Chowdhery et al., 2022; Brown et al., 2020) and text-to-image generation (Ramesh et al., 2021a). Powered by their abilities in modality-specific tasks, the recent incorporation of multimodality (Driess et al., 2023; OpenAI, 2023; Huang et al., 2023b) has opened up possibilities for generative models to serve as general-purpose learners in different formats of information.

However, generative models suffer from inevitable limitations, such as hallucinations (Ye and Durrett, 2022), arithmetic difficulties (Patel et al., 2021), and lack of interpretability. Thus, a promising solution for generative models is learning to interact with the external world and retrieve knowledge in different formats, thus augmenting their generation abilities (Mialon et al., 2023).

Recently, there have been emerging studies focusing on retrieval-based approaches, which aim to provide generative models with more information. Among them, most (Nakano et al., 2021; Guu et al., 2020) use textual information retrieved from the web or textual corpora. Although the textual format aligns with data used during pre-training and offers a natural medium for interaction, there is more world knowledge contained in other formats, such as images, videos, graphs, and audio. These types of information are often inaccessible, unavailable, or not describable in traditional textual corpora.

Recent advancements in Multimodal Large Language Models (MLLMs) (Huang et al., 2023b; OpenAI, 2023; Driess et al., 2023) have improved the capability to handle multi-format information of generative models, demonstrating the significant potential in augmenting generation with multimodal knowledge. This has resulted in an emerging trend of work that utilizes retrieval-based and multimodal techniques to effectively address the limitations such as hallucination and lack of interpretability.

In this survey, we review recent advancements in multimodal retrieval-augmented generation. Specifically, for each modality, there are often differences in retrieval and synthesis procedures, goals, and targeted tasks. Thus, we group relevant methods into different modalities, including image, code, structured knowledge, audio, and video.

For each modality, we review the previous work, the current state, and future challenges. For example, in the image domain, retrieval-augmented methods have been used to better ground visual question-answering (VQA) tasks (Chen et al., 2022a; Tiong et al., 2022) and generate more factual captions (Yang et al., 2023b; Yasunaga et al., 2022). In the code domain, retrieval-based works decouple logic and textual information, which results in more faithful and factual outputs (Lyu et al., 2023; Chen et al., 2022c). To enhance factuality, some methods (Thoppilan et al., 2022; Cheng et al., 2022) also retrieve grounding contexts from structured knowledge, such as tables and knowledge

*Work done while the author is on leave from NTU

graphs. Moreover, there are emerging works in combining audio and video retrieval in generative models (He et al., 2022b; Bogolin et al., 2022).

We believe that the emergence of multimodal retrieval-augmented generation contains the solution to many current challenges. To encourage more future research in this domain, we analyze several promising future directions, including retrieval-augmented multimodal reasoning, building a multimodal knowledge index, and combining retrieval with pre-training.

As the direction of multimodal retrieval-augmented generation is emerging, we will continue to add new works and expand the scope of our current survey.

## 2 Background

### 2.1 Multimodal Learning

Multimodal learning focuses on learning a unified representation for data from different modalities, e.g., text, images, audio, and video. Multimodal learning aims at extracting complementary information to facilitate compositional tasks (Baltrušaitis et al., 2018; Gao et al., 2020). With the fruitful progress made in computer vision (Dosovitskiy et al., 2021; Liu et al., 2021d), natural language processing (Lan et al., 2020; Lewis et al., 2020), and speech recognition (Baevski et al., 2020; Hsu et al., 2021), multimodal models that are capable of processing and integrating data from different modalities have been greatly improved.

Multimodal learning has numerous applications. For instance, multimodal learning can improve image recognition accuracy by analyzing images and videos in conjunction with textual descriptions in computer vision (Ju et al., 2022; Alayrac et al., 2022a; Jia et al., 2021; Radford et al., 2021b). Multimodal models can incorporate visual information from images or videos to enhance language understanding and generation (Zhou et al., 2020; Lei et al., 2021). It also has the potential to significantly enhance the performance of machine learning systems in different domains by allowing them to learn from and integrate multiple sources of information (Tsai et al., 2019; Acosta et al., 2022; Nagrani et al., 2021).

With the increasing availability of large-scale multimodal datasets (Elliott et al., 2016; Sheng et al., 2016; Duarte et al., 2021), multimodal pre-trained models have been developed and showed promising results in various applications (Gan et al., 2022; Uppal et al., 2022). Using the successful Transformer-based architecture, large multimodal pre-trained models, such as VL-Bert (Su et al., 2020), SimVLM (Wang et al., 2021d), ALBEF (Li et al., 2021), and CLIP (Radford et al., 2021b) are highly effective at learning complex patterns and relationships in multimodal data. These large models can then be transferred to different downstream tasks including VQA, image captioning, and object detection.

Additionally, there has been growing interest in developing models that can generate output that incorporates multiple modalities of data. For example, DALL-E (Ramesh et al., 2021b) is fed with pairs of textual descriptions and corresponding images to learn the joint representations. It can generate highly creative and diverse images from even very complex textual descriptions. Similarly, VQGAN-CLIP (Crowson et al., 2022) can generate new images based on textual prompts, where the textual description is used to guide the generation of the image. It combines the CLIP model for image-text understanding with the VQGAN model for image generation. There is also potential to improve the performance of natural language processing models by incorporating visual information in language generation tasks (Lin and Byrne, 2022; Chen et al., 2022a).

Multimodal generative models have a wide range of applications, such as text-image generation, creative writing generation, and multilingual translation. They can also be used to produce new product designs or textual content including website content and documents. However, there remain challenges for multimodal generation models, such as access to a large amount of multimodal data, the network design that produces semantically meaningful outputs, the interpretability of the models, and related ethical issues. It is critical to address these challenges to realize the full potential of multimodal generative models and ensure the proper use of these models.

### 2.2 Retrieval-Augmented Generation

The idea of retrieval-augmented generation is popular nowadays in natural language processing (NLP), which has been a longstanding challenge in the field of artificial intelligence (AI). In the past, the primary research focus was on developing specialized frameworks for specific tasks (Chiu and Nichols, 2016; Liu et al., 2016; Ding et al., 2020; Qin and

Joty, 2022a). In recent years, there has been a significant shift in approach towards utilizing powerful, general-purpose language models that can be fine-tuned or prompt-tuned for a wide range of applications (Devlin et al., 2019; Yang et al., 2019; Raffel et al., 2019; Lewis et al., 2019; Brown et al., 2020; Liu et al., 2021b; Qin and Joty, 2022b; Ding et al., 2022b; Qin et al., 2023a). Through pre-training on a large-scale unlabeled corpus, pre-trained language models have shown significant improvement in a wide range of NLP tasks (He et al., 2021b; Liu et al., 2021a; Ding et al., 2022a; Qin et al., 2023b; Zhou et al., 2023). While this approach showed great potential, it is mainly applied to simple tasks such as sentiment analysis, which humans can easily accomplish without requiring additional knowledge or expertise (Lewis et al., 2020).

In order to address the difficulties associated with resolving knowledge-intensive NLP tasks, there exist primarily two approaches. The first approach involves pre-training on a knowledge base and storing the acquired knowledge within a PLM (Zhang et al., 2019; Liu et al., 2020; Wang et al., 2021a; Liu et al., 2022; Zhou et al., 2022b; Jiang et al., 2022). The benefit of this approach is that it leverages a single model. However, it has two significant disadvantages: Firstly, it is difficult to control what knowledge has been learned by the models; Secondly, parameter updates are required when new knowledge comes in. The second approach is to develop retrieval-augmented generation methods (Gu et al., 2018; Weston et al., 2018; Cai et al., 2019b; Lewis et al., 2020) by combining a retrieval-based component and a generative component (e.g. PLM, LLM, etc.). Specifically, we denote the generative model by $f$ and input text by $x$. Traditional generative models focus on predicting output $y$ by $y = f(x)$. Denote the retriever module by $g$, and we could retrieve segments of information $c^r$ based on (parts of) the input $x^r \in x$. Thus, the retriever can predict $c^r = g(x^r)$. Then, the retrieval-augmented generation can be formulated as: $y = f(x, c)$, where $c = \{\langle x^r, c^r \rangle\}$ is a set of relevant instances retrieved from either the original training set or external datasets to improve response generation. The primary concept behind this approach is that $c^r$ can aid in generating a better response if it is similar or relevant to the input $x^r$. The retrieval memory can be obtained from three sources: the training corpus, external datasets, and large-scale unsupervised corpus (Li et al., 2022a).

Retrieval-augmented generation has been applied to a wide range of downstream NLP tasks, including machine translation (Gu et al., 2018; Zhang et al., 2018; Xu et al., 2020; He et al., 2021a), dialogue generation (Weston et al., 2018; Wu et al., 2019; Cai et al., 2019a), abstractive summarization (Peng et al., 2019), knowledge-intensive generation (Lewis et al., 2020; Izacard and Grave, 2021), etc. For text retrieval, there exist two types of retrievers that can be used to augment an LM: dense and sparse (Mialon et al., 2023). Sparse retrievers (Robertson et al., 2009) use sparse bag-of-words representations, while dense neural retrievers (Asai et al., 2022) use dense query and document vectors. Both types assess document relevance to a query, with sparse retrievers excelling at precise term overlap and dense retrievers being better at computing semantic similarity (Luan et al., 2021). Various works have proposed methods to jointly train a retrieval system with an encoder or sequence-to-sequence LM, achieving comparable performance to larger LMs that use significantly more parameters. These models include REALM (Guu et al., 2020), RAG (Lewis et al., 2020), and RETRO (Borgeaud et al., 2022), which integrate retrieval into existing pre-trained LMs, and Atlas (Izacard et al., 2022), which obtains a strong few-shot learning capability despite being much smaller than other large LMs. Recent works propose combining a retriever with chain-of-thought (CoT) prompting for reasoning to augment language models (He et al., 2022a; Trivedi et al., 2022). For example, Anonymous (2023) verifies the validity of CoT reasoning steps and retrieves relevant contexts to augment the generation of the uncertain ones. He et al. (2022a) generate reasoning paths using CoT prompts and retrieve knowledge to support the explanations and predictions. Trivedi et al. (2022) propose an information retrieval CoT approach for multi-step question answering, where retrieval guides CoT reasoning and vice versa.

## 3 Multimodal Retrieval-Augmented Generation

As there are different retrieval and synthesis procedures, targeted tasks, and challenges for each modality, we discuss relevant methods by grouping them in terms of modality, including image, code, structured knowledge, audio, and video.

### 3.1 Image

Incorporating image data with text information has long been a crucial research topic, as a considerable amount of world knowledge is stored exclusively in images.

Recent advances on pretrained models shed light on general image-text multi-modal models: Flamingo (Alayrac et al., 2022b) can generate comprehensive captions from input images. FIBER (Dou et al., 2022) proposes a two-stage vision-language (VL) pre-training strategy benefiting different levels of VL tasks. DALL-E (Ramesh et al., 2021a) and Parti (Yu et al., 2022) can generate images based on given text instructions. CM3 (Aghajanyan et al., 2022) models both text and image for its input and output. Blip-2 (Li et al., 2023) bootstraps language-image pre-training from off-the-shelf frozen vision and language models.

However, these models require huge computational resources for pre-training and large amounts of model parameters — as they need to memorize vast world knowledge, such as what chinchillas look like and where they commonly habitat. More critically, such models cannot efficiently deal with new or out-of-domain knowledge. To this end, multiple retrieval-augmented works have been proposed to better incorporate external knowledge from images and text documents.

Towards open-domain visual question answering (VQA), RA-VQA (Lin and Byrne, 2022) jointly trains the document retriever and answer generation module by approximately marginalizing predictions over retrieved documents. It first uses existing tools of object detection, image captioning, and optical character recognition (OCR) to convert target images to textual data. Then, it performs dense passage retrieval (DPR) (Karpukhin et al., 2020a) to fetch text documents relevant to target image in the database. Finally, each retrieved document is concatenated with the initial question to generate the final prediction, similar to RAG (Lewis et al., 2020). Besides external documents, PICa (Yang et al., 2022) and KAT (Gui et al., 2022) also consider LLMs as implicit knowledge bases and extract relevant implicit information from GPT-3. Plug-and-Play (Tiong et al., 2022) retrieves relevant image patches by using GradCAM (Selvaraju et al., 2017) to localize relevant parts based on the initial question. It then performs image captioning on retrieved patches to acquire augmented context. Beyond text-only augmented context, MuRAG (Chen et al., 2022a) retrieves both text and image data and incorporates images as visual tokens. RAMM (Yuan et al., 2023) retrieves similar biomedical images and captions, then encodes two modalities through different networks.

Apart from VQA, RA-transformer (Sarto et al., 2022) and Re-ViLM (Yang et al., 2023b) generate more factual captions by retrieving relevant captions. Beyond retrieving images and text documents before generating text, Re-Imagen (Chen et al., 2022b) leverages a multi-modal knowledge base to retrieve image-text pairs to facilitate image generation. RA-CM3 (Yasunaga et al., 2022) can generate mixtures of images and text. It shows that retrieval-augmented image generation performs much better on knowledge-intensive generation tasks and opens up new capabilities such as multimodal in-context learning.

### 3.2 Code

Software developers attempt to search for relevant information to improve their productivity from large amounts of available resources such as explanations for unknown terminologies, reusable code patches, and solutions to common programming bugs (Xia et al., 2017). Inspired by the progress of deep learning in NLP, a general retrieval-augmented generation paradigm has benefited a wide range of code intelligent tasks including code completion (Lu et al., 2022b), code generation (Zhou et al., 2022a), and automatic program repair (APR) (Nashid et al.). However, these approaches often treat programming languages and natural languages as equivalent sequences of tokens and ignore the rich semantics inherent to source code. To address these limitations, recent research work has focused on improving code generalization performance via multimodal learning, which incorporates additional modalities such as code comments, identifier tags, and abstract syntax trees (AST) into code pretrained models (Wang et al., 2021c; Guo et al., 2022; Li et al., 2022c). To this end, multimodal retrieval-augmented generation approach has demonstrated its feasibility in a variety of code-specific tasks, including:

**Text-to-Code Generation** Numerous research studies have investigated the utilization of relevant codes and associated documents to benefit code generation models. A prominent example is REDCODER (Parvez et al., 2021), which retrieves the top-ranked code snippets or summaries from an ex-

isting codebase, and aggregates them with source code sequences to enhance the generation or summarization capabilities. As another such approach, DocPrompting (Zhou et al., 2022a) uses a set of relevant documentation as in-context prompts to generate corresponding code via retrieval. In addition to these lexical modalities, RECODE (Hayati et al., 2018) proposes a syntax-based code generation approach to reference existing subtree from the AST as templates to direct code generation explicitly.

**Code-to-Text Generation** Retrieval-based code summarization methods have been studied extensively. For example, RACE (Shi et al., 2022) leverages relevant code differences and their associated commit message to enhance commit message generation. Besides, RACE calculates the semantic similarity between source code differences and retrieved ones to weigh the importance of different input modalities. Another retrieval-based neural approach is Rencos (Zhang et al., 2020), which retrieves two similar code snippets based on the aspects of syntactic-level similarity and semantic-level similarity of a given query code. These similar contexts are then incorporated into the summarization model during the decoding phase. This idea is further explored by Liu et al. (2021c), where retrieved code-summary pairs are used to augment the original code property graph (Yamaguchi et al., 2014) of source code via local attention mechanism. To capture the global semantics for better code structural learning, a global structure-aware self-attention mechanism (Zhu et al., 2019) is further employed.

**Code Completion** Recent advances in retrieval-based code completion tasks (McConnell, 2004) have gained increasing attention. Notably, Hashimoto et al. (2018) adapt the retrieve-and-edit framework to improve the model's performance in code auto-completion tasks. To address practical code completion scenarios, ReACC (Lu et al., 2022b) takes both lexical and semantic information of the unfinished code snippet into account, utilizing a hybrid technique to combine a lexical-based sparse retriever and a semantic-based dense retriever. First, the hybrid retriever searches for a relevant code from the codebase based on the given incomplete code. Then, the unfinished code is concatenated with the retrieval, and an auto-regressive code completion generator will generate the completed code based on them. In order to address project relations, CoCoMIC (Ding et al., 2022c) decomposes a code file into four components: *files*, *global variables*, *classes*, and *functions*. It constructs an in-file context graph based on the hierarchical relations among all associated code components, forming a project-level context graph by considering both in-file and cross-file dependencies. Given an incomplete program, CoCoMIC retrieves the most relevant cross-file entities from its project-level context graph and jointly learns the incomplete program and the retrieved cross-file context for code completion.

**Automatic Program Repair (APR)** Inspired by the nature that a remarkable portion of commits is comprised of existing code commits (Martinez et al., 2014), APR is typically treated as a search problem by traversing the search space of repair ingredients to identify a correct fix (Qi et al., 2014), based on a redundancy assumption (White et al., 2019) that the target fix can often be reconstructed in the search space. Recent studies have shown that mining relevant bug-fix patterns from existing search space (Jiang et al., 2018) and external repair templates from StackOverflow (Liu and Zhong, 2018) can significantly benefit APR models. Joshi et al. (2022) intuitively rank a collection of bug-fix pairs based on the similarity of error messages to develop few-shot prompts. They incorporate compiler error messages into a large programming language model Codex (Chen et al., 2021) for multilingual APR. CEDAR (Nashid et al.) further extends this idea to retrieval-based prompts design using relevant code demonstrations, comprising more modalities such as unit test, error type, and error information. Additionally, Jin et al. (2023) leverage a static analyzer Infer to extract error type, error location, and syntax hierarchies (Clement et al., 2021) to prioritize the focal context. Then, they retrieve semantically-similar fixes from an existing bug-fix codebase and concatenate the retrieved fixes and focal context to form the instruction prompts for program repair.

**Reasoning over Codes as Intermediate Steps** While large language models (LLMs) have recently demonstrated their impressive capability to perform reasoning tasks, they are still prone to logical and arithmetic errors (Gao et al., 2022; Chen et al., 2022c; Madaan et al., 2022). To mitigate

this issue, emerging research works have focused on using LLMs of code (e.g., Codex (Chen et al., 2021)) to generate the code commands for solving logical and arithmetic tasks and calling external interpreters to execute the commands to obtain the results. Notably, Gao et al. (2022) propose to generate Python programs as intermediate reasoning steps and offload the solution step to a Python interpreter. Additionally, Chen et al. (2022c) explore generating chain-of-thought (CoT) (Wei et al., 2022) for not only text but also programming language statements as reasoning steps to solve the problem. During the inference phase, answers are obtained via an external interpreter. Similarly, Lyu et al. (2023) propose Faithful CoT that first translates the natural language query to a symbolic reasoning chain and then solves the reasoning chain by calling external executors to derive the answer. Another example is Ye et al. (2023), which utilizes LLMs to decompose table-based reasoning tasks into subtasks, decouples logic and numerical computations in each step through SQL queries by Codex, and calls SQL interpreters to solve them (a process called "parsing-execution-filling").

LLMs of code are also known as good-structured commonsense reasoners, and even better-structured reasoners than LLMs (Madaan et al., 2022). As a result, prior studies have also investigated the idea of transforming structured commonsense generation tasks into code generation problems and employing LLMs of code as the solvers. One such work is CoCoGen (Madaan et al., 2022) which converts each training sample (consisting of textual input and the output structure) into a Tree class in Python. The LLMs of code then perform few-shot reasoning over the textual input to generate Python code, and the Python code is then converted back to the original structure for evaluation. Besides, the success of LLMs of code such as Codex in synthesizing computer code also makes them suitable for generating formal codes. Motivated by this, Wu et al. (2022) propose to prove mathematical theorems by adopting Codex to generate formalized theorems from natural language mathematics for the interactive theorem prover Isabelle (Wenzel et al., 2008).

### 3.3 Structured Knowledge

To increase factual grounding and reduce hallucinations, a promising direction is to incorporate more structured knowledge, such as knowledge graphs, tables, and databases. An open challenge in generative models is hallucination, where the model is likely to output seemingly plausible sentences that do not conform to the ground-truth facts. Researchers have denoted that language models, while only relying on internal knowledge (pre-trained weights), fail to recall accurate details when functioning as a knowledge base in question-answering tasks (Ye and Durrett, 2022; Creswell et al., 2022). Thus, A potential solution is to ground generation with retrieved structured knowledge. Structured knowledge, such as knowledge graphs, tables, and databases, often represents how knowledge from different domains is integrated. They could function as a reliant source of truth to enhance factuality.

As the format of structured knowledge departs from the natural texts seen by LLMs during pre-training, how to effectively retrieve and synthesize it for generation has been an open challenge. Xie et al. (2022) represent an early attempt, where all formats of knowledge, including tables, triplets, and ontology, are linearized into text format and fed into the LLM without retrieval. Such methods, however, are limited to the acceptable context length of the PLM and are often computationally expensive.

Some works design task-specific queries to retrieve structured knowledge by fine-tuning. For example, Large language models such as LaMDA (Thoppilan et al., 2022) have adopted such techniques. During fine-tuning, it learns to consult external knowledge sources before responding to the user, including an information retrieval system that can retrieve knowledge triplets and web URLs. Li et al. (2022b) propose a unified dialog model that learns to query pre-defined databases with belief states, which is a list of triplets.

Graph embeddings are used in works such as Pramanik et al. (2021), where a context graph is built on-the-fly to retrieve question-relevant evidence from RDF datasets, including knowledge graphs, using fine-tuned BERT models. Similarly, Heterformer (Jin et al., 2022) retrieves relevant nodes from text-rich networks, such as academic graphs, product graphs, and social media. By combining GNNs and PLMs, it handles tasks such as link prediction and query-based retrieval.

Some works treat the generative model (often large language models) as black-box and retrieve structured information without fine-tuning. For

example, BINDER (Cheng et al., 2022) uses in-context learning to output designed API calls that retrieve question-relevant columns from tables. He et al. (2022a) retrieve from knowledge graphs, such as Wikidata and Conceptnet, based on reasoning steps obtained from the chain-of-thought (CoT) prompting (Wei et al., 2022).

By retrieving from relevant sources, the model not only improves its factual grounding but also provides the grounding contexts while generating, thus addressing interpretability and robustness concerns.

With the potential to handle all types of information expanded by recent advances in LLMs (OpenAI, 2023), we believe that there is much work to be done in this modality, which offers efficient solutions to factuality concerns. There are still many future challenges to be addressed. For example, there should be new designs for better retrieval systems that could promote efficient interactions suitable for diverse knowledge bases. Synthesizing this information correctly into the models is also an open challenge, where it is hard to decide which parts need augmenting in the textual outputs.

### 3.4 Audio

There currently exist several works that use audio information to augment generation.

When audio information is the input for the generation task, retrieval augmentation is explored to learn the audio and lyrics alignment through contrastive learning (He et al., 2022b), which results in a higher-quality generation of captions for music. Moreover, retrieval of key/value pairs from the external knowledge catalog is used for automatic speech recognition tasks (Chan et al., 2023).

In cases where audio information is the output, retrieval is applied in a music generation system with deep neural hashing that encodes the music segments (Royal et al., 2020). Audio-text retrieval is also applied to produce candidates in the process of pseudo prompt enhancement for text-to-audio generation (Huang et al., 2023a). Although there is a limited amount of research work which focuses on retrieval augmented generation tasks involving the audio, it could be a promising future direction (Li et al., 2022a).

It is worth noting that the audio modality is closely intertwined with other modalities. Therefore, recent advancements in audio-text retrieval techniques (Hu et al., 2022; Lou et al., 2022; Koepke et al., 2022) and uses of audio features for text-video retrieval (Falcon et al., 2022; Mithun et al., 2018) can benefit retrieval augmented generation tasks involving other modalities.

### 3.5 Video

Currently, very few works have explored video retrieval for generative tasks, e.g., video captioning. However, the recent studies on dense video representation learning can be useful when developing video knowledge-enhanced generative approaches in the future. Bogolin et al. (2022) propose a query bank normalization method for cross-modal text-video retrieval. Besides, Cap4Video (Wu et al., 2023) and CLIP-ViP (Xue et al., 2022) are data augmentation frameworks that utilize the web-scale pre-trained knowledge to enhance text-video retrieval pre-training. Besides, some works also try to introduce fine-grained interaction between different modalities (Yang et al., 2023a; Wang et al., 2021b). However, these methods still own a significant gap to be the foundation of retrieval-augmented generation models due to the cost of building a video index for knowledge search.

## 4 Future Directions

### 4.1 Retrieval Augmented Multimodal Reasoning

> *The words of the language, as they are written or spoken, do not seem to play any role in my mechanism of thought. The psychical entities which seem to serve as elements in thought are certain signs and more or less clear images which can be "voluntarily" reproduced and combined. — Albert Einstein*

One potential application of multimodal information retrieval is multimodal reasoning. Lu et al. (2022a) first introduce ScienceQA, a large-scale multimodal science question dataset annotated with lectures and explanations. Based on this benchmark, Zhang et al. (2023) propose Multimodal Chain-of-Thought (Multimodal-CoT) which incorporates language and vision modalities into a two-stage (rationale generation and answer inference) framework, surpassing GPT-3.5 by a large margin with a much smaller fine-tuned model. Similar to Zhang et al. (2023), kosmos-1 (Huang et al., 2023b) breaks down multimodal reasoning into two steps. It first generates intermediate content as the rationale based on visual information, and then uses the generated rationale to induce the result. However, both methods may have difficulties in

understanding certain types of images (e.g., maps), which could be mitigated by retrieving relevant informative image-text pairs. We hope that future work can pay more attention to how to effectively and efficiently combine multimodal reasoning with multimodal retrieval.

### 4.2 Building a Multimodal Knowledge Index

In order to facilitate retrieval augmented generation, one of the most fundamental aspects is the building of a multimodal knowledge index. The goal of building a knowledge index is twofold: Firstly, the dense representation should support low storage, dynamic updating of the knowledge base, and accurate search. Secondly, it could enable faster search speed with the help of local sensitive hashing (Leskovec et al., 2014), which combats scaling and robustness concerns when the knowledge base is scaled up extremely.

Currently, the dense representation for text snippets has been widely studied for documents (Karpukhin et al., 2020b; Gao and Callan, 2021; Gao et al., 2021), entities (Sciavolino et al., 2021; Lee et al., 2021), and images (Radford et al., 2021a). Besides, there are also many studies optimizing dense representations in an end-to-end manner (Lewis et al., 2020). Nevertheless, few works (Chen et al., 2022a) have explored building a multimodal index at the same time for downstream generation, and are also limited in text and image. How to map a multimodal knowledge index into a unified space is still a long-term challenge.

### 4.3 Pre-training combined with multimodal retrieval

With the goal of better aligning the abilities to handle different modalities in a pre-trained model, there could be future work built on employing retrieval-based approaches during pre-training. Currently, there have been many methods that fine-tune the pre-trained generative model for retrieval. For example, LaMDA (Thoppilan et al., 2022) can call an external toolset for fine-tuning, including an information retrieval system, a calendar, and a calculator. Similarly, during fine-tuning, Toolformer (Schick et al., 2023) augments models with API calls to tools including a question-answering system and a Wikipedia search engine.

During pretraining, if similar retrieval abilities are leveraged, the generative model would be able to interact with retrieval tools better. Thus, it could output more grounded information, provide relevant contexts to users, and update their information accordingly. When new information comes in, the generative model would be able to effectively retrieve from an up-to-date external base instead of relying solely on pre-trained weights. This advantage also expands to handling robustness in out-of-domain questions.

To incorporate retrieval with pre-training, there remains the challenge of developing appropriate datasets labeled with retrieval-based API calls. To tackle this challenge, LaMDA (Thoppilan et al., 2022) uses labels developed by human annotators, which could be expensive to collect. Toolformer (Schick et al., 2023) uses a sampling and filtering approach for automatic labeling, which is inexpensive but could induce noise and bias. A potential solution is to use a neuro-symbolic approach such as Davoudi and Komeili (2021), which use prototype learning and deep-KNN to find nearest neighbors during training.

## 5 Conclusions

This survey reviews works that augment generative models by retrieving multi-modal information from external sources. Specifically, we categorize the current domain into enhancing with different modalities, including image, code, structured knowledge, speech, and video. As many pretained models call for an external module to handle different formats, they often require further tuning or a tuned external retriever to interact with. With the emergence of large multi-modal models, we believe that this survey could serve as a comprehensive overview of an emerging and promising field. Moreover, we hope it could encourage future research in the domain, including retrieval-augmented multimodal reasoning, building a multimodal knowledge index, and combining retrieval with pretraining.

## References

Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature Medicine*, 28(9):1773–1784.

Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, and Luke Zettlemoyer. 2022. CM3: A causal masked multimodal model of the internet. *CoRR*, abs/2201.07520.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022a. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022b. Flamingo: a visual language model for few-shot learning. *CoRR*, abs/2204.14198.

Anonymous. 2023. Verify-and-edit: A knowledge-enhanced chain-of-thought framework. Anonymous preprint under review.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2022. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Simion-Vlad Bogolin, Ioana Croitoru, Hailin Jin, Yang Liu, and Samuel Albanie. 2022. Cross modal retrieval with querybank normalisation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pages 5184–5195. IEEE.

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.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019a. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1219–1228, Minneapolis, Minnesota. Association for Computational Linguistics.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, and Shuming Shi. 2019b. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1866–1875, Hong Kong, China. Association for Computational Linguistics.

David M Chan, Shalini Ghosh, Ariya Rastrow, and Björn Hoffmeister. 2023. Using external off-policy speech-to-text mappings in contextual end-to-end automated speech recognition. *arXiv preprint arXiv:2301.02736*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022a. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. 2022c. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *arXiv preprint arXiv:2211.12588*.

Zhoujun Cheng, Tianbao Xie, Peng Shi, Chengzu Li, Rahul Nadkarni, Yushi Hu, Caiming Xiong, Dragomir Radev, Mari Ostendorf, Luke Zettlemoyer, et al. 2022. Binding language models in symbolic languages. *arXiv preprint arXiv:2210.02875*.

Jason P.C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling

language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Colin B Clement, Shuai Lu, Xiaoyu Liu, Michele Tufano, Dawn Drain, Nan Duan, Neel Sundaresan, and Alexey Svyatkovskiy. 2021. Long-range modeling of source code files with ewash: Extended window access by syntax hierarchy. *arXiv preprint arXiv:2109.08780*.

Antonia Creswell, Murray Shanahan, and Irina Higgins. 2022. Selection-inference: Exploiting large language models for interpretable logical reasoning. *arXiv preprint arXiv:2205.09712*.

Katherine Crowson, Stella Biderman, Daniel Kornis, Dashiell Stander, Eric Hallahan, Louis Castricato, and Edward Raff. 2022. Vqgan-clip: Open domain image generation and editing with natural language guidance. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 88–105. Springer.

Seyed Omid Davoudi and Majid Komeili. 2021. Toward faithful case-based reasoning through learning prototypes in a nearest neighbor-friendly space. In *International Conference on Learning Representations*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022a. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657, Dublin, Ireland. Association for Computational Linguistics.

Bosheng Ding, Linlin Liu, Lidong Bing, Canasai Kruengkrai, Thien Hai Nguyen, Shafiq Joty, Luo Si, and Chunyan Miao. 2020. DAGA: Data augmentation with a generation approach for low-resource tagging tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6045–6057, Online. Association for Computational Linguistics.

Bosheng Ding, Chengwei Qin, Linlin Liu, Lidong Bing, Shafiq Joty, and Boyang Li. 2022b. Is gpt-3 a good data annotator? *arXiv preprint arXiv:2212.10450*.

Yangruibo Ding, Zijian Wang, Wasi Uddin Ahmad, Murali Krishna Ramanathan, Ramesh Nallapati, Parminder Bhatia, Dan Roth, and Bing Xiang. 2022c. Cocomic: Code completion by jointly modeling in-file and cross-file context. *arXiv preprint arXiv:2212.10007*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, pages 0–7.

Zi-Yi Dou, Aishwarya Kamath, Zhe Gan, Pengchuan Zhang, Jianfeng Wang, Linjie Li, Zicheng Liu, Ce Liu, Yann LeCun, Nanyun Peng, et al. 2022. Coarse-to-fine vision-language pre-training with fusion in the backbone. *arXiv preprint arXiv:2206.07643*.

Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. 2023. Palm-e: An embodied multimodal language model. *arXiv preprint arXiv:2303.03378*.

Amanda Duarte, Shruti Palaskar, Lucas Ventura, Deepti Ghadiyaram, Kenneth DeHaan, Florian Metze, Jordi Torres, and Xavier Giro-i Nieto. 2021. How2sign: a large-scale multimodal dataset for continuous american sign language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2735–2744.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74.

Alex Falcon, Giuseppe Serra, and Oswald Lanz. 2022. A feature-space multimodal data augmentation technique for text-video retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4385–4394.

Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352.

Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation*, 32(5):829–864.

Luyu Gao and Jamie Callan. 2021. Condenser: a pre-training architecture for dense retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 981–993, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. Pal: Program-aided language models. *arXiv preprint arXiv:2211.10435*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiatao Gu, Yong Wang, Kyunghyun Cho, and Victor OK Li. 2018. Search engine guided neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *NAACL-HLT*, pages 956–968. Association for Computational Linguistics.

Daya Guo, Shuai Lu, Nan Duan, Yanlin Wang, Ming Zhou, and Jian Yin. 2022. Unixcoder: Unified cross-modal pre-training for code representation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 7212–7225. Association for Computational Linguistics.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR.

Tatsunori B Hashimoto, Kelvin Guu, Yonatan Oren, and Percy S Liang. 2018. A retrieve-and-edit framework for predicting structured outputs. *Advances in Neural Information Processing Systems*, 31.

Shirley Anugrah Hayati, Raphael Olivier, Pravalika Avvaru, Pengcheng Yin, Anthony Tomasic, and Graham Neubig. 2018. Retrieval-based neural code generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 925–930. Association for Computational Linguistics.

Hangfeng He, Hongming Zhang, and Dan Roth. 2022a. Rethinking with retrieval: Faithful large language model inference. *arXiv preprint arXiv:2301.00303*.

Qiuxiang He, Guoping Huang, Qu Cui, Li Li, and Lemao Liu. 2021a. Fast and accurate neural machine translation with translation memory. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3170–3180.

Ruidan He, Linlin Liu, Hai Ye, Qingyu Tan, Bosheng Ding, Liying Cheng, Jiawei Low, Lidong Bing, and Luo Si. 2021b. On the effectiveness of adapter-based tuning for pretrained language model adaptation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2208–2222, Online. Association for Computational Linguistics.

Zihao He, Weituo Hao, and Xuchen Song. 2022b. Recap: Retrieval augmented music captioner. *arXiv preprint arXiv:2212.10901*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Tao Hu, Xuyu Xiang, Jiaohua Qin, and Yun Tan. 2022. Audio-text retrieval based on contrastive learning and collaborative attention mechanism.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. 2023a. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. 2023b. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*.

Gautier Izacard and Edouard 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, Online. Association for Computational Linguistics.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv*, 2208.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

Jiajun Jiang, Yingfei Xiong, Hongyu Zhang, Qing Gao, and Xiangqun Chen. 2018. Shaping program repair

space with existing patches and similar code. In *ISSTA*, pages 298–309. ACM.

Xiaoze Jiang, Yaobo Liang, Weizhu Chen, and Nan Duan. 2022. Xlm-k: Improving cross-lingual language model pre-training with multilingual knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10840–10848.

Bowen Jin, Yu Zhang, Qi Zhu, and Jiawei Han. 2022. Heterformer: A transformer architecture for node representation learning on heterogeneous text-rich networks. *arXiv preprint arXiv:2205.10282*.

Matthew Jin, Syed Shahriar, Michele Tufano, Xin Shi, Shuai Lu, Neel Sundaresan, and Alexey Svyatkovskiy. 2023. Inferfix: End-to-end program repair with llms. *arXiv preprint arXiv:2303.07263*.

Harshit Joshi, José Cambronero, Sumit Gulwani, Vu Le, Ivan Radicek, and Gust Verbruggen. 2022. Repair is nearly generation: Multilingual program repair with llms. *arXiv preprint arXiv:2208.11640*.

Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. 2022. Prompting visual-language models for efficient video understanding. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV*, pages 105–124. Springer.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020a. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020b. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

A Sophia Koepke, Andreea-Maria Oncescu, Joao Henriques, Zeynep Akata, and Samuel Albanie. 2022. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Jinhyuk Lee, Mujeen Sung, Jaewoo Kang, and Danqi Chen. 2021. Learning dense representations of phrases at scale. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6634–6647, Online. Association for Computational Linguistics.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341.

Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. 2014. *Mining of Massive Datasets, 2nd Ed*. Cambridge University Press.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Huayang Li, Yixuan Su, Deng Cai, Yan Wang, and Lemao Liu. 2022a. A survey on retrieval-augmented text generation. *arXiv preprint arXiv:2202.01110*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705.

Miaoran Li, Baolin Peng, Jianfeng Gao, and Zhu Zhang. 2022b. Opera: Harmonizing task-oriented dialogs and information seeking experience. *arXiv preprint arXiv:2206.12449*.

Zhiyu Li, Shuai Lu, Daya Guo, Nan Duan, Shailesh Jannu, Grant Jenks, Deep Majumder, Jared Green, Alexey Svyatkovskiy, Shengyu Fu, and Neel Sundaresan. 2022c. Automating code review activities by large-scale pre-training. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, ESEC/FSE 2022, Singapore, Singapore, November 14-18, 2022*, pages 1035–1047. ACM.

Weizhe Lin and Bill Byrne. 2022. Retrieval augmented visual question answering with outside knowledge. In *EMNLP*, pages 11238–11254. Association for Computational Linguistics.

Linlin Liu, Bosheng Ding, Lidong Bing, Shafiq Joty, Luo Si, and Chunyan Miao. 2021a. MulDA: A multilingual data augmentation framework for low-resource cross-lingual NER. In *Proceedings of the*

*59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5834–5846, Online. Association for Computational Linguistics.

Linlin Liu, Xin Li, Ruidan He, Lidong Bing, Shafiq Joty, and Luo Si. 2022. Enhancing multilingual language model with massive multilingual knowledge triples. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6878–6890.

P Liu, W Yuan, J Fu, Z Jiang, H Hayashi, and G Neubig. 2021b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing (arxiv: 2107.13586). arxiv.

Pengfei Liu, Xipeng Qiu, and Xuanjing Huang. 2016. Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.

Shangqing Liu, Yu Chen, Xiaofei Xie, Jing Kai Siow, and Yang Liu. 2021c. Retrieval-augmented generation for code summarization via hybrid GNN. In *International Conference on Learning Representations*.

Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. 2020. K-bert: Enabling language representation with knowledge graph. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2901–2908.

Xuliang Liu and Hao Zhong. 2018. Mining stackoverflow for program repair. In *SANER*, pages 118–129. IEEE Computer Society.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021d. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Siyu Lou, Xuenan Xu, Mengyue Wu, and Kai Yu. 2022. Audio-text retrieval in context. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4793–4797. IEEE.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022a. Learn to explain: Multimodal reasoning via thought chains for science question answering. *arXiv preprint arXiv:2209.09513*.

Shuai Lu, Nan Duan, Hojae Han, Daya Guo, Seungwon Hwang, and Alexey Svyatkovskiy. 2022b. ReACC: A retrieval-augmented code completion framework. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6227–6240, Dublin, Ireland. Association for Computational Linguistics.

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, dense, and attentional representations for text retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. Faithful chain-of-thought reasoning. *arXiv preprint arXiv:2301.13379*.

Aman Madaan, Shuyan Zhou, Uri Alon, Yiming Yang, and Graham Neubig. 2022. Language models of code are few-shot commonsense learners. *arXiv preprint arXiv:2210.07128*.

Matias Martinez, Westley Weimer, and Monperrus Martin. 2014. Do the fix ingredients already exist? an empirical inquiry into the redundancy assumptions of program repair approaches. *Companion Proceedings of the 36th International Conference on Software Engineering*.

Steve McConnell. 2004. *Code complete*. Pearson Education.

Grégoire Mialon, Roberto Dessì, Maria Lomeli, Christoforos Nalmpantis, Ram Pasunuru, Roberta Raileanu, Baptiste Rozière, Timo Schick, Jane Dwivedi-Yu, Asli Celikyilmaz, et al. 2023. Augmented language models: a survey. *arXiv preprint arXiv:2302.07842*.

Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K Roy-Chowdhury. 2018. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 19–27.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Noor Nashid, Mifta Sintaha, and Ali Mesbah. Retrieval-based prompt selection for code-related few-shot learning.

OpenAI. 2023. Gpt-4 technical report.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Md Rizwan Parvez, Wasi Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2021. Retrieval augmented code generation and summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2719–2734, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are nlp models really able to solve simple math word problems? *arXiv preprint arXiv:2103.07191*.

Hao Peng, Ankur Parikh, Manaal Faruqui, Bhuwan Dhingra, and Dipanjan Das. 2019. Text generation with exemplar-based adaptive decoding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2555–2565, Minneapolis, Minnesota. Association for Computational Linguistics.

Soumajit Pramanik, Jesujoba Alabi, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Uniqorn: unified question answering over rdf knowledge graphs and natural language text. *arXiv preprint arXiv:2108.08614*.

Yuhua Qi, Xiaoguang Mao, Yan Lei, Ziying Dai, and Chengsong Wang. 2014. The strength of random search on automated program repair. In *ICSE*, pages 254–265. ACM.

Chengwei Qin and Shafiq Joty. 2022a. Continual few-shot relation learning via embedding space regularization and data augmentation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2776–2789, Dublin, Ireland. Association for Computational Linguistics.

Chengwei Qin and Shafiq Joty. 2022b. LFPT5: A unified framework for lifelong few-shot language learning based on prompt tuning of t5. In *International Conference on Learning Representations*.

Chengwei Qin, Shafiq Joty, Qian Li, and Ruochen Zhao. 2023a. Learning to initialize: Can meta learning improve cross-task generalization in prompt tuning? *arXiv preprint arXiv:2302.08143*.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023b. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021a. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021b. 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. *ArXiv*, abs/1910.10683.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021a. Zero-shot text-to-image generation. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021b. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

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

Brandon Royal, Kien Hua, and Brenton Zhang. 2020. Deep composer: Deep neural hashing and retrieval approach to automatic music generation. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2022. Retrieval-augmented transformer for image captioning. In *CBMI*, pages 1–7. ACM.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. *arXiv preprint arXiv:2302.04761*.

Christopher Sciavolino, Zexuan Zhong, Jinhyuk Lee, and Danqi Chen. 2021. Simple entity-centric questions challenge dense retrievers. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6138–6148, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, pages 618–626. IEEE Computer Society.

Shurong Sheng, Luc Van Gool, and Marie Francine Moens. 2016. A dataset for multimodal question answering in the cultural heritage domain. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 10–17.

Ensheng Shi, Yanlin Wang, Wei Tao, Lun Du, Hongyu Zhang, Shi Han, Dongmei Zhang, and Hongbin Sun. 2022. RACE: retrieval-augmented commit message generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5520–5530. Association for Computational Linguistics.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. 2022. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

Anthony Meng Huat Tiong, Junnan Li, Boyang Li, Silvio Savarese, and Steven C. H. Hoi. 2022. Plug-and-play VQA: zero-shot VQA by conjoining large pretrained models with zero training. In *EMNLP (Findings)*, pages 951–967. Association for Computational Linguistics.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv preprint arXiv:2212.10509*.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Shagun Uppal, Sarthak Bhagat, Devamanyu Hazarika, Navonil Majumder, Soujanya Poria, Roger Zimmermann, and Amir Zadeh. 2022. Multimodal research in vision and language: A review of current and emerging trends. *Information Fusion*, 77:149–171.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Jianshu Ji, Guihong Cao, Daxin Jiang, and Ming Zhou. 2021a. K-Adapter: Infusing Knowledge into Pre-Trained Models with Adapters. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1405–1418, Online. Association for Computational Linguistics.

Xiaohan Wang, Linchao Zhu, and Yi Yang. 2021b. T2VLAD: global-local sequence alignment for text-video retrieval. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5079–5088. Computer Vision Foundation / IEEE.

Yue Wang, Weishi Wang, Shafiq R. Joty, and Steven C. H. Hoi. 2021c. Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8696–8708. Association for Computational Linguistics.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021d. Simvlm: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Makarius Wenzel, Lawrence C. Paulson, and Tobias Nipkow. 2008. The isabelle framework. In *Theorem Proving in Higher Order Logics (TPHOLs 2008)*, volume 5170 of *LNCS*, pages 33–38. Springer.

Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and refine: Improved sequence generation models for dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 87–92, Brussels, Belgium. Association for Computational Linguistics.

Martin White, Michele Tufano, Matias Martinez, Monperrus Martin, and Denys Poshyvanyk. 2019. Sorting and transforming program repair ingredients via deep learning code similarities. *2019 IEEE 26th International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 479–490.

Wenhao Wu, Haipeng Luo, Bo Fang, Jingdong Wang, and Wanli Ouyang. 2023. Cap4video: What can auxiliary captions do for text-video retrieval? *CoRR*, abs/2301.00184.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2019. Response generation by context-aware prototype editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7281–7288.

Yuhuai Wu, Albert Qiaochu Jiang, Wenda Li, Markus Norman Rabe, Charles E Staats, Mateja Jamnik, and Christian Szegedy. 2022. Autoformalization with large language models. In *Advances in Neural Information Processing Systems*.

Xin Xia, Lingfeng Bao, David Lo, Pavneet Singh Kochhar, Ahmed E. Hassan, and Zhenchang Xing. 2017. What do developers search for on the web? *Empir. Softw. Eng.*, 22(6):3149–3185.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

Jitao Xu, Josep-Maria Crego, and Jean Senellart. 2020. Boosting neural machine translation with similar translations. In *Annual Meeting of the Association for Computational Linguistics*, pages 1570–1579. Association for Computational Linguistics.

Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2022. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *CoRR*, abs/2209.06430.

Fabian Yamaguchi, Nico Golde, Daniel Arp, and Konrad Rieck. 2014. Modeling and discovering vulnerabilities with code property graphs. In *2014 IEEE Symposium on Security and Privacy, SP 2014, Berkeley, CA, USA, May 18-21, 2014*, pages 590–604. IEEE Computer Society.

Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. 2023a. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. *CoRR*, abs/2302.14115.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of gpt-3 for few-shot knowledge-based vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3081–3089.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Neural Information Processing Systems*.

Zhuolin Yang, Wei Ping, Zihan Liu, Vijay Korthikanti, Weili Nie, De-An Huang, Linxi Fan, Zhiding Yu, Shiyi Lan, Bo Li, Ming-Yu Liu, Yuke Zhu, Mohammad Shoeybi, Bryan Catanzaro, Chaowei Xiao, and Anima Anandkumar. 2023b. Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning. *CoRR*, abs/2302.04858.

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

Xi Ye and Greg Durrett. 2022. The unreliability of explanations in few-shot prompting for textual reasoning. In *Advances in Neural Information Processing Systems*.

Yunhu Ye, Binyuan Hui, Min Yang, Binhua Li, Fei Huang, and Yongbin Li. 2023. Large language models are versatile decomposers: Decompose evidence and questions for table-based reasoning. *arXiv preprint arXiv:2301.13808*.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. 2022. Scaling autoregressive models for content-rich text-to-image generation. *CoRR*, abs/2206.10789.

Zheng Yuan, Qiao Jin, Chuanqi Tan, Zhengyun Zhao, Hongyi Yuan, Fei Huang, and Songfang Huang. 2023. RAMM: retrieval-augmented biomedical visual question answering with multi-modal pre-training. *CoRR*, abs/2303.00534.

Jian Zhang, Xu Wang, Hongyu Zhang, Hailong Sun, and Xudong Liu. 2020. Retrieval-based neural source code summarization. In *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, ICSE '20, page 1385–1397, New York, NY, USA. Association for Computing Machinery.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Graham Neubig, and Satoshi Nakamura. 2018. Guiding neural machine translation with retrieved translation pieces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1325–1335, New Orleans, Louisiana. Association for Computational Linguistics.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, Florence, Italy. Association for Computational Linguistics.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2023. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.

Ce Zhou, Qian Li, Chen Li, Jun Yu, Yixin Liu, Guangjing Wang, Kai Zhang, Cheng Ji, Qiben Yan, Lifang He, et al. 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419*.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13041–13049.

Shuyan Zhou, Uri Alon, Frank F Xu, Zhiruo Wang, Zhengbao Jiang, and Graham Neubig. 2022a. Docprompting: Generating code by retrieving the docs. *arXiv preprint arXiv:2207.05987*.

Wenxuan Zhou, Fangyu Liu, Ivan Vulić, Nigel Collier, and Muhao Chen. 2022b. Prix-LM: Pretraining for multilingual knowledge base construction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5412–5424, Dublin, Ireland. Association for Computational Linguistics.

Jie Zhu, Junhui Li, Muhua Zhu, Longhua Qian, Min Zhang, and Guodong Zhou. 2019. Modeling graph structure in transformer for better amr-to-text generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5458–5467. Association for Computational Linguistics.