A Combination-based Framework for Generative Text-image Retrieval: Dual Identifiers and Hybrid Retrieval Strategies

Kaipeng Li
Graduate School of Comprehensive Human Sciences
University of Tsukuba
Tsukuba, Ibaraki, Japan
keeey.kk@gmail.com

Haitao Yu*
Institute of Library, Information and Media Science
University of Tsukuba
Tsukuba, Ibaraki, Japan
yuhaitao@slis.tsukuba.ac.jp

Yubo Fang
Graduate School of Comprehensive Human Sciences
University of Tsukuba
Tsukuba, Ibaraki, Japan
s2430518@u.tsukuba.ac.jp

Chao Lei
Graduate School of Comprehensive Human Sciences
University of Tsukuba
Tsukuba, Ibaraki, Japan
lei.chao.tkb_ga@u.tsukuba.ac.jp

Abstract

Cross-modal retrieval plays a fundamental role in bridging distinct information sources, such as text, images, and videos. Different from the traditional methods that predominantly rely on discriminative matching via cross-attention or joint embedding spaces, *generative cross-modal retrieval* has recently emerged as a new paradigm. Despite the improvements achieved by recent studies, there are still many open questions. For instance, given the conventional setting that representing each candidate item by a unique identifier and expanding the prefix by one token at a time in a greedy manner during the constrained beam search process, once the prefix of a relevant item's identifier is pruned, it becomes impossible for that item to appear in the final result list, leading to the risk of getting stuck in a local optimum. In this paper, we develop a combination-based framework for generative cross-modal retrieval. Specifically, we not only explore the effectiveness of imposing dual identifiers during the constrained beam search process, but also investigate the benefits of combining different retrieval strategies so as to mitigate the information loss caused by the discrete representations. Based on two benchmark collections, our extensive empirical experiments reveal that: (1) Compared with the state-of-the-art generative text-image retrieval method, our proposed approach based on dual identifiers achieves substantially improved performance. Although generative retrieval with discrete identifiers offers higher efficiency, it still falls significantly short of dense embedding-based retrieval in terms of performance. To overcome this limitation, we propose a hybrid strategy that performs initial generative retrieval followed by reranking the top-k candidates using dense embeddings, resulting in notable improvements in both retrieval performance and efficiency. (2) The factors, such as the choice of base LLMs, the number of top candidate items to rerank, the beam size and the way of combining ranking strategies significantly influence the retrieval performance. Careful examinations of these factors are highly recommended in the development of generative text-image retrieval methods.

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**.

Keywords

Large Language Model, Generative Cross-Modal Retrieval, Dual Identifiers, Hybrid Retrieval

ACM Reference Format:
Kaipeng Li, Haitao Yu, Yubo Fang, and Chao Lei. 2025. A Combination-based Framework for Generative Text-image Retrieval: Dual Identifiers and Hybrid Retrieval Strategies. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region (SIGIR-AP 2025), December 7–10, 2025, Xi'an, China.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3767695.3769474

1 INTRODUCTION

Nowadays, information retrieval (IR) systems play a crucial role in bridging the ever-expanding World Wide Web (WWW) with diverse user information needs, ranging from simple queries to acquiring knowledge and decision-making. The traditional IR methods can be broadly categorized into three groups. The first group consists of *word-based retrieval methods*, such as TF-IDF [53] and BM25 [50], which rely on word matching between queries and documents, emphasizing statistical relationships of words. The second group comprises *learning-to-rank methods*, such as LambdaRank [2] and gradient-boosted decision trees (GBDT)-based ranking models [6, 16], which enhance retrieval accuracy by leveraging handcrafted features to train supervised ranking models. Building upon the boom of deep learning, the third group consists of the more recent *ranking approaches* [17, 43, 48] *based on pre-trained language models*, which have significantly advanced IR performance by densely embedding queries and retrieval items without relying on manually crafted features. Essentially, the above three groups

*Corresponding author.

SIGIR-AP 2025, Xi'an, China

ACM ISBN 979-8-4007-2218-9/2025/12
https://doi.org/10.1145/3767695.3769474

boil down to *the multi-stage index-retrieve-then-rank paradigm* [39]. Benefiting from the recent success of large language models (LLMs), the IR landscape is being reshaped, shifting the index-retrieve-then-rank paradigm to *generative information retrieval* (GIR). In GIR, the candidate items (e.g., documents and images) are first symbolized though unique identifiers, then, a generative retrieval model would be optimized based on a LLM to generate relevant item identifiers in response to a given query.

Towards effective GIR, many methods have been proposed. Based on the modalities involved, we can categorize relevant studies into three groups: (1) *generative unimodal retrieval* (GUR), which focuses on retrieval within a single modality, such as [1, 5, 7, 18, 32, 33, 37, 38, 44, 45, 51, 54–57, 61, 62, 65, 71, 74, 76]; (2) *generative cross-modal retrieval* (GCMR), which addresses retrieval across two distinct modalities, such as [13, 25, 30, 46]; and (3) *universal generative multimodal retrieval* (UGMR), which enables retrieval across multiple modalities within a unified framework, such as [20, 34, 63]. It is obvious that most prior studies focus on GUR, while GCMR and UGMR remain relatively underexplored. In this work, we focus on GCMR, especially the case of generative text-image retrieval. Despite the improvements made by the previous studies [13, 25, 30, 46], the answers to a number of fundamental questions are yet not well understood. For example, in the context of GUR applied to document retrieval, the state-of-the-art method by Zeng et al. [72] demonstrates the effectiveness of incorporating an additional set-based identifier to mitigate the risk of getting stuck in a local optimum when generating tokens one by one via constrained beam search. However, *how to address this issue in the context of GCMR remains an open question*. Furthermore, a review of prior studies (e.g., [30, 31, 73]) reveals that the proposed generative models significantly outperform the classical one-tower and two-tower frameworks. However, given that each framework has its own strengths and limitations, *it remains unclear whether combining them would lead to further improvements*. The aforementioned open questions motivate us to approach GCMR in a novel way. Specifically, inspired by the work [72] that simultaneously incorporates approximate document-level score with prefix-oriented expanding score during the decoding process for generative text-text retrieval, we impose a whole image-level score during the decoding process so as to mitigate the risk of pruning a relevant image solely based on the prefix. Furthermore, to circumvent the inherent limitation that the top-ranked items in the result list by generative models may not be well differentiated due to the discrete identifier presentation, we deploy the dense embedding-based strategy to rerank the top-ranked items. To summarize, the main contributions can be listed as follows:

(1) In the context of GCMR, we view the intermediate discrete code representations of images obtained via the algorithm by Li et al. [30] as *pseudo textual documents*, then we tailor the algorithm by Zeng et al. [72] for constructing sequential identifiers and order-invariant identifiers. This strategy not only allows to validate the reproducibility of these two studies, but also enables us to investigate the effectiveness of imposing dual identifiers to mitigate the risk of getting stuck in a local optimum when generating tokens one by one via constrained beam search for GCMR.

(2) Based on two widely used benchmark datasets (i.e., Flickr30K and MS-COCO), we not only conduct an in-depth intra-paradigm comparison among generative retrieval methods, along with a detailed ablation study on the impact of individual components within the proposed framework, but also perform a comprehensive inter-paradigm comparison of representative methods spanning different retrieval paradigms and architectures, evaluating both performance and efficiency. The intra-paradigm comparison demonstrates that our proposed method achieves significantly superior performance among generative approaches. The inter-paradigm results further show that our hybrid retrieval strategy (namely combining generative retrieval with dense embedding-based methods) preserves the efficiency advantage of generative retrieval while substantially narrowing the performance gap to strong dense baselines (e.g., BLIP). Together, these analyses offer valuable insights into the targeted problem of text-image retrieval.

The remainder of the paper is structured as follows. In the next section, we briefly survey the prior studies on traditional methods for cross-modal retrieval and generative information retrieval. In Section 3, we describe the notations and the formulation of GCMR. In Section 4, we detail the proposed framework. A series of experiments that we conducted are discussed in Sections 5 and 6. Finally, we conclude the paper in Section 7.

2 RELATED WORK

In this section, we first briefly introduce the traditional methods for cross-modal retrieval (CMR). Due to space constraints, we refer the reader to the work [3] for a detailed overview. Then we describe the recent studies on GIR.

2.1 The Traditional Methods for CMR

The conventional methods for CMR rely on learning a shared embedding space where items from different modalities (e.g., text and image) are projected for similarity comparison. The typical lines of early efforts include *exploring different network architectures* [10–12, 22, 60, 73] and *designing various loss functions* [8, 15, 41, 64, 70, 73]. Later on, attention-based architectures [12, 24, 36, 42] and transformer-based methods [13, 27, 40, 47, 67] are introduced, generally falling into the frameworks of *two-tower, two-leg and one-tower*. Specifically, the two-tower framework [4, 47, 75] encodes textual and visual inputs with two separate encoders respectively. The two-leg framework [52, 69] deploys an additional multi-modal fusion layer aiming for feature interaction between image and text modalities. The one-tower [14, 59] aims to unify vision and language learning with a single encoder in order to facilitate efficient communications across data modalities. Nevertheless, the above frameworks face fundamental scalability bottlenecks, because their reliance on extensive indexing structures and the non-linear increase in computation as candidate sets grow hinder their deployment in real-world, large-scale retrieval tasks.

2.2 Generative Information Retrieval

In this section, we introduce representative studies on GIR and discuss the relevance of our work in relation to closely related methods. We refer the reader to the work [29] for a detailed survey on GIR.

2.2.1 Generative Unimodal Retrieval. Numerous methods [1, 5, 7, 18, 32, 33, 37, 38, 44, 45, 51, 54–57, 61, 62, 65, 71, 73, 74, 76] have been

proposed to address generative unimodal retrieval from various perspectives, particularly in the context of document retrieval. To the best of our knowledge, the work by Tay et al. [57] is the first to initiate this line of research. Later on, different methods [51, 54, 62] are proposed focusing the more effective ways of constructing the document identifiers, such as interpretable document identifiers and learning-based tokenization. The studies like [37, 74] explore how to cope with corpus changes over time. Tang et al. [56] investigate the scenario with multi-graded relevance. The study by Qiao et al. [45] investigate the effectiveness of incorporating the diffusion model with constrained decoding, showing boosted retrieval precision. The methods known as RIPOR [71] and PAG [72] are the most related to our work. By tailoring their designs, we construct both sequential identifiers and order-invariant identifiers to represent images.

2.2.2 Generative Cross-modal Retrieval. Due to modality differences, the above methods for generative unimodal retrieval can not be directly used for GCMR. In the context of generative text-image retrieval, Li et al. [31] investigate the effectiveness of different ways for tokenizing images. The recent method [30], referred to as AVG, proposes a new pipeline. First, it tokenizes an image into a sequence of visual tokens. Then, it formulates the text-image retrieval as a token-to-visual-token generation problem. In our work, we follow AVG to obtain intermediate discrete code representations of images. Additionally, exploring the reproducibility of AVG constitutes a key part of our study.

2.2.3 Universal Generative Multimodal Retrieval. UGMR is an emerging research topic that aims to enable modality-agnostic generative retrieval. Typically, a multimodal large language models (MLMM) serves as the core component of such models. For example, Wei et al. [63] establish a benchmark for UGMR by assembling 10 diverse datasets. Kim et al. [20] propose modality-decoupled semantic quantization, transforming multimodal items into discrete identifiers. Lin et al. [34] show that zero-shot MLLM-based rerankers can improve ranking accuracy over strong retrievers.

3 Preliminaries

Let Q and M be the textual query space and the image space, respectively. For a given query $q \in Q$, the task of text-image retrieval is to retrieve relevant images from the image space M.

In the context of generative text-image retrieval, the framework typically consists of two core components: a *converter* denoted as C and a *generative retriever* denoted as G. The converter assigns each image a unique identifier represented as $\mathbf{t} = [t_1, ..., t_i, ..., t_n]$, where t_i is the i-th token, and n is the length. We note that the number of identifiers assigned to each image is not restricted to one; thus, assigning multiple identifiers to a single image is a valid and permissible choice. The generative retriever is optimized to autoregressively generate image identifiers corresponding to a given query q. Different generative text-image retrieval frameworks can be constructed by varying the converter design, adopting different token selection strategies during the autoregressive generation process, and devising alternative loss functions for optimization.

4 The Proposed Framework

Figure 1 briefly shows the proposed *combination-based framework for generative text–image retrieval*, which is referred to as **ComGTIR**. Different from the traditional methods that symbolize each candidate image by a single identifier, we propose to represent an image by combining two different types of identifiers. Thanks to this strategy, it provides us the flexibility of integrating image-level relevance scores as guiding priors during the autoregressive generation process. Next, we elaborate on the details on constructing the image identifiers and the generative retrieval process.

4.1 Constructing Dual Image Identifiers

Though many effective methods [51, 54, 62] for identifier construction have been proposed, especially for generative query-document retrieval, most of them can not be directly used for our case due to modality differences. Specifically, text is inherently discrete and sequential, consisting of well-defined tokens that align naturally with autoregressive generation. In contrast, images are continuous and high-dimensional, lacking clear token boundaries. To enable token-based generative processing, images must first be explicitly discretized. To enable token-based generative text-image retrieval, we first obtain the intermediate discrete representation of each image, then we apply different methods to generate dual identifiers for an image (Figure 1-a).

4.1.1 Intermediate Discrete Representation. Although many image tokenizers [9, 23, 49, 58, 66] have been proposed, particularly in the context of image generation, directly applying them to retrieval tasks is not straightforward. Several shortcomings hinder their effectiveness: (S-1) these tokenizers prioritize reconstruction quality over semantic alignment, often failing to capture features that are sufficiently discriminative for retrieval; (S-2) the generated token sequences may include redundant or low-information tokens, which reduce retrieval efficiency and introduce noise into the matching process; and (S-3) these tokenizers are not trained with retrieval-specific objectives, such as contrastive loss, and therefore their outputs are not optimized for item rankings.

After an in-depth comparison of two closely related studies [30, 73], we finally follow the approach proposed by Li et al. [30] to obtain the intermediate discrete code representations of images. Specifically, for a query-image pair (q, m), we obtain their embedding vectors based on a pretrained transformer-based visual encoder $Encoder_v$ and text encoder $Encoder_t$, namely $\mathbf{q} = Encoder_v(q)$ and $\mathbf{m} = Encoder_v(m)$. Next, we perform the *residual quantization* over each image so as to obtain the intermediate discrete representation $\mathbf{v} = [v_1, ..., v_j, ..., v_J]$. To cope with the aforementioned shortcomings (i.e., S-1, S-2 and S-3), the whole loss function includes three parts. Specifically, the reconstruction loss by Equation-1 ensures that discrete visual tokens keep the original visual information as much as possible. The commit loss by Equation-2 aims to make $\mathbf{v}^{(j)}$ sequentially decrease the quantization error of $\mathbf{m}$ as j increases, where $sg()$ is the stop-gradient operator. The alignment loss by Equation-3 aims to guarantee the semantic similarity between the reconstructed image vector $\hat{\mathbf{m}}$ and the query's vector $\mathbf{q}$.

$$\mathcal{L}_{recon} = \| \mathbf{m} - \mathbf{v} \|_2^2 \tag{1}$$

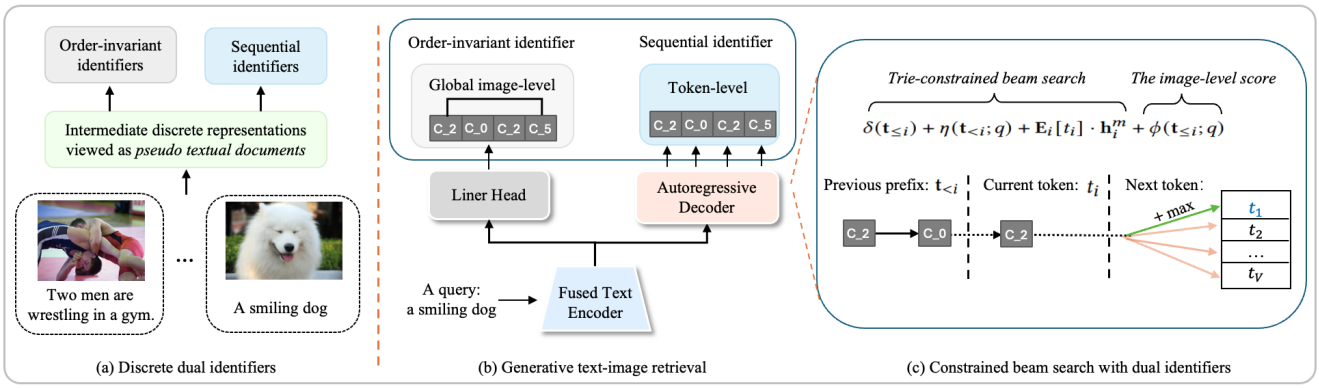

Figure 1: An overview of the proposed framework.

$$\mathcal{L}_{commit} = \sum_{j=1}^{J} \| \mathbf{m} - sg(\mathbf{v}^{(j)}) \|_2^2 \tag{2}$$

$$\mathcal{L}_{align} = \| \mathbf{m} - \mathbf{q} \|_2^2 \tag{3}$$

4.1.2 Sequential Identifier. Given the intermediate discrete representation of an image, we first construct a sequential identifier for the image. Following the study of RIPOR [71], we construct relevance-oriented, order-sensitive sequential identifiers by viewing the intermediate discrete representations of images as *pseudo textual documents*. Specifically, based on the encoder-decoder architecture of the generative retriever G, we feed the discrete code representation $\mathbf{v}$ of an image m into the encoder and a special start token t_0 as the input to the decoder. Then we can get the intermediate representation as

$$\overline{\mathbf{m}} = Decoder(t_0, Encoder(\mathbf{v})) \in \mathbb{R}^D \tag{4}$$

Subsequently, we deploy the residual quantization algorithm again. We compile L embedding matrices $\{\mathbf{E}_l\}_{l=1}^{L}$ to obtain the sequential identifier $\mathbf{w} = [w_1, ..., w_l, ..., w_L]$ for each image. Finally, the optimization objective is to approximate each intermediate representation of a relevant image as the sequence of token embeddings $\overline{\mathbf{m}} \approx \sum_{l=1}^{L} \mathbf{E}_l[w_l]$, where [] denotes the indexing operation.

4.1.3 Order-invariant Identifier. Given the intermediate discrete representation of an image, we also construct an order-invariant identifier for the image. By following the study of PAG [30], we construct order-invariant identifiers by viewing the discrete code representations of images as *pseudo textual documents* again. Specifically, we first compute the contextualized representations $\mathbf{Q}$ by feeding the query text into the generative retriever G as

$$\mathbf{Q} = Decoder(Encoder(q), q) \in \mathbb{R}^{|q| \times |D|} \tag{5}$$

where $|\ q\ |$ and $|\ D\ |$ denotes the query length and the embedding size for each token of the order-invariant identifiers. Let $|\ U\ |$ be the vocabulary size for order-invariant identifiers and $\mathbf{E}_{oi}$ be the associated embedding matrix, we can compute a weight for each token by performing log-saturation and maxpooling operations as follows:

$$\mathbf{h}^q = MaxPool(\log(1 + Relu(\mathbf{E}_{oi} \cdot \mathbf{Q}^\top))) \in \mathbb{R}^{|U|} \tag{6}$$

Then the image-level relevance score with respect to a query q can be given as:

$$s^{sim}(q, m) = s^{sim}(q, \mathbf{u}) = \sum_{k=1}^{K} \mathbf{h}^q[u_k] \tag{7}$$

Finally, we appeal to the triple-based fine-tuning strategy [72] for high-quality order-invariant identifiers. Let (q, m^+, m^-) be a training triple where m^+ and m^- represent a relevant image and a non-relevant image, the loss function is computed as:

$$\mathcal{L}_{oi}(q, m^+, m^-) = \| s^{sim}(q, m^+) - s^{sim}(q, m^-) - T(q, m^+, m^-) \|_2^2 \tag{8}$$

where $T(q, m^+, m^-)$ denotes the teacher margin.

4.2 Generative Retrieval

During the autoregressive generation process (Figure 1-b), we also appeal to constrained beam search based on a prefix tree. We build the prefix tree using the sequential image identifiers. The validation function $\delta(\mathbf{t}_{\leq i})$ that indicates whether a newly generated prefix belongs to at least one valid image identifier or not is defined as:

$$\delta(\mathbf{t}_{\leq i}) = \begin{cases} 0 & \text{if } [t_1,...,t_i] \text{ is a valid prefix in the prefix tree} \\ -\infty & \text{otherwise} \end{cases} \tag{9}$$

Given the previously generated tokens $\mathbf{t}_{<i} = [t_1, ..., t_{i-1}]$ before the i-th step, the hidden representation $\mathbf{h}_i^m$ for the i-th token can be obtained as:

$$\mathbf{h}_i^m = Decoder(Encoder(q), \mathbf{t}_{<i}) \in \mathbb{R}^D \tag{10}$$

Let $\mathbf{E}_i \in \mathbb{R}^{V \times D}$ be the token embedding table at position i, where V is the identifier vocabulary size. Thus we can use $\mathbf{E}_i[t_i]$ to represent the D-dimensional embedding of the token t_i. Following RIPOR [71], we quantify the query-image relevance based on the sequential identifier as:

$$\eta(\mathbf{t} = [t_1, ..., t_L]; q) = \sum_{i=1}^{L} \mathbf{E}_i[t_i] \cdot \mathbf{h}_i^m \tag{11}$$

Given the the i-th token and the previously generated tokens $\mathbf{t}_{<i} = [t_1, ..., t_{i-1}]$, we particularly compute an image-level relevance score as the guiding prior, which is defined as:

$$\phi(\mathbf{t}_{\leq i}; q) = \max_{m \in M_{\mathbf{t}_{\leq i}}} s^{sim}(q, m) \tag{12}$$

where $M_{\mathbf{t}_{\leq i}}$ represents the subset of all images from M with the prefix of $\mathbf{t}_{\leq i}$, and the $\max(\cdot)$ operator is used.

In accordance with the autoregressive principle that the generation of the i-th token t_i is conditioned on the given query q and the previously generated tokens $\mathbf{t}_{<i} = [t_1, ..., t_{i-1}]$, we quantify the gain of expanding the prefix $\mathbf{t}_{<i}$ with the token t_i by jointly combing the aforementioned three factors and define the scoring function as:

$$f(\mathbf{t}_{\leq i}; q) = \delta(\mathbf{t}_{\leq i}) + \eta(\mathbf{t}_{<i}; q) + \mathbf{E}_i[t_i] \cdot \mathbf{h}_i^m + \phi(\mathbf{t}_{\leq i}; q) \tag{13}$$

The underlying rational of Equation-13 is that: The validation score $\delta(\mathbf{t}_{\leq i})$ based on the prefix tree guarantees that the current prefix belongs to at least one valid sequential identifier. Given the current prefix, the score based on the sequential identifier provides fine-grained and order-sensitive score (i.e., $\eta(\mathbf{t}_{<i}; q) + \mathbf{E}_i[t_i] \cdot \mathbf{h}_i^m$), but may suffer from the potential risk of a local optimum. The score $\phi(\mathbf{t}_{\leq i}; q)$ based on the order-invariant identifier offers global image-level semantic cues. Together, they complement each other and enables a better generative control (Figure 1-c).

4.3 Hybrid Retrieval Strategies

A closer examination of the previous steps in Sections 4.1 and 4.2 reveals that discrete quantization is a crucial component of generative retrieval, which inevitably incurs information loss. To mitigate this issue, we propose to combine different retrieval strategies. Specifically, we first perform an initial generative retrieval, then rerank the top-k candidates using an alternative retrieval strategy based on dense embedding vectors, allowing for finer-grained differentiation among top-k candidate images. Given that the number of candidates to rerank is relatively small, the additional computational cost remains negligible.

5 Experimental Setup

In this section, we first introduce the adopted datasets and the evaluation metrics. We then describe the configuration of each method to be evaluated, and the implementation details[1].

5.1 Datasets and Metrics

In this work, we used two widely-used datasets: Flickr30K [68] and MS-COCO [35]. Flickr30K comprises 31,783 images, each paired with five human-annotated sentences. We utilized the data split employed by Li et al. [28], with 29,783 images for training, 1,000 for validation, and 1,000 for testing. MS-COCO consists of 123,287 images, each accompanied by five annotated sentences. We followed the dataset split proposed in [24], using 113,287 images for training, 5,000 for validation, and 5,000 for testing. In our experiments, *we used the training data to learn the model parameters, use the validation data to select the model setting based on R@1, and use the testing data for performance evaluation.*

In line with previous studies [4, 68], we conducted the evaluation by using the standard recall metrics R@k (k=1, 5, 10), as well as rSum [19], which is the sum of R@1, R@5, and R@10. In particular, we view R@1 as the primary metric, since it best captures the retrieval system's ability to rank the most relevant item first in a ranking context.

5.2 Baseline Methods and Configuration

In this work, we compare three groups of methods, each representing a distinct paradigm or a representative neural architecture for text-image retrieval.

(1) **CLIP**[2] [47] is adopted to represent the two-tower framework. For each query-image pair, the query and image are encoded into dense vectors using two separate encoders, and the matching score is computed based on the resulting vectors.

(2) **BLIP** [26] is another representative baseline. In particular, three objectives are jointly optimized during pre-training. The *image-text contrastive* loss (ITC) aims to align the feature space of the visual transformer and the text transformer by encouraging positive image-text pairs to have similar representations in contrast to the negative pairs. The *image-text matching* loss (ITM) aims to learn image-text multimodal representation that captures the fine-grained alignment between vision and language. It is worth noting that solely using the ITC component corresponds to the two-tower framework, while solely using the ITM component corresponds to the one-tower framework. To ensure an accurate and fair comparison, we use $\textbf{BLIP}_{\textbf{itc}}$, $\textbf{BLIP}_{\textbf{itm}}$ and $\textbf{BLIP}_{\textbf{itc+itm}}$ to distinguish the underlying framework differences when deploying BLIP for text-image retrieval. In fact, the reported performance by Li et al. [26] corresponds to $\text{BLIP}_{\text{itc+itm}}$, namely first selecting k candidates based on the image-text feature similarity s_{itc}, and then rerank the selected candidates based on their pairwise ITM scores s_{itm}. Unfortunately, **the previous study [30] inaccurately cited the reported performance of BLIP [26] as representative of the one-tower framework**. In the experiments, we use the fine-tuned ViT-L checkpoint[3].

(3) **GRACE** [31] and **AVG** [30] are adopted to represent the up-to-date generative text-image retrieval methods. AVG represents the state-of-the-art model for generative text-image retrieval. GRACE explores various types of identifiers for images. In this work, we only compare with the setting of Structured ID since it achieves the best performance in their work. Following AVG, we also exclude the "Atomic ID" variant as it falls outside the generative retrieval paradigm.

To clearly show the contribution of each component in our proposed framework, we use **ComGTIR-D** to denote the setting that fuses dual identifiers. Building upon this, **ComGTIR-DH** represents the configuration of further deploying hybrid ranking strategies, namely an initial generative retrieval step followed by reranking the top-k candidates using an alternative retrieval strategy.

For the detailed configuration, we adopt T5-base [48] as the backbone for generative retrieval. We set the default codebook size to 1024 with 4 quantization levels. The codebook is initialized via

[1]The core code for reproducing the reported experimental results is available at: https://github.com/ii-research/2025-SIGIRAP-T2IGIR

[2]https://github.com/openai/CLIP

[3]https://github.com/salesforce/BLIP

k-means clustering on the first training batch. The entire optimization consists of 3 stages. Stage-1 (sequential identifier) and stage-2 (order-invariant identifier) are optimized using AdamW [21] with a learning rate of 1×10^{-3}. Stage-1 is trained for 100 epochs and stage-2 for 30 epochs with a linear decay. For stage-3 (the unified joint training), we use AdamW with an initial learning rate of 5×10^{-4} and linear decay, and train the model for 30 epochs. The batch size is set to 128 for stage 1 and 3, and 256 for stage 2. All models are trained on four NVIDIA A100 40GB GPUs.

6 Results and Analysis

In this section, we present the experimental results and conduct a detailed analysis. We begin with an intra-paradigm comparison of representative generative retrieval methods, then move on to an inter-paradigm comparison across different retrieval frameworks. Finally, we perform an ablation study to assess the contribution of each component of the proposed framework.

6.1 Intra-paradigm Comparison

Regarding the reproducibility of GRACE and AVG, we found the following issues: (1) Both methods only use the training data and test data. Put another way, during each epoch, the models are trained on the training data and directly evaluated on the test data without using the validation data. The best performance on the test set is reported after a certain number of training epochs, which results in the risk of cherry-picking and undermines reproducibility. (2) Both methods involve an augmentation strategy (please refer to Section A in GRACE [31] for more details). Specifically, a specific multimodal large language model (MLLM) is trained over the training data by taking images as input and generating corresponding queries. The trained model is then used to generate pseudo-queries for images in the test set. These pseudo pairs—comprising test images and their generated queries—are then combined with the original training data during the training phase. However, *we argue that this augmentation strategy is potentially problematic. In particular, there is a risk of test data leakage, as some generated queries may closely resemble—or even exactly match—the original test queries, thereby compromising the validity of the evaluation.*

To mitigate the impact of the aforementioned issues, we utilize the datasets as described in Section 5.1. Table 1a presents the performance results without applying the augmentation strategy, while Table 1b reports the results with the augmentation strategy deployed. The symbol † indicates that the results are directly cited from the original paper, as we were unable to reproduce them.

From Table 1a, we can observe that ComGTIR-D shows obviously superior performance than the state-of-the-art method AVG. AVG is a typical general retrieval method that represents each candidate item by a unique identifier and expands the prefix by one token at a time in a greedy manner during the constrained beam search process. In contrast, ComGTIR-D mitigates the risk of getting stuck in a local optimum by incorporating an additional order-invariant identifier. This key difference is the primary reason for its superior performance.

Building upon ComGTIR-D, ComGTIR-DH further reranks the top-k results. Specifically, ComGTIR-DH$_{\text{clip}}$ and ComGTIR-DH$_{\text{itm}}$ denote the reranking variants using the CLIP and ITM components, respectively. As shown in Table 1a, the performance improvement introduced by the reranking strategy is substantial, with the ITM-based reranking proving particularly effective. This significant gain can be primarily attributed to the following reason. Representing images with discrete identifiers is a double-edged sword: while it enables efficient generative retrieval by substantially reducing the search space, it also introduces a degree of information loss due to discrete quantization. By applying a reranking step based on dense embedding vectors, the system can achieve finer-grained differentiation among the top-k candidate images, effectively compensating for the limitations introduced by discrete representations.

Regarding the questionable augmentation strategy, a joint comparison of Table 1a and Table 1b reveals that: the augmentation strategy contributes to performance improvements not only for GRACE and AVG, but also for our proposed methods. However, due to the potential risk of test data leakage, we argue that the results in Table 1a should be more reliable.

6.2 Inter-paradigm Comparison

To comprehensively evaluate the strengths and weaknesses of various text-image retrieval frameworks, Table 2 presents the results of the three representative groups of methods we focus on for text-image retrieval, namely one-tower (OT), two-tower (TT), and generative (G). The best result is highlighted in bold, while the second-best is underlined. The symbol "+" denotes a combination of different frameworks. *Size* refers to the number of parameters. *QPS* refers to the number of queries that a system can process per second, which reflects the system's throughput. Correspondingly *LS* indicates whether the corresponding method is suitable for large-scale retrieval. Furthermore, Figure 2 presents the inter-paradigm efficiency comparison, where the x-axis represents the size of the image set, and the y-axis indicates efficiency in terms of QPS.

A joint analysis of Table 2 and Figure 2 yields several key insights: (1) The two-tower framework encodes texts and images independently. A key advantage of this design is that the embedding vectors of all candidate images can be precomputed and stored in advance, enabling efficient retrieval at query time and significantly reducing computational overhead. This is why the QPS of CLIP and BLIP$_{\text{itc}}$ is substantially higher. However, this efficiency comes at the cost of reduced accuracy due to the absence of deep cross-modal interactions between text and image representations. (2) In contrast, the one-tower framework uses cross-attention between text and image tokens, enabling deeper understanding of local interactions. From Table 2, we can note that BLIP$_{\text{itm}}$ achieves the highest performance, since it is able to distinguish between candidates that are globally similar but differ in details. However, the one-tower framework requires computing the similarity between an input query and all candidate images at runtime, which substantially hinders its scalability and real-time applicability in large retrieval systems. This inefficiency is further evidenced by the sharp decline of the green curve in Figure 2. (3) Unlike the one-tower and two-tower frameworks, generative retrieval methods produce results through an autoregressive decoding process regardless of the dataset size. This explains why the QPS of AVG and ComGTIR-D remains constant in Figure 2. Due to the use of dual identifiers, ComGTIR-D exhibits slightly lower efficiency compared to AVG. A closer look

Method	R@1	R@5	R@10	rSum
Flickr30K				
AVG	40.8	75.1	84.2	200.1
ComGTIR-D	42.6	75.7	85.3	203.6
ComGTIR-DH$_{clip}$	68.4	86.3	90.7	245.4
ComGTIR-DH$_{itm}$	**84.4**	**93.5**	**94.4**	**272.3**
MS-COCO (5k)				
AVG	19.3	45.7	59.7	124.7
ComGTIR-D	20.6	46.7	61.0	128.3
ComGTIR-DH$_{clip}$	45.2	64.2	72.4	181.8
ComGTIR-DH$_{itm}$	**60.8**	**79.7**	**84.3**	**223.8**

(a) The performance without using the augmentation strategy.

Method	R@1	R@5	R@10	rSum
Flickr30K				
GRACE (Structured ID)†	37.4	59.5	66.2	163.1
AVG	49.4	81.6	89.0	220.0
ComGTIR-D	53.8	83.4	89.9	227.1
ComGTIR-DH$_{clip}$	68.7	87.2	91.9	247.8
ComGTIR-DH$_{itm}$	**85.6**	**94.9**	**95.8**	**276.3**
MS-COCO (5k)				
GRACE (Structured ID)†	16.7	39.2	50.3	106.2
AVG	23.2	51.1	64.1	138.4
ComGTIR-D	24.5	52.6	65.2	142.3
ComGTIR-DH$_{clip}$	45.7	64.7	72.8	183.2
ComGTIR-DH$_{itm}$	**61.8**	**80.9**	**85.6**	**228.3**

(b) The performance with the augmentation strategy applied.

Table 1: The intra-paradigm comparison of representative generative retrieval methods.

at Table 2 reveals that pure generative retrieval methods still fall significantly short of dense embedding-based retrieval in terms of accuracy. To address this gap, we propose combining the strengths of different retrieval methods. Specifically, the computationally expensive BLIP$_{itm}$ can be used to rerank a small set of top candidates, where fine-grained distinctions are quite critical. ComGTIR-D, on the other hand, excels at performing efficient initial retrieval. As shown in Table 2, ComGTIR-DH preserves the efficiency advantage of generative retrieval while substantially narrowing the accuracy gap to the strong dense baseline, BLIP. By incorporating a reranking stage based on dense embeddings, ComGTIR-DH strikes a more favorable balance between effectiveness and efficiency. Notably, ComGTIR-D also has the fewest parameters among all compared methods.

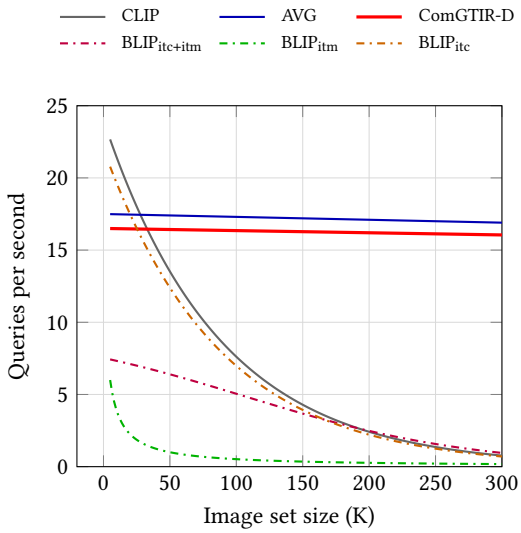

Figure 2: The inter-paradigm efficiency comparison.

	Size	Method	R@1	R@5	R@10	QPS	LS
Flickr30K							
TT	427M	CLIP	66.9	88.9	93.4	24.2/s	✓
TT	430M	BLIP$_{itc}$	82.5	96.4	98.3	21.7/s	✓
OT	430M	BLIP$_{itm}$	**87.4**	**97.6**	**99.0**	6.2/s	✗
TT+OT	430M	BLIP$_{itc+itm}$	87.3	97.4	98.3	8.4/s	✓
G	220M	ComGTIR-D	42.6	75.7	85.3	16.8/s	✓
G+TT	647M	ComGTIR-DH$_{clip}$	68.4	86.3	90.7	15.3/s	✓
G+OT	650M	ComGTIR-DH$_{itm}$	84.4	93.5	94.4	13.8/s	✓
MS-COCO (5k)							
	Size	Method	R@1	R@5	R@10	QPS	LS
TT	427M	CLIP	36.9	61.6	71.5	13.1/s	✓
TT	430M	BLIP$_{itc}$	58.7	82.3	89.2	11.4/s	✓
OT	430M	BLIP$_{itm}$	**65.1**	**86.2**	**91.8**	1.3/s	✗
TT+OT	430M	BLIP$_{itc+itm}$	64.9	85.2	89.2	6.8/s	✓
G	220M	ComGTIR-D	20.6	46.7	61.0	16.3/s	✓
G+TT	647M	ComGTIR-DH$_{clip}$	45.2	64.2	72.4	15.2/s	✓
G+OT	650M	ComGTIR-DH$_{itm}$	60.8	79.7	84.3	13.6/s	✓

Table 2: The inter-paradigm performance comparison.

6.3 Ablation Study

To better understand the impact of each component and setting choice of our proposed framework, such as the selection of base LLMs, the number of top candidates for reranking, and the beam size, we conduct a thorough ablation study in Table 3. Unless otherwise specified, our default setting is highlighted with a gray background.

Table 3a presents the impact of key components in ComGTIR-D, from which we derive the following observations. (1) Without the complementary order-invariant identifier, the performance drops significantly. This suggests that the order-invariant identifier

Method	R@1	R@5	R@10	rSum
ComGTIR-D	42.6	75.7	85.3	203.6
w/o max(·) operator in Eq-12	41.2	74.7	84.9	200.8
w/o order-invariant identifier	40.3	74.3	84.8	199.4
w/o $\mathcal{L}_{align}$ in Eq-3	38.1	70.7	79.9	188.7

(a) The impacts of key components in ComGTIR-D.

Beam Size	Flickr30K R@1	Flickr30K R@5	Flickr30K R@10	Flickr30K rSum	COCO R@1	COCO R@5	COCO R@10	COCO rSum
1	37.4	-	-	37.4	19.1	–	–	19.1
5	52.1	76.3	-	128.4	24.3	49.1	–	73.4
10	53.3	81.3	85.8	220.4	24.4	51.8	62.9	139.1
20	53.5	83.1	89.0	225.6	24.4	52.4	65.1	141.9
30	53.7	83.3	**89.9**	226.9	24.4	52.5	**65.2**	142.1
40	**53.8**	83.3	89.8	226.9	**24.5**	52.4	**65.2**	142.1
50	**53.8**	**83.4**	**89.9**	**227.1**	**24.5**	**52.6**	**65.2**	**142.3**

(b) The impact of beam size.

Model	# Params	R@1	R@5	R@10	rSum
T5-small	60M	36.2	68.5	80.1	184.8
T5-base	220M	42.6	75.7	85.3	203.6
T5-large	800M	44.1	76.2	85.3	205.6

(c) The effect of different base models based on Flickr30K.

$D \times L$	R@1	R@5	R@10	rSum
1024×4	42.6	75.7	85.3	203.6
2048×4	38.2	72.7	84.0	194.9
1024×8	44.7	77.2	85.1	207.0
2048×8	37.9	70.1	80.4	188.4

(d) The effect of various codebook settings based on Flickr30K.

Table 3: A comprehensive ablation study on the proposed framework.

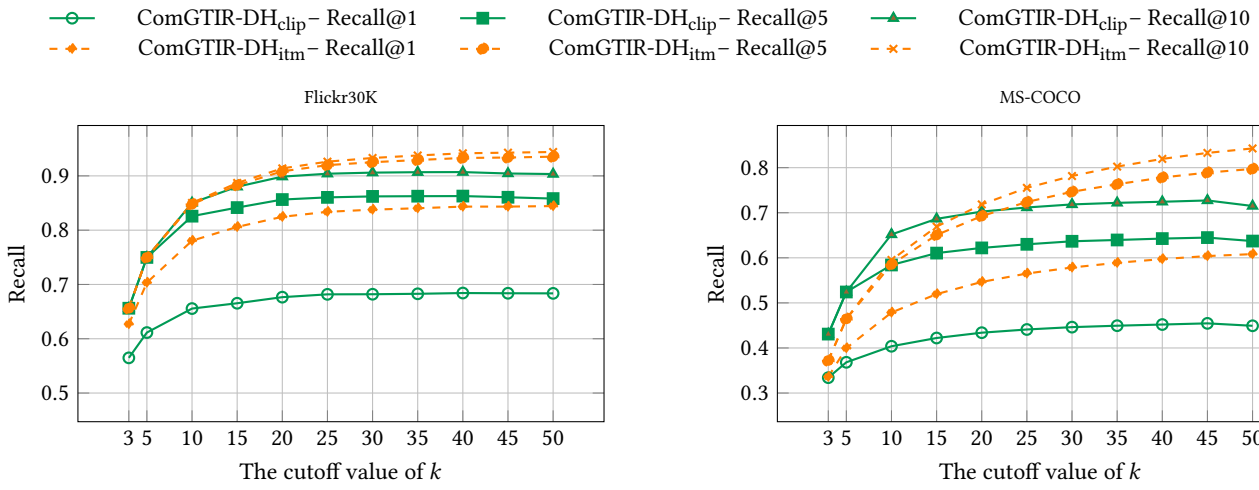

Figure 3: An in-depth analysis of the hybrid retrieval setting: impact of reranking size and method choice.

provides valuable guidance, helping the decoder avoid drifting toward locally plausible but globally irrelevant sequences. (2) Using the order-invariant identifier without the max(·) operator yields slight performance improvement, but remains inferior to the full ComGTIR-D setup. This indicates that the max(·) operator plays a crucial role in mitigating the risk of over-pruning due to locally optimal prefix scores. (3) Removing the cross-modal semantic alignment loss leads to a substantial performance drop, highlighting the importance of semantic alignment in getting intermediate discrete representations.

Table 3b shows the impact of beam size on two evaluated datasets. We can see that increasing the beam size leads to notable performance improvements, as a larger beam enables a broader search space. However, once the beam size approaches 50, the performance gains become marginal.

Table 3c shows the effect of different base models on Flickr30K. We observe that increasing the model size within the T5 series leads to a consistent performance improvement. This can be reasonably attributed to the fact that larger language models typically exhibit stronger semantic understanding capabilities. To ensure a consistent comparison with AVG, T5-base is used in our experiments. We leave the detailed exploration of T5-large as a future work.

Table 3d reports the impact of various codebook settings on Flickr30K. With a fixed length of four tokens, enlarging the codebook from 1024 to 2048 lowers R@1 from 42.6 % to 38.2 %, indicating that the additional visual granularity is offset by the harder optimisation of a twice-as-large softmax space. By setting D as 1024 and doubling the sequence to eight tokens yields a modest gain (R@1 = 44.7 %), but when both parameters are simultaneously increased (2048×8) performance drops sharply (R@1 = 37.9 %), suggesting

The ground-truth image	An input query	Step-1	Step-2	Step-3	Step-4	The ground-truth identifier	Hit
	A black and white dog is jumping over a hurdle	**c_58** 0.98	**c_232** 0.99	c_404 0.33	**c_32** 0.39	c_58 c_232 c_517 c_32	✗
	Fans of a football team look out onto the field	**c_185** 0.51	c_241 0.52	c_968 0.55	c_332 0.12	c_185 c_959 c_206 c_326	✗
	A lady with dark hair is playing a harp	**c_437** 0.99	**c_329** 0.78	**c_990** 0.78	**c_388** 0.33	c_437 c_329 c_990 c_388	✓

Table 4: The case study of prefix-guided decoding. The bold tokens indicate matches with the ground-truth identifiers, while the red tokens denote the first deviation. The numbers below each token represent the softmax probabilities after applying image-level global guidance based on the order-invariant identifiers.

that an overly fine-grained representation becomes difficult for the model to learn. Because decoding latency grows linearly with L, the speed cost of the 1024×8 variant outweighs its small accuracy benefit in large-scale settings. We therefore adopt D=1024, L=4 as the default, as it achieves the best balance between retrieval accuracy and inference efficiency.

Figure 3 illustrates the impact of the number of top candidates used for reranking. On both datasets, increasing the number of reranked candidates leads to substantial performance gains. However, once the number exceeds 20, the improvements become marginal. Notably, ComGTIR-DH$_{\text{itm}}$ consistently outperforms its CLIP-based counterpart across all metric values and at all cut-off points.

6.4 Case study

To qualitatively analyze the effect of our guided decoding strategy, we conduct some case studies on three representative queries with varying performance outcomes. Each case presents the input query, ground-truth code sequence, and the predicted identifier sequence generated by our model using prefix-guided decoding. We also trace the top-3 token predictions at each decoding step to provide insight into the decision process.

Table 4 illustrates three typical outcomes of prefix-guided decoding based on dual identifiers. For the first query—*A black and white dog is jumping over a hurdle*—the decoder follows the global guidance closely, getting 3/4 codes correct; it drifts only at Step-3, where `c_404` (0.33) replaces the ground-truth `c_517`, yet confidence for the remaining tokens remains high (`c_58` 0.98, `c_232` 0.99). The second query—*Fans of a football team look out onto the field*—shows a failure case: early selections such as `c_241` (0.52) and `c_968` (0.55) steer the beam off track, while the correct code `c_959` is ignored, suggesting that visually crowded or ambiguous scenes may dilute the effect of global guidance. In contrast, the third query—*A lady with dark hair is playing a harp*—features distinct and unambiguous objects, so all ground-truth tokens are correctly predicted with a high certainty (`c_437` 0.99 → `c_388` 0.33), resulting in a perfect 4/4 match. These examples demonstrate that prefix-guided decoding is highly effective when the query and image are well-aligned, while semantically dense or ambiguous scenes remain challenging.

7 CONCLUSION AND FUTURE WORK

In this paper, we propose a combination-based framework for generative text-image retrieval. Recognizing that using a single identifier in generative retrieval potentially leads to suboptimal results due to the risk of getting stuck in a local optimum, we propose to represent an image by combining two different types of identifiers, which provides us the flexibility of integrating image-level relevance scores as guiding priors during the autoregressive generation process. In addition, we propose a hybrid retrieval strategy that first performs generative retrieval and then reranks the top-k candidates using dense embeddings for enhanced accuracy. The experimental results show that the proposed method using dual identifiers achieves superior performance than the state-of-the-art generative baseline. Furthermore, the introduced reranking strategy yields notable performance gains. As generative text-image retrieval remains an underexplored problem, we believe our method offers valuable insights for advancing this topic.

For future work, the following practical issues are worthy to be investigated. First, as shown in Tables 3c and 3d, the performance of our proposed framework can be further improved by using a larger base model or increasing the codebook size. Therefore, it is worthwhile to explore the upper bound of the framework's capabilities through additional experiments. Moreover, our current approach relies on two separate steps for obtaining image identifiers: generating intermediate discrete representations followed by target identifier generation. An important direction would be to investigate how these steps can be integrated and jointly optimized. Second, it is interesting to explore the effectiveness of the proposed framework on reversed image-text retrieval. Third, we also plan to study how to accelerate generative retrieval by trying non-autoregressive sequence-to-sequence structures.

References

[1] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. *Advances in Neural Information Processing Systems* 35 (2022), 31668–31683.

[2] Christopher J C Burges, Robert Ragno, and Quoc Viet Le. 2006. Learning to Rank with Nonsmooth Cost Functions. In *Proceedings of NeurIPS*. 193–200.

[3] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text Retrieval: A Survey on Recent Research and Development. In *Proceedings of*

the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22. International Joint Conferences on Artificial Intelligence Organization, 5410–5417. Survey Track.

[4] Jiacheng Chen, Hexiang Hu, Hao Wu, Yuning Jiang, and Changhu Wang. 2021. Learning the best pooling strategy for visual semantic embedding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15789–15798.

[5] Jiangui Chen, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Wei Chen, Yixing Fan, and Xueqi Cheng. 2023. Continual learning for generative retrieval over dynamic corpora. In *Proceedings of the 32nd ACM international conference on information and knowledge management*. 306–315.

[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd SIGKDD Conference*. 785–794.

[7] Xiaoyang Chen, Yanjiang Liu, Ben He, Le Sun, and Yingfei Sun. 2023. Understanding Differential Search Index for Text Retrieval. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics, Toronto, Canada, 10701–10717.

[8] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. 2021. Training vision transformers for image retrieval. *arXiv preprint arXiv:2102.05644* (2021).

[9] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. 2022. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*. Springer, 89–106.

[10] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*. Springer, 529–545.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[12] Mariya Hendriksen, Svitlana Vakulenko, Ernst Kuiper, and Maarten de Rijke. 2023. Scene-centric vs. object-centric image-text cross-modal retrieval: a reproducibility study. In *European Conference on Information Retrieval*. Springer, 68–85.

[13] Mariya Hendriksen, Shuo Zhang, Ridho Reinanda, Mohamed Yahya, Edgar Meij, and Maarten de Rijke. 2025. Benchmark Granularity and Model Robustness for Image-Text Retrieval: A Reproducibility Study. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 3183–3193.

[14] Jiho Jang, Chaerin Kong, Donghyeon Jeon, Seonhoon Kim, and Nojun Kwak. 2023. Unifying vision-language representation space with single-tower transformer. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 980–988.

[15] HeeJae Jun, Byungsoo Ko, Youngjoon Kim, Insik Kim, and Jongtack Kim. 2019. Combination of multiple global descriptors for image retrieval. *arXiv preprint arXiv:1903.10663* (2019).

[16] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *NIPS*. 3149–3157.

[17] Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT. In *Proceedings of SIGIR*. 39–48.

[18] Chaeeun Kim, Soyoung Yoon, Hyunji Lee, Joel Jang, Sohee Yang, and Minjoon Seo. 2024. Exploring the Practicality of Generative Retrieval on Dynamic Corpora. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 13616–13633.

[19] Dongwon Kim, Namyup Kim, and Suha Kwak. 2023. Improving cross-modal retrieval with set of diverse embeddings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 23422–23431.

[20] Sungyeon Kim, Xinliang Zhu, Xiaofan Lin, Muhammet Bastan, Douglas Gray, and Suha Kwak. 2025. GENIUS: A generative framework for universal multimodal search. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 19659–19669.

[21] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[23] Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. 2022. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11523–11532.

[24] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*. 201–216.

[25] Haoxuan Li, Yi Bin, Yunshan Ma, Guoqing Wang, Yang Yang, See-Kiong Ng, and Tat-Seng Chua. 2025. SemCORE: A Semantic-Enhanced Generative Cross-Modal Retrieval Framework with MLLMs. *arXiv preprint arXiv:2504.13172* (2025).

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[27] 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 (2021), 9694–9705.

[28] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4654–4662.

[29] Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2025. From matching to generation: A survey on generative information retrieval. *ACM Transactions on Information Systems* 43, 3 (2025), 1–62.

[30] Yongqi Li, Hongru Cai, Wenjie Wang, Leigang Qu, Yinwei Wei, Wenjie Li, Liqiang Nie, and Tat-Seng Chua. 2025. Revolutionizing text-to-image retrieval as autoregressive token-to-voken generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 813–822.

[31] Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024. Generative Cross-Modal Retrieval: Memorizing Images in Multimodal Language Models for Retrieval and Beyond. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 11851–11861.

[32] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023. Multiview Identifiers Enhanced Generative Retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 6636–6648.

[33] Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2024. Learning to rank in generative retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 8716–8723.

[34] Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2025. Mm-embed: Universal multimodal retrieval with multimodal llms. *The Thirteenth International Conference on Learning Representations* (2025).

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*. Springer, 740–755.

[36] Chunxiao Liu, Zhendong Mao, An-An Liu, Tianzhu Zhang, Bin Wang, and Yongdong Zhang. 2019. Focus your attention: A bidirectional focal attention network for image-text matching. In *Proceedings of the 27th ACM international conference on multimedia*. 3–11.

[37] Sanket Vaibhav Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Q Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. 2022. DSI++: Updating transformer memory with new documents. *arXiv preprint arXiv:2212.09744* (2022).

[38] Kidist Amde Mekonnen, Yubao Tang, and Maarten de Rijke. 2025. Lightweight and Direct Document Relevance Optimization for Generative Information Retrieval. *Proceedings of SIGIR* (2025).

[39] Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. In *Acm sigir forum*, Vol. 55. ACM New York, NY, USA, 1–27.

[40] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2022. Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision*. Springer, 529–544.

[41] Naoki Muramoto and Hai-Tao Yu. 2020. Deep Metric Learning Based on Rank-sensitive Optimization of Top-k Precision. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2161–2164.

[42] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 299–307.

[43] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv:1901.04085v4* (2019).

[44] Ronak Pradeep, Kai Hui, Jai Gupta, Adam Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Tran. 2023. How Does Generative Retrieval Scale to Millions of Passages?. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 1305–1321.

[45] Shanbao Qiao, Xuebing Liu, and Seung-Hoon Na. 2023. DiffusionRet: Diffusion-Enhanced Generative Retriever using Constrained Decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. 9515–9529.

[46] Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. 2025. Tiger: Unifying text-to-image generation and retrieval with large multimodal models. In *The Thirteenth International Conference on Learning Representations*.

[47] 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*. PmLR, 8748–8763.

[48] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* 21, 140 (2020), 1–67.

[49] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[50] Stephen E Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at TREC-3. In *Proceedings of TREC*.

[51] Zihua Si, Zhongxiang Sun, Jiale Chen, Guozhang Chen, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, Jun Xu, and Kun Gai. 2024. Generative Retrieval with Semantic Tree-Structured Identifiers and Contrastive Learning. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 154–163.

[52] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15638–15650.

[53] Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* 28, 1 (1972), 11–21.

[54] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2023. Learning to tokenize for generative retrieval. *Advances in Neural Information Processing Systems* 36 (2023), 46345–46361.

[55] Yubao Tang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Shihao Liu, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. 2025. Generative Retrieval for Book Search. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1 (KDD '25)*. Association for Computing Machinery, 2606–2617.

[56] Yubao Tang, Ruqing Zhang, Jiafeng Guo, Maarten Rijke, Wei Chen, and Xueqi Cheng. 2024. Generative retrieval meets multi-graded relevance. *Advances in Neural Information Processing Systems* 37 (2024), 72790–72817.

[57] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems* 35 (2022), 21831–21843.

[58] Marvin Teichmann, Andre Araujo, Menglong Zhu, and Jack Sim. 2019. Detect-to-retrieve: Efficient regional aggregation for image search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5109–5118.

[59] Michael Tschannen, Basil Mustafa, and Neil Houlsby. 2023. Clippo: Image-and-language understanding from pixels only. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11006–11017.

[60] Aaron Van Den Oord, Oriol Vinyals, et al. 2017. Neural discrete representation learning. *Advances in neural information processing systems* 30 (2017).

[61] Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. 2022. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems* 35 (2022), 25600–25614.

[62] Zihan Wang, Yujia Zhou, Yiteng Tu, and Zhicheng Dou. 2023. Novo: Learnable and interpretable document identifiers for model-based ir. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2656–2665.

[63] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *European Conference on Computer Vision*. Springer, 387–404.

[64] Christian Wengert, Matthijs Douze, and Hervé Jégou. 2011. Bag-of-colors for improved image search. In *Proceedings of the 19th ACM international conference on Multimedia*. 1437–1440.

[65] Shiguang Wu, Zhaochun Ren, Xin Xin, Jiyuan Yang, Mengqi Zhang, Zhumin Chen, Maarten de Rijke, and Pengjie Ren. 2025. Constrained Auto-Regressive Decoding Constrains Generative Retrieval. *Proceedings of SIGIR* (2025).

[66] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems* 32 (2019).

[67] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783* (2021).

[68] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics* 2 (2014), 67–78.

[69] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research* (2022). https://openreview.net/forum?id=Ee277P3AYC

[70] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* 2, 3 (2022), 5.

[71] Hansi Zeng, Chen Luo, Bowen Jin, Sheikh Muhammad Sarwar, Tianxin Wei, and Hamed Zamani. 2024. Scalable and effective generative information retrieval. In *Proceedings of the ACM Web Conference 2024*. 1441–1452.

[72] Hansi Zeng, Chen Luo, and Hamed Zamani. 2024. Planning ahead in generative retrieval: Guiding autoregressive generation through simultaneous decoding. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 469–480.

[73] Yidan Zhang, Ting Zhang, Dong Chen, Yujing Wang, Qi Chen, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, Fan Yang, et al. 2024. Irgen: Generative modeling for image retrieval. In *European Conference on Computer Vision*. Springer, 21–41.

[74] Zhen Zhang, Xinyu Ma, Weiwei Sun, Pengjie Ren, Zhumin Chen, Shuaiqiang Wang, Dawei Yin, Maarten de Rijke, and Zhaochun Ren. 2025. Replication and Exploration of Generative Retrieval over Dynamic Corpora. *Proceedings of SIGIR* (2025).

[75] Zhedong Zheng, Liang Zheng, Michael Garrett, Yi Yang, Mingliang Xu, and Yi-Dong Shen. 2020. Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16, 2 (2020), 1–23.

[76] Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. 2022. Bridging the gap between indexing and retrieval for differentiable search index with query generation. *arXiv preprint arXiv:2206.10128* (2022).