GENIUS: A Generative Framework for Universal Multimodal Search

Sungyeon Kim[1,2] Xinliang Zhu[1] Xiaofan Lin[1] Muhammet Bastan[1] Douglas Gray[1] Suha Kwak[2]

[1] Amazon [2] POSTECH

{sungyeon.kim,suha.kwak}@postech.ac.kr {xlzhu, xiaofanl, mbastan, douggray}@amazon.com

Abstract

Generative retrieval is an emerging approach in information retrieval that generates identifiers (IDs) of target data based on a query, providing an efficient alternative to traditional embedding-based retrieval methods. However, existing models are task-specific and fall short of embedding-based retrieval in performance. This paper proposes GENIUS, a universal generative retrieval framework supporting diverse tasks across multiple modalities and domains. At its core, GENIUS introduces modality-decoupled semantic quantization, transforming multimodal data into discrete IDs encoding both modality and semantics. Moreover, to enhance generalization, we propose a query augmentation that interpolates between a query and its target, allowing GENIUS to adapt to varied query forms. Evaluated on the M-BEIR benchmark, it surpasses prior generative methods by a clear margin. Unlike embedding-based retrieval, GENIUS consistently maintains high retrieval speed across database size, with competitive performance across multiple benchmarks. With additional reranking, GENIUS often achieves results close to those of embedding-based methods while preserving efficiency.

1. Introduction

Information Retrieval (IR) is a fundamental task of finding relevant information from a large database [33, 43]. With the rapid growth of data, efficient and accurate IR is more essential than ever. Conventional IR approaches commonly follow the embed-and-retrieve paradigm, known as embedding-based retrieval (Fig. 1(a)). They embed the query and the database into a high-dimensional embedding space, which is learned by metric learning [20, 35, 44, 45, 49, 56], and then find the nearest neighbors of the query. As the database expands, however, a scalability issue arises due to the rapidly increasing cost of index building, maintenance, and nearest neighbor search, even with approximate nearest neighbor search like HNSW [32] and Faiss [9].

Recently, generative retrieval has emerged as a promising alternative. Inspired by Differentiable Search Index [48] and SPLADE [11], this approach generates identifiers (IDs) of target data directly from a query, bypassing the nearest neighbor search. However, existing methods in this line of research have limited capability due to their task-specific designs. Most of them are dedicated to text retrieval [48, 50], and only a few recent works address images [58] and cross-modal retrieval [25] (Fig. 1(b)). Hence, these methods fail to meet the diverse, multimodal demands of users in real-world applications. Moreover, existing generative methods underperform in cross-modal retrieval compared to embedding-based retrieval methods [25, 58].

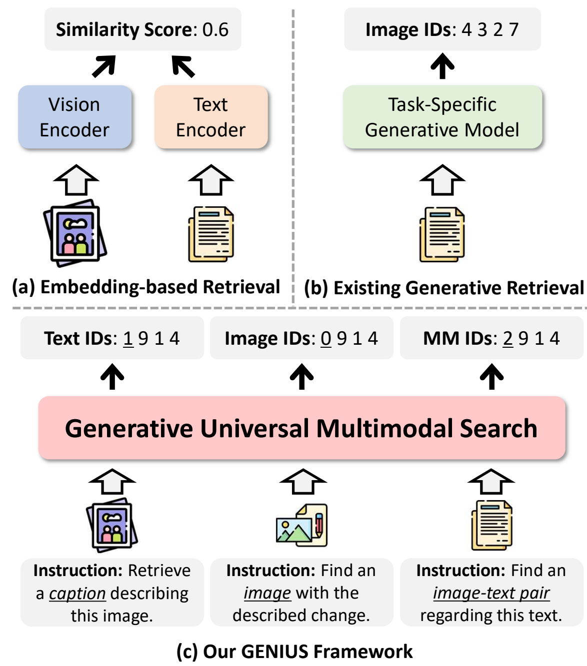

Figure 1. Illustrations of three Information Retrieval paradigms. (a) Embedding-based retrieval, where queries and candidates are embedded, and similarity is measured. (b) Existing generative retrieval generates task-specific identifiers. (c) The GENIUS framework generates identifiers across modalities based on queries and instructions, where the first-level code indicates modality.

In this paper, we propose **GENeratIve Universal multimodal Search** (GENIUS), the first generative retrieval framework that handles diverse retrieval tasks across multiple modalities. In GENIUS, each task is defined as finding data of a specified type, based on a multimodal query with an instruction that clarifies the user's intention. Our framework uses the instructions to retrieve data of the appropriate format and domain among diverse data within the database. Unlike prior generative methods restricted to specific modalities or tasks, GENIUS generates IDs of relevant data across heterogeneous modalities, effectively addressing a wide range of retrieval scenarios. GENIUS consists of a multimodal encoder that processes the query and instruction, coupled with a decoder that generates target IDs based on this input, as illustrated in Fig. 2.

A key contribution of GENIUS is **modality-decoupled semantic quantization** to assign a target ID to multimodal data. It transforms multimodal data into compact, layered representations capturing both semantic content and modality. Fig. 1(c) illustrates this concept, with each target ID represented as a sequence of discrete codes comprising two components. The first code of the target ID indicates the data modality (*e.g.*, 0 for images, 1 for text, and 2 for image-text pairs). This is achieved by training a quantization model with instructions that specify the modality of the target, allowing GENIUS to separate different modalities of the target. The subsequent codes capture the semantic content of the data while ensuring compatibility across modalities. For example, when image and text have similar contents, their IDs should be similar, particularly in their leading codes (except the first one which is kept for modality encoding), regardless of their modality. This is achieved through contrastive learning combined with residual quantization, which clusters semantically related items, enabling a nuanced representation from coarse to fine granularity.

Next, we train the decoder to generate target IDs from a given query. While these compact IDs are effective, they inherently contain less information than dense embeddings. As a result, the model may struggle to generalize to new or varied queries, especially with limited query-target pairs. To address this, we introduce **Query Augmentation** strategy. This strategy generates augmented queries by linearly interpolating between the embeddings of a query and its corresponding target. Including these augmented queries in training enriches the data with diverse query examples that retain the same semantics. This augmentation allows the decoder to learn a more generalized mapping from queries to target IDs, making it robust to variations in query formulations at test time.

We train and evaluate GENIUS on a large-scale multimodal benchmark, M-BEIR [52], which includes instructions for multimodal retrieval tasks. GENIUS outperforms the best generative retrieval method by 28.6 points in Recall@5 on the COCO dataset [26] for text-to-image retrieval. Unlike prior generative models, GENIUS supports a broader range of tasks and significantly narrows the performance gap to embedding-based retrieval methods across multiple tasks. It maintains a nearly constant retrieval speed across database sizes, and operates faster than previous generative methods. Moreover, by re-ranking predicted candidates based solely on their embeddings, GENIUS often achieves results close to those of embedding-based baselines in several tasks while preserving high efficiency. This combination of *versatility*, *performance*, and *efficiency* marks a big step forward for generative multimodal retrieval.

2. Related Work

2.1. Multimodal Information Retrieval

Multimodal Information Retrieval (IR) has advanced significantly, particularly in cross-modal tasks like text-to-image retrieval. Traditional methods are divided into two main approaches: multi-encoder and single-encoder with cross-attention. Multi-encoder models [14, 17, 19, 39, 57, 59] efficiently map visual and textual features and other format features into a shared embedding space. Single-encoder models [21–23, 53] provide more detailed modality interactions but incur a higher computational cost. Recent advances in IR have introduced composed image retrieval (CIR) tasks, which integrate image and text inputs based on user feedback [2, 3, 42]. Fine-grained retrieval also requires models to handle complex multimodal queries, posing additional challenges [6, 31]. Moreover, benchmarks like WebQA [5] and frameworks such as UniIR [52] extend IR capabilities to retrieve diverse data types, supporting unified retrieval across multiple datasets for broader generalization. Most retrieval methods follow the embed-to-retrieve paradigm, while recent efforts [25, 58] have started to explore generative approaches for handling multi-modal tasks, which remain largely unexplored.

2.2. Generative Retrieval

Generative retrieval has recently emerged as an innovative paradigm, primarily targeting text-based document retrieval. Early works explored generating concise identifiers (IDs), such as entity names or passage titles, to represent documents effectively [4, 8]. These approaches have evolved into more generalized methods, such as NCI [51] and DSI [48], which use hierarchical clustering of document embeddings and pretrained language models to assign document identifiers effectively. Recent studies have further refined these concepts [10, 34, 36, 47], with some proposing end-to-end methods to directly learn IDs [18, 46]. While text retrieval benefits from the inherent discreteness of language, extending generative retrieval to multiple modalities introduces challenges in addressing modality gaps. GRACE [25] is one of the few studies that has explored cross-modal generative retrieval by introducing se-

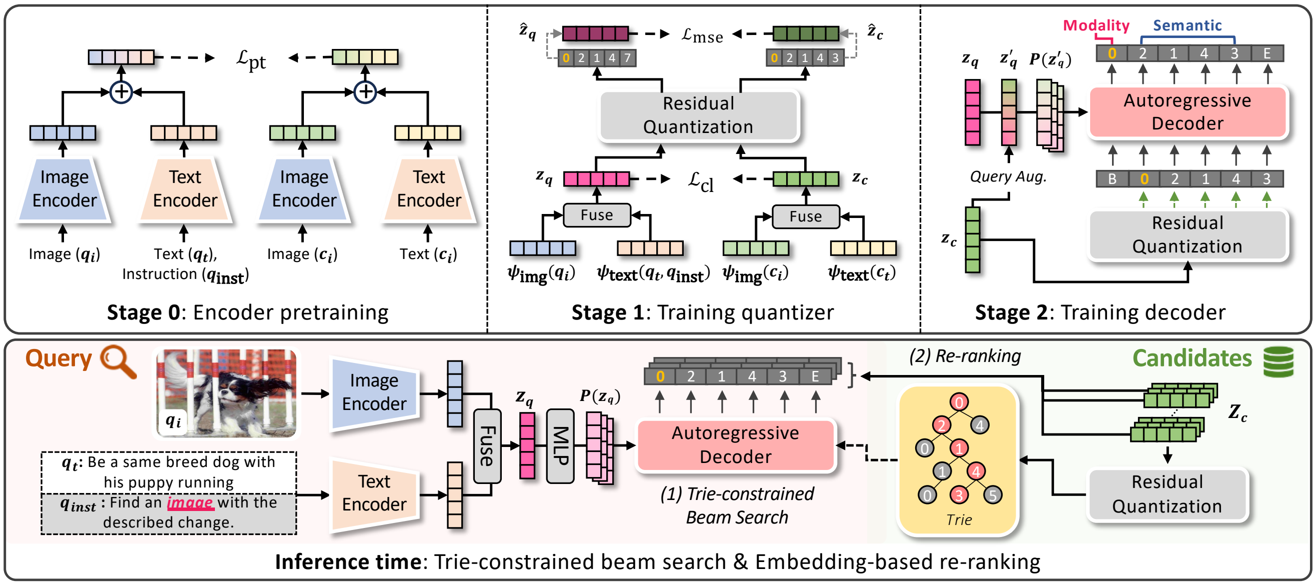

Figure 2. **Overview of the GENIUS framework.** GENIUS includes three components: image and text encoders, a modality-decoupled quantization module, and an autoregressive decoder. The framework follows three stages in training. First, the image-text encoders are pre-trained to enhance instruction comprehension and representation abilities. Next, residual quantization is trained to assign discrete IDs to candidate embeddings, where the first quantization level captures modality information and subsequent levels encode semantic details. Finally, the decoder learns to generate modality-decoupled semantic IDs. At inference, GENIUS generates candidate IDs from a query using Trie-constrained beam search, additionally followed by embedding-based re-ranking to further enhance retrieval accuracy.

mantic IDs for images, while IRGen [58] focuses solely on image-based retrieval and struggles with tasks beyond single-modality scenarios. These models are designed for a specific scenario and show significantly lower performance than embedding-based retrieval methods, highlighting their limitations in real-world applications. Our work addresses these limitations by introducing a universal framework that dynamically generates IDs across text and images, supporting a broader range of retrieval tasks.

3. Problem Formulation

Universal multimodal search [52] aims to enable users to query and retrieve targets across diverse tasks based on user instruction q_{inst}. In this setup, we define a query $\mathbf{q}$ as a combination of the query content and the instruction, represented as $(q_{\text{con}}, q_{\text{inst}})$, where q_{con} can take various forms, including an image q_i, text q_t, or an interleaved image-text pair (q_i, q_t). The target candidate $\mathbf{c}$ can be represented as an image c_i, text c_t, or an interleaved image-text pair (c_i, c_t).

We formalize universal generative multimodal search as the process of generating an ID T_c for the relevant target $\mathbf{c}$, conditioned on the query $\mathbf{q}$:

$$
\begin{aligned}
&T_c := (t_1^c, \ldots, t_M^c) \\
&\text{where } \; t_k^c = \underset{t \in \mathcal{T}}{\operatorname{argmax}} \left[\log p \left(t \mid \mathbf{q}, t_{<k}^c; \theta \right) \right],
\end{aligned} \tag{1}
$$

where θ denotes the parameters of both the encoder and decoder, $t_{<k}^c$ is the previously generated tokens, and $p(\cdot)$ is the probability distribution over the next token given the context. That is, the model generates the ID T_c by sequentially predicting tokens t_k^c that maximize the conditional probability. This generative approach eliminates the need for similarity computations, indexing, and ranking across the entire target dataset, making retrieval efficient and scalable.

4. Proposed Method

To address the universal generative retrieval problem, we propose GENeratIve Universal multimodal Search, dubbed GENIUS, which aims to generate target IDs across various modalities, guided by multimodal queries and instructions.

As shown in Fig. 2, GENIUS involves three distinct training stages. First, in Sec. 4.1, we describe multimodal encoder pretraining, which enables the encoder to effectively comprehend instructions and extract meaningful image-text features, aligning query intent with target semantics. Next, Sec. 4.2 introduces the modality-decoupled quantization module, which quantizes multimodal embeddings into discrete IDs, explicitly encoding modality and semantic information. These discrete IDs then serve as target outputs for decoder training. Finally, Sec. 4.3 presents the autoregressive decoder training process, enabling the decoder to generate modality-decoupled semantic IDs directly from the query. In Sec. 4.4, we detail the inference pipeline of GENIUS.

4.1. Encoder Pretraining

To handle diverse retrieval tasks, a model should understand the relations between queries and targets by comprehending both query content and instructions. We achieve this through encoder pretraining, which enables the multimodal encoder to understand query semantics and instructive information. For image and text encoders, we leverage CLIP [39]. Specifically, we use the text encoder ψ_{text} to process text-based query contents q_t and instructions q_{inst}, while the image encoder ψ_{image} is used for image inputs q_i.

To ensure strong alignment between queries and their corresponding positive targets, we employ contrastive learning. When both modalities are present in a query or target, we combine their features using simple element-wise addition [30, 52] to create a unified embedding: $\phi(\mathbf{q}) = \psi_{\text{image}}(q_i) + \psi_{\text{text}}(q_t, q_{\text{inst}}) \in \mathbb{R}^d$ for query, and $\phi(\mathbf{c}) = \psi_{\text{image}}(c_i) + \psi_{\text{text}}(c_t) \in \mathbb{R}^d$ for targets, where d is the embedding dimension. The contrastive loss between the query and target embeddings is defined as:

$$\mathcal{L}_{\text{pt}} = -\log \frac{\exp\left(\langle \phi(\mathbf{q}), \phi(c^+)\rangle / \tau\right)}{\sum_{c' \in \mathcal{C}} \exp\left(\langle \phi(\mathbf{q}), \phi(c')\rangle / \tau\right)}, \tag{2}$$

where $\phi(c^+)$ is the embedding of a target corresponding to the query $\mathbf{q}$, $\mathcal{C}$ is the set of all candidates, $\langle \cdot, \cdot \rangle$ denotes cosine similarity, and τ is a temperature parameter. This training follows the CLIP-based learning framework of UniIR [52]. For implementation simplicity, we directly utilize its pre-trained weights. After this phase, both the image and text encoders are frozen.

4.2. Modality-Decoupled Semantic Quantization

In generative retrieval, targets are represented as discrete IDs forming the output structure of the decoder model. Quantizing targets into these IDs is crucial, directly impacting retrieval performance. Unlike existing methods, GENIUS retrieves target data across modalities, and thus, it is essential to distinguish different modalities while accurately capturing semantic content.

To this end, we propose a quantization method that represents *modality* and *semantic* information separately. Our key idea is to provide an embedding space that captures both modality and semantic information using *contrastive learning with queries including instructions* and to systematically separate these features through *residual quantization* (RQ). Leveraging the unique property of residual quantization allows us to produce structured code sequences, where modality is explicitly encoded at the first level and semantic details are progressively refined in subsequent levels.

4.2.1. Fusion Module for Quantization Input

To facilitate effective quantization that captures both modality and semantics, we construct unified multimodal embeddings as inputs to the quantization. For this purpose, we introduce a lightweight, learnable module that combines image and text features into a unified representation. Inspired by previous work [3], the fusion module is defined as:

$$h(x, y) = \lambda \cdot x + (1 - \lambda) \cdot y + \texttt{MLP}([x; y]),$$

where $\texttt{MLP}([x; y])$ introduces additional bimodal information through a multi-layer perceptron (MLP) applied to the concatenation of x and y. The balance parameter λ is dynamically determined via another MLP with a sigmoid activation over the concatenated image-text features. The fused query embedding is computed as $\mathbf{z}_q = h\left(\psi_{\text{image}}(q_i), \psi_{\text{text}}(q_t, q_{\text{inst}})\right)$ and the fused target embedding as $\mathbf{z}_c = h\left(\psi_{\text{image}}(c_i), \psi_{\text{text}}(c_t)\right)$. The fusion module is optimized with the quantization module by the objectives of the quantization process.

4.2.2. Contrastive Learning with Instruction

We construct an embedding space that integrates both modality and semantic information to prepare input embeddings for the modality-decoupled quantization. Using queries including instructions that specify the desired modality of the target, we apply a contrastive loss to align between these queries and their corresponding targets. This loss encourages data with the same semantics and modality to close together in the embedding space while pushing apart data that differ in either aspect. The contrastive loss is defined as:

$$\mathcal{L}_{\text{cl}} = -\log \frac{\exp\left(\langle \mathbf{z}_q,\ \mathbf{z}_{c^+}\rangle / \tau\right)}{\sum_{c' \in \mathcal{C}} \exp\left(\langle \mathbf{z}_q,\ \mathbf{z}_{c'}\rangle / \tau\right)}, \tag{3}$$

where $\mathbf{z}_q$ and $\mathbf{z}_{c^+}$ is the query and the corresponding target embedding, $\mathcal{C}$ is the set of all candidate targets. Through this loss, clusters form in the embedding space, where modality-based groups naturally form due to the larger sample size within each modality, while semantically similar data are closely aligned within these clusters.

4.2.3. Residual Quantization

Residual quantization (RQ) is a recursive process that approximates an embedding by quantizing its residuals at each level. This process enables a progressive information decomposition, allowing distinct levels to capture modality-specific and semantic elements separately. The RQ process converts the embedding $\mathbf{z}$ into a sequence of discrete codes, represented as:

$$T := \mathcal{RQ}(\mathbf{z}) = (t_1, \ldots, t_M), \tag{4}$$

where M is the number of quantization levels. Starting with the initial residual vector $r_0 = \mathbf{z}$, we perform quantization recursively. At each step i, we find the nearest neighbor within the i-th codebook $E_i = \{\mathbf{e}_i^k \in \mathbb{R}^d \mid k = 1, \ldots, K_i\}$, where K_i is the size of the i-th codebook, selecting the clos-

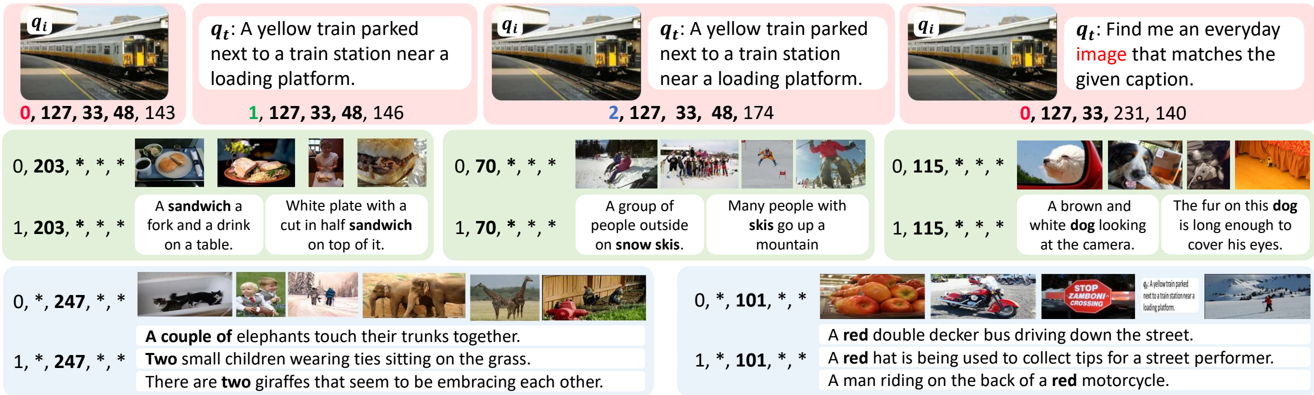

Figure 3. **Examples of modality-decoupled semantic quantization.** For simplicity, we use a quantization scheme with five levels of codes, where each code (except the first) has a value of up to 256. The first code (top) indicates modality: 0 for image, 1 for text, and 2 for image-text pairs. If an instruction is provided, this code adapts to the modality specified by the instruction. The second code (middle) represents primary objects or dominant semantics shared across modalities, while the third code (bottom) captures key attributes of the main object, such as "two" or "red", which are consistent across objects or data types. Beyond these levels, finer and additional information is incorporated to enrich the representation. This visualization is based on examples from the COCO dataset [26].

est code embedding $\mathbf{e}_i^{t_i}$ to the current residual vector:

$$t_i = \underset{k \in K_i}{\text{argmin}} \, ||\mathbf{r}_{i-1} - \mathbf{e}_i^k||^2, \tag{5}$$

and then update the residual for the next level:

$$\mathbf{r}_i = \mathbf{r}_{i-1} - \mathbf{e}_i^{t_i}, \tag{6}$$

The original embedding is approximated by summing the code embeddings up to level M, and we define this approximation as the quantized vector, $\hat{\mathbf{z}} = \sum_{i=1}^{M} \mathbf{e}_i^{t_i}$. Our key idea is to exploit the inherent property of residual quantization, where code embeddings at each level represent the residual information specific to that level. This property enables the progressive separation of information across levels. We utilize this property to distinguish modality and semantic information at each level. The first code in each ID explicitly represents modality, with a codebook of size $K_1 = 3$ to indicate images, text, and image-text pairs. Subsequent residuals exclude modality information, allowing the remaining levels to encode semantics solely in a coarse-to-fine manner.

4.2.4. Training Objectives

For training the codebooks and the fusion module h, we adopt three losses as follows. To ensure alignment between the assigned codes and the original residuals, we apply a residual quantization loss:

$$\mathcal{L}_{\text{rq}} = \sum_{i=1}^{M} ||\mathbf{r}_{i-1} - \text{sg}\left(\mathbf{e}_i^{t_i}\right)||^2, \tag{7}$$

where $\text{sg}(\cdot)$ denotes the stop-gradient operator, preventing gradients from directly updating codebook entries. Instead, they are updated via an exponential moving average (EMA) [41] over training steps to ensure stable updates. In addition, to further reinforce semantic similarity in the quantized space, we introduce a mean squared error (MSE) loss between the quantized vector of the query and target as $\mathcal{L}_{\text{mse}} = ||\hat{\mathbf{z}}_q - \hat{\mathbf{z}}_c||^2$, where $\hat{\mathbf{z}}_q$ and $\hat{\mathbf{z}}_c$ are the quantized query and target vectors, respectively. The training loss is a linear combination of the three aforementioned losses:

$$\mathcal{L}_{\text{combined}} = \mathcal{L}_{\text{cl}} + \beta \mathcal{L}_{\text{rq}} + \gamma \mathcal{L}_{\text{mse}}, \tag{8}$$

where β and γ are weighting parameters. Unlike prior methods focused on reconstructing original embeddings [25], our optimization aims to encode contrastive relations into the codebook. As a result, the quantizer produces the initial code representing modality, as shown in Fig. 3. The second code captures dominant semantics, while later codes add finer attributes, creating a structured representation that preserves rich, interpretable semantics and enhances retrieval performance across modalities.

4.3. Autoregressive Decoder for Retrieval

4.3.1. Decoder Training

The last step is to train an autoregressive decoder model that produces an ID of the target given a query. We adopt T5 decoder architecture [40], which generates the target ID autoregressively. To condition the decoder on the query embedding, we employ a lightweight network with an MLP that maps the query embedding $\mathbf{z}_q$ into N prefix embeddings, reshaping it as follows:

$$\mathbf{P}(\mathbf{z}_q) = \texttt{Reshape}(\texttt{MLP}(\mathbf{z}_q)) \in \mathbb{R}^{N \times d'}, \tag{9}$$

where d' represents the hidden dimension of the decoder. These prefix embeddings $\mathbf{P}(\mathbf{z}_q)$ are fed to the decoder through cross-attention, enabling it to generate target IDs based on the semantic information embedded in the query. The training loss for this generative model is a cross-entropy loss applied over the generated ID as follows:

$$\mathcal{L}_{\text{GR}}(\mathbf{P}(\mathbf{z}_q), T_c) = -\sum_{k=1}^{M} \log p\left(t_k^c \mid \mathbf{P}(\mathbf{z}_q), t_{<k}^c\right). \quad (10)$$

This encourages the model to generate a target code sequence conditioned on the query, which can be considered mapping a query embedding to the target ID.

However, due to the inherently limited representation capacity in these discrete IDs compared to embeddings, the model may struggle to generalize effectively, particularly in scenarios with few query-target pairs for training. In text document generative retrieval, this challenge arises but is often addressed by generating diverse queries from documents using methods like Doc2Query [37, 38]; however, such methods are not feasible in multimodal retrieval.

4.3.2. Query Augmentation via Interpolation

To address the above issue, we propose a *Query Augmentation* based on query-target interpolation. This technique enriches the training data by generating diverse augmented queries that remain semantically aligned with their target. The interpolated query embedding $\mathbf{z}_q'$ is computed as:

$$\mathbf{z}_q' = \mu \cdot \mathbf{z}_q + (1 - \mu) \cdot \mathbf{z}_c, \quad (11)$$

where μ is randomly sampled from a Beta distribution, Beta(α, α). The decoder is trained with the same cross-entropy loss with the augmented query, $\mathcal{L}_{\text{GR}}(\mathbf{P}(\mathbf{z}_q'), T_c)$. This strategy generates varied augmented queries, each maintaining relevance to the target, helping the decoder to learn a generalized mapping from query embeddings to target IDs. This makes the model more robust to variations in the query, improving its generalization.

4.4. Inference

Constrained beam search. GENIUS retrieves relevant targets for inference by generating IDs based on a given query. To produce a ranked list of candidates, we use beam search, which explores multiple ID sequences and ranks them by the sum of the log probabilities for each level in the sequence. However, to prevent the risk of generating invalid IDs, we use *constrained beam search* [8] with a Trie structure [12] that restricts the model to only valid prefixes matching actual test set IDs. The Trie is pre-constructed from all candidate IDs, allowing the decoder to ensure that generated IDs are valid. The time complexity for searching using Trie is $O(M)$, depending only on the length M of the IDs, which can significantly enhance scalability.

Embedding-based re-ranking. Despite this efficiency, generative retrieval with discrete IDs often lags behind embedding-based retrieval in performance due to the limitations of discrete representations, as observed in prior work [25]. To address this, we present a re-ranking method: after predicting B candidate IDs via beam search, we measure the similarity between the embeddings of these candidates and the query embedding. Since the number of comparisons is small, this method incurs negligible computational cost while greatly improving retrieval accuracy.

5. Experiments

To evaluate the effectiveness of our generative universal retrieval framework, we conducted comprehensive experiments across various retrieval tasks and domains, comparing our model against state-of-the-art baselines in both embedding-based and generative retrieval paradigms.

5.1. Dataset and metrics

Dataset. We use M-BEIR dataset [52], a combination of multiple datasets. It includes datasets like MS-COCO [26] for image-caption retrieval, Fashion200K [15] and FashionIQ [54] for fashion, VisualNews [27] for news images, and NIGHTS [13] for image similarity. Complex retrieval tasks are addressed by OVEN [16], EDIS [28], and CIRR [29], with InfoSeek [7] and WebQA [5] for VQA-based retrieval. These datasets cover 8 multimodal tasks and have a total of 5.6 million candidates.

Evaluation metrics. Following prior work [52], we report Recall@5 (R@5) as the main metric, using Recall@10 (R@10) for Fashion200K and FashionIQ.

5.2. Implementation Details

Network architectures. Following UniIR [52], we use the pre-trained CLIP ViT-L/14 model [39] as the vision and text encoder. For the decoder, we use T5-small [40], with hidden dimension $d' = 512$, which is initialized randomly.

Network optimization. Our model is optimized with AdamW, using a learning rate of 1×10^{-4} for both the RQ and decoder training. Residual quantization is trained for 20 epochs, while the decoder is trained for 30 epochs with cosine scheduling. We use a batch size of 256 for training.

Hyperparameters. The contrastive learning temperature τ in Eq. 2 is set to 0.01. Parameters β and γ are both fixed at 100 Eq. 8, and α parameter in Eq. 11 is set to 2. For the prefix embeddings in Eq. 9, we use a fixed length of 30.

Codebook configurations of RQ. Our default setting uses a codebook size of 4096 with 9 levels, except for the first codebook, which has a fixed size of 3. The codebook is initialized using k-means clustering on the first training batch.

Inference. As described in Section 4.4, we evaluate GENIUS in two ways: (i) constrained beam search and (ii) re-ranking beam search candidates based on their embeddings and that of query, both using a default beam size of 50 unless otherwise specified. The embedding-based methods are evaluated using nearest neighbor search by Faiss [9].

Method	$q_t \to c_i$			$q_t \to c_t$	$q_t \to (c_i, c_t)$		$q_i \to c_t$			$q_i \to c_i$	$(q_i, q_t) \to c_t$		$(q_i, q_t) \to c_i$		$(q_i, q_t) \to (c_i, c_t)$	
	COCO R@5	VN R@5	F200K R@10	WebQA R@5	EDIS R@5	WebQA R@5	COCO R@5	VN R@5	F200K R@10	NIGHTS R@5	OVEN R@5	InfoS R@5	FIQ R@10	CIR R@5	OVEN R@5	InfoS R@5
Embedding-based Retrieval																
CLIP-SF [52]	**81.1**	**42.6**	18.0	**84.7**	**59.4**	78.7	**92.3**	**43.1**	18.3	32.0	**45.5**	**27.9**	24.4	44.6	**67.6**	**48.9**
BLIP-FF [52]	79.7	23.4	**26.1**	80.0	50.9	**79.8**	89.9	22.8	**28.9**	**33.0**	41.0	22.4	**29.2**	**52.2**	55.8	33.0
Generative Retrieval																
IRGen [58]	50.7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
GRACE [25]	39.5	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
GENIUS	68.1	18.5	13.7	32.5	37.0	49.7	83.2	18.7	12.8	8.2	36.6	11.2	13.2	20.7	36.4	14.6
GENIUS$^{\mathcal{R}}$	**78.0**	**27.4**	**16.2**	**44.6**	**44.3**	**60.6**	**91.1**	**28.4**	**16.3**	**30.2**	**41.9**	**20.7**	**19.3**	**39.5**	**52.5**	**30.1**

Table 1. **Task-specific Information Retrieval.** Performance of methods on the M-BEIR dataset, retrieved from a task-specific pool. $\mathcal{R}$ denotes re-ranking using embedding vectors within the set of predicted candidates. Some datasets are denoted by abbreviations: VN–VisualNews, F200K–Fashion200K, InfoS–InfoSeek, and FIQ–FashionIQ.

5.3. Baselines

Training strategies. We evaluate models under two different training strategies: (i) *single-task fine-tuning*, where models are independently trained and evaluated on each specific task, and (ii) *unified instruction fine-tuning*, where models leverage multi-task learning with instructional guidance on M-BEIR [52], enabling a single model to handle retrieval tasks across multiple domains and modalities.
Embedding-based retrieval baselines. We compare GENIUS with fine-tuned variants of CLIP [39] and BLIP [22] proposed in UniIR [52]. These baselines employ two fusion strategies: score-level fusion (SF), which fuses information at the output embedding level, and feature-level fusion (FF), which uses transformers to achieve feature fusion.
Generative retrieval baselines. We benchmark against GRACE [25] and IRGen [58] which is originally for image-to-image retrieval, adapted for text-to-image retrieval by replacing image inputs with text, reported in [24]. Note that previous generative methods are designed for a single task.

5.4. Experimental Results

We evaluate multimodal retrieval models in three scenarios: (i) *task-specific information retrieval*, using original datasets to ensure a fair comparison with single-task methods; (ii) *universal information retrieval*, leveraging the full M-BEIR candidate pool of 5.6M items to assess models' capability in instruction-following and cross-modal retrieval tasks, a setting unsupported by existing generative approaches; and (iii) *text-to-image generative retrieval*, evaluated on standard generative retrieval benchmarks (Flickr30K and MS-COCO), with models trained and evaluated separately on each dataset.
Task-specific information retrieval. In Table 1, GENIUS is compared against embedding-based retrieval methods (CLIP-SF and BLIP-FF) and existing generative retrieval baselines (GRACE and IRGen) on various datasets from M-BEIR. Generative retrieval methods show significantly lower performance compared to embedding-based approaches, even on single-task retrieval. Notably, GENIUS significantly outperforms previous generative methods on COCO text-to-image retrieval by 28.6 points in R@5, substantially narrowing the gap with embedding-based methods. GENIUS demonstrates competitive performance across multiple datasets, with embedding-based re-ranking further enhancing its effectiveness, enabling it to surpass BLIP-FF in several tasks. However, GENIUS underperforms on knowledge-intensive retrieval tasks (*e.g.*, WebQA, InfoSeek) compared to embedding-based retrieval. This limitation is likely due to the inherent capacity of discrete IDs, which should be addressed in future research.

Task	Dataset	Embedding-based		Generative	
		CLIP$_{\text{SF}}$	BLIP$_{\text{FF}}$	GENIUS	GENIUS$^{\mathcal{R}}$
$q_t \to c_i$	VisualNews	**42.6**	23.0	18.5	**27.3**
	MSCOCO	**77.9**	75.6	55.1	**68.0**
	Fashion200K	17.8	**25.4**	13.7	**16.2**
$q_t \to c_t$	WebQA	**84.7**	79.5	31.1	**42.9**
$q_t \to (c_i, c_t)$	EDIS	**59.4**	50.3	36.6	**44.1**
	WebQA	78.8	**79.7**	49.0	**59.7**
$q_i \to c_t$	VisualNews	**42.8**	21.1	18.4	**26.8**
	MSCOCO	**92.3**	88.8	82.7	**90.6**
	Fashion200K	17.9	**27.6**	12.8	**16.2**
$q_i \to c_i$	NIGHTS	**33.4**	33.0	8.1	**30.2**
$(q_i, q_t) \to c_t$	OVEN	**39.2**	34.7	34.6	**38.0**
	InfoSeek	**24.0**	19.7	10.4	**18.0**
$(q_i, q_t) \to c_i$	FashionIQ	26.2	**28.5**	18.9	**19.2**
	CIRR	43.0	**51.4**	20.1	**38.3**
$(q_i, q_t) \to (c_i, c_t)$	OVEN	**60.2**	57.8	36.5	**48.6**
	InfoSeek	**44.6**	27.7	14.2	**28.6**
Average		**48.9**	45.5	28.8	**38.3**

Table 2. **Universal Information Retrieval.** Recall@5 results of methods except Fashion200K and FashionIQ, where Recall@10 is reported. Retrieval is performed from a global pool spanning diverse modalities. $\mathcal{R}$ denotes re-ranking using embedding vectors within the set of predicted candidates.

Universal information retrieval. Table 2 presents results for a range of retrieval tasks on entire candidates in M-BEIR dataset. Unlike prior settings, this universal

Method	Flickr30K			COCO		
	R@1	R@5	R@10	R@1	R@5	R@10
GRACE [25]	37.4	59.5	66.2	16.7	39.2	50.3
IRGen [58]	49.0	68.9	72.5	29.6	50.7	56.3
GENIUS	60.6	84.0	90.5	40.1	66.2	75.8
GENIUS$^{\mathcal{R}}$	**74.1**	**92.0**	**94.8**	**46.1**	**74.0**	**82.7**

Table 3. Text-to-image retrieval performance comparison on standard generative retrieval benchmark (Flickr30K and MS-COCO). $\mathcal{R}$ denotes re-ranking. Note that all models, including GENIUS, are trained and evaluated separately on each dataset.

	COCO		WebQA		CIRR
Method	T → I	I → T	T → T	T → (I,T)	(I,T) → I
GENIUS	**55.4**	**82.7**	**28.3**	**47.1**	20.5
w/o Modality-decoupled	20.2	73.2	25.9	34.3	18.3
w/o Query augmentation	47.8	67.7	19.6	38.8	11.7
w/o $\mathcal{L}_{\text{cl}}$ in Eq. 8	0.0	0.1	0.0	0.0	0.0
w/o $\mathcal{L}_{\text{mse}}$ in Eq. 8	45.5	83.1	27.1	35.2	**21.6**

Table 4. Ablation study for key components of GENIUS (universal information retrieval, R@5) with 30 beams. I and T denote image and text modalities, respectively, and (I,T) is image-text pair.

scenario requires models to identify target modalities precisely based solely on given instructions. GENIUS demonstrates competitive performance and versatility across multimodal tasks, though it typically achieves lower results than embedding-based retrieval baselines.

Text-to-image generative retrieval. Table 3 compares GENIUS against recent generative retrieval models on the Flickr30K [55] and MS-COCO [26] datasets. GENIUS significantly outperforms existing generative baselines such as GRACE and IRGen, showing substantial improvements across all metrics on both datasets. Further performance gains are achieved through embedding-based re-ranking, which yields state-of-the-art results in generative retrieval.

5.5. Analysis

Ablation study on key components. Table 4 presents an ablation study of key components under retrieval from a global pool. Removing modality-decoupling severely harms modality discrimination, notably in COCO text-to-image retrieval. Excluding query augmentation leads to decreased accuracy, highlighting its contribution to robustness. The contrastive loss ($\mathcal{L}_{\text{cl}}$) is crucial for aligning modality-decoupled representations; without it, query and target features become misaligned, leading to near-zero performance. Excluding MSE loss ($\mathcal{L}_{\text{mse}}$) weakens alignment in the codespace, reducing performance in certain datasets.

Analysis on efficiency. We compare retrieval efficiency between embedding-based (CLIP) and generative methods (GRACE, GENIUS) by measuring queries per second, as shown in Fig. 4. For a fair comparison with GRACE, we use text queries with image candidates. As the candidate dataset size increases, the efficiency of CLIP declines due to the growing cost of the nearest neighbor search, while generative methods remain nearly constant. GENIUS is a lightweight equipped with a T5-small decoder and CLIP encoder, and thus achieves roughly 4 times higher efficiency than GRACE with Flamingo-3B model [1]. The efficiency advantage becomes more significant as the dataset grows, maintaining high retrieval speed at scale without the expensive index building typical in embedding-based methods.

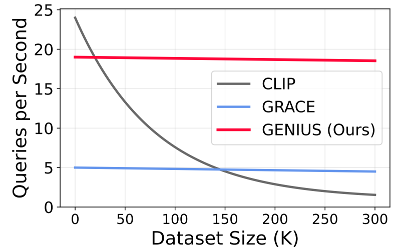

Figure 4. Efficiency in processed queries per second across varying dataset sizes, measured with a single RTX3090 GPU.

	COCO		WebQA		CIRR
$K \times M$	T → I	I → T	T → T	T → (I,T)	(I,T) → I
<u>4096</u>× <u>9</u> (Default)	65.3	**83.4**	**28.8**	**47.4**	21.0
<u>4096</u>× 7	65.2	82.9	25.3	40.8	23.1
<u>4096</u>× 5	62.4	81.5	17.3	33.1	20.4
1024× <u>9</u>	**66.4**	82.0	24.7	39.4	**24.5**
256× <u>9</u>	61.2	76.6	18.3	33.5	18.3
1024× 7	64.3	82.2	24.6	42.7	16.4
256× 5	53.4	72.4	9.7	22.8	13.0

Table 5. Ablation over codebook size K (except for the first level) and code level M (task-specific information retrieval, R@5) with 30 beams. The default codebook size and level are underlined.

Codebook configuration. Table 5 shows that larger codebook sizes and higher levels generally increase expressive power, and thus improve performance, especially in knowledge-intensive tasks such as WebQA. However, excessively large codebooks can disperse clusters, weakening representations in some datasets. This highlights the need to balance codebook size according to dataset characteristics.

6. Conclusion

We have introduced GENIUS, a universal generative retrieval framework that addresses the limitations of existing generative models by handling diverse tasks across modalities. Leveraging a novel modality-decoupled quantization technique for ID generation, GENIUS ensures consistent semantic information across modalities. Our query augmentation enhances generalization through diverse query-target mappings. Experiments show that GENIUS outperforms prior generative methods and narrows the performance gap with embedding-based methods across benchmarks. Moreover, GENIUS sustains high retrieval speed, laying the groundwork for scalable multimodal search.

Acknowledgement. Part of this work was done while Sungyeon Kim was an intern at Amazon. Sungyeon Kim and Suha Kwak were supported by NRF grants (RS-2021-NR059830–30%, RS-2022-II220290–30%, RS-2022-II220926–30%) and IITP grants (RS-2019-II191906–10%, AI Graduate School - POSTECH) funded by Ministry of Science and ICT, Korea.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Proc. Neural Information Processing Systems (NeurIPS)*, 2022. 8

[2] Yang Bai, Xinxing Xu, Yong Liu, Salman Khan, Fahad Khan, Wangmeng Zuo, Rick Siow Mong Goh, and Chun-Mei Feng. Sentence-level prompts benefit composed image retrieval. In *Proc. International Conference on Learning Representations (ICLR)*, 2023. 2

[3] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 2, 4

[4] Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive search engines: Generating substrings as document identifiers. *Proc. Neural Information Processing Systems (NeurIPS)*, 2022. 2

[5] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 6

[6] Soravit Changpinyo, Jordi Pont-Tuset, Vittorio Ferrari, and Radu Soricut. Telling the what while pointing to the where: Multimodal queries for image retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[7] Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. 6

[8] N De Cao, G Izacard, S Riedel, and F Petroni. Autoregressive entity retrieval. In *ICLR 2021-9th International Conference on Learning Representations*. ICLR, 2020. 2, 6

[9] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024. 1, 6

[10] Xin Du, Lixin Xiu, and Kumiko Tanaka-Ishii. Bottleneck-minimal indexing for generative document retrieval. In *Forty-first International Conference on Machine Learning*, 2024. 2

[11] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021. 1

[12] Edward Fredkin. Trie memory. *Communications of the ACM*, 1960. 6

[13] Stephanie Fu, Netanel Yakir Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 6

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[15] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. Automatic spatially-aware fashion concept discovery. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 6

[16] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[17] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proc. International Conference on Machine Learning (ICML)*, 2021. 2

[18] Bowen Jin, Hansi Zeng, Guoyin Wang, Xiusi Chen, Tianxin Wei, Ruirui Li, Zhengyang Wang, Zheng Li, Yang Li, Hanqing Lu, Suhang Wang, Jiawei Han, and Xianfeng Tang. Language models as semantic indexers. In *Forty-first International Conference on Machine Learning*, 2024. 2

[19] Dongwon Kim, Namyup Kim, and Suha Kwak. Improving cross-modal retrieval with set of diverse embeddings. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[20] Sungyeon Kim, Dongwon Kim, Minsu Cho, and Suha Kwak. Proxy anchor loss for deep metric learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[21] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proc. AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 2

[22] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 7

[23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proc. International Conference on Machine Learning (ICML)*, 2023. 2

[24] Yongqi Li, Hongru Cai, Wenjie Wang, Leigang Qu, Yinwei Wei, Wenjie Li, Liqiang Nie, and Tat-Seng Chua. Revolutionizing text-to-image retrieval as autoregressive token-to-voken generation. *arXiv preprint arXiv:2407.17274*, 2024. 7

[25] Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. Generative cross-modal retrieval:

Memorizing images in multimodal language models for retrieval and beyond. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. 1, 2, 5, 6, 7, 8

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: common objects in context. In *Proc. European Conference on Computer Vision (ECCV)*, 2014. 2, 5, 6, 8

[27] Fuwen Liu, Xiaojie Wang, Jianping Shi, Alan L Huang, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, and Alan Yuille. Visual news: Benchmark and challenges in news image captioning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010. 6

[28] Siqi Liu, Weixi Feng, Tsu-jui Fu, Wenhu Chen, and William Yang Wang. Edis: Entity-driven image search over multimodal web content. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4877–4894, 2023. 6

[29] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pretrained vision-and-language models. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2021. 6

[30] Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 4

[31] Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. End-to-end knowledge retrieval with multimodal queries. In *61st Annual Meeting of the Association for Computational Linguistics, ACL 2023*, pages 8573–8589. Association for Computational Linguistics (ACL), 2023. 2

[32] Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2018. 1

[33] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, USA, 2008. 1

[34] Sanket Vaibhav Mehta, Jai Gupta, Yi Tay, Mostafa Dehghani, Vinh Q Tran, Jinfeng Rao, Marc Najork, Emma Strubell, and Donald Metzler. Dsi++: Updating transformer memory with new documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8198–8213, 2023. 2

[35] Yair Movshovitz-Attias, Alexander Toshev, Thomas K Leung, Sergey Ioffe, and Saurabh Singh. No fuss distance metric learning using proxies. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017. 1

[36] Thong Nguyen and Andrew Yates. Generative retrieval as dense retrieval. *arXiv preprint arXiv:2306.11397*, 2023. 2

[37] Rodrigo Nogueira and Jimmy Lin. From doc2query to docttttquery. *Online preprint*, 2019. 6

[38] Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*, 2019. 6

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning (ICML)*, 2021. 2, 4, 6, 7

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research (JMLR)*, 2020. 5, 6

[41] Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2019. 5

[42] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2word: Mapping pictures to words for zero-shot composed image retrieval. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[43] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 2001. 1

[44] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2016. 1

[45] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[46] Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. Learning to tokenize for generative retrieval. *Proc. Neural Information Processing Systems (NeurIPS)*, 2024. 2

[47] Yubao Tang, Ruqing Zhang, Jiafeng Guo, Jiangui Chen, Zuowei Zhu, Shuaiqiang Wang, Dawei Yin, and Xueqi Cheng. Semantic-enhanced differentiable search index inspired by learning strategies. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2023. 2

[48] Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, et al. Transformer memory as a differentiable search index. In *Proc. Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[49] Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R Scott. Multi-similarity loss with general pair weighting for deep metric learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[50] Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. A neural corpus indexer for document retrieval. *Proc. Neural Information Processing Systems (NeurIPS)*, 2022. 1

[51] Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, et al. A neural corpus indexer for document retrieval. *Proc. Neural Information Processing Systems (NeurIPS)*, 2022. 2

[52] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 2, 3, 4, 6, 7

[53] Jiwei Wei, Xing Xu, Yang Yang, Yanli Ji, Zheng Wang, and Heng Tao Shen. Universal weighting metric learning for cross-modal matching. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[54] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 6

[55] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2014. 8

[56] Baosheng Yu and Dacheng Tao. Deep metric learning with tuplet margin loss. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2019. 1

[57] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2023. 2

[58] Yidan Zhang, Ting Zhang, Dong Chen, Yujing Wang, Qi Chen, Xing Xie, Hao Sun, Weiwei Deng, Qi Zhang, Fan Yang, et al. Irgen: Generative modeling for image retrieval. In *Proc. European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3, 7, 8

[59] Xinliang Zhu, Sheng-Wei Huang, Han Ding, Jinyu Yang, Kelvin Chen, Tao Zhou, Tal Neiman, Ouye Xie, Son Tran, Benjamin Yao, et al. Bringing multimodality to amazon visual search system. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024. 2