Generative Text-to-Image Retrieval via Hierarchical Identifiers and Semantic Internalization

Jie Huang[1,2,3], **Junjie Wang**[1,2,3*], **Xin Liao**[4], **Ziyou Jiang**[1,2,3], **Wenshuo Wang**[1,2,3], **Shoubin Li**[1,2,3] and **Qing Wang**[1,2,3*]
[1]State Key Laboratory of Complex System Modeling and Simulation Technology, Beijing, China [2]Science and Technology on Integrated Information System Laboratory Institute of Software Chinese Academy of Sciences, Beijing, China
[3]University of Chinese Academy of Sciences [4]Hunan University
{huangjie, junjie, ziyou2019, wangwenshuo2024, shoubin, wq}@iscas.ac.cn, xinliao@hnu.edu.cn

Abstract

Generative Retrieval (GR) has emerged as a promising text-to-image paradigm, yet it suffers from limited semantic discriminability, alignment bias, and closed-set restrictions. To address these challenges, we propose SIGMA, a novel framework for Semantic Internalization for Generative Multimodal Alignment. SIGMA constructs multi-granularity hierarchical identifiers to ensure unique, semantically consistent image representations. We further introduce a progressive semantic internalization training strategy augmented with semantic soft labels, which captures fine-grained text-image affinities and enables inductive identifier assignment for unseen samples realizing open-set dynamic indexing capabilities. Experiments on the Flickr30K and MS-COCO datasets demonstrate that SIGMA outperforms state-of-the-art baselines, achieving average Recall@1, Recall@5, and Recall@10 improvements of 10.65%, 8.50%, and 7.00%, respectively.

1 Introduction

In recent years, with the explosive growth of multimedia data, cross-modal retrieval (Xia et al., 2023; Wang et al., 2024; Fang et al., 2025) has gradually become a research hotspot in the field of information retrieval. Among these, text-to-image retrieval (Li et al., 2024; Shen et al., 2025; Chen et al., 2025) has attracted widespread attention due to its significant value in application scenarios such as information querying, visual search and human-computer interaction. This task aims to retrieve the most semantically relevant images from large-scale image databases based on input natural language queries. However, due to the inherent representation gap and structural heterogeneity between modalities, effectively bridging the semantic gap remains a core challenge.

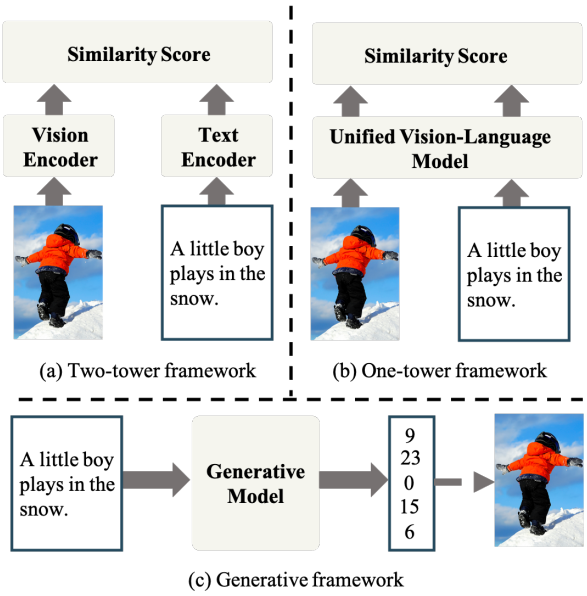

Figure 1: Illustrations of three paradigms for text-to-image retrieval. (a) The two-tower framework matches embeddings in a shared space; (b) the one-tower framework relies on deep cross-modal interaction; and (c) the generative framework (e.g., SIGMA) directly predicts image identifiers from a text query.

Existing text-to-image retrieval methods can be mainly categorized into two-tower and one-tower frameworks. Two-tower frameworks (Chen et al., 2020; Radford et al., 2021; Qu et al., 2023; Cherti et al., 2023) independently encode modalities and align representation spaces through contrastive learning, enabling efficient vectorized retrieval (Figure 1 (a)). However, the independence of the tow encoders precludes deep cross-modal interactivity, making them incapable of capturing fine-grained word-region alignments. In contrast, one-tower frameworks (Li et al., 2022, 2023b; Chen et al., 2024; Zhang et al., 2025) leverage deep cross-attention for superior performance (Figure 1 (b)), but their heavy computational overhead makes them prohibitive for large-scale real-time

*Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 12972–12986

applications. Despite their respective strengths, effective solutions that achieve deep cross-modal understanding without compromising retrieval efficiency remain an open challenge.

Empowered by recent advances in generative models (Brown et al., 2020; Touvron et al., 2023; Lu et al., 2024; Bai et al., 2025), generative retrieval (Li et al., 2024; Qu et al., 2024; Li et al., 2025a,b; Kim et al., 2025) has emerged as a paradigm shift, reformulating retrieval as direct identifier prediction (Figure 1 (c)). Despite its promise, it still faces three key challenges:

(1) Trade-off between semantic discriminability and generability of identifiers. Generative retrieval requires identifiers to possess sufficient semantic discriminability to distinguish different images while maintaining good generability for stable model prediction.

(2) Semantic alignment bias caused by "one image, multiple captions". In practical retrieval scenarios, different users often formulate diverse queries to describe the same image, focusing on varying perspectives or granularities. Existing methods (Li et al., 2024, 2025a) typically enforce mapping of all query texts to the same discrete identifier. This rigid alignment strategy ignores the varying degrees of semantic affinity between text-image pairs, easily leading to gradient conflicts during training and causing confusion in semantic understanding (Lavoie et al., 2024).

(3) Closed-set assumption limits scalability. Mainstream generative methods (Li et al., 2025b; Kim et al., 2025; Fang et al., 2025) typically operate under a closed-set setting, where all images to be retrieved must be known during training. When new samples are added, the entire codebook or model parameters often need to be retrained, which severely limits their ability to handle dynamic and incremental data, rendering them unsuitable for open-world scenarios where new images are continuously generated.

Although generative cross-modal retrieval demonstrates great potential, these challenges restrict its practical effectiveness. To address these challenges, we propose SIGMA (**S**emantic **I**nternalization for **G**enerative **M**ultimodal **A**lignment), a novel generative text-to-image retrieval framework. Specifically, we first construct a five-level hierarchical identifier system via structured clustering and subspace decomposition, enabling each image to have a unique and semantically traceable representation (addressing Challenge 1). Subsequently, we develop a progressive semantic internalization training strategy coupled with a semantic soft label mechanism. This not only enables the model to capture the nuanced semantic discrepancies across multiple textual descriptions of the same image (addressing Challenge 2), but also transforms identifier assignment into a learnable mapping function that supports the dynamic indexing of unseen samples (addressing Challenge 3). This empowers the model with open-set retrieval capabilities, a feature largely absent in prior generative arts. We conduct extensive experiments on Flickr30K and MS-COCO datasets, comparing with discriminative methods and existing generative approaches. Experimental results demonstrate that SIGMA achieves average improvements of 10.65%, 8.50%, and 7.00% on Recall@1, Recall@5, and Recall@10.

The main contributions of this paper are as follows:

- We propose SIGMA, a novel generative text-to-image retrieval framework that redefines identifier design and alignment strategies to enhance fine-grained cross-modal understanding.
- We develop a progressive semantic internalization training strategy augmented with semantic soft labels. Crucially, this enables inductive identifier assignment for unseen samples, equipping the model with open-set dynamic indexing capabilities that overcomes the closed-set limitations inherent in prior generative methods.
- We conduct comprehensive experiments on multiple benchmark datasets to validate the effectiveness of our method. Furthermore, we provide ablation studies to analyze the specific contribution of each module.

2 Related Works

2.1 Cross-modal Retrieval

Existing approaches predominately fall into two categories. Two-tower architectures (Faghri et al., 2017; Chen et al., 2020; Radford et al., 2021; Qu et al., 2023; Cherti et al., 2023) map distinct modalities into a shared latent space via independent encoders. While facilitating large-scale indexing, they inherently lack fine-grained cross-

modal interaction. In contrast, One-tower architectures (Li et al., 2022, 2023b; Chen et al., 2024; Zhang et al., 2025) employ cross-attention mechanisms to achieve deep alignment, yielding superior semantic understanding at the cost of prohibitive inference latency. Both paradigms frame retrieval as a metric learning problem, inevitably encountering bottlenecks related to either memory-intensive indexing or exhaustive pairwise computations.

2.2 Generative Retrieval

Generative retrieval reformulates retrieval as a sequence generation task, directly predicting the identifier of a target item. This paradigm originated in document retrieval with DSI (Tay et al., 2022) and NCI (Wang et al., 2022). Recently, this paradigm has been extended to text-to-image retrieval. Early studies (Li et al., 2024, 2025b) utilized textual or atomic IDs, while recent approaches (e.g., TIGER (Qu et al., 2024), CART (Fang et al., 2025)) adopt residual quantization to incorporate visual semantics. Despite these advancements, current methods remain constrained by three critical limitations: (1) Semantic Decoupling: Identifiers often lack direct grounding in the visual representation learning process; (2) Alignment Bias: Rigid "hard ID" assignments conflict with the intrinsic "one-image–multiple-captions" nature of the data; and (3) Closed-set Assumption: Incorporating new samples typically necessitates model retraining. Distinguishing itself from prior arts, SIGMA introduces hierarchical semantic identifiers and progressive semantic internalization training strategy, thereby enabling efficient, inductive retrieval tailored for open-world scenarios.

3 Methodology

In this section, we provide a comprehensive overview of the proposed SIGMA framework. As illustrated in Figure 2, the framework consists of two core modules: (1) Multi-granularity Identifier Construction Module: constructs a coarse-to-fine hierarchical identifier system for candidate images; (2) Progressive Semantic Internalization Training Module: enhances the model's fine-grained cross-modal understanding through identifier memorization, identifier comprehension, and retrieval alignment based on images, image identifiers, and image descriptions.

3.1 Multi-granularity Identifier Construction Module

The design of identifiers directly impacts the learning efficiency and retrieval performance of generative retrieval models. In existing research, image identifiers take various forms, including string identifiers, numeric identifiers, and atomic identifiers. However, regardless of the identifier form adopted, discretizing image representations inevitably introduces information loss. To mitigate this and enhance the semantic discriminability and model learnability, we propose a coarse-to-fine hierarchical construction method. Inspired by existing literature (Fang et al., 2025), we find that identifiers with a length of 5 achieve an optimal balance between expressiveness and learning efficiency. Accordingly, this process decomposes an image into a five-level identifier sequence: $z_i = [id_1, id_2, id_3, id_4, id_5]$. It begins by establishing a broad semantic anchor for each image, then iteratively disentangles and refines its distinctive components, and concludes by ensuring global uniqueness. The overall construction pipeline is illustrated in Figure 2 (Stage 0).

(1) Establishing Global Semantic Anchors (id_1). This step constructs the coarsest layer of the identifier, id_1 (semantic anchor). Its objective is to partition the image corpus into K_1 broad, semantically coherent regions. We apply K-Means clustering on the visual features $\mathbf{v}_i \in \mathbb{R}^D$ extracted from a frozen vision encoder. Each image is assigned the index of its nearest centroid $\mathbf{c}_k$:

$$id_1 = \arg \min_{k \in \{1,\dots,K_1\}} \|\mathbf{v}_i - \mathbf{c}_k\|^2 \qquad (1)$$

This provides a high-level contextual framework for subsequent fine-grained encoding.

(2) Extracting Dominant Semantic Components (id_2). Within each global cluster, we perform a finer-grained semantic disentanglement using Semi-Nonnegative Matrix Factorization (Semi-NMF) (Ding et al., 2008). For images within the same id_1 cluster, we stack their features into a matrix V_{local} and decompose it as $V_{\text{local}} \approx WH$, where H represents the latent semantic basis. This stage identifies the most salient feature directions assigning id_2 based on the primary basis vector contribution.

(3) Discerning Fine-grained Semantic Groups (id_3). To capture subtle intra-cluster variations,

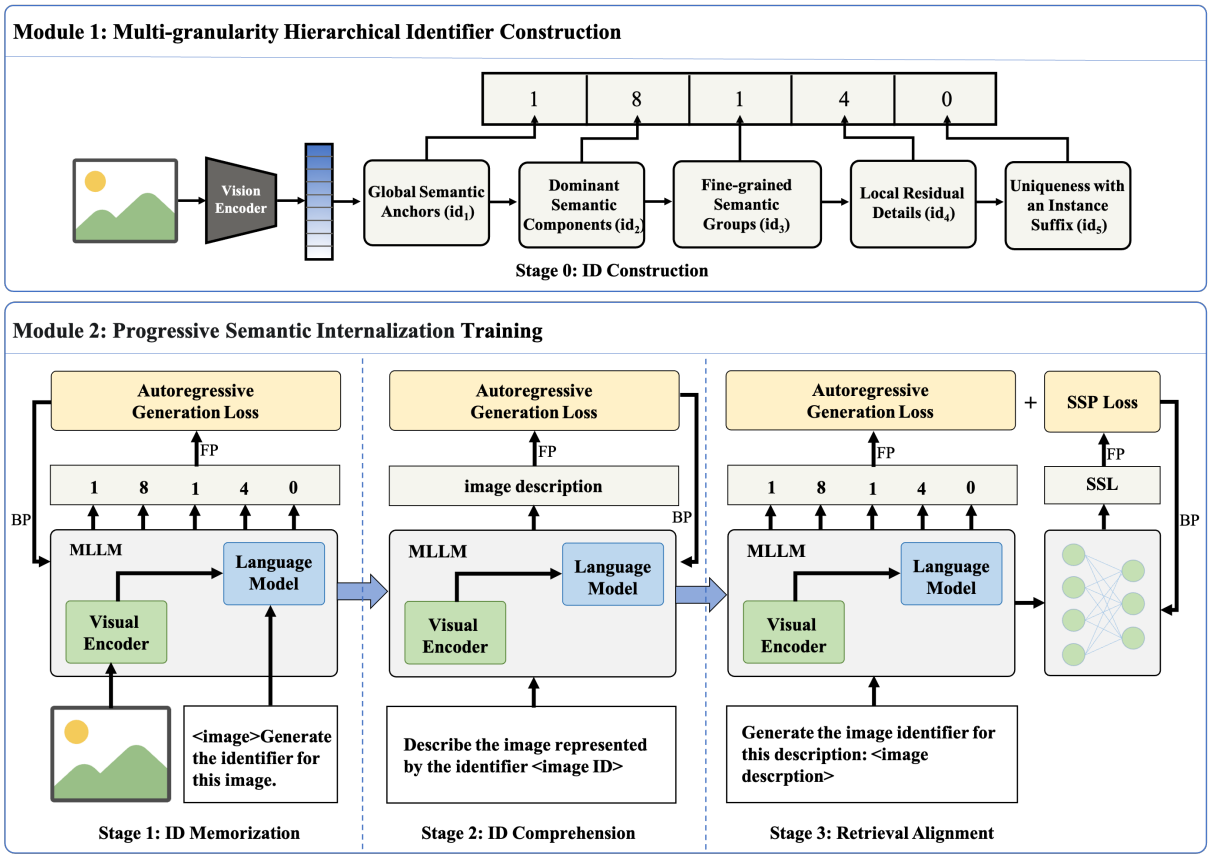

Figure 2: Overview of SIGMA.

we apply a second-order Semi-NMF to further refine the subspace. This stage operates on the residuals or specialized basis of the previous step, dividing images into tighter semantic sub-groups. By discerning these categorical nuances, we assign id_3, which serves to bridge the gap between high-level semantics and instance-specific details.

(4) Encoding Local Residual Details (id_4). To capture the unique, often non-semantic visual details within the same semantic group, we introduce id_4 (Residual Detail). Instead of quantizing the original features, we first compute the residual vector relative to its semantic anchor:

$$\mathbf{r}_i = \mathbf{v}_i - \mathbf{c}_{id_1} \tag{2}$$

where $\mathbf{c}_{id_1}$ is the centroid assigned.

To achieve high-fidelity encoding with a compact index, $\mathbf{r}_i$ is evenly divided into M subspaces. For each subspace m, we maintain a learnable codebook $C^{(m)} = \{c_1^{(m)}, \ldots, c_N^{(m)}\}$. By performing a nearest-neighbor search within each codebook, we obtain a sequence of quantization indices $(q_i^{(1)}, \ldots, q_i^{(M)})$, which collectively form the id_4 component of the identifier sequence.

(5) Guaranteeing Uniqueness with an Instance Suffix (id_5). The final step ensures global uniqueness by appending an instance_suffix (id_5). This serves as a fail-safe mechanism to distinguish between images that are semantically and visually nearly identical (e.g., near-duplicates). By mapping each unique image instance to a terminal ID, we eliminate collisions in the generative space, ensuring instance-level precision during retrieval.

3.2 Progressive Semantic Internalization Training Module

To enable the model to truly understand the visual semantic information encoded in image identifiers rather than merely memorizing correspondences, we draw inspiration from the human cognitive strategy for learning new symbols—"memorize the form first, then understand the semantics". Based on this principle, we design a progressive semantic internalization training module that leverages a multimodal large language model (MLLM) to sequentially complete three training stages: identifier memorization, identifier comprehension, and retrieval alignment, as illustrated in Figure 2 (Stage1, Stage2, Stage3). The MLLM first learns to mem-

orize identifiers through "image–identifier" mappings, then comprehends the semantic information of identifiers through "identifier–description" mappings, and finally perceives subtle semantic differences among descriptions using the semantic soft label mechanism during retrieval alignment satge.

3.2.1 Identifier Memorization: Mapping Images to Identifiers

The objective of this stage is to establish an accurate mapping from visual content to its corresponding structured identifier, enabling the model to master the "syntax" of identifier generation. Given an input image v_i, the model autoregressively generates the target identifier sequence label z_i:

$$z_i = MLLM(v_i, t_{prompt}) \tag{3}$$

where t_{prompt} is a guiding text prompt in the form of "<image> Generate the identifier for this image.", instructing the model to output the corresponding identifier.

In this stage, .We adopt the standard autoregressive generation loss, i.e., the cross-entropy loss:

$$\mathcal{L}_{\text{mem}} = -\sum_{t=1}^{T} \log P(z_i^t \mid z_i^{<t}, v_i, t_{\text{prompt}}) \tag{4}$$

where z_i^t is the t-th token of the identifier sequence $\mathbf{z}_i$, $z_i^{<t}$ represents the first $t-1$ generated tokens, and T is the length of the identifier sequence. This loss function optimizes the model's ability to accurately generate the corresponding identifiers given input images.

3.2.2 Identifier Comprehension: Instilling Semantics into Identifiers

Building upon the model's ability to generate the "form" of identifiers, this stage aims to instill concrete visual semantics into these abstract symbols, achieving "understanding of meaning". The training task is reversed:

$$d_i = MLLM(t'_{prompt}) \tag{5}$$

where t'_{prompt} is the input guiding prompt, e.g., "Describe the image represented by the identifier <image ID>", where "<image ID>" is the image identifier z_i. d_i is the image description generated by the model.

Through training in this stage, the model learns to associate abstract identifier symbols with concrete visual semantic content. The training loss employs the cross-entropy loss:

$$\mathcal{L}_{\text{com}} = -\sum_{t=1}^{T'} \log P(d_i^t \mid d_i^{<t}, \mathbf{z}_i, t'_{\text{prompt}}) \tag{6}$$

where d_i^t is the t-th token of the description text, and T' is the length of the description.

During this training stage, we integrate a subset of samples from the Identifier Memorization phase into the training process. This strategy preserves the model's identifier generation capability, laying the groundwork for the subsequent assignment of identifiers to newly added samples.

3.2.3 Retrieval Alignment: Learning Discriminative Semantic Matching

The final stage aims to equip the model with discriminative semantic matching abilities to address the real-world challenge where a single image corresponds to multiple descriptions of varying relevance. To address this issue, we introduce a semantic soft label (SSL) mechanism that enables the model to perceive semantic differences among various descriptions of the same image.

Semantic Soft Label Construction Single-encoder models achieve deep text-image interaction through joint encoding, excelling at fine-grained semantic matching. Therefore, we employ a pre-trained cross-modal encoder model as a teacher to pre-compute semantic similarity scores for each "image-caption" pair in the training set as soft labels. for image v_i and its descriptions $\{d_i^1, d_i^2, \ldots, d_i^K\}$, we compute similarity using a cross-modal encoder (CE):

$$s_{i,k} = CE(v_i, d_i^k) \tag{7}$$

where the output $s_{i,k} \in [0, 1]$ represents the semantic similarity between the input text d_i^k and the image v_i.

Generation and Similarity Prediction In this stage, given text query d_i^k, the model performs two tasks: (1) generate the target identifier $\mathbf{z}_i$; (2) predict the query-image semantic similarity $\hat{s}_{i,k}$. We leverage this for similarity prediction: first extracting the last-layer hidden states $\{\mathbf{h}_1, \ldots, \mathbf{h}_l\}$ of all identifier tokens and applying average pooling:

$$\mathbf{h}_{\text{id}} = \frac{1}{L} \sum_{l=1}^{L} \mathbf{h}_l \tag{8}$$

then predicting similarity through a lightweight MLP:

$$\hat{s}_{i,k} = \sigma(\text{MLP}(\mathbf{h}_{\text{id}})) \in [0, 1] \tag{9}$$

where MLP is a three-layer perceptron and σ is the Sigmoid function.

Joint Training Objective In the retrieval alignment stage, the model's training objective comprises two components, optimizing both identifier generation capability and semantic similarity perception.

For identifier generation loss, we adopt the standard autoregressive cross-entropy loss to optimize the model's ability to generate correct identifiers given text queries:

$$\mathcal{L}_{\text{gen}} = -\sum_{l=1}^{L} \log P\left(z_i^l \mid z_i^{<l}, d_i^k\right) \tag{10}$$

where $z_i^{<l}$ denotes the first $l-1$ generated identifier tokens. This loss ensures that the model can accurately map text queries to the corresponding image identifiers.

For semantic similarity prediction (SSP) loss, we employ the mean squared error (MSE) loss to optimize the consistency between the model's predicted similarity and the soft labels provided by the teacher model:

$$\mathcal{L}_{\text{ssp}} = \frac{1}{K}\sum_{k=1}^{K}\left(\hat{s}_{i,k} - s_{i,k}\right)^2 \tag{11}$$

where K denotes the number of descriptions associated with image v_i. This loss enables the model to learn to distinguish semantic relevance between different text queries and images.

The final joint loss is a weighted combination of both components:

$$\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{gen}} + \lambda_{\text{ssp}} \cdot \mathcal{L}_{\text{ssp}} \tag{12}$$

where λ_{ssp} is the weight coefficient for the similarity loss. It is crucial to emphasize that accurate identifier generation remains the primary objective. Therefore, we set a relatively small λ_{ssp}, allowing similarity prediction to serve as an auxiliary supervisory signal that guides the model to perceive subtle semantic differences between query-image pairs without interfering with the main task.

The detailed inference process for model retrieval is described in Appendix B.

4 Experiments

To comprehensively evaluate the effectiveness of SIGMA, we design the following research questions (RQs).

- **RQ1:** How does SIGMA perform compared to existing methods on text-to-image retrieval?
- **RQ2:** What is the contribution of each module in SIGMA to retrieval performance?
- **RQ3:** How efficient is SIGMA in large-scale retrieval scenarios?

4.1 Experimental Setup

This section describes the experimental setup, including the datasets used, baseline methods for comparison, evaluation metrics and implementation details.

Datasets We conduct experiments on two widely-used benchmarks: Flickr30K (Young et al., 2014) and MS-COCO (Lin et al., 2014). More details are provided in the Appendix C.

Baselines To comprehensively evaluate the performance of SIGMA, we select representative methods from both the discriminative paradigm (Dual-path (Zheng et al., 2020), SGM (Wang et al., 2020), IMRAM (Chen et al., 2020), DIME (Qu et al., 2023), CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023)) and the generative paradigm (Grace (Li et al., 2024), IRGen (Zhang et al., 2024), TIGeR (Qu et al., 2024), AVG (Li et al., 2025b), SemCORE (Li et al., 2025a), GENIUS (Kim et al., 2025)) as baselines for systematic comparison. Further details are provided in Appendix D.

Evaluation Metrics In generative retrieval tasks, we follow prior studies (Young et al., 2014; Chen et al., 2021) and adopt R@K (K = 1, 5, and 10) as the evaluation metric for retrieval performance.

Implementation Details The proposed SIGMA consists of two core components: multi-granularity hierarchical identifier construction and progressive semantic internalization training. The detailed parameter settings and implementation details of these two components are provided in Appendix E.

Table 1: Experimental results of all baselines and our proposed SIGMA on Flickr30K and MS-COCO for text-to-image retrieval. The "Avg. Improvement" row denotes the average performance gain of SIGMA over six generative retrieval baselines (GRACE-atomic, IRGen, TIGeR, AVG, SemCORE, and GENIUS). The best results are in **bold**.

Type	Model	Flickr30K			MS-COCO		
		Recall@1	Recall@5	Recall@10	Recall@1	Recall@5	Recall@10
Two-tower	Dual-path	39.1	69.2	80.9	25.3	53.4	66.4
	SGM	53.5	79.6	86.5	35.3	64.9	76.5
	IMRAM	53.9	79.4	87.2	39.6	69.1	79.8
	DIME	59.1	85.5	91.0	39.7	70.3	81.0
	CLIP	58.4	81.5	88.1	37.8	62.4	72.7
	OpenCLIP	63.9	87.3	93.2	39.4	65.4	75.6
GR	GRACE-numeric	22.5	28.9	29.4	0.03	0.14	0.28
	GRACE-string	30.5	39.0	40.4	0.12	0.37	0.88
	GRACE-semantic	22.9	34.9	37.4	13.3	30.4	35.9
	GRACE-structure	37.4	59.5	66.2	16.7	39.2	50.3
	GRACE-atomic	68.4	88.9	93.7	41.5	69.1	79.1
	IRGen	49.0	68.9	72.5	29.6	50.7	56.3
	TIGeR	71.7	91.8	95.4	46.1	69.0	76.1
	AVG	62.8	85.4	-	31.3	58.0	-
	SemCORE	69.0	83.0	-	42.4	57.5	-
	GENIUS	74.1	92.0	94.8	46.1	**74.0**	**82.7**
Our	SIGMA	**77.4**	**92.5**	**97.2**	**49.3**	72.6	79.5
	Avg. Improvement	↑11.5	↑7.5	↑8.1	↑9.8	↑9.5	↑5.9

4.2 Performance of SIGMA (RQ1)

To answer RQ1, we conduct extensive experiments on Flickr30K and MS-COCO datasets, comprehensively comparing SIGMA with existing discriminative retrieval methods (two-tower frameworks) and generative cross-modal retrieval methods.

Overall Performance Comparison Table 1 reports the retrieval performance on Flickr30K and MS-COCO datasets. SIGMA consistently outperforms both discriminative and generative baselines. Compared with existing generative methods, SIGMA achieves average improvements of 10.65%, 8.50%, and 7.00% in R@1, R@5, and R@10, across the two datasets. Performance on MS-COCO dataset is generally lower than Flickr30K dataset across all methods, a natural consequence of the larger search space (5,000 vs 1,000 test images). Nevertheless, SIGMA maintains robust superiority, validating its effectiveness in enhancing accuracy while retaining the efficiency benefits of the generative paradigm.

Analysis of Generative Baselines Comparing SIGMA with state-of-the-art generative approaches highlights the critical role of identifier semantics. While GRACE suffers from interference caused by pre-existing token semantics and SemCORE relies on rigid hard alignment, SIGMA excels via its "memorization-understanding" strategy. By deeply comprehending the semantic connotations of identifiers rather than merely memorizing mappings, SIGMA achieves superior fine-grained alignment. Although GENIUS shows marginally higher R@5 and R@10 scores on MS-COCO dataset due to its computationally expensive post-hoc re-ranking, SIGMA retains the lead on the most critical Recall@1 metric. This underscores SIGMA's precise semantic sensitivity without the latency overhead of additional re-ranking steps.

Table 2: Ablation study of key modules in SIGMA on Flickr30K.

Method	Recall@1	Recall@5	Recall@10
w/o ID Mem	73.2	87.6	92.0
w/o ID Com	75.1	90.1	94.3
w/o SSL	73.5	88.2	93.4
SIGMA	**77.4**	**92.5**	**97.2**

4.3 Ablation Study (RQ2)

To answer RQ2, we conduct ablation experiments on the Flickr30K dataset by selectively removing the ID Memorization (w/o ID Mem), ID Comprehension (w/o ID Com), and Semantic Soft Label (w/o SSL) modules to analyze the contribution of each module in SIGMA to retrieval performance.

As shown in Table 2, removing any single mod-

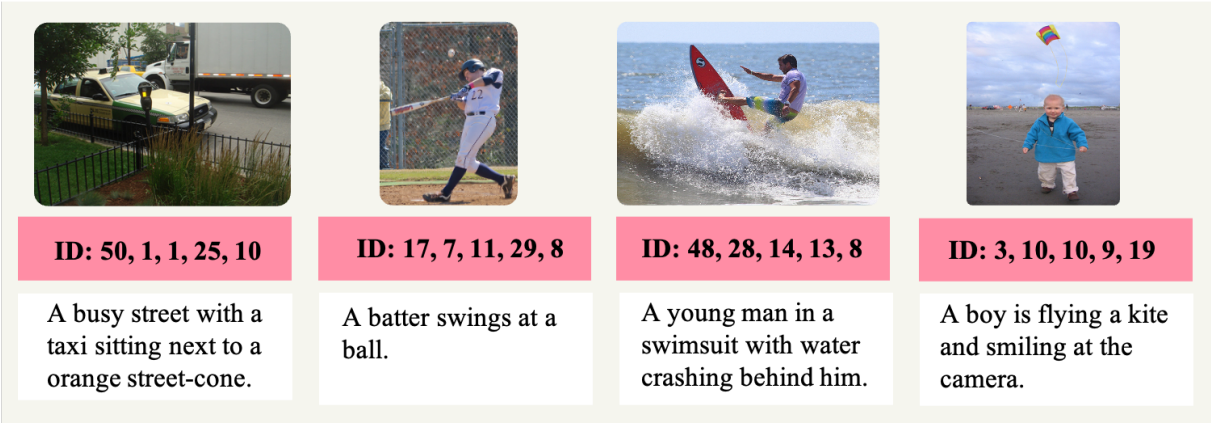

Figure 3: Examples of descriptions generated for previously unseen images.

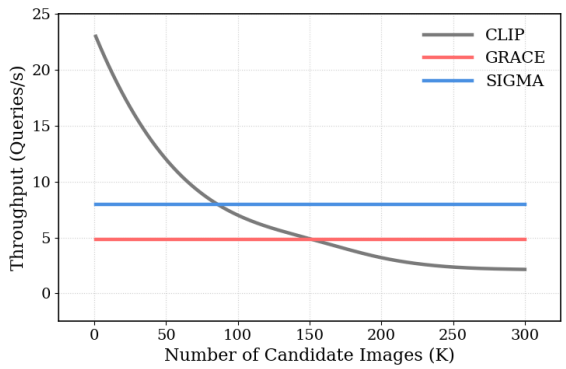

Figure 4: Throughput Comparison. This shows throughput performance across different candidate set sizes.

ule leads to performance degradation, confirming their individual necessity. Specifically, ID Memorization proves most critical, with its removal causing the sharpest drop (4.2% in R@1), indicating its fundamental role in establishing visual-identifier mappings. The ID Comprehension stage contributes a 2.3% gain by internalizing identifier semantics, transforming them from shallow index keys into meaningful units. SSL outperforms hard-label training by 3.9%, effectively capturing fine-grained text-image affinities and mitigating the alignment bias inherent in one-to-many mappings.

4.4 Efficiency Analysis (RQ3)

To answer RQ3, we compare the efficiency of generative frameworks (SIGMA, GRACE) and two-tower frameworks (CLIP) from the throughput perspective (Queries Per Second) , as shown in Figure 4. Due to the linear cost of exhaustive similarity computation, CLIP exhibits significant efficiency degradation as the scale increases, with an approximately threefold drop from 1K to 100K. In contrast, SIGMA exhibits remarkable scale-invariance, maintaining stable throughput regardless of database scale. This advantage stems from the generative paradigm, which decouples inference complexity from dataset size, depending solely on identifier generation length. Moreover, benefiting from the efficient Qwen backbone, SIGMA achieves nearly twice the throughput of the Flamingo-based GRACE. It is important to note that these results are obtained without any inference acceleration techniques (e.g., vLLM (Kwon et al., 2023)).

To provide a fairer comparison between SIGMA and CLIP in a real-world retrieval system, we further report a more practical efficiency evaluation under deployment settings. In this setup, CLIP uses pre-extracted image features with a vector index, while SIGMA is accelerated using the vLLM inference framework. The single-query retrieval latency (ms) under different database scales is reported subsequently.

Model	50K	100K	200K	300K
CLIP	0.95	1.42	1.85	2.27
SIGMA	20	20	20	20

Table 3: Single-query retrieval latency (ms) under different database scales.

The experimental results reveal a fundamental trade-off between the two retrieval paradigms: CLIP achieves lower absolute latency on small-scale databases, but its latency increases steadily as the database grows. In contrast, SIGMA's retrieval latency is independent of database size, since generative retrieval directly decodes the target identifier without traversing the gallery or computing similar-

ities. This scale-invariant property makes SIGMA more predictable and scalable for large-scale retrieval scenarios. As the database size increases, the latency gap between the two paradigms gradually narrows.

While retrieval efficiency is important, retrieval quality is equally critical. As shown in Table 1, SIGMA achieves substantial improvements in retrieval accuracy compared to CLIP: Recall@1 increases by 19% on the Flickr30K dataset and by 11.5% on the MS-COCO dataset. SIGMA trades a tolerable increase in latency for significantly better retrieval performance, which is particularly advantageous in large-scale deployment scenarios that demand both high accuracy and strong scalability.

5 Discussion

We investigate SIGMA's inductive capability in cold-start scenarios, where new samples must be indexed without retraining. Using a model trained on Flickr30K dataset, we evaluate 100 unseen MS-COCO images. SIGMA achieves 56.4%, 71.2%, and 84.0% on R@5, R@10, and R@20, indicating strong cross-dataset generalization under moderate distribution shift.

In contrast, when the same model is applied to a cross-domain e-commerce fruit image dataset, performance drops substantially, with accuracies of only 22.5%, 29.2%, and 32.0% on R@5, R@10, and R@20. This degradation is expected and interpretable: because the Flickr30K training distribution lacks fruit-related visual concepts, the semantic content encoded in the identifier labels cannot adequately cover this domain, leading to unreliable identifier assignments. These two experiments reveal how retrieval performance degrades as the degree of domain shift increases.

To probe the nature of these assignments, we perform inverse semantic decoding, as illustrated in Figure 3. Surprisingly, the model can reconstruct accurate captions conditioned solely on the assigned identifiers. This suggests that identifiers function as semantic compression codes, where visual semantics are encapsulated into discrete symbols. This finding implies that SIGMA does not merely memorize mappings but internalizes identifiers as meaningful semantic units, offering a novel perspective on the representational mechanism of generative retrieval.

6 Conclusion

In this paper, we propose SIGMA, a novel generative text-to-image retrieval framework that synergizes multi-granularity hierarchical identifiers with a progressive semantic internalization training strategy. This framework supports inductive identifier assignment for unseen samples, enabling open-set dynamic indexing. Experiment results demonstrate that SIGMA achieves an average improvement of 10.65% on Recall@1. Furthermore, efficiency analyses confirm the scale-invariance of this paradigm. Future work will explore optimizing identifier generation and enhancing retrieval performance on newly added samples.

Limitations

Despite SIGMA's promising performance, several limitations remain. First, the discrete nature of hierarchical identifiers introduces inevitable quantization errors compared to continuous embeddings, potentially limiting the capture of ultra-fine-grained visual nuances. Second, due to the auto-regressive decoding mechanism, inference latency scales linearly with the number of retrieved items, which may pose bottlenecks in high-recall scenarios. Finally, while SIGMA supports inductive assignment for unseen samples, the long-term impact of massive-scale dynamic updates on retrieval recall requires further validation. In future work, we aim to address these constraints by exploring learnable codebooks to reduce quantization loss, investigating more efficient decoding mechanisms to accelerate retrieval, and developing robust incremental indexing strategies to ensure long-term stability.

Ethical Considerations

We acknowledge potential ethical risks associated with large-scale cross-modal retrieval. First, since our model is trained on public datasets (Flickr30K, MS-COCO) that may contain inherent societal biases (e.g., gender or racial stereotypes), the retrieved results could inadvertently reflect or amplify these biases. Second, regarding privacy and copyright, we strictly adhere to the usage licenses of all datasets and use them solely for academic research purposes. Finally, although SIGMA is designed for information retrieval, we condemn any misuse of the framework for retrieving harmful or illegal content, and future work should explore

safety alignment mechanisms within the generative process.

Acknowledgement

We sincerely appreciate all the reviewers for their constructive suggestions. This work was supported by the National Key Research and Development Program of China (No.2024YFF0618800),

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. Imram: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12655–12663.

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*, pages 15789–15798.

Jiahui Chen, Xiaoze Jiang, Zhibo Wang, Quanzhi Zhu, Junyao Zhao, Feng Hu, Kang Pan, Ao Xie, Maohua Pei, Zhiheng Qin, and 1 others. 2025. Unisearch: Rethinking search system with a unified generative architecture. *arXiv preprint arXiv:2509.06887*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829.

Chris HQ Ding, Tao Li, and Michael I Jordan. 2008. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Minghui Fang, Shengpeng Ji, Jialong Zuo, Hai Huang, Yan Xia, Jieming Zhu, Xize Cheng, Xiaoda Yang, Wenrui Liu, Gang Wang, and 1 others. 2025. Cart: A generative cross-modal retrieval framework with coarse-to-fine semantic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15120–15133.

Edward Fredkin. 1960. Trie memory. *Communications of the ACM*, 3(9):490–499.

Michael Günther, Saba Sturua, Mohammad Kalim Akram, Isabelle Mohr, Andrei Ungureanu, Bo Wang, Sedigheh Eslami, Scott Martens, Maximilian Werk, Nan Wang, and 1 others. 2025. jina-embeddings-v4: Universal embeddings for multimodal multilingual retrieval. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 531–550.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

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*, pages 19659–19669.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

Samuel Lavoie, Polina Kirichenko, Mark Ibrahim, Mahmoud Assran, Andrew Gordon Wilson, Aaron Courville, and Nicolas Ballas. 2024. Modeling caption diversity in contrastive vision-language pretraining. In *Proceedings of the 41st International Conference on Machine Learning*, pages 26070–26084.

Haoxuan Li, Yi Bin, Junrong Liao, Yang Yang, and Heng Tao Shen. 2023a. Your negative may not be true negative: Boosting image-text matching with false negative elimination. In *Proceedings of the 31st ACM international conference on multimedia*, pages 924–934.

Haoxuan Li, Yi Bin, Yunshan Ma, Guoqing Wang, Yang Yang, See-Kiong Ng, and Tat-Seng Chua. 2025a. Semcore: A semantic-enhanced generative cross-modal retrieval framework with mllms. *arXiv preprint arXiv:2504.13172*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

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*, pages 12888–12900. PMLR.

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*, pages 4654–4662.

Yongqi Li, Hongru Cai, Wenjie Wang, Leigang Qu, Yinwei Wei, Wenjie Li, Liqiang Nie, and Tat-Seng Chua. 2025b. 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*, pages 813–822.

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. *arXiv preprint arXiv:2402.10805*.

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 *European conference on computer vision*, pages 740–755. Springer.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, and 1 others. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

Leigang Qu, Haochuan Li, Tan Wang, Wenjie Wang, Yongqi Li, Liqiang Nie, and Tat-Seng Chua. 2024. Tiger: Unifying text-to-image generation and retrieval with large multimodal models. *arXiv preprint arXiv:2406.05814*.

Leigang Qu, Meng Liu, Wenjie Wang, Zhedong Zheng, Liqiang Nie, and Tat-Seng Chua. 2023. Learnable pillar-based re-ranking for image-text retrieval. In *Proceedings of the 46th international ACM SIGIR conference on research and development in information retrieval*, pages 1252–1261.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Zhixiao Shen, Jianfei Yu, Wenya Wang, and Rui Xia. 2025. Global question-aware multimodal retrieval-augmented generation for multimedia multi-hop question answering. In *Proceedings of the 7th ACM International Conference on Multimedia in Asia*, pages 1–8.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Yi Tay, Vinh Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, and 1 others. 2022. Transformer memory as a differentiable search index. *Advances in Neural Information Processing Systems*, 35:21831–21843.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sijin Wang, Ruiping Wang, Ziwei Yao, Shiguang Shan, and Xilin Chen. 2020. Cross-modal scene graph matching for relationship-aware image-text retrieval. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1508–1517.

Yabing Wang, Le Wang, Qiang Zhou, Zhibin Wang, Hao Li, Gang Hua, and Wei Tang. 2024. Multimodal llm enhanced cross-lingual cross-modal retrieval. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8296–8305.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, and 1 others. 2022. A neural corpus indexer for document retrieval. *Advances in Neural Information Processing Systems*, 35:25600–25614.

Yan Xia, Hai Huang, Jieming Zhu, and Zhou Zhao. 2023. Achieving cross modal generalization with multimodal unified representation. *Advances in Neural Information Processing Systems*, 36:63529–63541.

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*, 2:67–78.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yidan Zhang, Ting Zhang, Dong Chen, Yujing Wang, Qi Chen, Xing Xie, Hao Sun, Weiwei Deng,

Qi Zhang, Fan Yang, and 1 others. 2024. Irgen: Generative modeling for image retrieval. In *European Conference on Computer Vision*, pages 21–41. Springer.

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):1–23.

A Preliminary

Generative Retrieval (GR) refers to the process of directly outputting target identifiers through autoregressive generation conditioned on a given query, thereby completing the retrieval task. In the generative retrieval paradigm, retrieval no longer relies on explicit similarity computation. Instead, each retrieval target is treated as a generatable discrete sequence and assigned a unique identifier.

In text-to-image retrieval, let V denote the image retrieval corpus. For any image $v_i \in V$, its corresponding target identifier is represented as a discrete sequence of length L:

$$z_i = (z_i^1, z_i^2, ..., z_i^l) \tag{13}$$

where z_i^l denotes the l-th identifier token, and L is the length of the identifier sequence.

Given any text query t, the generative model M autoregressively generates the corresponding image identifier in sequence. Specifically, when generating the l-th identifier token, the model M performs conditional prediction based on the text query t and the previously generated $l-1$ tokens. This generation process can be formalized as:

$$p(\mathbf{z}_i \mid t) = \prod_{l=1}^{L} p\big(z_i^l \mid t, z_i^1, \ldots, z_i^{l-1}\big) \tag{14}$$

During inference, the model generates the most probable identifier sequence by maximizing the above conditional probability, and maps the generated identifier back to its corresponding image instance, thereby completing the retrieval process. To ensure the validity and interpretability of generated identifiers, constrained search strategies (e.g., constrained decoding) are typically introduced during the decoding phase to ensure that generated results always reside within the valid identifier space.

B Inference

During inference, given a query text q, SIGMA feeds it into the trained MLLM, which autoregressively generates the target identifier sequence $z = (z^1, z^2, \ldots, z^L)$ token by token, as illustrated in Figure 5. To ensure the generated identifier is always valid, we employ trie-based constrained decoding (Fredkin, 1960).

We construct a prefix tree based on all image identifiers in the training set, where each root-to-leaf path represents a valid identifier. During generation, the model uses beam search (Sutskever et al., 2014), but the search space is constrained to consider only valid next tokens allowed by the prefix tree, rather than all tokens in the vocabulary. This constraint mechanism guarantees that the generation result always corresponds to a valid image, significantly improving retrieval efficiency and avoiding invalid generations.

C Datasets

We evaluate SIGMA on two widely-used cross-modal retrieval benchmarks: Flickr30K (Young et al., 2014) and MS-COCO (Lin et al., 2014). Flickr30K contains 31,783 images, each paired with five human-annotated captions. Following the standard split (Li et al., 2019), we use 29,783 images for training, 1,000 for validation, and 1,000 for testing. MS-COCO comprises 123,287 images, each also accompanied by five captions. Following existing studies(Fang et al., 2025), we adopt the Karpathy split (Karpathy and Fei-Fei, 2015) with 113,287 training images, 5,000 validation images, and 5,000 test images.

D Bselines

We compare SIGMA against representative methods from both discriminative and generative paradigms.

Discriminative Methods We select Dual-path (Zheng et al., 2020), SGM (Wang et al., 2020), IMRMA (Chen et al., 2020), DIME (Qu et al., 2023), CLIP (Radford et al., 2021), and OpenCLIP (Cherti et al., 2023) as discriminative baselines. Note that one-tower architectures, while excelling in semantic understanding, incur substantial inference overhead that makes them more suitable for re-ranking rather than large-scale retrieval, therefore following existing studies (Li et al., 2025b), we exclude them from comparison.

Generative Methods We compare against the following generative retrieval approaches: (1) GRACE (Li et al., 2024) pioneered the application of generative paradigms to cross-modal re-

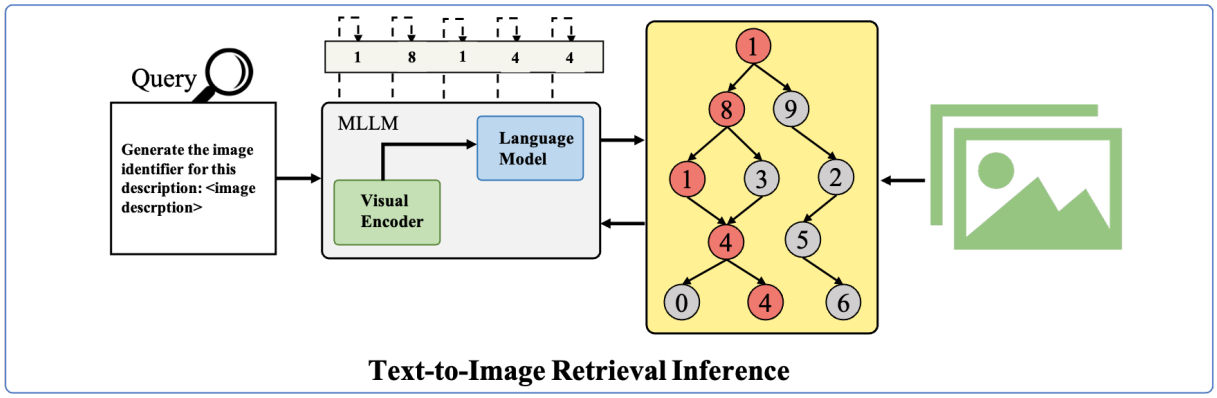

Figure 5: SIGMA Inference Procedure.

trieval, exploring various forms of image identifiers; (2) IRGen (Zhang et al., 2024) was originally proposed for image-to-image retrieval and was later adapted for text-to-image retrieval (Li et al., 2025b); (3) TIGeR (Qu et al., 2024) unifies text-to-image generation and retrieval within multimodal large language models; (4) AVG (Li et al., 2025b) reformulates text-to-image retrieval as a generation process from text tokens to visual tokens; (5) SemCORE (Li et al., 2025a) enhances semantic understanding through structured natural language identifiers and generative semantic verification strategies; (6) GENIUS (Kim et al., 2025) employs modality-decoupled semantic quantization and query augmentation strategies for cross-modal retrieval.

E Implementation Details

Multi-granularity Identifier Construction module We utilize the pre-trained jina-embeddings-v4 (Günther et al., 2025) to encode images into 2,048-dimensional vectors. Hierarchical clustering is then performed using Mini-Batch K-Means ($K = 64$) from scikit-learn, followed by Semi-NMF for intra-cluster decomposition with refinement factors of $K = 32$ and $K = 16$. Finally, Residual Quantization ($K = 32$) is applied for fine-grained classification. This hierarchical structure supports a total capacity of 1,048,576 unique identifiers. To scale to larger capacities, higher K values can be configured at each stage as needed.

Progressive Semantic Internalization Training Module We perform full-parameter fine-tuning of the Qwen2.5-VL (Bai et al., 2025) backbone on four NVIDIA A6000 GPUs (48GB memory per GPU). The training process consists of three stages: ID Memorization, ID Comprehension, and Retrieval Alignment. In the soft label construction for retrieval alignment, we use the jina-reranker-m0 model. We use the AdamW optimizer with a cosine learning rate schedule across all stages. Specifically, the ID Memorization stage runs for 5 epochs with a learning rate (lr) of 5×10^{-5} and a batch size of 64. The ID Comprehension stage follows for 5 epochs with lr=3×10^{-5} (batch size 64). Finally, the Retrieval Alignment stage employs a reduced batch size of 10 and lr=2×10^{-5} to refine semantic alignment.

F Analysis on comprehension ability

Figure 6 visualizes the Top-5 retrieval results generated by SIGMA. We observe a high degree of semantic fidelity between the retrieved images and the textual queries, accompanied by notable structural similarity in their corresponding identifiers. This suggests that our hierarchical identifiers effectively cluster semantically related samples. For instance, given the query "Little girl in arm floaties exploring the coast line", despite variations in composition and posture, all retrieved images accurately align with the fine-grained constraints (e.g., specific objects and settings), demonstrating the model's robust capability in fine-grained semantic capturing.

To further probe the representational nature of identifiers, we conducted an identifier-to-caption generation experiment. As shown in Figure 7, the model generates highly accurate descriptions conditioned solely on the identifiers. This phenomenon confirms that identifiers are not arbitrary index keys but function as "semantic compression codes". Through the ID Memorization and ID Comprehension stages, the model learns to encode visual semantics into discrete tokens and subsequently

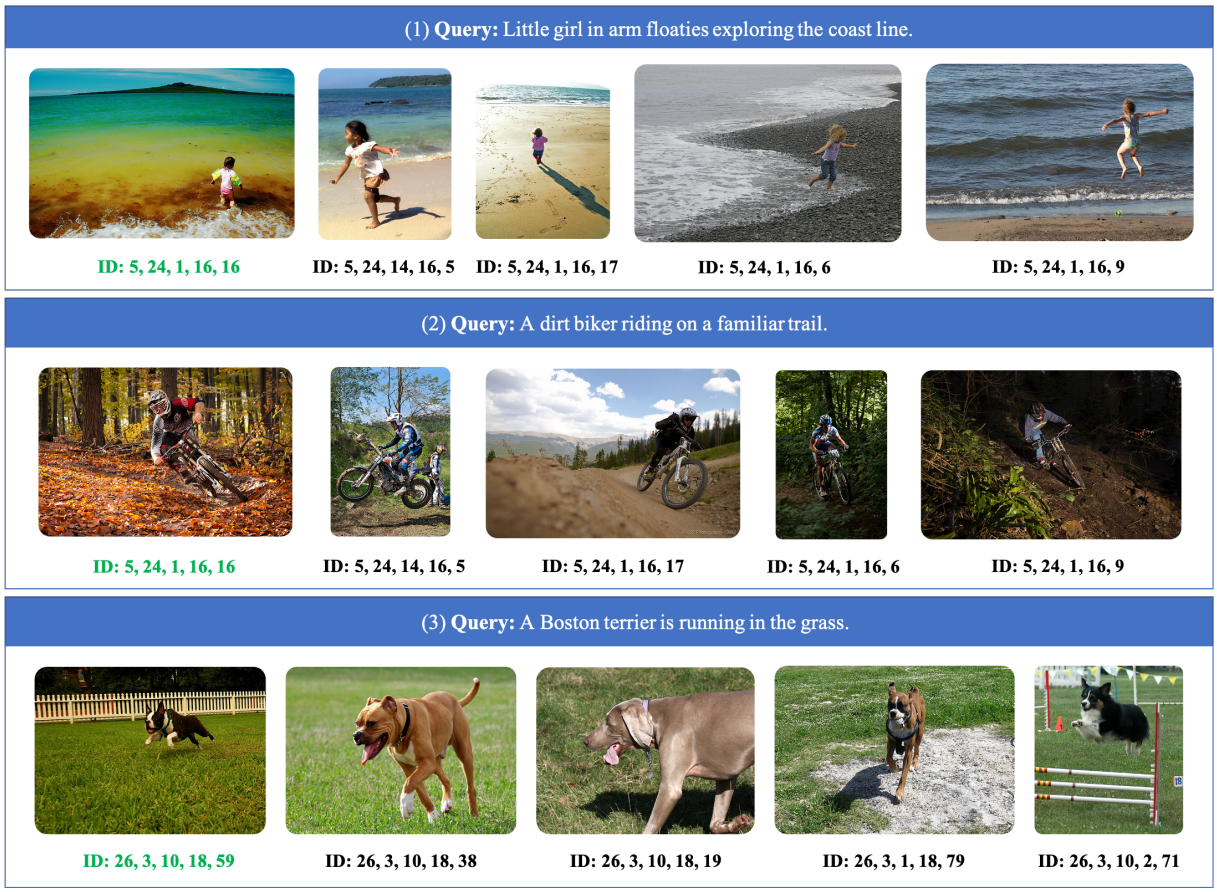

Figure 6: Cases of the top-5 retrieval results from SIGMA. The green text indicates the target image.

decode them. This "compression-decompression" mechanism validates that identifiers have been internalized as meaningful symbolic units, capable of preserving and transmitting rich semantic information within the generative framework.

G Error Analysis

To investigate the failure modes of SIGMA, we conduct a qualitative analysis of representative "failure" cases. As illustrated in Figure 8, we observe instances where the ground-truth image ranks lower (e.g., 3rd) or fails to appear in the Top-3 results. Notably, despite missing the specific ground truth, the retrieved candidates exhibit high semantic congruence with the query text, sharing significant visual and scene characteristics with the target.

This phenomenon essentially reflects the inherent limitations of annotation in text-to-image retrieval datasets. In many-to-many image-text matching scenarios, a single caption often corresponds to multiple semantically similar images, yet existing dataset annotations typically provide only one or a few pairings. Consequently, other semantically correct images are erroneously treated as negatives during evaluation. This aligns closely with the "false negative" problem identified by Li et al. (Li et al., 2023a).

From an application perspective, this phenomenon actually highlights SIGMA's strength. In practical retrieval scenarios, users typically prioritize semantic relevance over exact matching to specific annotated images. SIGMA's ability to return semantically relevant results demonstrates its strong generalization and semantic comprehension capabilities, which hold significant value in real-world deployment.

Figure 7: Examples of image descriptions generated based on image IDs. Blue text denotes the model-generated image descriptions, and bold text indicates the image identifiers.

Figure 8: Qualitative analysis of retrieval errors.