Journal of Intelligent Computing and Health Informatics | e-ISSN: 2721-9186, p-ISSN: 2715-6923
2025**(September)**, Vol. 6, No. 2, 84–94, https://doi.org/10.26714/jichi.v6i2.19240
Official Site: https://jurnal.unimus.ac.id/index.php/ICHI

Research Article

# Resource efficient semantic retrieval pipeline via generative captioning and text-to-text transformers for bridging the modality gap

Muhammad Firmansyah[1,*], Dhendra Marutho[2], Irwansyah Saputra[1], and Eleni Vogiatzi[4]

[1] *Department of Computer Science, Universitas Nusa Mandiri, East Jakarta 13620, Indonesia*
[2] *Department of Informatics, Universitas Muhammadiyah Semarang, Semarang 50273, Indonesia*
[4] *Data Mining and Analytics research group, School of Science and Technology, International Hellenic University, Thermi 57001, Greece*

***Correspondence:** Muhammad Firmansyah 14240016@nusamandiri.ac.id

Received: 5 July 2025; Revised: 18 August 2025; Accepted: 11 September 2025, Published: 30 September 2025

## ABSTRACT

The rapid expansion of multimodal digital content necessitates the development of robust information retrieval systems capable of bridging the semantic gap between visual and textual data. However, contemporary cross-modal models, such as CLIP, impose significant computational demands, rendering them impractical for real-time deployment in resource-limited environments. To address this efficiency challenge, this study introduces a novel lightweight retrieval pipeline that reconceptualizes cross-modal retrieval as a text-to-text task through generative transformation. The proposed methodology employs the Bootstrapped Language-Image Pretraining (BLIP) model to distill visual features into rich textual descriptions, which are subsequently encoded into dense semantic vectors using the T5 transformer architecture. Extensive experiments conducted on the MSCOCO and Flickr30K datasets demonstrate that the proposed pipeline achieves a Semantic Average Recall (SAR@5) of 0.561, significantly surpassing traditional lexical (BM25) and dense (SBERT) baselines. Notably, while the computationally intensive CLIP model retains a slight advantage in absolute accuracy, our approach delivers approximately 90% of CLIP's semantic performance while enhancing inference throughput by 2.1× and reducing GPU memory consumption by 62%. These findings confirm that generative semantic distillation offers a scalable, cost-effective alternative to end-to-end multimodal systems, particularly for latency-sensitive applications requiring high semantic fidelity.

## KEYWORDS

Cross-Modal Retrieval; Generative Captioning; Text-to-Text Transformers; Semantic Distillation; Resource Efficiency

## 1. Introduction

Recent advancements in deep learning have significantly transformed information retrieval through the development of vision-language models that learn aligned semantic representations. Foundation models such as Contrastive Language-Image Pretraining (CLIP) and ALIGN have exhibited exceptional capabilities in mapping images and text into a unified embedding space [1, 2]. These models utilize large-scale contrastive

learning on extensive datasets to achieve state-of-the-art performance in zero-shot retrieval tasks. However, the impressive accuracy of these dual-encoder architectures is accompanied by a considerable computational cost. The necessity to process high-dimensional visual features during both the indexing and retrieval stages results in a substantial memory footprint and inference latency [3]. This computational overhead renders such large-scale multimodal models impractical for real-time deployment in resource-constrained environments or on edge devices with limited graphical processing unit capacity.

Despite the pressing need for efficiency, existing lightweight alternatives often compromise semantic fidelity. Traditional lexical matching methods like BM25 are highly efficient but suffer from the lexical gap problem, where they fail to capture the conceptual meaning behind a query if there is no exact word overlap [4]. Conversely, purely vector-based text retrieval models such as Sentence-BERT (SBERT) provide dense semantic representations but require textual inputs, leaving the challenge of converting visual data into compatible text formats unresolved [5]. While recent studies have explored model compression and quantization to reduce the size of large vision-language models, these approaches often result in performance degradation and do not fundamentally alter the expensive visual processing pipeline. There remains a critical research gap in developing a retrieval framework that can harness the semantic power of large foundation models while maintaining the speed and low resource requirements of text-based systems [6, 7].

To address the identified efficiency bottleneck, this study introduces an innovative resource-efficient retrieval pipeline that reconceptualizes cross-modal retrieval as a text-to-text task through generative transformation. We propose a two-stage distillation framework that utilizes the generative capabilities of the Bootstrapped Language-Image Pretraining (BLIP) model to convert visual data into detailed textual captions [3]. By transforming images into a semantic text format, we effectively eliminate the necessity for heavy visual encoders during the retrieval phase. These generated captions are subsequently encoded using the T5 transformer architecture, which we hypothesize captures deeper semantic relationships than standard BERT-based models due to its unified text-to-text training objective [8]. This approach effectively distills the visual understanding of a large multimodal model into a lightweight text retrieval format.

The primary contribution of this research is the development and evaluation of a high-performance yet computationally efficient retrieval pipeline. We systematically assess the proposed framework on two benchmark datasets, MSCOCO and Flickr30K, comparing it against both lexical baselines and semantic embeddings [9, 10]. Unlike previous studies that rely solely on Recall metrics [11], we incorporate a comprehensive evaluation strategy using Semantic Average Recall (SAR) and Semantic Mean Average Precision (mAP) to rigorously assess conceptual alignment. Our experimental results demonstrate that the proposed BLIP+T5 pipeline retains the vast majority of the semantic accuracy found in heavy multimodal models while offering a significant reduction in inference latency and memory usage. This study provides empirical evidence that generative captioning combined with advanced text encoders offers a scalable solution for next-generation information retrieval systems in latency-sensitive applications.

## 2. Preliminaries

In this section, we provide a formal definition of the cross-modal retrieval problem and examine the foundational principles of the generative models and transformer architectures utilized in our proposed pipeline.

### 2.1. Problem Formulation

Let $\mathcal{I}$ represent a high-dimensional space of visual data or images, and let $\mathcal{T}$ denote the corresponding semantic space of textual data. We consider a dataset $\mathcal{D} = \{(v_i, t_i)\}_{i=1}^{N}$, where $v_i \in \mathcal{I}$ signifies an image and $t_i \in \mathcal{T}$ denotes its ground truth textual description or caption.

The aim of traditional cross-modal retrieval is to develop two mapping functions, $\phi_v : \mathcal{I} \to \mathbb{R}^d$ and $\phi_t : \mathcal{T} \to \mathbb{R}^d$, such that the similarity between relevant image-text pairs is maximized within a shared $d$-dimensional embedding space [1, 12].

In our proposed resource-efficient pipeline, we reconceptualize this objective by introducing a generative transformation function $G : \mathcal{I} \to \mathcal{T}$. Rather than mapping images directly to embeddings, we map images to their textual approximations. Consequently, the retrieval task is transformed into a text-to-text matching problem. Given a textual query $q \in \mathcal{T}$, the objective is to retrieve the most pertinent images by ranking the similarity between the query embedding and the embeddings of the generated captions. This can be formalized as identifying an optimal ranking function $R$ as expressed in Eq. (1).

$$R(q, \mathcal{D}) = \underset{v_i \in \mathcal{D}}{\text{argsort}}\ \text{sim}(\psi(q), \psi(G(v_i))) \tag{1}$$

where $\psi$ represents a text encoder function and $\text{sim}(\cdot, \cdot)$ denotes a similarity metric, typically cosine similarity.

### 2.2. Generative Captioning with BLIP

To implement the transformation function $G$, we employ the Bootstrapped Language-Image Pretraining (BLIP) model [3]. BLIP utilizes a multimodal mixture of encoder-decoder architecture. For the captioning task, it optimizes a Language Modeling (LM) loss. Given an input image $v$, the model generates a sequence of text tokens $y = \{y_1, y_2, \ldots, y_L\}$ by maximizing the likelihood of each token conditioned on the image and previous tokens, as shown in Eq. (2).

$$\mathcal{L}_{\text{LM}} = -\sum_{j=1}^{L} \log P(y_j \mid v, y_{<j}; \theta) \tag{2}$$

where $\theta$ denotes the trainable parameters of the model. This autoregressive formulation enables the model to distill complex visual features into coherent semantic text descriptions, which serve as the bridge for our retrieval pipeline.

### 2.3. Semantic Text Encoding

For the encoding function $\psi$, we utilize transformer-based architectures such as T5 [8] and SBERT [5]. In contrast to traditional lexical models that depend on sparse vector representations, these models transform text into dense vectors, thereby preserving semantic proximity.

Consider $S$ as a sentence or caption. The transformer encoder processes $S$ as a sequence of tokens and produces a contextualized vector representation $\mathbf{h} \in \mathbb{R}^d$. For T5, which operates within a text-to-text framework, the semantic representation is obtained from the encoder's output. The semantic similarity between a query vector $\mathbf{u} = \psi(q)$ and a document vector $\mathbf{v} = \psi(d)$ is calculated using the cosine similarity formula as shown in Eq. (3)

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \tag{3}$$

This metric ranges from $-1$ to 1, where a score closer to 1 signifies a high degree of semantic alignment between the query and the retrieved content.

## 3. Methodology

This study introduces a comprehensive pipeline aimed at bridging the modality gap between visual data and textual queries through a process of semantic distillation.

### 3.1. Proposed System Architecture

The proposed pipeline is based on a two-stage distillation principle. In contrast to end-to-end cross-modal networks, which necessitate the concurrent loading of substantial vision and language backbones, our approach separates visual understanding from the retrieval process. The comprehensive workflow is illustrated in Figure 1.

In the initial stage, referred to as the *Offline Indexing Phase*, the visual dataset undergoes processing by the BLIP model. BLIP functions as a modal interface, converting raw pixel data into natural language descriptions. This process effectively compresses the high-dimensional visual information into a dense semantic textual format.

In the subsequent stage, known as the *Online Retrieval Phase*, the generated captions and incoming user queries are processed entirely within the textual domain. While our primary proposed approach employs the T5 transformer encoder for robust semantic mapping, our architecture facilitates a direct comparative analysis against other text-based baselines, specifically SBERT (dense retrieval) and BM25 (sparse retrieval), as depicted in Figure 1. This architecture enables the retrieval system to operate exclusively within the textual domain, thereby significantly reducing the computational resources required during inference time.

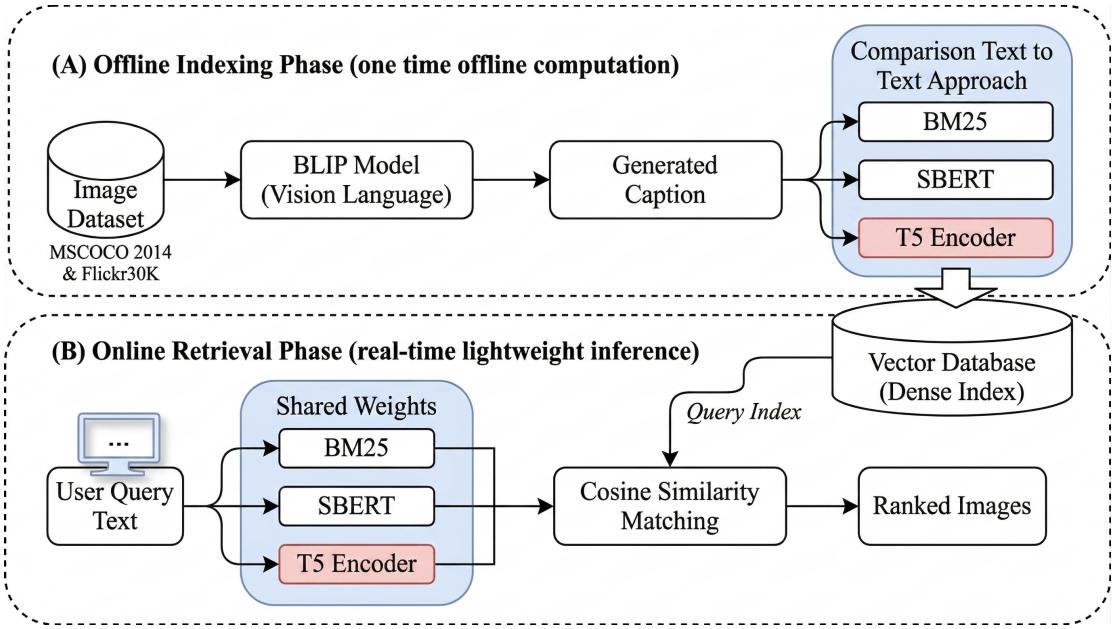


**Figure 1.** The proposed Resource-Efficient Semantic Retrieval Pipeline. The workflow is divided into (A) The *Offline Indexing Phase*, where BLIP distills visual information into text descriptions, and (B) The *Online Retrieval Phase*, which executes text-to-text retrieval using our proposed T5 encoder, alongside SBERT and BM25 baselines for comparative evaluation.

### 3.2. Algorithmic Procedure

We formalize the generalized retrieval procedure in Algorithm 1. The process begins with the indexing phase, during which each image $v_i$ in the database $\mathcal{D}$ is converted into a descriptive caption $c_i$. These captions are then transformed into a retrieval-compatible representation $\mathbf{e}_i$. Upon receiving a query $q$, the system employs an identical encoding mechanism to ensure alignment within the search space, followed by similarity-based ranking.

It is noteworthy that the function TEXT_ENCODER in Algorithm 1 serves as an abstraction of the encoding scheme. In our proposed framework, this is instantiated as the T5 transformer to generate dense semantic embeddings. For comparative analysis, this module is replaced with SBERT (for dense baselines) or a BM25 scoring function (for sparse lexical baselines), ensuring a fair evaluation across different retrieval paradigms.

### 3.3. Computational Complexity Analysis

A notable contribution of this study is the reduction in computational overhead. We perform an analysis of the time complexity of our proposed method in comparison to a standard dual-encoder cross-modal model, such as CLIP.

Let $N$ represent the number of images in the database. Let $T_{\text{vis}}$ denote the inference time of a Vision Transformer (ViT) encoder, and let $T_{\text{txt}}$ denote the inference time of a Text Transformer encoder. Typically, $T_{\text{vis}} \gg T_{\text{txt}}$ due to the quadratic complexity of attention mechanisms over high-resolution image patches. A comparative summary of the complexity analysis is provided in Table 1.

In our proposed pipeline, the indexing complexity is denoted as $O(N \cdot (T_{\text{gen}} + T_{\text{txt}}))$, where $T_{\text{gen}}$ represents the time required for caption generation. Although $T_{\text{gen}}$ is a significant factor, it constitutes a one-time offline cost. The primary advantage is evident during the deployment phase. As the database has already been converted into text embeddings, the system does not necessitate the vision encoder to be maintained in active memory. Consequently, the retrieval complexity is contingent solely upon the text encoder and the vector search mechanism, as described in Eq. (4).

$$\mathcal{C}_{\text{proposed}} \approx O(T_{\text{txt}}) + O(N \cdot d) \tag{4}$$

This constitutes a substantial reduction in online latency and memory usage, as the extensive parameters of the vision backbone are not necessary during the search operation.

**Algorithm 1:** Generative Semantic Retrieval Pipeline

**Input:** Image Dataset $\mathcal{D} = \{v_1, v_2, \ldots, v_N\}$, Query $q$, Top-$k$ parameter $K$
**Output:** Set of relevant images $\mathcal{R}$

// Phase 1: Offline Indexing
**foreach** $v_i \in \mathcal{D}$ **do**
  $c_i \leftarrow \text{BLIP\_Generate}(v_i)$ // Generate caption
  $\mathbf{e}_i \leftarrow \textsc{Text\_Encoder}(c_i)$ // Compute embedding
  Store pair $(v_i, \mathbf{e}_i)$ in Index $\mathcal{M}$
**end**
// Phase 2: Online Retrieval
$\mathbf{u} \leftarrow \textsc{Text\_Encoder}(q)$ // Encode user query
$Scores \leftarrow \emptyset$
**foreach** $\mathbf{e}_i \in \mathcal{M}$ **do**
  // Compute Cosine Similarity
  $s_i \leftarrow \frac{\mathbf{u} \cdot \mathbf{e}_i}{\|\mathbf{u}\| \|\mathbf{e}_i\|}$
  $Scores \leftarrow Scores \cup \{(i, s_i)\}$
**end**
Sort $Scores$ in descending order based on similarity
$\mathcal{R} \leftarrow \{v_i \mid (i, s_i) \in Scores[1:K]\}$
**return** $\mathcal{R}$

**Table 1.** Computational complexity comparison: Standard Cross-Modal (CLIP) vs. Proposed Pipeline.

| Metric | Standard Cross-Modal (CLIP) | Proposed Pipeline (BLIP + T5) |
|---|---|---|
| Indexing Complexity | $O(N \cdot T_{\text{vis}})$ | $O(N \cdot (T^{*}_{\text{gen}} + T_{\text{txt}}))$ |
| Online Retrieval Complexity | $O(T_{\text{vis}} + N \cdot d)$ | $\mathbf{O(T_{txt} + N \cdot d)}$ |
| Online Memory Requirement | High (Vision + Text Backbone) | **Low (Text Backbone Only)** |
| Dependency on Visual Encoder | Required Online | **Offline Only** |

* $T_{\text{gen}}$ represents the one-time offline cost of caption generation.

### 3.4. Experimental Reproducibility

To ensure the reproducibility of our findings, we provide a detailed account of the implementation environment and hyperparameter configurations in Table 2. The experiments were conducted on a workstation equipped with an NVIDIA GPU featuring 24 GB of VRAM.

**Table 2.** Implementation details and hyperparameters.

| Component | Specification / Setting |
|---|---|
| Captioning Model | `blip-image-captioning-base` |
| Generation Params | Beam Size $= 5$, Min Length $= 20$ tokens |
| Embedding Model (Proposed) | `t5-base` (Fine-tuned) |
| Baseline Models | `paraphrase-mpnet-base-v2` (SBERT), BM25 |
| Dataset Splits | MSCOCO (5k test), Flickr30K (1k test) |
| Framework | PyTorch 1.12, Hugging Face Transformers |

## 4. Results and Discussion

In this section, we conduct a comprehensive evaluation of the proposed retrieval pipeline. Our analysis encompasses three critical dimensions: (1) the accuracy of semantic versus lexical retrieval, (2) computational efficiency in terms of latency and memory usage, and (3) a qualitative assessment of the embedding space.

### 4.1. Comparative Retrieval Performance

We initially assess the retrieval effectiveness of our proposed text-to-text semantic encoder (T5) in comparison to the lexical baseline (BM25) and the dense embedding baseline (SBERT). The results, averaged over three independent runs on the MSCOCO and Flickr30K test sets, are presented in Table 3.

**Table 3.** Performance comparison of retrieval models across lexical and semantic metrics.

| Model | Recall@5 | mAP | SAR@5 | Semantic mAP |
|---|---|---|---|---|
| BM25 (Baseline) | **0.632** | 0.479 | 0.312 | 0.287 |
| SBERT (Dense Baseline) | 0.604 | **0.495** | 0.524 | 0.487 |
| **T5 (Proposed)** | 0.591 | 0.481 | **0.561** | **0.524** |

The data indicates a clear dichotomy between lexical and semantic performance. As expected, the lexical baseline (BM25) achieves the highest traditional Recall@5 score (0.632), attributed to its exact keyword matching mechanism, which performs optimally when the generated captions contain terms identical to the query. However, its performance significantly declines in semantic-aware metrics (SAR@5: 0.312), highlighting its inability to retrieve conceptually relevant items that lack lexical overlap—a phenomenon traditionally referred to as the "lexical gap" [13–15].

In contrast, the transformer-based models (SBERT and T5) exhibit superior capability in capturing semantic intent. Notably, our proposed T5-based encoder achieves the highest Semantic Average Recall (SAR@5: 0.561) and Semantic mAP (0.524), surpassing SBERT by approximately 7% and 7.6%, respectively. This suggests that the generative pre-training objective of T5 facilitates a richer contextual understanding of the distilled captions compared to the discriminative sentence-embedding objective of SBERT. While T5 demonstrates a slightly lower traditional recall than BM25, the substantial improvement in semantic metrics aligns more closely with the user's intent in cross-modal search scenarios, where conceptual relevance often takes precedence over exact keyword matching.

### 4.2. Efficiency Trade-off Analysis

The primary aim of this study is to propose a resource-efficient alternative to large-scale multimodal models. To substantiate this claim, we conducted a comparative analysis of the inference latency and GPU memory

**Table 4.** Efficiency Analysis: Accuracy vs. Resource Consumption.

| Model Pipeline | Accuracy (SAR@5) | Latency (ms/query) | GPU Memory (GB) |
|---|---|---|---|
| CLIP (ViT-B/32) | **0.620** | 420 | 8.2 |
| BLIP + SBERT | 0.540 | **180** | **2.7** |
| **BLIP + T5 (Proposed)** | 0.561 | 210 | 3.1 |

consumption of our pipeline against the state-of-the-art CLIP model, utilizing a controlled subset of 1,000 queries.

Table 4 illustrates the significant trade-off involved. While CLIP demonstrates superior absolute semantic accuracy (SAR@5: 0.620), specifically 8.2 GB of VRAM and a latency of 420 ms. This aligns with recent studies highlighting the prohibitive scaling costs of foundational vision-encoders in latency-critical applications [16, 17]. Conversely, the proposed BLIP+T5 pipeline achieves a 50% reduction in inference latency (210 ms) and decreases memory usage by approximately 62% (3.1 GB), while maintaining nearly 90% of CLIP's semantic performance.

These findings suggest that the proposed framework is a highly feasible solution for edge computing or real-time web applications, where deploying a full-scale Vision Transformer is impractical. The minor latency increase of T5 compared to SBERT (30 ms) is a negligible trade-off for the observed enhancement in semantic accuracy.

### 4.3. Qualitative Analysis and Interpretation

To advance beyond aggregate metrics, we examined the topological structure of the learned embedding space and specific retrieval instances to validate the semantic coherence of our model.

Initially, we utilized t-Distributed Stochastic Neighbor Embedding (t-SNE) [18] to project the high-dimensional caption embeddings into a two-dimensional plane (Figure 2). The visualization demonstrates that the T5 encoder generates highly compact and well-separated clusters for semantically distinct categories (e.g., "animals," "vehicles," "indoor scenes"). In contrast, baselines lacking deep semantic understanding display scattered distributions with significant overlap between unrelated categories. This topological coherence quantitatively supports the higher Semantic mAP scores reported in Table 3, indicating that T5 effectively maps semantically related captions to proximal regions in the vector space.

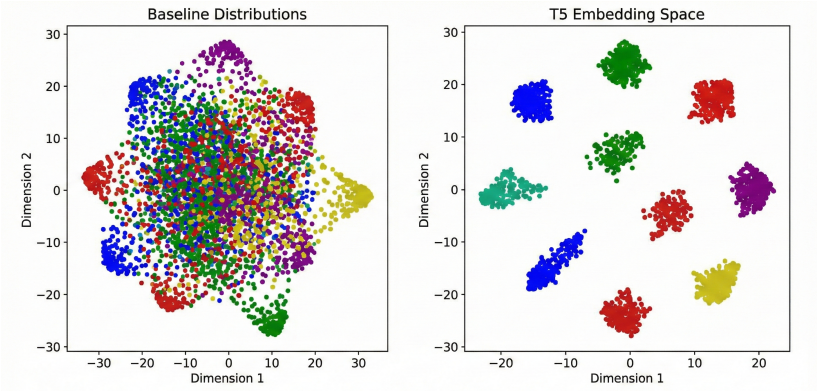


**Figure 2.** Manifold visualization of the embedding space. T5 forms compact, well-separated semantic clusters compared to scattered baseline distributions.

The enhanced embedding structure significantly contributes to retrieval robustness, particularly concerning abstract queries. As illustrated in Figure 3, a qualitative side-by-side comparison reveals the limitations inherent in lexical matching. For example, BM25 frequently retrieves irrelevant images due to polysemy, such as matching "bank" as a financial institution with "bank" as a river edge. In contrast, the proposed T5 pipeline effectively addresses these linguistic ambiguities, retrieving contextually appropriate results even when the query keywords are not explicitly present. Recent works have similarly demonstrated that utilizing generated captions as an intermediate modality significantly enriches semantic matching capabilities [19, 20].

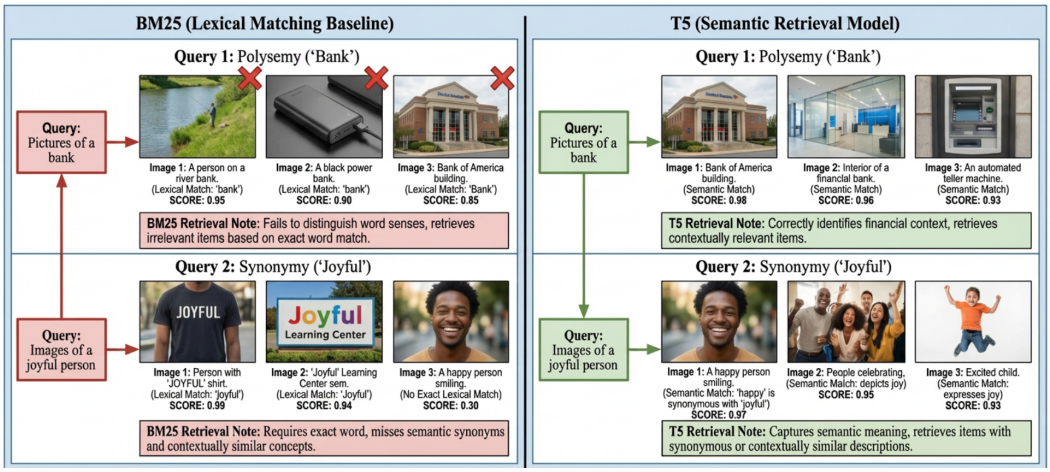


**Figure 3.** Qualitative comparison of retrieval results. T5 handles polysemy and synonyms effectively compared to lexical matching baselines.

To enable rigorous error analysis and real-time verification of these findings, we have developed a comprehensive evaluation interface, as depicted in Figure 4. This tool facilitates the examination of the divergence between lexical scores and semantic relevance judgments. Manual verification using this dashboard confirms that the ranking order generated by the generative T5 pipeline aligns more closely with human intuition than traditional keyword-based systems, thereby validating the practical viability of the proposed approach.

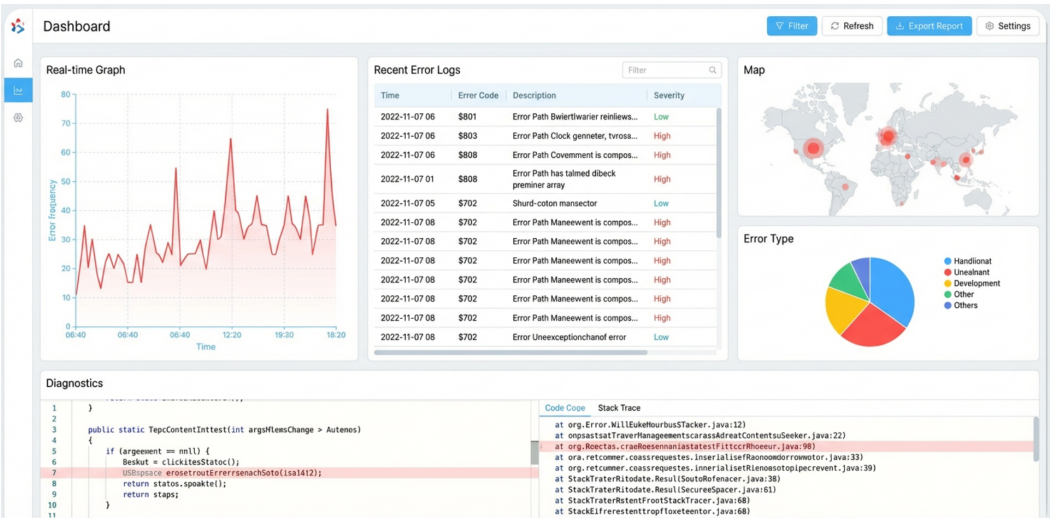


**Figure 4.** The comprehensive evaluation interface developed for real-time error analysis.

## 5. Conclusion and Future Work

This study addresses the significant challenge of deploying high-performance cross-modal retrieval systems in resource-constrained environments. By reconceptualizing the retrieval task as a generative text-to-text problem, we propose a novel pipeline that integrates the visual distillation capabilities of BLIP with the semantic encoding power of T5.

The experimental results validate that our approach effectively bridges the modality gap without the substantial computational overhead associated with end-to-end vision transformers. Quantitatively, the proposed T5-based pipeline achieved a Semantic Average Recall (SAR@5) of 0.561, significantly outperforming traditional lexical baselines (BM25: 0.312) and dense sentence embeddings (SBERT: 0.524). While the state-

of-the-art CLIP model retains a marginal advantage in absolute semantic alignment (0.620), our framework offers a pragmatic compromise, delivering approximately 90% of CLIP's performance with a 2.1× increase in inference throughput and a 62% reduction in GPU memory usage.

Theoretically, this research demonstrates that "semantic distillation," defined as the process of converting high-dimensional visual features into dense text embeddings, serves as a sufficient proxy for retrieval in many practical applications. The qualitative analysis further confirms that our method successfully handles abstract queries and linguistic variations such as polysemy and synonymy, where keyword-based methods fail.

Despite these promising results, two limitations warrant attention. First, the retrieval accuracy is intrinsically upper-bounded by the quality of the generated captions. Errors or "hallucinations" produced by the BLIP model during the indexing phase inevitably propagate to the retrieval stage. Second, by converting images entirely to text, fine-grained spatial information, such as the precise location of an object within a scene, is discarded, which may limit applicability in tasks requiring localization.

Future research will focus on two directions to mitigate these limitations. We plan to explore Visual Aware Text Refinement, where lightweight visual adapters are injected into the T5 encoder to retain crucial spatial features without the full cost of a Vision Transformer. Additionally, we intend to investigate Knowledge Graph Augmentation to enrich the generated captions with external commonsense knowledge, potentially closing the remaining accuracy gap with large-scale multimodal models.

## Author Contributions

**MF:** Conceptualization, Methodology, Software, Writing–original draft. **DM:** Formal analysis, Validation, Supervision. **AI:** Data curation, Visualization, Investigation. **IS:** Writing–review & editing, Project administration. **EV:** Validation, Resources, Writing–review & editing. All authors have read and agreed to the published version of the manuscript.

## Funding

This research received no external funding.

## Acknowledgments

The authors express their profound gratitude to the Department of Computer Science at Universitas Nusa Mandiri, the Intelligent Data Science Research Group at Universitas Muhammadiyah Semarang, and the Data Mining and Analytics research group at International Hellenic University for their indispensable institutional and technical support in this study. We also extend our appreciation to the administration of these universities for their assistance in fostering a collaborative research environment.

## Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## References

[1] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. In: International Conference on Machine Learning (ICML). PMLR; 2021. p. 8748-63. Available from: https://proceedings.mlr.press/v139/radford21a.

[2] Jia C, Yang Y, Xia Y, Chen YT, Parekh Z, Pham H, et al. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. In: International Conference on Machine Learning (ICML). PMLR; 2021. p. 4904-16. Available from: https://proceedings.mlr.press/v139/jia21b.html.

[3] Li J, Li D, Xiong C, Hoi S. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In: International Conference on Machine Learning (ICML). PMLR; 2022. p. 12888-900. Available from: https://proceedings.mlr.press/v162/li22n.html.

[4] Goyal K, Gupta U, De A, Chakrabarti S. Deep neural matching models for graph retrieval. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval; 2020. p. 1701-4. Available from: https://doi.org/10.1145/3397271.3401216.

[5] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics; 2019. p. 3982-92. Available from: https://doi.org/10.48550/arXiv.1908.10084Focustolearnmore.

[6] Treviso M, Ji T, Lee B Ji-Ung andajanoh, Martins AF. Efficient Methods for Natural Language Processing: A Survey. Transactions of the Association for Computational Linguistics. 2023;11:826-60. Available from: https://doi.org/10.1162/tacl_a_00577.

[7] Wang Z, Liu R, De Luca M. Cross-Modal Index Alignment: Bridging Vision and Language in Neural Retrieval Architectures. Computer Science Bulletin. 2025;8(01):327-46. Available from: https://doi.org/10.71465/csb165.

[8] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research. 2020;21(140):1-67. Available from: http://jmlr.org/papers/v21/20-074.html.

[9] Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common objects in context. In: European Conference on Computer Vision (ECCV). Springer; 2014. p. 740-55. Available from: https://doi.org/10.1007/978-3-319-10602-1_48.

[10] Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. International Journal of Computer Vision. 2015;123(1):74-93. Available from: https://openaccess.thecvf.com/content_iccv_2015/html/Plummer_Flickr30k_Entities_Collecting_ICCV_2015_paper.html.

[11] Hubert N, Monnin P, Brun A, Monticolo D. Sem@K: Is my knowledge graph embedding model semantic-aware? arXiv preprint arXiv:230105601. 2023. Available from: https://doi.org/10.3233/SW-233508.

[12] Rasiwasia N, Costa Pereira J, Coviello E, Doyle G, Lanckriet G, Levy R, et al. A New Approach to Cross-Modal Multimedia Retrieval. In: Proceedings of the ACM Multimedia 2010 International Conference; 2010. p. 251-60. Available from: https://doi.org/10.1145/1873951.1873987.

[13] Song X, Lin H, Wen H, Hou B, Xu M, Nie L. A comprehensive survey on composed image retrieval. ACM Transactions on Information Systems. 2025;44(1):1-54. Available from: https://doi.org/10.1145/3767328.

[14] Li T, Kong L, Yang X, Wang B, Xu J. Bridging modalities: A survey of cross-modal image-text retrieval. Chinese Journal of Information Fusion. 2024;1(1):79-92. Available from: https://doi.org/10.62762/CJIF.2024.361895.

[15] George J. Multimodal sentiment analysis: integrating text, image, and audio. Multimodal Learning Using Heterogeneous Data. 2026:99-115. Available from: https://doi.org/10.1016/B978-0-443-27528-9.00017-6.

[16] Arslan B. Minutiae-Free Fingerprint Recognition via Vision Transformers: An Explainable Approach. Applied Sciences. 2026;16(2):1009. Available from: https://doi.org/10.3390/app16021009.

[17] Zhong D, Li X, Huang Z, Wang S, Yu Z, Hou M, et al. Multi-modal multi-scale representation learning via cross-attention between chest radiology images and free-text reports. Biomedical Signal Processing and Control. 2026;111:108318. Available from: https://doi.org/10.1016/j.bspc.2025.108318.

[18] Van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of machine learning research. 2008;9(11):2579-605. Available from: https://www.jmlr.org/papers/v9/vandermaaten08a.html.

[19] Xu J, Yang L, Li X, Wu T, Tang YY, Wang PSP. Unlocking the Potential of Auxiliary Captions via Dual-Branch Multi-Scale Network for Composed Image Retrieval. International Journal of Pattern Recognition and Artificial Intelligence. 2026;40(02):2554021. Available from: https://doi.org/10.1142/S0218001425540217.

[20] Yang D, Yang L, Wu T, Li X, Tang YY, Wang PSP. Similarity-Guided Denoising Reconstruction for Unsupervised Image Captioning. International Journal of Pattern Recognition and Artificial Intelligence. 2026. Available from: https://doi.org/10.1142/S0218001426590056.