SemRank: Semantic Image Retrieval via Knowledge Graph Filtering and CLIP Embeddings

Sourav Pandey and Ashish Singh Patel

Department of Computer Science and Engineering,
National Institute of Technology Mizoram, Mizoram, India

Content

- Introduction
- Problem Statement
- Related work
- Methodology
- Experimental Setup
- Results and Discussion
- Limitations
- Conclusion

Introduction

- Semantic image retrieval aims to find images based on their meaning and context, rather than simple keyword matching.
- Traditional image retrieval methods often fail to understand **relationships between multiple objects** in a query.
- Vision-language models such as CLIP improve semantic matching but struggle with complex relational reasoning.
- Large-scale image retrieval requires both **high retrieval accuracy** and fast search performance.
- To address these challenges, **SemRank** combines Knowledge Graph (KG) filtering, CLIP embeddings, and FAISS similarity search for efficient and context-aware image retrieval.
- Improves semantic understanding, reduces false positives, and enables scalable image retrieval for complex natural language queries.

Problem Statement

- **Scale vs. Accuracy Trade-off:**
  Real-time retrieval on millions of images demands efficiency, while user intent often involves complex, multi-object relationships.
- **Weak Relational Reasoning:**
  Independent-encoder methods (VSE++, CLIP dual-encoder) are fast but ignore explicit multi-entity constraints, causing false positives.
- **Cross-Attention Doesn't Scale:**
  Transformer-based cross-modal models (SCAN, ViLBERT) reason richly but can't precompute embeddings for large-scale ANN indexing.

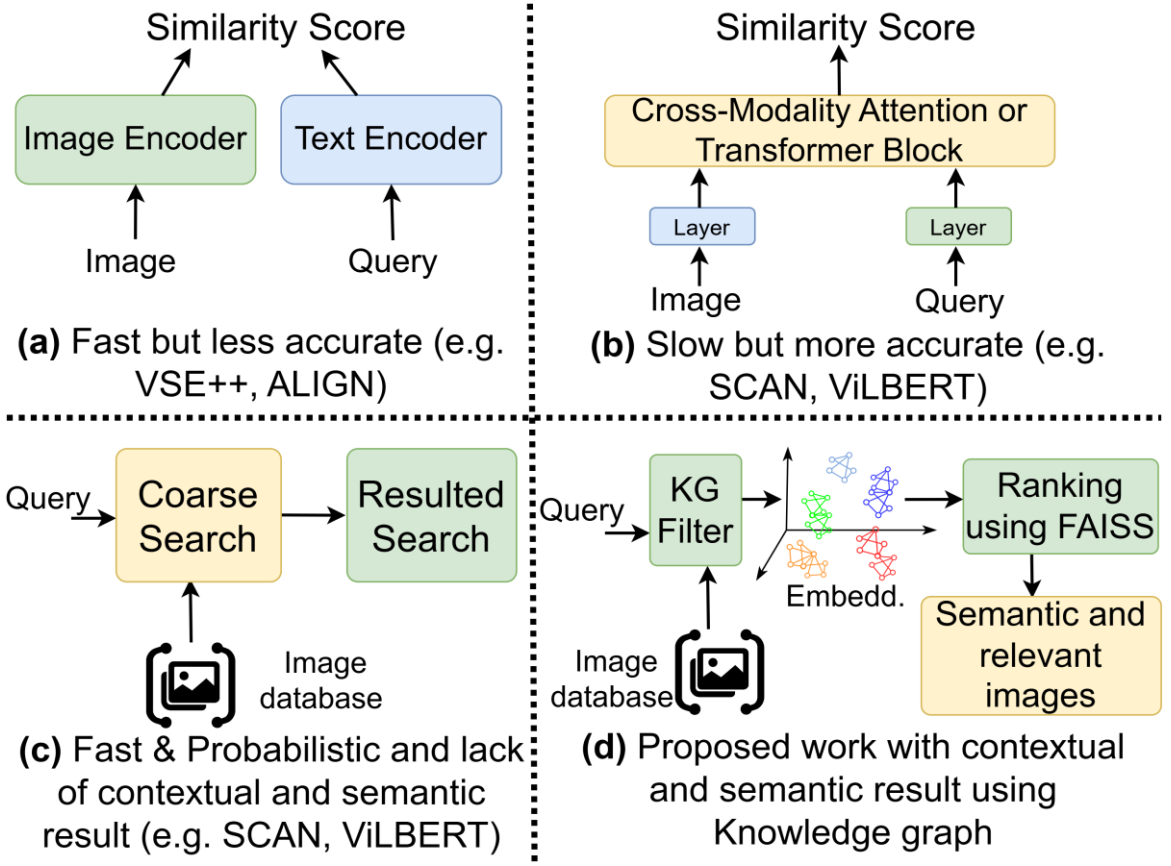


Fig. 1: Comparison of proposed work with the baseline and other related work

Related Work

Category	Representative models	Key Findings	Limitations
Independent Encoders	VSE++ [17], ALIGN [16], CLIP Dual-Encoder, Florence-2	Fast retrieval, shared embedding space, supports offline embedding computation and ANN indexing	Weak relational reasoning; struggles with compositional and multi-object queries
Cross-modal Attention Models	SCAN, ViLBERT [15], LXMERT [13], UNITER [14], OSCAR, BLIP [18]	Fine-grained image-text interaction and strong semantic reasoning	Computationally expensive and cannot efficiently use precomputed embeddings for large-scale retrieval
Probabilistic (Coarse-to-Fine) Retrieval Models	Keyword/Hash Filtering + Cross-attention Re-ranking [19]	Fast initial candidate filtering and better balance between efficiency and accuracy.	Initial filtering is contextually weak. Misses relational semantics
Proposed SemRank	Knowledge Graph + CLIP + FAISS	KG-based semantic filtering, preserves entity relationships, scalable ANN search, improved retrieval accuracy	Requires offline Knowledge Graph construction

Methodology

1. Text Preprocessing

- Accepts a natural language query as input.
- Performs tokenization and lemmatization using **spaCy**.
- Extracts keywords, named entities, POS tags, and dependency relations.
- Identifies important objects and actions from the query.

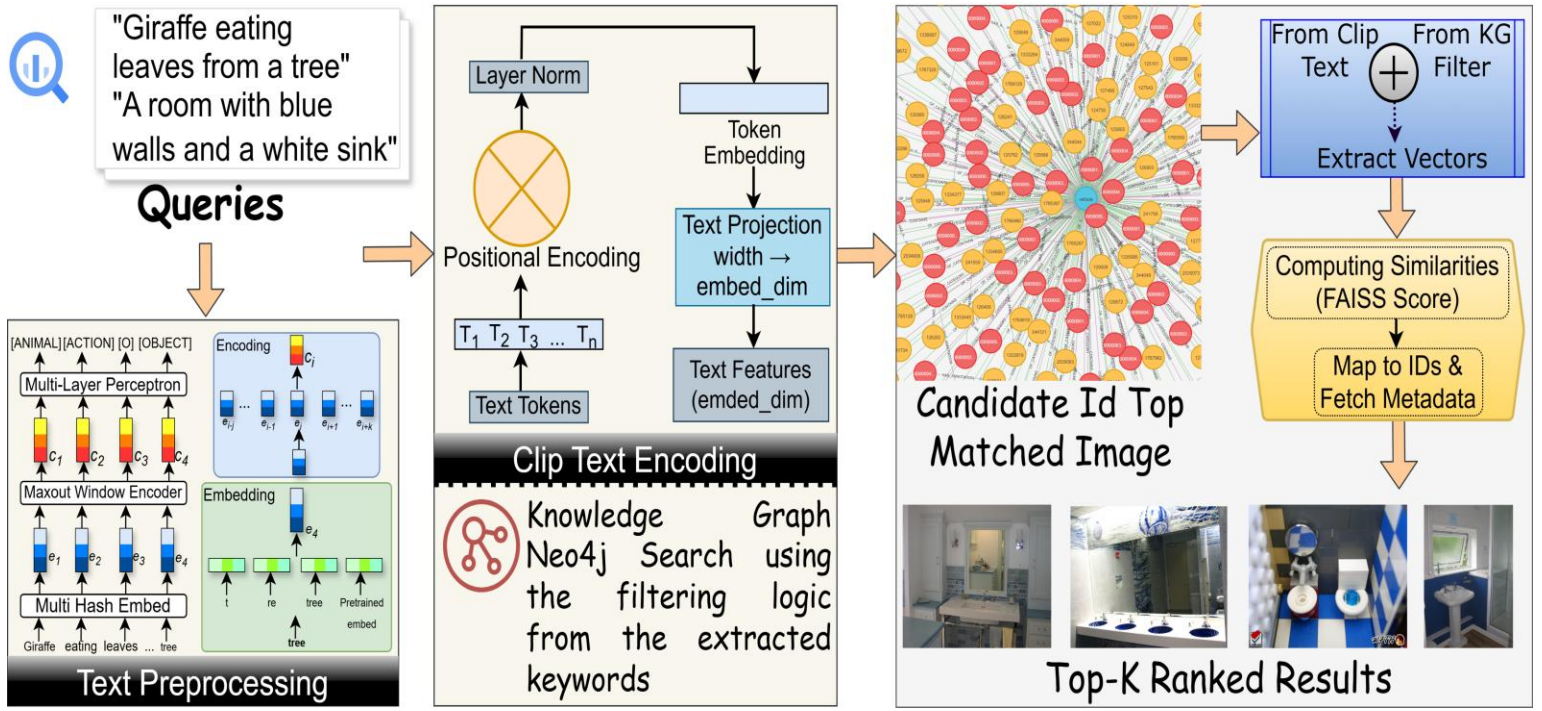


Fig. 2: The Proposed SemRank Architecture for the semantic search using the KG

2. Knowledge Graph

- KG is constructed from the image dataset annotations.
- Represents images, categories, and actions as graph nodes.
- Relationships such as *has_category* and *has_action* connect the nodes.
- The complete graph is stored in the Neo4j Graph Database.
- This graph provides semantic reasoning before neural retrieval.

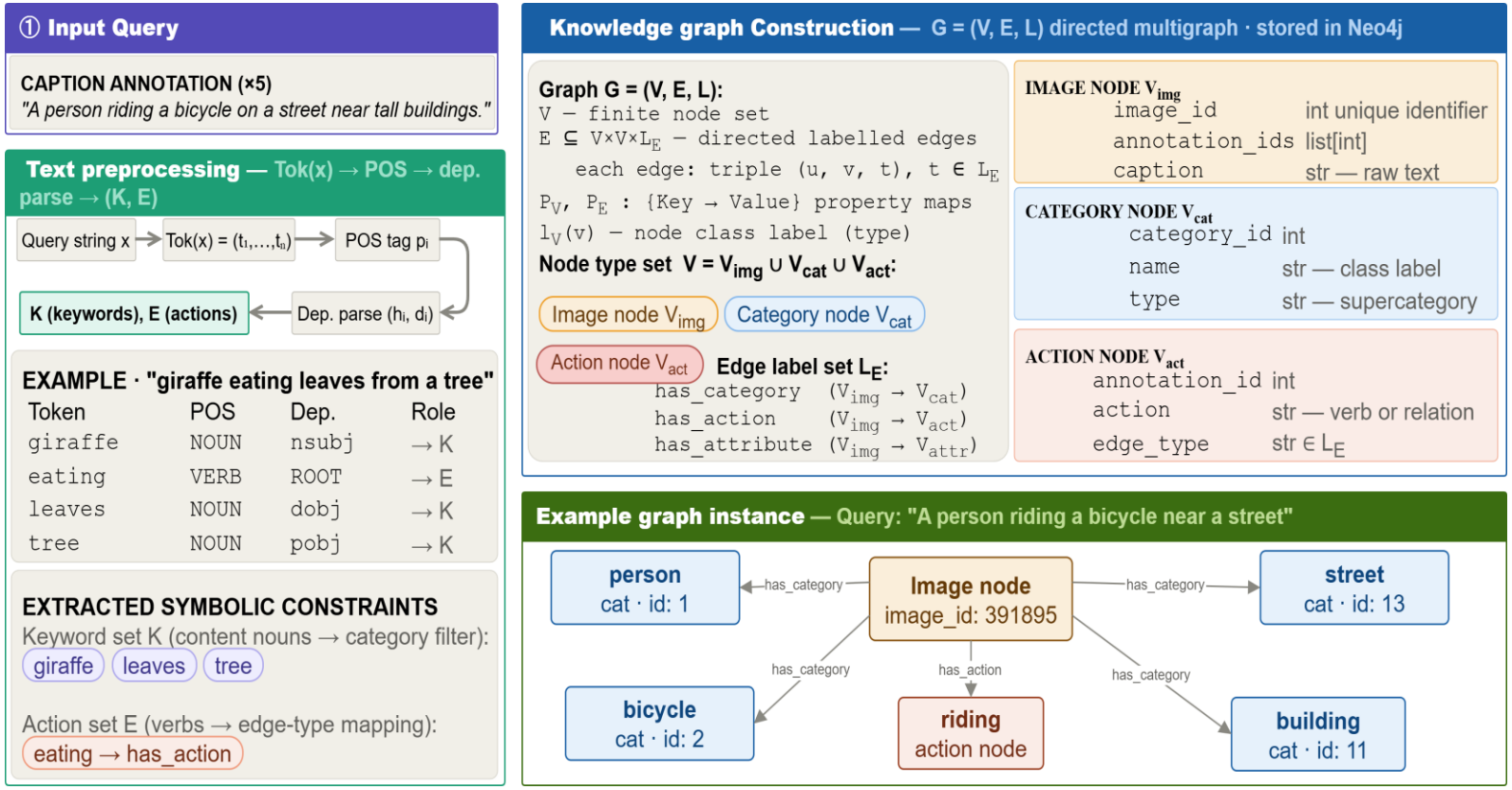


Fig. 3: Knowledge Graph construction framework in SemRank

3. Query Subgraph

- The extracted keywords are converted into a **query subgraph**.
- Nodes represent important entities present in the query.
- Edges represent semantic relationships between these entities.
- The query subgraph captures both **objects** and **their interactions**.
- The system searches for matching subgraphs inside the KG.
- Only structurally matching images are selected as candidates.

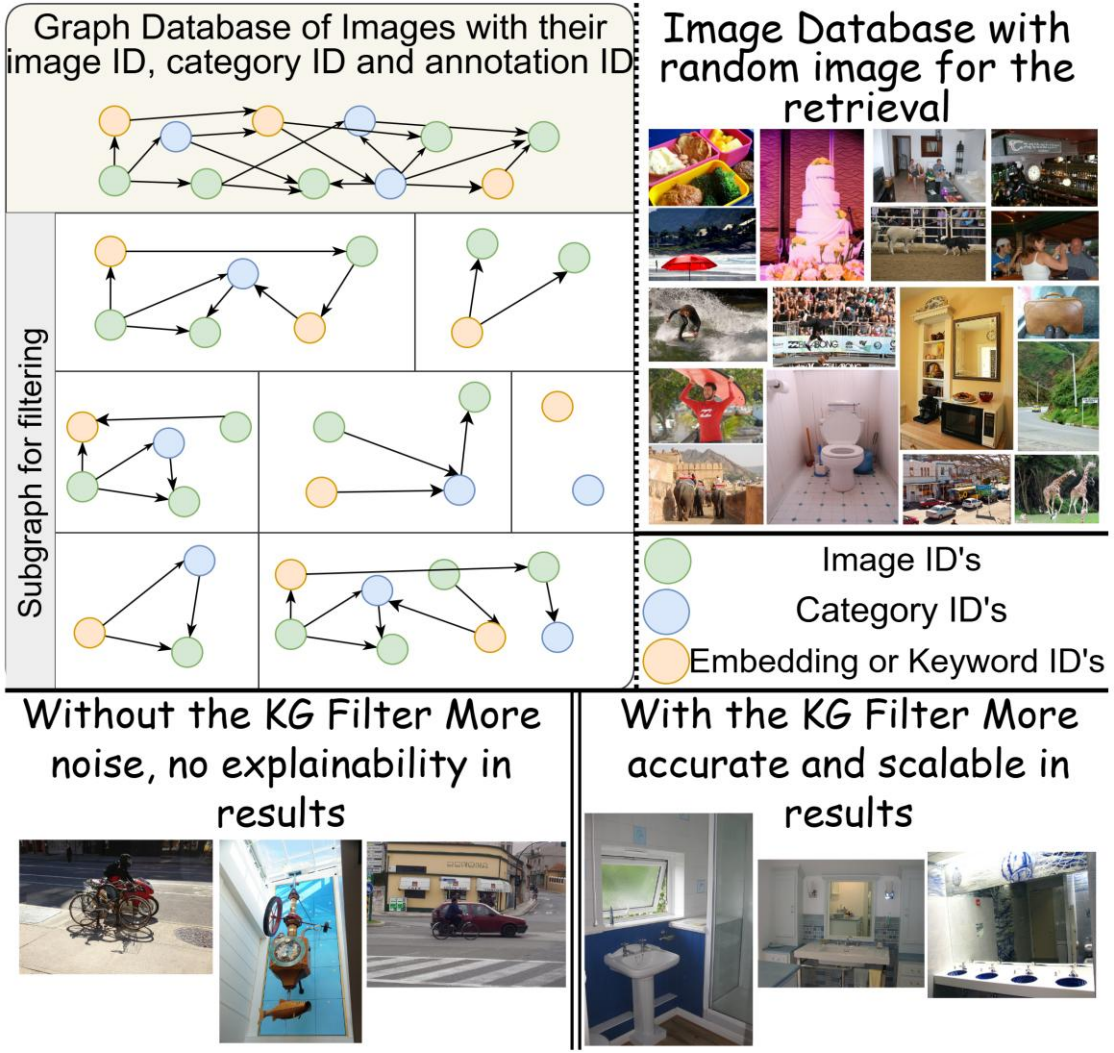


Fig. 4: Query subgraph representation within the KG for semantic image search

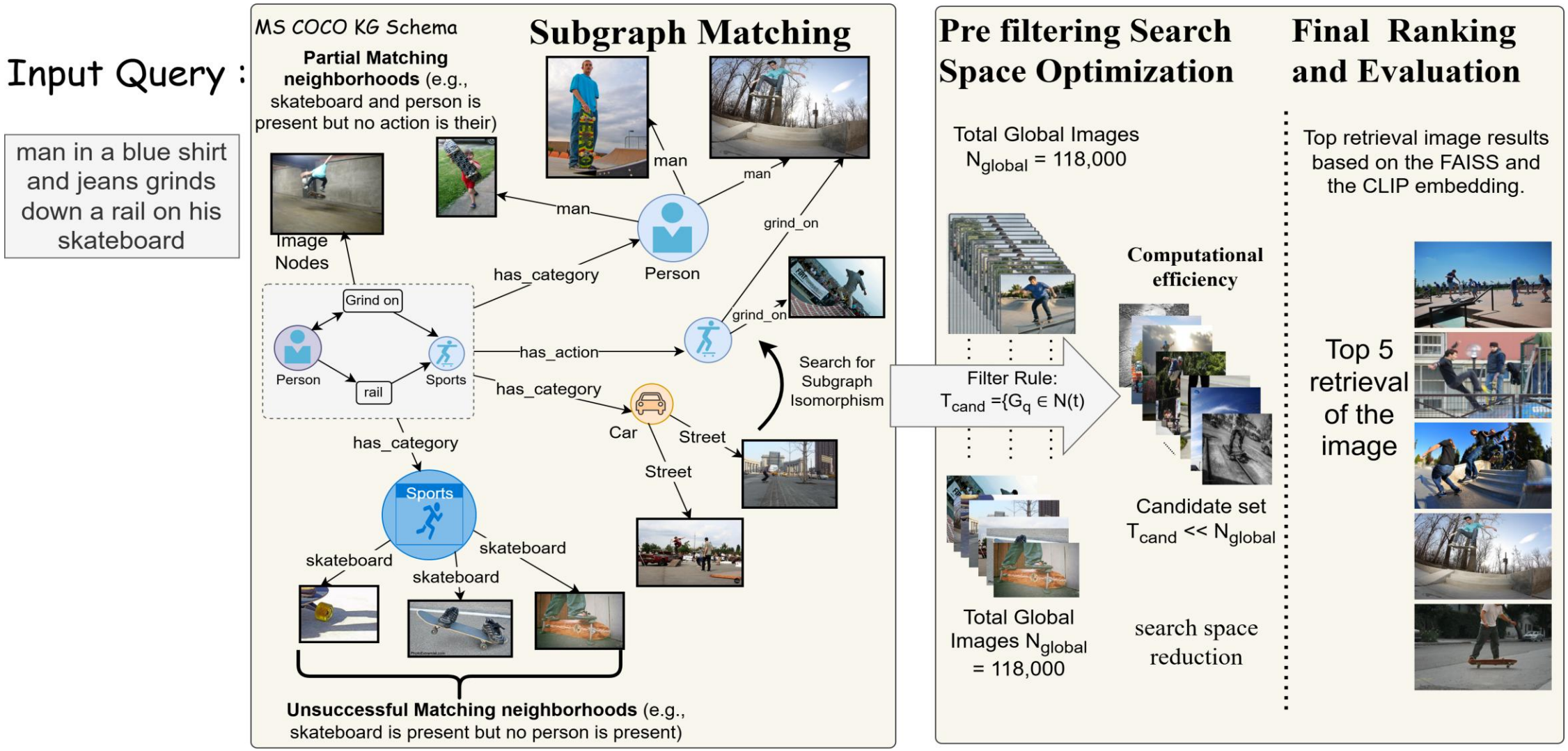


Fig. 5: Query Subgraph Workflow: illustrating the complete operation and foundational importance of the Query Subgraph (G_q) in proposed SemRank image retrieval framework

4. Semantic Search

- Candidate images obtained from the Knowledge Graph are semantically ranked.
- CLIP generates embedding vectors for both the query and candidate images.
- All embeddings are normalized before similarity computation.
- **FAISS** performs Approximate Nearest Neighbor search for efficient retrieval.
- Cosine similarity scores are calculated between query and image embeddings.
- Images are ranked according to similarity scores.
- The framework returns the **Top-K most semantically relevant images**.

Experiment

➢ Dataset

- Flickr30K [1] and MS-COCO [2] are widely used for semantic search because they pair real images with natural, human written captions.
- Flickr30K contains about 31k everyday pictures with five short captions each. MS-COCO is much larger, with over 118k diverse and complex scenes, each captioned five times.

➢ Experimental Setup

- *Evaluation Metrics:*

  To evaluate the performance of the proposed work, the Recall@K (R@K = 1, 5 and 10) is employed, which measures how well a semantic image retrieval system can find the correct image within its top-K search results.

- Neo4j serves as the symbolic filter, reducing the search space by selecting images connected through relevant objects, attributes, and relationships.

Results and Discussion

1. Comparison with State-of-the-art Methods

- Compared SemRank with **12 state-of-the-art text-to-image retrieval methods** on the **Flickr30K** and **MS-COCO** datasets shown in Table 1.
- Performance is evaluated using **Recall@1 (R@1, 5, 10)**.
- SemRank consistently achieves the **highest retrieval accuracy** across all evaluation metrics.
- On **Flickr30K**, SemRank achieved **65.4% (R@1), 88.7% (R@5), and 93.7% (R@10)**.
- On **MS-COCO**, SemRank achieved **68.8% (R@1), 93.5% (R@5), and 97.5% (R@10)**.
- The improvement demonstrates the effectiveness of KG filtering combined with CLIP embeddings for semantic image retrieval.

Table 1: Comparison of text to image retrieval performance on Flickr30K and MS-COCO datasets.

Methods	Flickr30K			MS-COCO		
	R@1	R@5	R@10	R@1	R@5	R@10
IMRAM [3]	16.6	41.8	54.1	39.6	69.1	79.8
VSRN [4]	25.5	47.6	58.6	40.5	70.6	81.1
SCAN [5]	35.5	65.0	75.2	38.6	69.3	80.4
SGR [6]	40.2	66.8	75.3	23.5	58.9	75.1
SAF [6]	49.7	73.6	78.0	57.8	86.4	91.9
NAAF [7]	52.5	79.3	86.2	58.5	86.0	91.3
GraDual [8]	60.4	86.7	92.0	65.3	91.9	96.4
CVSE [9]	61.9	87.1	92.0	59.9	89.4	95.2
GPO [10]	61.4	85.9	91.5	42.4	72.7	83.1
ESC [11]	59.1	83.8	89.1	64.8	90.7	96.0
MKG [12]	63.2	87.4	92.3	65.6	91.5	96.2
MKVSE [12]	63.2	87.4	92.3	66.4	92.1	96.6
Proposed Work (SemRank)	**65.4**	**88.7**	**93.7**	**68.8**	**93.5**	**97.5**

2. Ablation Study

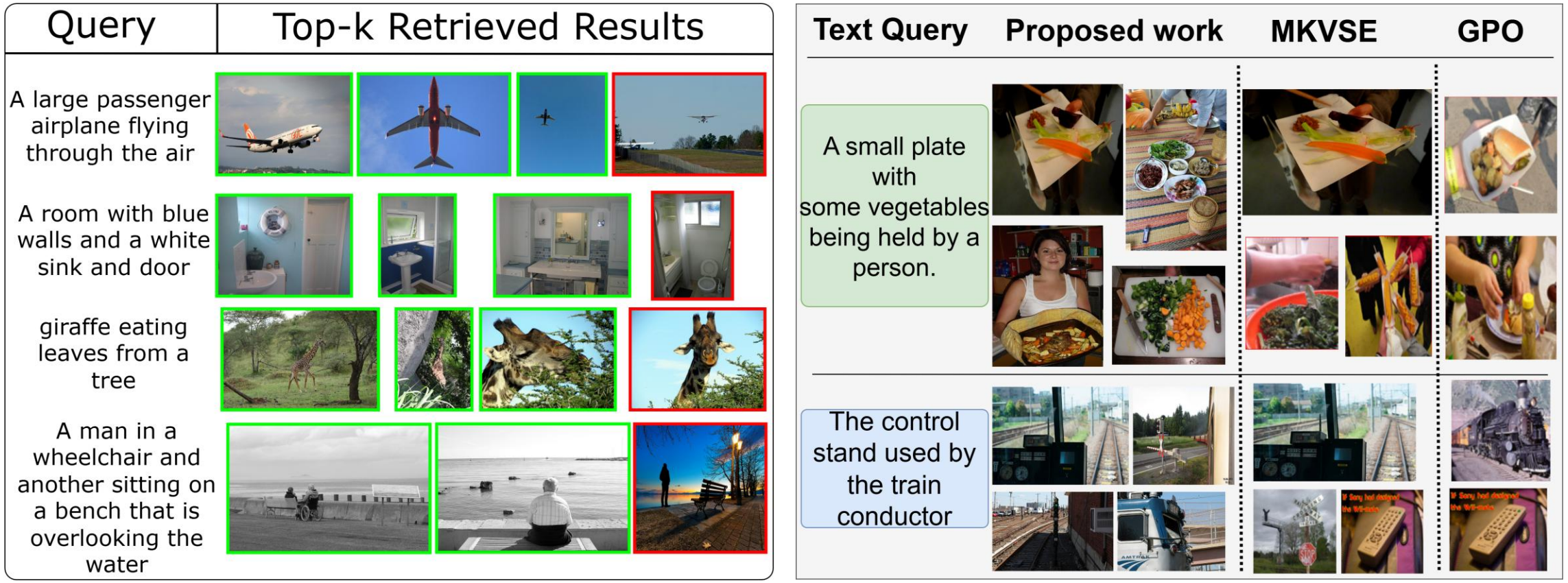
1. Parameter Analysis and Impact of KG:

- Baseline models using **GloVe + ResNet18** & **MiniLM + EfficientNet** achieve relatively lower retrieval performance.
- Integrating KG filtering with CLIP significantly improves retrieval accuracy across all Recall@K metrics.
- Larger CLIP backbones provide richer semantic representations, resulting in more accurate image retrieval.

Table 2: Performance comparison of proposed work showing improvements from integrating the KG over the Baseline.

Methods	**Architecture**	**MS-COCO**		
		R@1	**R@5**	**R@10**
	(Without KG Filter)			
Baseline	GloVe + ResNet18	25.0	30.1	42.2
Baseline	MiniLM + EfficientNet	37.8	53.5	62.3
	(With KG Filter)			
Proposed work	CLIP ViT-B/32 + FAISS	63.7	87.0	90.1
Proposed work	CLIP ViT-L/14 + FAISS	68.8	93.5	97.5
Proposed work	CLIP ViT-B/32	35.0	70.0	74.4

2. Qualitative Results:



(a) (b)

Fig. 6: (a) Semantic search results of the proposed method with the most relevant image and semantically/ contextually rich images. (b) Qualitative comparison of the proposed work with MKVSE [12] and GPO [10] showing semantically and contextually relevant results on MS-COCO dataset.

3. Pairwise Cosine Similarity Analysis of Candidate Images:

- Figure 7 presents a **pairwise cosine similarity heatmap** of the Top-K retrieved images.
- All candidate images first satisfy the KG **constraints**, ensuring semantic relevance to the query.
- The heatmap visualizes the **similarity relationships** among the retrieved image embeddings in the CLIP latent space.
- Distinct similarity clusters indicate that the retrieved images represent **different semantic sub-concepts** while remaining relevant to the query.
- The results demonstrate that combining KG **filtering** with **CLIP embeddings** improves both retrieval diversity and semantic ranking quality.

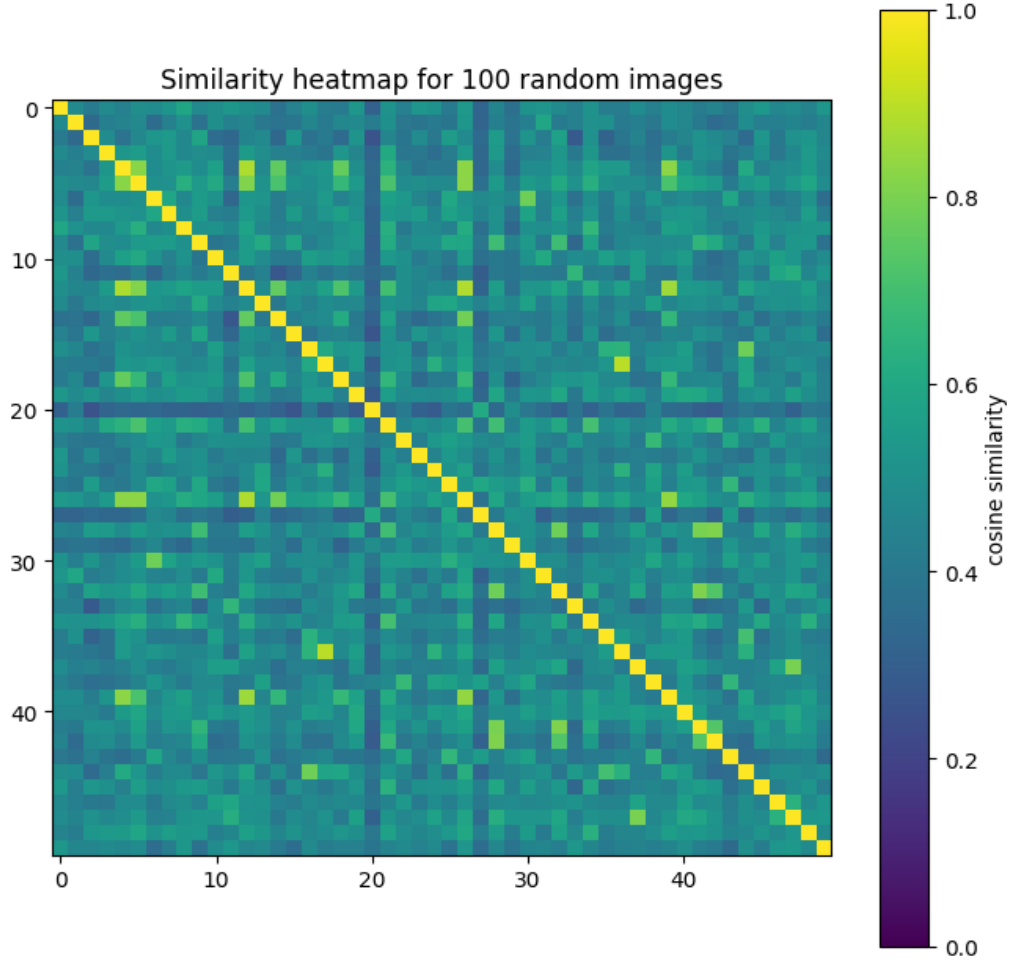


Fig. 7: Results of the proposed work with the most relevant image with semantic and contextually rich images

Limitations

- Complex or ambiguous queries may be incorrectly parsed by the NLP pipeline, leading to overly restrictive or incomplete candidate filtering.
- The system relies on CLIP embeddings, which may still miss fine-grained relationships, spatial details, or domain-specific semantics.
- Evaluation is limited to **MS-COCO** and **Flickr30K**, which may not fully represent real-world or domain-specific image collections.
- Overly restrictive or rare queries can produce empty candidate sets, requiring a fallback strategy that may reduce semantic precision.

Conclusion

- **SemRank** introduces a hybrid semantic image retrieval framework by integrating KG filtering, CLIP embeddings, and FAISS similarity search.
- KG effectively captures **multi-entity relationships** and removes semantically irrelevant images before ranking.
- Experimental results on **Flickr30K** and **MS-COCO** demonstrate **higher Recall@K scores** and reduced false positives compared to existing methods.
- The proposed framework achieves an effective balance between **semantic accuracy** and **retrieval efficiency**, making it suitable for large-scale semantic image retrieval applications.

Future Work

- Develop dynamic Knowledge Graph construction for continuously evolving datasets.
- Improve multimodal grounding to better understand complex image-text relationships.
- Explore adaptive retrieval mechanisms for handling more complex and context-aware user queries.

References

1. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015)
2. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
3. Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., Han, J.: 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. pp. 12655–12663 (2020)
4. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4654–4662 (2019)
5. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV). pp. 201–216 (2018)
6. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI conference on artificial intelligence. vol. 35, pp. 1218–1226 (2021)
7. Zhang, K., Mao, Z., Wang, Q., Zhang, Y.: Negative-aware attention framework for image-text matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15661–15670 (2022)
8. Long, S., Han, S.C., Wan, X., Poon, J.: Gradual: Graph-based dual-modal representation for image-text matching. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 3459–3468 (2022)
9. Wang, H., Zhang, Y., Ji, Z., Pang, Y., Ma, L.: Consensus-aware visual-semantic embedding for image-text matching. In: European conference on computer vision. pp. 18–34. Springer (2020)
10. Chen, J., Hu, H., Wu, H., Jiang, Y., Wang, C.: Learning the best pooling strategy for visual semantic embedding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15789–15798 (2021)
11. Shi, H., Liu, M., Mu, X., Song, X., Hu, Y., Nie, L.: Breaking through the noisy correspondence: A robust model for image-text matching. ACM Transactions on Information Systems 42(6), 1–26 (2024)

Continue…

12. Feng, D., He, X., Peng, Y.: Mkvse: Multimodal knowledge enhanced visual-semantic embedding for image-text retrieval. ACM Transactions on Multimedia Computing, Communications and Applications 19(5), 1–21 (2023)
13. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv preprint arXiv:1908.07490 (2019)
14. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: European conference on computer vision. pp. 104–120. Springer (2020)
15. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic vision linguistic representations for vision-and-language tasks. Advances in neural information processing systems 32 (2019)
16. Wehrmann, J., Kolling, C., Barros, R.C.: Adaptive cross-modal embeddings for image-text alignment. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 12313–12320 (2020)
17. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612 (2017)
18. 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. pp. 12888–12900. PMLR (2022)
19. Han, J., Ngan, K.N., Li, M., Zhang, H.: Learning semantic concepts from user feedback log for image retrieval. In: 2004 IEEE International Conference on Multimedia and Expo (ICME). vol. 2, pp. 995–998. IEEE (2004)

Supplemental Material Statement:

The source code for the architecture is available on the repository at the following link:
https://github.com/Sourav544/SemRank-Semantic-search.git

Acknowledgment

Travel Support

This presentation at ICPR was made possible with travel support from the **Anusandhan National Research Foundation (ANRF)**, Government of India, under its **International Travel Support (ITS) Scheme**.

Thank you!