Proceedings of the 4th International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/2026.36913

Benchmarking Pre-Trained Vision-Language Models for Bidirectional Image-Text Retrieval on MS-COCO: BERT+ResNet, CLIP, and BLIP

Tixun Wang

School of Mechanical Engineering, Beijing Institute of Technology, Beijing, China
lingfeng199005@gmail.com

Abstract. Bidirectional image-text retrieval evaluates whether a model can align visual and textual representations for both text-to-image and image-to-text search. This paper presents a controlled benchmark on the Microsoft Common Objects in Context (MS-COCO) Karpathy split, using the same 5,000-image test gallery, 25,000 captions, similarity-matrix layout, and Recall@K/Recall Sum (RSUM) evaluation for four settings: a frozen Bidirectional Encoder Representations from Transformers (BERT)+ Residual Network (ResNet) dual encoder, Contrastive Language-Image Pre-training (CLIP) Vision Transformer Base (ViT-B)/32 zero-shot, CLIP ViT-B/32 fine-tuned on COCO, and Bootstrapping Language-Image Pre-training (BLIP) image-text contrastive only (ITC-only). The results show a clear gap between independently combined unimodal encoders and pre-trained vision-language models. BERT+ResNet obtains RSUM 148.740, while CLIP zero-shot reaches 361.868. Fine-tuning CLIP improves RSUM to 423.496, while the COCO-trained BLIP checkpoint evaluated with ITC-only scoring achieves the strongest result in this experiment, with RSUM 489.588. Image-to-text retrieval scores higher under COCO's asymmetric protocol, while longer captions are associated with higher text-to-image recall. Qualitative errors center on attributes, counting, spatial relations, and plausible false negatives. These results characterize the evaluated checkpoints and training conditions on this benchmark; they do not establish a universal model hierarchy.

Keywords: bidirectional image-text retrieval, vision-language models, MS-COCO, contrastive learning, retrieval benchmarking

1. Introduction

Bidirectional image-text retrieval has two directions. For a text query, the system ranks candidate images; in the reverse direction, it ranks descriptions for an image. Unlike classification or caption generation, the output is a candidate list, and success depends on placing matched visual and textual representations near each other while separating mismatched pairs. This shared setup makes it possible to compare models with substantially different architectures and training objectives [1]. MS-COCO was selected for this study because its everyday scenes are paired with several human-written captions, allowing both retrieval directions to be evaluated under controlled conditions [2].

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.36913

Image-text retrieval has moved from separately trained visual and text encoders toward models pre-trained jointly on large image-text collections. Earlier systems extracted features in each modality and then learned a common space for ranking. CLIP instead learns the two representations through contrastive supervision and can be evaluated without task-specific training [3]. BLIP adds bootstrapped caption learning and combines contrastive retrieval with an image-text matching objective [4]. Comparing these approaches makes pre-training strategy, as well as network architecture, a likely source of differences in retrieval behavior.

Scores reported in separate studies cannot be compared without accounting for their experimental setup. They vary with the data split, candidate gallery size, checkpoint, fine-tuning procedure, and the use of contrastive scoring alone or matching-based re-ranking. This study therefore asks two questions: (RQ1) under the same MS-COCO Karpathy split, candidate gallery, and Recall@K/RSUM evaluation, how do BERT+ResNet, CLIP zero-shot, CLIP fine-tuned, and BLIP ITC-only differ in bidirectional retrieval performance? (RQ2) How do fine-tuning status, retrieval direction, caption length, inference cost, and possible false negatives affect the interpretation of these differences? The design controls these choices so that the study does not claim a universal best model for every retrieval scenario.

Three contributions follow from the experiment. First, this paper presents a single Karpathy-split pipeline so that BERT+ResNet, CLIP, and BLIP score against the same gallery and similarity-matrix implementation, with identical Recall@K and RSUM calculations. Second, the pipeline evaluates BERT+ResNet, zero-shot CLIP, COCO-fine-tuned CLIP, and BLIP with ITC scoring only. Within these runs, BLIP produced the largest RSUM, while adapting CLIP to COCO improved it markedly over zero-shot use. Finally, the analysis examines the sources of the scores by separating retrieval direction, caption length, runtime, and qualitative error type, and by considering limitations of the benchmark. Together, these contributions constitute a controlled MS-COCO comparison, not a general ranking of all vision-language models.

2. Related work

2.1. Image-text retrieval and benchmarks

Image-text retrieval evaluates whether paired visual and textual items can be ranked correctly in a shared semantic space. This study uses the task as a controlled comparison setting rather than as a direct comparison of published numbers from different protocols. The retrieval task is evaluated in both text-to-image and image-to-text directions. MS-COCO provides everyday images with multiple human-written captions. This paper follows the widely used Karpathy train, validation, and test partition; the same 113,287/5,000/5,000 image allocation is adopted in recent journal research on image-text retrieval [2, 5]. Recall@K and RSUM are retained because they are widely used in cross-modal retrieval, but they are treated as benchmark-specific indicators rather than complete evidence of visual-language understanding. Keeping the split, gallery, and metrics fixed is therefore central to the paper's design.

2.2. Pre-trained vision-language models

BERT+ResNet is used here as a transparent dual-encoder baseline. The image side relies on ResNet-50 features, while the text side relies on BERT-base representations; both belong to established Convolutional Neural Network and transformer-based language-model families discussed in the corresponding architectural reviews [6, 7]. This baseline is not presented as a state-of-the-art system.

DOI: 10.54254/2755-2721/2026.36913

Its role is to make the effect of explicit vision-language pre-training easier to observe when compared with CLIP and BLIP under the same retrieval pipeline.

CLIP is the main contrastive pre-trained baseline in this paper. Because its visual and textual encoders are trained jointly, the same checkpoint can be evaluated by directly comparing image and text embeddings [3]. The paper separates zero-shot CLIP from COCO fine-tuned CLIP because they answer different questions: zero-shot evaluation tests transfer, while fine-tuning tests benchmark adaptation. Broader retrieval literature treats adaptation settings as an important source of score variation, and CLIP adaptation work further shows that lightweight tuning can change downstream behavior [8, 9].

BLIP is included because it makes the scoring design more explicit. Its retrieval model can provide image-text contrastive similarity and can also support image-text matching, but these are not the same evaluation condition [4]. For a fair main comparison with BERT+ResNet and CLIP, this paper reports BLIP with ITC-only cosine similarity. Image-Text Matching (ITM) re-ranking would add pairwise matching computation after the embedding search and would therefore be better treated as a secondary setting, not as the same baseline. This distinction also matters for efficiency: prior CLIP-based cross-modal retrieval work shows that feature dimensionality and retrieval design can materially affect storage and retrieval cost [10].

2.3. Benchmark limitations and robustness

The captions attached to an MS-COCO image cannot enumerate every image-text relation that a person might consider acceptable. Consequently, a caption may describe a retrieved image reasonably well even when that image is counted as a negative. Existing reviews note that dataset construction and metric choice affect image-text retrieval benchmarks [1, 8]. Experiments with partially mismatched pairs also demonstrate sensitivity to imperfect alignment [11].

To limit this problem, the split, candidates, and evaluation code were kept unchanged between models. The comparison also did not rely on RSUM alone: direction-specific scores, zero-shot versus fine-tuned CLIP, timing records, and retrieved examples were considered together. These choices follow concerns raised in prior work, while the interpretation itself comes from the results observed here.

3. Methodology

The executed workflow is summarized in Fig. 1. Starting with the selected MS-COCO metadata, separate processors generated image and text embeddings for each model setting. Each pair of embedding matrices was then passed to the common Recall@K/RSUM evaluator, and the returned rankings supplied the cases used in the error review.

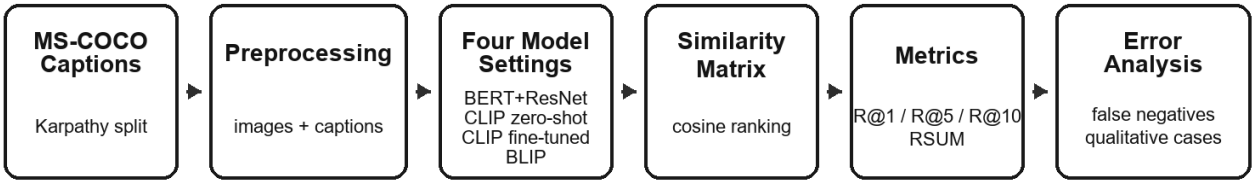

Figure 1. Experimental workflow for the bidirectional image-text retrieval comparison (photo credit: original)

DOI: 10.54254/2755-2721/2026.36913

3.1. Data preprocessing

Experiments used COCO 2014 Captions / MS-COCO Captions and the Karpathy split. Combining train and restval yielded 113,287 training images; after applying the five-caption rule, they were paired with 566,435 captions. Validation and test each contained 5,000 images and 25,000 captions. Some source records provided more than five captions, so the metadata script kept the first five per image and left out 332 additional entries in the JavaScript Object Notation (JSON) file. This gave 25,000 text queries and 5,000 image queries on the test set. The same image IDs, caption IDs, candidate lists, and evaluation implementation were used for every model.

Data preparation was executed with the project metadata script. It parsed the Karpathy JSON file, resolved COCO train2014 and val2014 image paths, and generated image_index.parquet, caption_index.parquet, pairs.parquet, split_summary.json, missing_files_report.csv, and duplicate_pairs_report.csv. The final split summary reported zero missing image files, zero empty captions, zero non-UTF-8 captions, zero duplicate image-caption pairs, and all checks passed. Each caption retained a stable caption_id and positive_image_id so that ground-truth matches could be recovered during evaluation.

Images were loaded as RGB. For BERT+ResNet, the image transform followed torchvision's ResNet50 IMAGENET1K_V2 preprocessing, including model-specific resizing and cropping, followed by ImageNet normalization. For CLIP and BLIP, Hugging Face CLIPProcessor and BlipProcessor were used so that image preprocessing and text tokenization matched the selected checkpoints. Text inputs were padded and truncated through the corresponding tokenizer or processor. For each model setting, image and text embeddings were cached, L2-normalized, and saved with model name, split, checkpoint, and embedding dimension.

3.2. Model settings

Four model settings were evaluated. The BERT+ResNet baseline used torchvision ResNet-50 pretrained on ImageNet and bert-base-uncased. CLIP zero-shot used openai/clip-vit-base-patch32 without COCO training. CLIP fine-tuned started from the same CLIP checkpoint and adapted selected parameters on the COCO training split. BLIP used the COCO-trained Salesforce/blip-itm-base-coco checkpoint. No additional BLIP training was performed in this experiment. The checkpoint was evaluated using only its ITC image and text embeddings with cosine similarity, making its scoring procedure directly comparable with the other embedding-based retrieval settings. The ITM head and ITM re-ranking were disabled and are treated as possible secondary extensions rather than part of the main comparison.

For the BERT+ResNet baseline, ResNet-50 and BERT-base were frozen. Two single-layer linear projection heads mapped the image and text representations to the shared embedding space: 2048 -> 512 for image features and 768 -> 512 for text features. No hidden layer or nonlinear activation was used, and the projected outputs were L2-normalized before cosine-similarity retrieval.

3.3. Training and retrieval strategy

For BERT+ResNet, only the two projection heads were optimized. Each image contributed one deterministically sampled caption per epoch rather than all five captions simultaneously. With batch size 64 and drop_last enabled, each epoch processed 113,280 examples in 1,770 steps; the final seven shuffled images were omitted from that epoch. Training used symmetric Information Noise Contrastive Estimation (InfoNCE) with temperature 0.07, Adam with decoupled Weight decay

DOI: 10.54254/2755-2721/2026.36913

(AdamW), learning rate 1e-4, weight decay 1e-4, mixed precision on Compute Unified Device Architecture, and seed 42. Validation RSUM was checked after every epoch with an early-stopping patience of 2. Training reached the maximum of 10 epochs without triggering early stopping, and the epoch-9 checkpoint achieved the highest validation RSUM and was used for final testing.

CLIP fine-tuning used symmetric InfoNCE with the learned CLIP logit scale, batch size 64, AdamW, mixed precision, seed 42, and validation-based checkpoint selection. As in the BERT+ResNet training procedure, each image contributed one deterministically sampled caption per epoch, and drop_last produced 113,280 training examples in 1,770 steps per epoch. Most CLIP parameters were frozen; the trainable groups were visual_projection, text_projection, logit_scale, and the last two Transformer layers of both the vision and text encoders. This made 21,135,873 of 151,277,313 parameters trainable. The learning rate for the unfrozen encoder layers was 1e-6, while the projection and logit-scale learning rate was 1e-5. Weight decay 1e-4 was applied to the encoder and projection groups, whereas logit_scale used no weight decay. Training reached the maximum of five epochs without triggering the early-stopping rule with a patience of 2, and epoch 5 produced the highest validation RSUM. CLIP zero-shot was evaluated without COCO adaptation. BLIP also received no additional training in this experiment, although its selected checkpoint had already been trained on COCO.

At test time, every model produced one image embedding matrix and one caption embedding matrix. Both CLIP settings and BERT+ResNet produced 512-dimensional embeddings, while BLIP ITC-only produced 256-dimensional embeddings. The full test similarity matrix had shape 25,000 x 5,000, with rows corresponding to captions and columns corresponding to images. Cosine similarity was computed after L2 normalization. Text-to-image retrieval ranked 5,000 images for each caption query, and image-to-text retrieval ranked 25,000 captions for each image query. Embedding extraction batch sizes were 64 images / 256 texts for CLIP zero-shot and BLIP, and 128 images / 256 texts for BERT+ResNet and fine-tuned CLIP. Experiments were run on Colab/Linux with Python 3.12.13, PyTorch 2.11.0+cu128, torchvision 0.26.0+cu128, transformers 5.12.1, and an NVIDIA L4 GPU with 22.03 GB memory.

3.4. Evaluation metrics and analysis

The main metrics were Recall@1, Recall@5, and Recall@10 for both retrieval directions. RSUM was computed as the sum of the six Recall@K scores. Mean rank and median rank were also saved because Recall@K only indicates whether a positive item appears within fixed cutoffs. The shared evaluation script used the caption_by_image similarity layout. For text-to-image retrieval, rank was computed as one plus the number of candidate images with similarity strictly greater than the annotated positive image. For image-to-text retrieval, the highest similarity among an image's five ground-truth captions was first selected, and rank was computed using the same strictly-greater rule. Candidates tied with the positive score therefore did not worsen the positive item's rank. Before the main runs, the evaluation code passed a synthetic unit test covering rank-1, rank-5, outside-top-10, and five-caption image-to-text cases.

This study did not treat Recall@K as a complete measure of semantic understanding. Since MS-COCO annotations are incomplete, a retrieved item may be plausible even when it is not the designated positive. The quantitative scores were therefore read alongside caption-length groups, inference costs, and retrieved examples. In the qualitative review, cases were separated into object-level matches, attribute errors, counting or spatial-relation errors, and plausible unannotated matches. This separation helps distinguish benchmark artifacts from clearer model failures.

DOI: 10.54254/2755-2721/2026.36913

4. Results and analysis

4.1. Overall retrieval performance

The analysis begins with the four overall scores and then considers retrieval direction, CLIP adaptation, caption length, runtime, and the inspected errors. Fig. 2 places RSUM and direction-specific Recall@1 in one view.

All four similarity matrices have 25,000 caption rows and 5,000 image columns, and one evaluation script produced Recall@K, RSUM, mean rank, and median rank from each of them. This common setup allows a direct within-experiment comparison. Fig. 2 summarizes the principal metrics used in the following comparisons.

Fig. 2 shows a clear performance ordering. BERT+ResNet is the weakest setting, with RSUM 148.740 and the highest mean rank. CLIP zero-shot raises RSUM to 361.868, showing the benefit of large-scale vision-language pre-training without COCO fine-tuning. Fine-tuned CLIP reaches RSUM 423.496, while BLIP ITC-only achieves the highest overall score, RSUM 489.588, in this specific checkpoint and scoring setting.

Fig. 2 also separates Recall@1 by retrieval direction. Every model achieves a higher Image-to-Text (I2T) R@1 than Text-to-Image (T2I) R@1. This directional gap partly reflects the asymmetric evaluation structure: I2T uses the best-ranked match among five positive captions within a 25,000-caption pool, whereas T2I uses one positive image within a 5,000-image pool. Because both the number of positives and the candidate-pool size differ, the observed gap should not be attributed to a single factor. Therefore, RSUM is useful as a compact summary, but the two directions should still be interpreted separately.

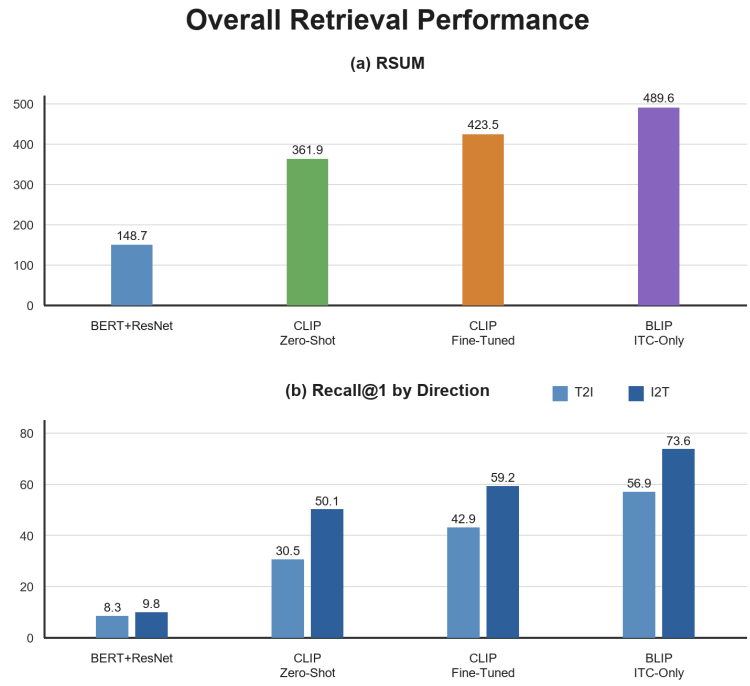

Figure 2. Combined overall retrieval performance: (a) RSUM comparison and (b) Recall@1 by retrieval direction (photo credit: original)

DOI: 10.54254/2755-2721/2026.36913

4.2. Effect of CLIP fine-tuning

The two CLIP settings are compared next because they isolate the effect of supervised adaptation from the choice of model family. Both use the same OpenAI/clip-vit-base-patch32 checkpoint and the same evaluation script. Fine-tuning improves T2I R@1 from 30.480 to 42.944 and I2T R@1 from 50.060 to 59.160. It also improves the wider cutoffs, increasing T2I R@10 from 66.852 to 79.804 and I2T R@10 from 83.520 to 89.180. These gains show that COCO training changes the ranking behavior substantially, even when the retrieval pipeline and gallery are unchanged.

The improvement does not remove the gap between fine-tuned CLIP and BLIP ITC-only: BLIP exceeds fine-tuned CLIP by 66.092 RSUM points. However, the BLIP checkpoint had already been trained on COCO, whereas CLIP was adapted using the training procedure implemented in this experiment. The difference, therefore, reflects checkpoint training history, model family, scoring mechanism, and evaluation protocol together and cannot be attributed to architecture alone. This evidence is specific to the evaluated checkpoints and protocol and does not establish that BLIP is universally superior.

4.3. Caption length analysis

Caption length was analyzed for text-to-image retrieval because each caption is a query. Caption length was measured from the shared caption_text field using the same regular-expression tokenizer for all models. The tokenizer lowercased each caption and applied the pattern, counting both lexical items and punctuation marks as tokens. The thresholds were derived from the one-third and two-thirds quantiles of the 25,000 test-caption lengths rather than being fixed in advance. These quantiles yielded boundaries of 10 and 12 tokens: the short group contained 10,820 captions with a mean length of 9.344 tokens, the medium group contained 8,425 captions with a mean of 11.421, and the long group contained 5,755 captions with a mean of 14.865. Longer groups were associated with higher T2I R@1 for all four models. For example, BLIP ITC-only increased from 50.379 in the short group to 67.559 in the long group, while fine-tuned CLIP increased from 38.272 to 50.964. One plausible reading is that longer captions supply more object, scene, or attribute cues, making queries easier to distinguish; the grouping alone, however, does not establish a causal effect. Fig. 3 displays the trend for all four models.

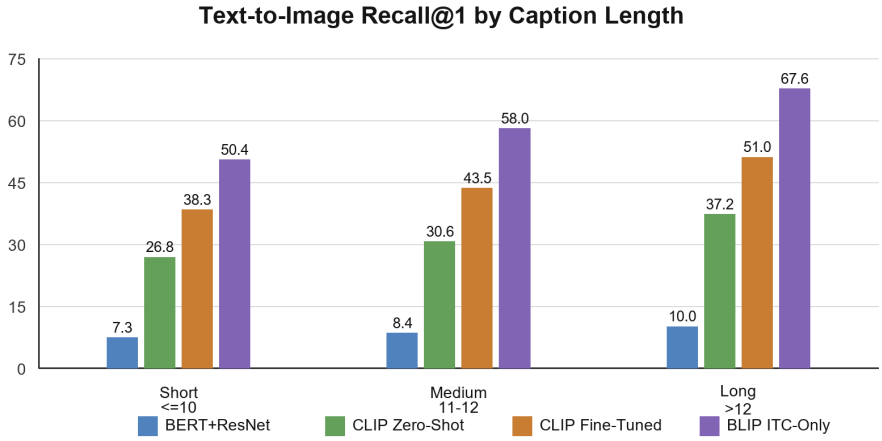

Figure 3. Text-to-image Recall@1 grouped by caption length (photo credit: original)

DOI: 10.54254/2755-2721/2026.36913

4.4. Runtime accounting

After retrieval accuracy, runtime is reported as a secondary practical metric. The values were recorded on the same Colab/Linux environment with an NVIDIA L4 GPU, but they do not all represent the same type of cost. BERT+ResNet and fine-tuned CLIP include training and validation, whereas CLIP zero-shot and BLIP ITC-only are inference-only retrieval runs.

Wall-clock values are reported as records for different run types rather than as a direct four-model inference benchmark. The complete inference-only workflows required 52.58 seconds for CLIP zero-shot and 63.50 seconds for BLIP ITC-only. Within these runs, CLIP used approximately 35.0 seconds for image embedding extraction and 5.0 seconds for text embedding extraction, while BLIP used 43.0 and 6.0 seconds for the same stages. In contrast, the recorded totals of 4,765.26 seconds for BERT+ResNet and 2,212.50 seconds for fine-tuned CLIP include all training epochs, full validation retrieval after every epoch, checkpoint selection, final test evaluation, and qualitative-output generation. Stage-specific inference times were not recorded under equivalent boundaries for the two trained settings; therefore, no four-way inference-speed ranking is made. A direct efficiency comparison would require evaluating the four final checkpoints with identical timing boundaries, batch settings, warm-up procedures, and repeated runs. No main result uses ITM re-ranking, so the four reported settings remain comparable as embedding-based retrieval evaluations.

4.5. Qualitative error analysis

The casebook was used to look behind the aggregate metrics. Top-1 retrievals were often correct when the query referred to a distinctive object or setting, such as skis, a bathroom, a giraffe, a motorcycle, or a plate of food. When retrieval failed, the recurring difficulties were attributes, object counts, and spatial relations. A different issue also appeared: the retrieved scene sometimes matched the query even though COCO did not label it as the positive.

Each model contributed five successful queries and five candidates for every error category. The BERT+ResNet examples were strongest when object and scene were easy to identify. Zero-shot CLIP recovered more specific food, animal, and scene content; fine-tuned CLIP and BLIP ITC-only generally followed captions with more detail. None of the models was consistently reliable for exact attributes, object counts, and spatial relations. Looking across repeated cases, rather than one example, made both these errors and the annotation ambiguity visible.

Recall@K treats only the annotated Karpathy positive as correct, so it cannot distinguish a clear semantic mistake from an unlabeled but reasonable retrieval. This is why the interpretation uses the casebook together with Figs. 2 and 3 and the timing records. Across this evidence, ranking behavior varies with the checkpoint, training condition, similarity mechanism, amount of detail in the query, and coverage of COCO's annotations.

5. Conclusion

RQ1 can be answered for the fixed Karpathy-split experiment. With one gallery and one Recall@K/RSUM implementation, BERT+ResNet ranked last among the four settings. Zero-shot CLIP was substantially stronger, and COCO fine-tuning improved CLIP in both retrieval directions. BLIP with ITC-only scoring returned the largest aggregate score. This ordering is specific to the evaluated checkpoints and protocol.

The evidence for RQ2 is less well captured by a single RSUM value. Image-to-text retrieval scores were higher for every model, but its protocol selects the best of five positive captions. In text-

DOI: 10.54254/2755-2721/2026.36913

to-image retrieval, longer captions were associated with higher Recall@1, suggesting that added object and scene details can make a query more distinctive. The timing logs also need separate interpretation: 52.58 seconds for zero-shot CLIP and 63.50 seconds for BLIP cover inference-only workflows, whereas the BERT+ResNet and fine-tuned CLIP totals include training and repeated validation. The qualitative cases further distinguish clear attribute, counting, and spatial errors from plausible matches omitted by the annotations. Taken together, the experiment indicates that checkpoint history, training, scoring, and benchmark design all contribute to the results; architecture alone does not explain them. A useful next step is to test ITM re-ranking as a separate condition and repeat the study with additional datasets and checkpoints.

References

[1] Li, T., Kong, L., Yang, X., Wang, B., & Xu, J. (2024). Bridging modalities: A survey of cross-modal image-text retrieval. *Chinese Journal of Information Fusion, 1*(1), 79–92.

[2] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision (ECCV), Lecture Notes in Computer Science 8693*, 740–755.

[3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139*, 8748–8763.

[4] Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML), PMLR 162*, 12888–12900.

[5] Li, Z., Guo, C., Wang, X., Zhang, H., & Wang, Y. (2024). Integrating listwise ranking into pairwise-based image-text retrieval. *Knowledge-Based Systems, 287*, 111431.

[6] Khan, A., Sohail, A., Zahoora, U., & Qureshi, A. S. (2020). A survey of the recent architectures of deep convolutional neural networks. *Artificial Intelligence Review, 53*(8), 5455–5516.

[7] Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842–866.

[8] Wang, T., Li, F., Zhu, L., Li, J., Zhang, Z., & Shen, H. T. (2024). Cross-modal retrieval: A systematic review of methods and future directions. *Proceedings of the IEEE, 112*(11), 1716–1754.

[9] Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., & Qiao, Y. (2024). CLIP-Adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision, 132*(2), 581–595.

[10] Zhou, K., Hassan, F. H., & Gan, K. H. (2024). Pretrained models for cross-modal retrieval: Experiments and improvements. *Signal, Image and Video Processing, 18*(5), 4915–4923.

[11] Hu, P., Huang, Z., Peng, D., Wang, X., & Peng, X. (2023). Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(8), 9595–9610.