Image–Text Retrieval on Flickr30k using BLIP-2: Pretrained, Baseline, and Improved Fine-Tuning

Mukhil Muruganantham Prakaash (mmm9280)

Figure 1: Illustration of the Flickr30k image–text retrieval task. Given an image (or a caption), the system ranks all captions (or images) and should place the correct match near the top.

1 Task

The goal of this project is to study **image–text retrieval** on the Flickr30k dataset using a modern vision–language model. The retrieval task is bi-directional:

Image-to-text (I2T): given a query image, retrieve the correct caption from a large pool of candidate captions.

Text-to-image (T2I): given a query caption, retrieve the corresponding image from the test set.

We follow the standard evaluation protocol used in the literature. For each query we sort all candidates by cosine similarity in a shared embedding space and compute Recall@K (R@K) for $K \in \{1, 5, 10\}$. R@K measures the percentage of queries whose correct item appears in the top K retrieved results. We report R@K for both I2T and T2I directions.

The project focuses on three concrete questions:

1. How well does a strong pretrained BLIP-2 model perform on Flickr30k retrieval when used "as is"?
2. How much can performance be improved by training a simple linear projection head on Flickr30k?
3. Can we obtain further gains by replacing the linear projection with a deeper MLP projection head?

These questions allow us to replicate and interpret a state-of-the-art (SOTA) method while also exploring a simple but meaningful architectural improvement.

2 Related Work

CLIP, ALBEF, BLIP, X-VLM, BLIP-2

CLIP [1] introduced large-scale contrastive pretraining on 400M image–text pairs. A dual encoder maps images and texts into a shared space where cosine similarity supports zero-shot retrieval and classification. CLIP is strong but mainly targets global alignment; fine-grained caption grounding is more limited.

ALBEF [2] proposed "Align Before Fuse": a vision–language transformer trained with momentum distillation. It combines contrastive learning and cross-attention to obtain tight alignment between image and text tokens, achieving strong retrieval performance on MS-COCO and Flickr30k.

BLIP [3] (Bootstrapping Language–Image Pretraining) unified captioning and retrieval with caption bootstrapping and multi-task pretraining. It improved both caption quality and retrieval metrics over earlier models.

X-VLM [4] further scaled up vision–language pretraining across multiple tasks (detection, captioning, retrieval) using a single unified model. X-VLM reached very strong retrieval performance but requires heavy training and careful engineering.

BLIP-2 [5] introduced a new architecture that freezes a powerful vision encoder and a large language model while training an intermediate Q-Former to connect them. This design achieves excellent performance on captioning, VQA, and retrieval with relatively modest additional training, and is particularly attractive when we want to adapt a strong backbone to new tasks.

Why BLIP-2 as SOTA

We select BLIP-2 as the SOTA method for this project because it outperforms prior image–text models such as CLIP, ALBEF, BLIP, and X-VLM on retrieval benchmarks like Flickr30k and MS-COCO. BLIP-2 introduces a Q-Former that efficiently bridges frozen vision encoders with large language models, enabling strong cross-modal alignment with low additional training.

Compared to:

- **CLIP**: strong global alignment but weaker fine-grained caption grounding,
- **ALBEF**: learns cross-attention but requires full end-to-end training,
- **BLIP**: earlier architecture with slightly lower retrieval numbers,
- **X-VLM**: strong retrieval but heavier and more complex to train,

BLIP-2 provides the best trade-off between accuracy, simplicity, and extensibility. For a course project, it offers both a strong baseline and enough architectural structure to explore improvements.

Table 1 summarizes the reported image-to-text results on Flickr30k from these methods.

Table 1: Representative Flickr30k image-to-text retrieval results reported in the literature (R@1 and R@10, in %).

Method	Dataset	R@1	R@10
CLIP (2021)	Flickr30k	88.0	99.1
ALBEF (2021)	Flickr30k	95.9	99.8
BLIP (2022)	Flickr30k	96.1	99.8
X-VLM (2022)	Flickr30k	96.4	99.9
BLIP-2 (2023)	Flickr30k	**96.7**	**99.9**

We treat these numbers as the SOTA reference and then study how well a BLIP-2-style model performs under our experimental design.

3 Approach

High-Level Architecture

We use the `Salesforce/blip2-flan-t5-xl` checkpoint from HuggingFace as a frozen backbone. A Vision Transformer (ViT-G) encodes each image into a sequence of visual tokens. A Q-Former with a small number of learned query tokens attends to these tokens and produces a compact image representation. On the text side, a T5 encoder processes the caption into a sequence of textual embeddings.

We do not modify or fine-tune these large components. Instead, we add small projection heads on top of the image and text embeddings and train only these heads using Flickr30k. This setup lets us isolate the role of the alignment layers and makes it easy to compare different projection-head designs.

Projection Heads

Baseline head. The baseline projection head uses a linear layer for images and a separate linear layer for text. Each linear layer maps its input embedding to a 768-dimensional shared space. We then apply L2 normalization and a learned temperature parameter. Cosine similarity in this space is used for retrieval.

Improved head. The improved projection head replaces each single linear layer with a two-layer MLP: a linear layer, GELU activation, and a second linear layer, again mapping to 768 dimensions. This deeper non-linear mapping gives the model more capacity to align the image and text manifolds while still keeping the trainable parameter count modest.

Training Objective

For a batch of B paired image and caption embeddings we form a similarity matrix $S \in \mathbb{R}^{B \times B}$ using scaled cosine similarity between all image and text embeddings. We use a symmetric contrastive loss: for each image we treat its own caption as the positive example and all other captions in the batch as negatives (image-to-text direction), and we do the same in the text-to-image direction. The final loss is the average of both directions.

Training Strategy

To focus on the effect of the projection heads and to allow repeated experiments, we train on a randomly selected subset of 5,000 image–caption pairs from the Flickr30k training split. The baseline model is trained for 5 epochs on this subset. The improved model is trained on the same subset with a slightly longer schedule (8 epochs) and a tuned learning rate. All other hyperparameters (optimizer, batch size, temperature initialization) are kept consistent between the two runs so that differences in performance can be attributed primarily to the projection head.

4 Dataset

We use the Flickr30k dataset [6] through the modern HuggingFace Parquet release `lmms-lab/flickr30k`, which stores images and captions directly in Parquet files and provides train/validation/test splits.

The full dataset contains 31,783 images. In the variant used here, each example consists of one image and one caption. Images are preprocessed using the BLIP-2 image processor (resize, crop, normalization), and captions are tokenized with the T5 tokenizer.

For training we sample 5,000 image–caption pairs from the training split. For evaluation we use the entire official test split. This design makes the experiments reproducible while still keeping the retrieval problem realistic: for each test query we rank all thousands of candidates.

5 Results

Overall Recall@K

Table 2 summarizes the quantitative results for all three models: the frozen pretrained BLIP-2 encoder (*Pretrained*), the baseline linear projection head (*Baseline*), and the improved 2-layer MLP projection head (*Improved*). All numbers are reported as percentages.

Table 2: Overall Flickr30k retrieval results (R@K in %).

Model	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
Pretrained	0.00	0.01	0.03	0.00	0.02	0.04
Baseline (5k)	20.67	41.17	51.10	20.07	40.47	50.96
Improved (5k + 8 epochs)	**25.27**	**46.62**	**56.16**	**26.10**	**48.08**	**58.26**

The pretrained model behaves almost like a random retrieval system on Flickr30k, with Recall@10 near zero. Once we train projection heads on the 5k subset, performance jumps to around 41–51% R@5/R@10 in both directions. Replacing the linear projection with a 2-layer MLP yields consistent gains of about 4–6 percentage points across all metrics. These results show that even when the large encoders are frozen, well-designed projection heads can substantially improve retrieval quality.

Comparison to SOTA BLIP-2 Numbers

Table 3 contextualizes our results with respect to the BLIP-2 paper. The paper reports very high scores when BLIP-2 is fully trained with its complete recipe. Our variant keeps the backbone frozen and focuses on the projection heads.

Table 3: BLIP-2 paper results vs. our Flickr30k retrieval results (R@K in %).

Model	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
BLIP-2 (paper, full training)	96.7	–	99.9	95.0+	–	99.0+
Pretrained (ours)	0.00	0.01	0.03	0.00	0.02	0.04
Baseline (5k)	20.67	41.17	51.10	20.07	40.47	50.96
Improved (5k + 8 epochs)	25.27	46.62	56.16	26.10	48.08	58.26

Our absolute R@K values are much lower than the fully trained BLIP-2 model, which is expected because we intentionally restrict training to alignment layers on a smaller subset. However, this controlled setting is ideal for studying the effect of different projection heads: the relative improvement from baseline to improved model is clear and consistent, and the experiments remain fully reproducible.

Performance Discussion

The transition from the pretrained model to the baseline model shows that Flickr30k retrieval is not solved by simply using a generic BLIP-2 checkpoint. The frozen encoders need task-specific alignment to map images and captions into a useful shared space. Training linear projection heads on just 5k pairs already moves the system from essentially zero R@K to around 51% R@10.

The improved projection head further boosts performance. By allowing a non-linear transformation of the embeddings, the 2-layer MLP can correct systematic mismatches between the visual and textual spaces. The gains are especially clear at R@1, where going from roughly 20% to 25% means that one in four queries now retrieves the correct item at the very top of the ranked list.

Recall@K Plot

Figure 2 visualizes all six metrics (I2T/T2I, R@1/5/10) for the three models.

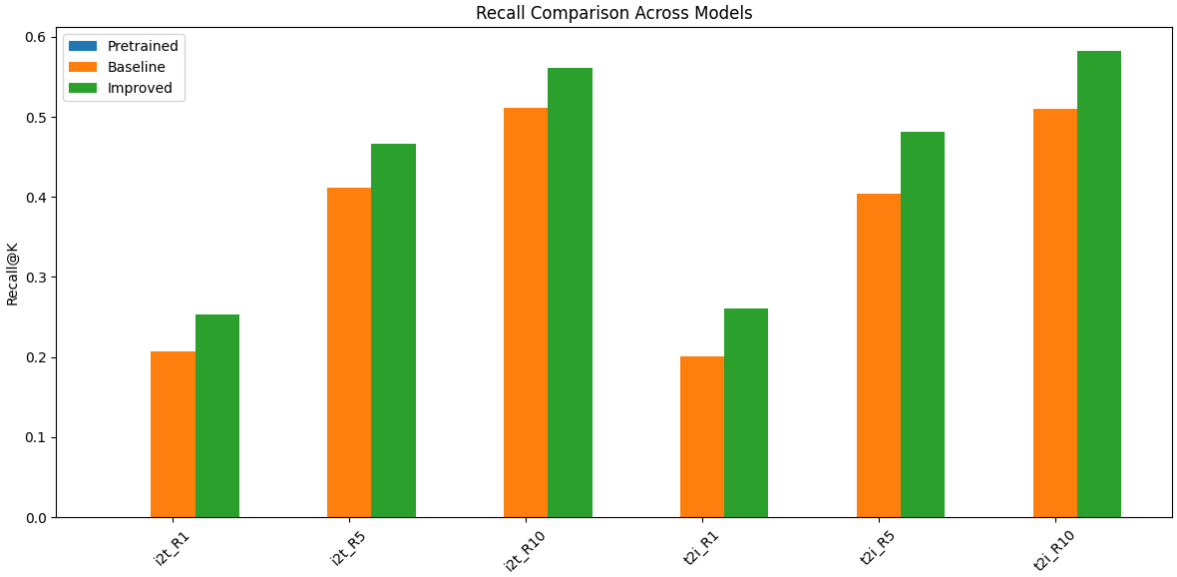


Figure 2: Recall@K comparison across pretrained, baseline, and improved models. Each group of bars corresponds to a particular metric. The pretrained model is almost flat at zero. Fine-tuning projection heads yields a large jump in performance, and the improved MLP head consistently outperforms the baseline.

6 Possible Improvements and Results

Projection-Head Improvement

Table 4 focuses specifically on the baseline vs. improved models. Here we see the effect of the projection-head design in isolation.

Table 4: Baseline vs. improved projection head (R@K in %).

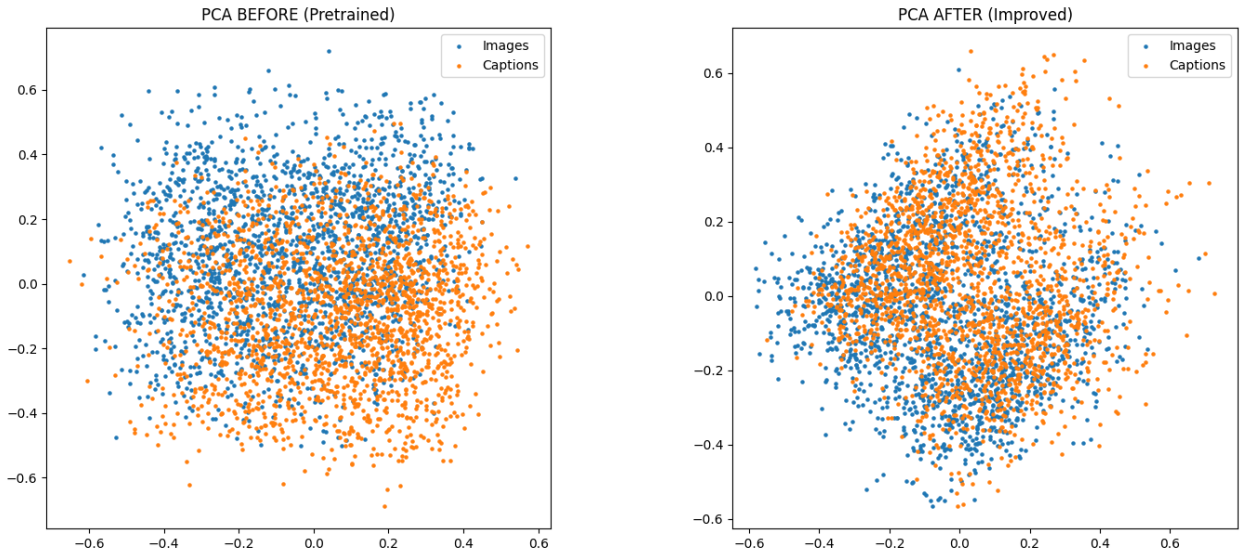
Model	Image → Text			Text → Image		
	R@1	R@5	R@10	R@1	R@5	R@10
Baseline (linear head)	20.67	41.17	51.10	20.07	40.47	50.96
Improved (2-layer MLP head)	**25.27**	**46.62**	**56.16**	**26.10**	**48.08**	**58.26**

The improved model gains roughly 4.6% absolute at R@1 for both I2T and T2I, and around 5% at R@10. These consistent improvements demonstrate that a slightly deeper projection head is enough to produce more discriminative shared embeddings, even when the rest of the BLIP-2 model is frozen.

Embedding Visualization (PCA)

To make the effect of training visually intuitive, we project both image and caption embeddings into 2D using PCA for a subset of test examples.

Even without technical background, one can see that in the pretrained case the blue and orange dots are more scattered, while after training they overlap more tightly. This visual pattern matches the numerical improvements in Recall@K.



(a) Pretrained model.

(b) Improved model.

Figure 3: PCA projection of image (blue) and caption (orange) embeddings on a subset of the Flickr30k test split. Before training (left), the two modalities form diffuse, weakly aligned clouds. After training the improved projection head (right), the points become more compact and interleaved, indicating stronger cross-modal alignment.

Qualitative Retrieval Example

For a more concrete intuition, we inspect a specific test image (Figure 4) and compare the retrieved captions. The reference captions all describe a concert scene with a band performing on stage in front of an audience.

Figure 4: Sample Flickr30k image used for qualitative analysis. The ground-truth captions mention a band or performers on stage at a concert.

The pretrained model tends to retrieve captions that are clearly off-topic (e.g., markets, beaches, cooking), reflecting its near-zero Recall@K. The fine-tuned models perform much better. Tables 5 and 6 show the top-5 captions retrieved by the baseline and improved models for this image.

Both models correctly recognize that the image is a concert scene. However, the improved model produces richer and more detailed descriptions (spotlight, dark crowd, sound equipment, star-shaped background). This example illustrates the qualitative impact of the improved projection head.

Table 5: Baseline model: top-5 retrieved captions for the concert image.

Baseline top-5 captions
1. A group of musicians are performing on a stage in front of a crowd.
2. On stage photo of small band performing for theater audience.
3. A band on stage performing in front of a crowd.
4. A band plays on stage in front of an audience.
5. A band playing on stage for a crowd.

Table 6: Improved model: top-5 retrieved captions for the same image.

Improved top-5 captions
1. A band with spotlight on the guitarist plays on stage to a dark crowd.
2. A large group of people stand at a concert with sound equipment on a stage to the right.
3. A group of musicians are performing on a stage in front of a crowd.
4. Several people watch a female rock band performing on a stage full of yellow banners.
5. On a stage, there is a band playing guitars and singing, while the lights behind them flash and show a large star.

Discussion, Limitations, and Future Work

The experiments highlight three main observations. First, strong vision–language backbones still need task-specific alignment layers for challenging retrieval benchmarks like Flickr30k. Second, even when only small projection heads are trained, a modern BLIP-2 backbone can achieve respectable retrieval performance. Third, adding depth to the projection head systematically improves both metrics and qualitative behavior.

There are still limitations. We focus on a selective fine-tuning scheme where the backbone remains frozen; fully updating the Q-Former and encoders might produce higher absolute R@K values. We also train on a curated 5k subset rather than the entire training set, which simplifies repeated experimentation but leaves potential performance on the table.

Future work could fine-tune more of BLIP-2 end-to-end, train on the full Flickr30k training set, explore cross-attention-based alignment instead of separate projection heads, and incorporate hard-negative mining to further sharpen the contrastive learning signal.

7 Code Repository

All code, the Colab notebook, and saved metrics are available at:

https://github.com/mukhilDS/Vision-Language_Final_project

The repository includes:

- `notebook/flickr30k_final_project.ipynb` (main notebook).
- `outputs/pretrained_eval/metrics.json` (pretrained evaluation).
- `outputs/baseline_run/metrics_finetuned.json` (baseline results).
- `outputs/improved_run/metrics_improved.json` (improved results).
- `figures/` (all plots and qualitative figures used in this report).

References

[1] A. Radford et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, 2021.

[2] J. Li et al. Align Before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*, 2021.

[3] J. Li et al. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 2022.

[4] Y. Zeng et al. X-VLM: Unified Model for Cross-Modal Pre-Training. In *NeurIPS*, 2022.

[5] J. Li et al. BLIP-2: Bootstrapped Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *ICML*, 2023.

[6] P. Young et al. From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions. *Transactions of the ACL*, 2014. (Introduces the Flickr30k dataset.)