A study on Composed Image Retrieval based on BLIP-CoVR

Arthur Bresnu [1,2] Guillaume Henon-Just [1,2]

[1] Ecole des Ponts et Chaussées, {surname.name}@eleves.enpc.fr [2] ENS Paris-Saclay

Topic:A Date: September 21, 2025

1. Introduction

Composed Image Retrieval (CoIR) is an emerging task in computer vision that involves retrieving a target image using a query image and a text description specifying desired modifications [3]. CoIR extends to Composed Video Retrieval (CoVR), where the goal is to retrieve a target video based on a reference video and transformation text.

Our work builds on [4], which introduced the notion of CoVR and proposed the BLIP-CoVR model. This study leveraged multimodal data to combine visual and textual information effectively for video retrieval, providing the foundation for our analysis.

The CIRR dataset [3] provides triplets of (reference image, modification text, target image) for evaluating CoIR tasks, while WebVid-CoVR dataset enables pretraining for CoVR tasks using video triplets.

Our primary objectives were to understand the BLIP and BLIP-CoVR models while establishing baseline performance in our setup, evaluate our method of combining embeddings from multimodal components, and investigate the impact of pretraining with a filtered WebVid-CoVR dataset on retrieval performance.

2. Presentation and Reproduction of BLIP-CoVR

2.1. Overview of the BLIP and BLIP-CoVR Models

BLIP (Bootstrapped Language-Image Pretraining) [2] is a multimodal model aligning textual and visual representations using contrastive learning. Its architecture, detailed in Appendix Figure 2, integrates an image encoder, a text encoder/decoder, and a multimodal encoder with shared parameters. Key loss functions include contrastive loss for embedding alignment, image-text matching loss for multimodal validation, and language modeling loss for text reconstruction, enabling robust embeddings for visual-textual tasks.

BLIP-CoVR [4] extends the BLIP architecture for the CoVR task by processing triplets of (reference video, transformation text, target video). As shown in Appendix Figure 3, key adaptations include encoding the reference video's middle frame as a query embedding with BLIP's frozen image encoder, combining it with transformation text via an image-grounded text encoder for a multimodal query, and aggregating target video frame embeddings using weighted averaging. The model leverages contrastive loss to align multimodal query and target video embeddings, enabling effective video retrieval.

2.2. Reproducing the Baseline on the CIRR Dataset

Using the BLIP-CoVR checkpoints provided by [4], we evaluated their performance on the CIRR dataset using our setup. The results are reported in Table 1, and, as expected, they are very close to the original paper's results. Interestingly, they are slightly better, but this is not straightforward to interpret since we are using the same checkpoints, dataset, and model for embedding. These results will now serve as our new baseline for comparison.

P.D.	T.D.	O.R.	R@1	R@5	R@10	R@50	Rsub@1	Rsub@2	Rsub@3
-	-	T	19.76	41.23	50.89	71.64	63.04	81.01	89.37
		O	20.02	42.00	52.34	73.57	64.77	83.42	91.98
	CIRR	T	48.84	78.05	86.10	94.19	75.78	88.22	92.80
		O	-	-	-	-	-	-	-
WebVid	-	T	38.48	66.70	77.25	91.47	69.28	83.76	91.11
		O	39.28	68.24	78.94	94.65	71.28	86.22	94.02
	CIRR	T	49.69	78.60	86.77	94.31	75.01	88.12	93.16
		O	50.04	80.96	89.33	97.64	77.25	91.04	96.34

Table 1. Evaluation results on CIRR, comparing the origin of the evaluation (O.R.), with personal (O: Ours) and paper (T: Their) results, for various pretraining data (P.D.) and training data (T.D.) configurations on our setup.

2.3. Training the BLIP-CoVR model on the CIRR dataset

We report in Table 2 our trainning reproduction on a single L4-GPU with a batch size of 128. In first part, we reproduced the fine-tuning results from no pretraining data and we achieve slightly better results. However, this is not due to the training method but rather because we are comparing our results to the paper's values instead of our new baseline, which we do not have for this model because we didn't have the chekpoint. This is confirmed by the other configuration where we pre-trained on WebVid-CoVR and fine-tuned on CIRR using our batch size. In this case, the results are slightly worse than the new baseline, which is expected since our batch size is almost 20 times smaller than theirs (2048 for pretraining and fine-tuning). We now establish our new baseline for training reproduction, which will be used as our ground truth moving forward.

P.D.	T.D.	O.R.	R@1	R@5	R@10	R@50	Rsub@1	Rsub@2	Rsub@3
-	CIRR	T	48.84	78.05	86.10	94.19	75.78	88.22	92.80
		T.R	49.325	80.723	88.916	97.614	76.699	90.289	95.831
WebVid-CoVR	CIRR	O	50.337	80.964	89.325	97.639	77.253	91.036	96.337
		T.R	49.133	80.699	89.229	97.494	76.120	90.024	95.735

Table 2. Comparison of our training reproduction (T.R.) and the baseline results, including our evaluation (O: Our evaluation), with their results (T) on CIRR.

We evaluated our model trained on CIRR (without WebVid-CoVR pretraining) through qualitative examples (Appendix):

- Positive Example: The model correctly retrieves the target image by applying two transformations—removing a dog and adding a girl—showcasing its ability to handle multiple changes. (Figure 6)
- Negative Example: The model partially succeeds, identifying an outdoor vending machine but failing to include the requested green box. The retrieved image, split into two vending machines, may have caused confusion, highlighting challenges with complex transformations and outliers. (Figure 7)

3. Combination Method

3.1. Motivation and Approach

We hypothesized that combining the multimodal embedding $f(q, t)$ with individual image and text embeddings q_{emb} and t_{emb} could improve retrieval performance by capturing additional contextual information. To achieve this, we:

- Averaged the three embeddings (AVG) : $\frac{f(q,t)+q_{\text{emb}}+t_{\text{emb}}}{3}$.
- Trained a Multi-Layer Perceptron (MLP) to dynamically weight the embeddings: $w_{\text{multimodal}} f(q, t) + w_{\text{image}} q_{\text{emb}} + w_{\text{text}} t_{\text{emb}}$.

3.2. Performance of the method

As shown in Table 3, averaging the embeddings resulted in suboptimal performance, with accuracy falling below the baseline when tested, so we only did it in a single configuration. However, the introduction of the MLP led to a slight improvement, surpassing our original baseline across all pretraining data configurations, which we will discuss in detail later. These results are promising, as they suggest that the MLP configuration enables all models to achieve better performance. Additionally, as illustrated in Appendix Figure 4, the loss function also benefits from the MLP configuration, showing improvement in convergences speed and final results.

Mode	Pretraining Data	C.M.	R@1	R@5	R@10	R@50	Rsubset@1	Rsubset@2	Rsubset@3
Train	-	NA	49.325	80.723	88.916	97.614	76.699	90.289	95.831
	-	AVG	49.325	80.193	88.554	97.566	75.349	89.349	95.590
	-	MLP	49.325	80.988	88.892	97.590	77.229	90.386	95.831
	WebVid-CoVR	NA	49.133	80.699	89.229	97.494	76.120	90.024	95.735
	WebVid-CoVR	MLP	50.145	80.554	88.843	97.590	76.651	90.193	95.807
	WebVid-CoVR_F	NA	48.964	80.747	89.253	97.663	76.386	90.217	95.735
	WebVid-CoVR_F	MLP	50.554	80.482	88.892	97.614	76.726	90.024	95.687
	WebVid-CoVR_F_S.I	NA	49.542	80.554	88.964	97.494	76.361	89.976	95.831
	WebVid-CoVR_F_S.I	MLP	50.096	80.867	88.916	97.663	76.554	90.096	95.735
Zero-shot	WebVid-CoVR	-	36.169	66.578	78.337	94.193	68.747	84.747	93.133
	WebVid-CoVR_F	-	37.870	67.349	78.651	94.434	70.819	86.410	94.048
	WebVid-CoVR_F_S.I	-	36.625	66.434	77.807	94.096	70.024	85.205	93.590

Table 3. Evaluation results on CIRR for different pretraining data, combination methods (C.M.) with NA: only the multimodal query; WebVid-CoVR_F : filtered WebVid; WebVid-CoVR_F_S.I: filtered WebVid trained on the same number of iterations

3.3. Quantitative Study of Distribution of the MLP Weights

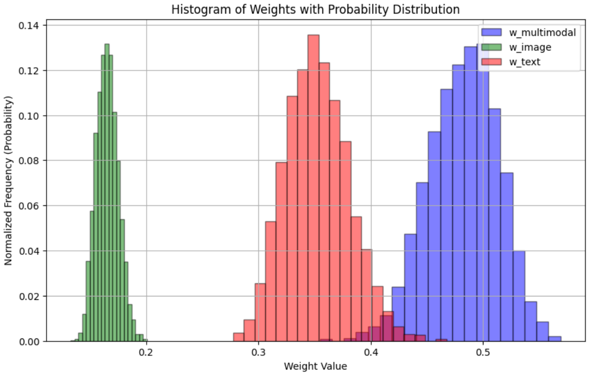

Figure 1. Distribution of the MLP-generated weights $w_{\text{multimodal}}$, w_{image}, and w_{text} across the CIRR dataset.

Figure 1 illustrates the distribution of the weights $w_{\text{multimodal}}$, w_{image}, and w_{text} across the CIRR dataset. All three weights exhibit Gaussian distributions with tight variances, demonstrating minimal overlap.

This observation suggests that these weights are relatively invariant to specific input characteristics, such as the reference image or transformation text. Instead, they appear to be intrinsic to the problem or dataset itself.
Interestingly, the average value of $w_{\text{multimodal}}$ is approximately 0.5, close to 0.35 for w_{text}, while w_{image} averages only around 0.1. This aligns with the nature of the task, as the transformation text often provides the most critical information for locating the target image, whereas the reference image may be far from this target.

3.4. Qualitative Study of MLP Weights

For each weight ($w_{\text{multimodal}}$, w_{image}, and w_{text}), we visualized the image from the CIRR dataset with the highest respective weight to analyze the role of these weights. The corresponding figures are presented in the Appendix.

- **Highest $w_{\text{multimodal}}$ (Figure 8):** The reference image is far from the target image, and the transformation text is ambiguous. The model failed to return the correct target image as the top output. This suggests that in complex cases where both the reference image and transformation text contribute equally, the model emphasizes the multimodal embedding.
- **Highest w_{image} (Figure 9):** The reference image closely resembles the target image, while the transformation text is unclear. Emphasizing the reference image is logical in such cases, as it provides clearer guidance for retrieval.
- **Highest w_{text} (Figure 10):** The transformation text is precise and accurately describes the target image, but the reference image is far from the target. The model relies more on the transformation text when it is the dominant source of information.

4. Pretraining on a Filtered WebVid-CoVR Dataset

4.1. Dataset Filtering and Expectations

We filtered the WebVid-CoVR dataset to retain video triplets with common authors or categories, aiming to reduce noise and improve visual consistency. The filtration utilized metadata from the CleanVid-15M map [1].
The most relevant number in Table 4 is the one of different videos in the dataset since this is the object we iterate on during training. We can see that the filtered WebVid-CoVR dataset is 2.7 times smaller than the original one.

Dataset	# Target Videos	# Triplets	# Query Videos
WebVid-CoVR	129,954	1,644,276	129,921
WebVid-CoVR Same Author	20,727	99,861	20,727
WebVid-CoVR Same Category	47,334	325,320	47,331
WebVid-CoVR Same Author or Category	48,097	339,265	48,093

Table 4. Dataset statistics showing the number of target videos, triplets, and query videos for WebVid-CoVR and its filtered subsets.

4.2. Performance Evaluation

We pretrained BLIP-CoVR on the filtered dataset using two strategies: WebVid-CoVR_F which is keeping the Same number of epochs as the original dataset and WebVid-CoVR_F_S.I (Same Iterrations) that is train on adjusted epochs to match the total number of training iterations on the original dataset.

As shown in Table 5, the model pretrained on the full WebVid-CoVR dataset achieves the best performance on the WebVid-CoVR test dataset, followed closely by the model pretrained on the filtered dataset with the same number of iterations, and then the one with the same number of epochs, which lags significantly behind. This behavior is understandable, as the model sees far fewer examples from COVR, leading to less adaptation to this dataset.

Pretraining Data	R@1	R@5	R@10	R@50
-	15.85	32.79	40.30	58.37
WebVid-CoVR	55.75	81.77	89.44	98.32
WebVid-CoVR_F	52.82	80.05	87.64	97.93
WebVid-CoVR_F_S.I	55.20	81.06	89.01	98.20

Table 5. Result for different pre-training data on the WebVid-CoVR test dataset.

However, of particular interest is the observation in Table 3 regarding the Zero-shot performance for CIRR, where the WebVid-CoVR_F model, despite using almost three times less computational resources in training compared to the other two models, achieves the best results. This finding is further validated when comparing the fine-tuned versions of this model with the fine-tuned versions of the other two models using the same combination method, as the WebVid-CoVR_F model produces results within the same range as the others.

These results highlight the advantage of using filtered datasets with fewer iterations during training, which can lead to competitive performance with significantly lower computational costs.

4.3. Qualitative Study of Filtered Data

We analyzed three pairs of images (Appendix), corresponding to middle frames of videos in WebVid-CoVR:

- Same Author (Figure 11): Images share a highly similar photographic style and background, reflecting strong stylistic coherence.
- Same Category (Figure 12): Images differ in style but represent the same subject (birds), showing thematic consistency.
- Different Author and Category (Figure 13): Images appear visually similar despite being unrelated, questioning the filtering criteria's precision.

While filtering improves coherence, the similarity in unrelated pairs suggests limitations, warranting further quantitative evaluation.

5. Conclusion

In this project, we gained a deeper understanding of the BLIP-CoVR model and its performance on the CIRR dataset, exploring the challenges and opportunities of multimodal retrieval tasks. We proposed an MLP-based combination method that improved retrieval performance and filtered the WebVid-CoVR dataset, achieving computational efficiency without compromising results. Additionally, this project provided valuable insights into multimodal models and enhanced our proficiency with cloud computing tools such as Google Cloud, further broadening our technical expertise.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021. 3

[2] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapped language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1

[3] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6701–6710, 2021. 1

[4] Lucas Ventura, Antoine Yang, Cordelia Schmid, and Gul Varol. Covr: Learning composed video retrieval from web video captions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 1

A. Appendix

A.1. BLIP and BLIP-CoVR Models

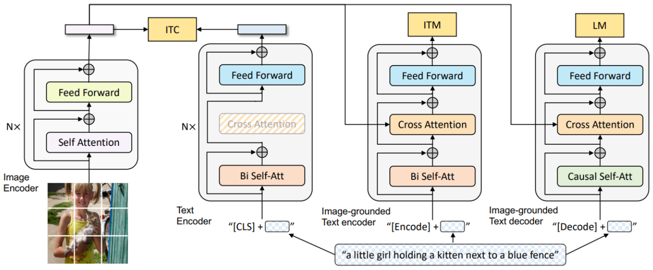

Figure 2. Overview of the BLIP model architecture.

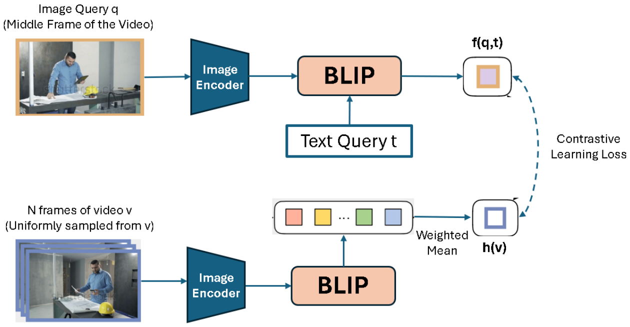

Figure 3. Overview of the BLIP-CoVR model architecture.

A.2. Loss Functions

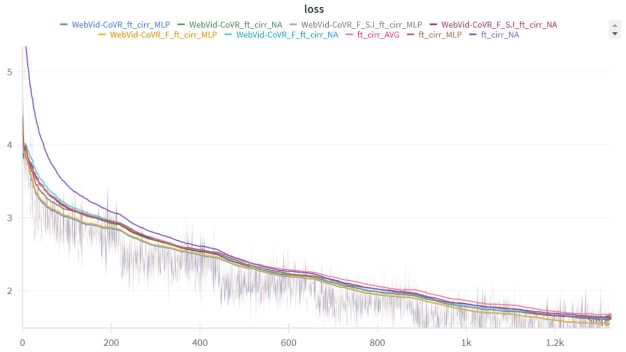

Figure 4. Fine-tuning on CIRR loss functions for all the differents configurations.

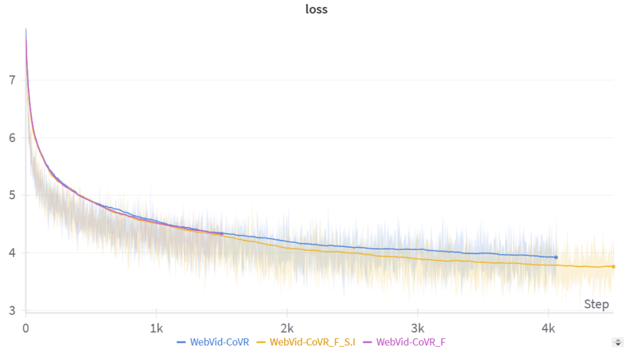

Figure 5. Pre-training WebVid-CoVR and its filtered versions with or without the same iterations.

A.3. CoIR Inference after Training on CIRR - Qualitative Study

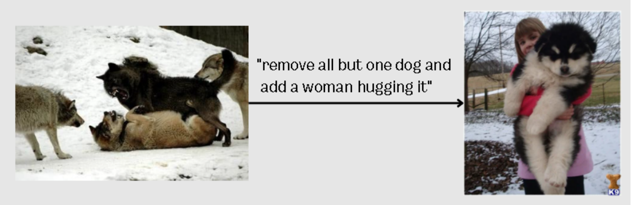

Figure 6. Positive example: The model successfully retrieves the target image by applying two transformations—removing a dog and adding a girl.

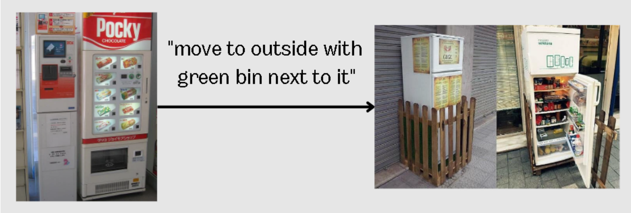

Figure 7. Negative example: The model identifies an outdoor vending machine but fails to include the green box, highlighting challenges with complex transformations.

A.4. MLP Weights for Queries Combination

A.4.1. Investigation over MLP Weights Role - Qualitiative Study

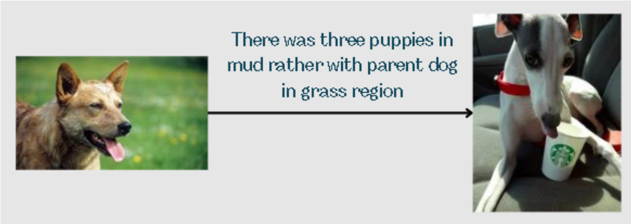
There was three puppies in mud rather with parent dog in grass region

Figure 8. Example with the highest $w_{\text{multimodal}}$: The reference image is far from the target image, and the transformation text is ambiguous. The model failed to return the correct target image as the top output, suggesting an emphasis on multimodal embedding in complex cases.

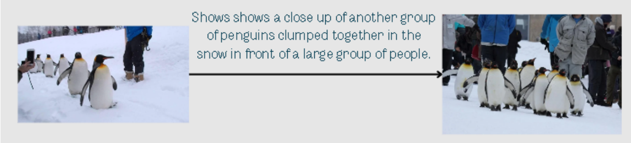
Shows shows a close up of another group of penguins clumped together in the snow in front of a large group of people.

Figure 9. Example with the highest w_{image}: The reference image closely resembles the target image, while the transformation text is unclear. The model emphasizes the reference image for guidance.

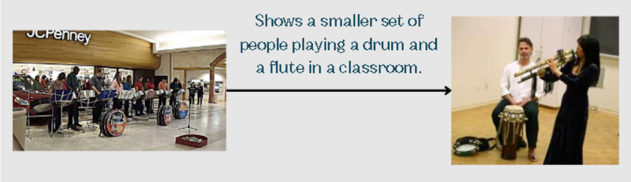
Shows a smaller set of people playing a drum and a flute in a classroom.

Figure 10. Example with the highest w_{text}: The transformation text is precise and accurately describes the target image, while the reference image is far from the target. The model relies on the transformation text as the dominant source of information.

A.5. Visualizing Pairs of the WebVid-CoVR Dataset

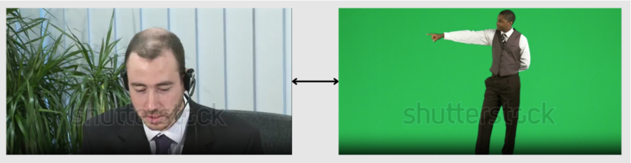

Figure 11. Example of images from the same author: These images share a highly similar photographic style and background, reflecting strong stylistic coherence.

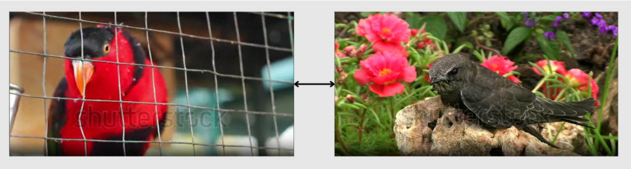

Figure 12. Example of images from the same category: These images differ in photographic style but represent the same subject (birds), showing thematic consistency.

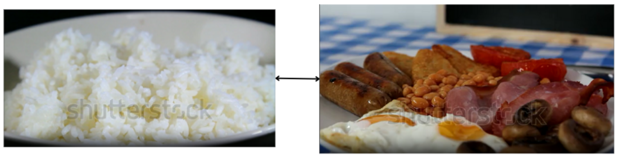

Figure 13. Example of images from different authors and categories: These images appear visually similar despite being unrelated, questioning the filtering criteria's precision.