Knowledge Distillation Effects on Zero-Shot CLIP Retrieval Across Flickr30k and MSCOCO

Assignee Research

May 31, 2026

Abstract

This report synthesises findings from 11 peer-reviewed papers addressing the following research question: How does knowledge distillation impact the zero-shot image-text retrieval accuracy of CLIP variants on Flickr30k and MSCOCO datasets. We present Distill CLIP (DCLIP), a fine-tuned variant of the CLIP model that enhances multimodal image-text retrieval while preserving the original model's strong zero-shot classification capabilities. CLIP models are typically constrained by fixed image resolutions and limited. 11 claims were extracted from source literature; 0 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: Distill CLIP (DCLIP): Enhancing Image-Text Retrieval via Cross-Modal Transformer Distillation. Research question: How does knowledge distillation impact the zero-shot image-text retrieval accuracy of CLIP variants on Flickr30k and MSCOCO datasets?.

2 Methodology

Systematic literature search across multiple databases yielded 11 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

3 Results

11 papers retrieved. 11 claims extracted; 0 independently verified. Quality review score: 4.2/10.

4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

5 Extracted Claims

Claim	Verified	Confidence
DCLIP is evaluated using the Karpathy testing set for MSCOCO and FLICKR30K, which consists of 5,000 images from MSCOCO's	×	0.05
For zero-shot evaluation, DCLIP uses the entire 50,000 images of the ImageNet dataset.	×	0.04
CIFAR-10 and CIFAR-100 datasets are utilized in evaluation for zero-shot classification to show the robustness of DCLIP.	×	0.05
Recall@K measures the proportion of queries for which the correct match appears within the top-K retrieved results, with	×	0.10
Mean Average Precision (MAP) evaluates the overall ranking quality by computing the average precision for each query and	×	0.01
Zero-Shot Classification assesses the model's ability to generalize to unseen classes without additional training, with	×	0.05
DCLIP preserves the generalization capacity of CLIP while enriching the visual representations for improved retrieval pe	×	0.06
ViT-L/14 tends to overfit rapidly to embedding distributions during distillation but DCLIP retains 91% of the original C	×	0.07
DCLIP achieves high retrieval performance on specific datasets at the cost of generalization if not carefully controlled	×	0.03
The DCLIP architecture consists of a teacher-student framework where a frozen cross-modal teacher receives region-level	×	0.11
The teacher model in DCLIP applies YOLOv8x to extract bounding boxes corresponding to salient image regions and weights	×	0.09

References

- http://arxiv.org/abs/2505.21549v4
- http://arxiv.org/abs/2307.09233v3
- http://arxiv.org/abs/2402.18400v2