Contrastive and Non-Contrastive Self-Supervised Alignment in Multimodal Retrieval Benchmarks

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 16 peer-reviewed papers addressing the following research question: What is the comparative performance of contrastive versus non-contrastive self-supervised alignment methods on multimodal retrieval benchmarks like Flickr30k and MS-COCO. 15 claims were extracted from source literature; 2 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: FiCo-ITR: bridging fine-grained and coarse-grained image-text retrieval for comparative performance analysis. Research question: What is the comparative performance of contrastive versus non-contrastive self-supervised alignment methods on multimodal retrieval benchmarks like Flickr30k and MS-COCO?.

2 Methodology

Systematic literature search across multiple databases yielded 16 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

16 papers retrieved. 15 claims extracted; 2 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
FG methods aim to map visual and textual information to a joint space where relevant samples are aligned at the instance	×	0.05
FG search uses continuous embeddings to find the retrieval sample that specifically corresponds to the query sample.	×	0.03
Under FG evaluation conditions, finding a relevant sample involves finding the retrieval sample with the same ID as the	×	0.03
CG search employs bitwise hash codes to find retrieval samples that are broadly relevant to the query instead of exact m	×	0.03
During CG evaluation, a match is defined as finding any retrieval sample with at least one matching category label relat	×	0.04
The broader search criteria of CG search allows for more efficient computational costs compared to FG search.	×	0.05
IMRAM is a representative Fine-Grained (FG) model.	×	0.13
UCCH is a representative Coarse-Grained (CG) model.	✓	0.15
Flickr30K has a 1K sample test set.	×	0.02
MS-COCO has a 5K sample test set.	×	0.02
Traditional ITR benchmark datasets like Flickr30K and MS-COCO are small compared to real-world applications.	×	0.04
Direct empirical comparative evaluations of recent representative FG and CG models are lacking in the literature.	✓	0.22
The FiCo-ITR library is available on the Python Package Index (PyPI).	×	0.08
The FiCo-ITR library is available on GitHub.	×	0.11
Experiments using incrementally larger retrieval sets reveal trade-offs between retrieval performance and computational	×	0.12

References

- `http://arxiv.org/abs/2308.10045v2`

- http://arxiv.org/abs/2206.02574v3
- http://arxiv.org/abs/2407.20114v3