MELTR-Enhanced Flamingo Cross-Modal Alignment Stability Under Frame Dropout Perturbations

Assignee Research

June 8, 2026

Abstract

This report synthesises findings from 8 peer-reviewed papers addressing the following research question: How does the cross-modal alignment stability of MELTR-enhanced Flamingo perform under frame dropout perturbations relative to standard Flamingo on MSR-VTT. 17 claims were extracted from source literature; 3 were independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.7/10. This report is a machine-generated literature synthesis and does not constitute original research.

1 Introduction

This paper examines: CMAL: A Novel Cross-Modal Associative Learning Framework for Vision-Language Pre-Training. Research question: How does the cross-modal alignment stability of MELTR-enhanced Flamingo perform under frame dropout perturbations relative to standard Flamingo on MSR-VTT?.

2 Methodology

Systematic literature search across multiple databases yielded 8 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.7/10.

3 Results

8 papers retrieved. 17 claims extracted; 3 independently verified. Quality review score: 4.7/10.

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
The pre-training dataset consists of 5.337M train and 131k validation image-text pairs from COCO Captions and Visual Gen	×	0.09
The initial COCO dataset processed by UNITER contains 106K images and 533K sentences, among which 27K images and 256K se	×	0.01
After removing duplicated data, 79K images and 277K texts are obtained for training.	×	0.03
CMAL adopts 12-layer Transformers as backbone.	×	0.03
The size of the tokenized word set W is 28997, the size of the object tag set T is 1600, and the size of the compute	×	0.03
The initial learning rate is set as 5-5, and the weight decay is set as 0.01.	×	0.03
The proportion hyperparameters 1-5 are set to [0.4, 0.2, 0.2, 0.1, 0.1] respectively.	×	0.02
The hyperparameters are all tuned with grid-search over the validation set.	×	0.01
AdamW optimizer is used to optimize the model.	×	0.03
CMAL achieves a VQA dev score of 71.96 and test score of 75.62.	×	0.01
CMAL achieves an NLVR2 dev score of 77.35 and test score of 78.28.	×	0.01
CMAL achieves a VE dev score of 74.20 and test score of 74.60.	×	0.01
CMAL achieves a REC dev score of 72.39 and test score of 74.21.	×	0.01
CMAL includes six pre-training tasks: ITM, MLM, MRM, VTC, AMC, and a combination of all.	×	0.05
The hierarchical semantic encoder projects visual objects and textual words into separate semantic spaces to learn intra	✓	0.17
The cross-modal associative prompt layer performs anchor point masking and swap feature filling for obtaining more fine-	✓	0.27
The cross-modal associative classification layer learns potential associative mappings by leveraging a fresh associative	✓	0.26

References

- http://arxiv.org/abs/2502.01547v3
- http://arxiv.org/abs/2410.12595v1
- http://arxiv.org/abs/2506.18985v3