



UGANDA CHRISTIAN UNIVERSITY

A Centre of Excellence in the Heart of Africa

SCHOOL OF RESEARCH & POSTGRADUATE STUDIES DISSERTATION CORRECTION COMPLIANCE FORM (POST VIVA FORM)

Date: 16th September, 2026

Name of Candidate: **Mugimba Kakure Jude**: Reg.No: **J25M19/034** Access No: **B31388**

Title of Dissertation:

An Explainable Transformer-Based Vision-Language Model for Multimodal Consistency Verification

S/N	COMMENTS BY EXTERNAL EXAMINER	ACTION TAKEN	INDICATOR
1	Strengthen empirical validation by moving beyond the proxy benchmark toward human-annotated and/or real-world data.	The study is now stated throughout as a proxy-benchmark evaluation on Flickr30K. The Abstract closes by restricting the work to "a public multi-caption photograph benchmark with algorithmically constructed negatives" and states that the findings should not be generalised to domains whose images, documents or contradiction types differ from this setting. A new Delimitations section records that no new human-annotated corpus of contradiction types was collected and no domain-specific image or document collection was evaluated. Human-annotated consistency datasets are carried forward as a named priority under Recommendations for Future Research. No additional data were collected and no new experiments were run for this revision.	Abstract, pp. iv-v Section 1.8, p. 5 Section 2.8, p. 11 Section 6.6.2, p. 45 Section 6.7, p. 46 Section 6.8.1, p. 46
2	Improve methodological rigour, particularly around construction of negative examples and	Section 3.3.2 now specifies the construction of the inconsistent class: random negatives pair an image with a multi-caption document drawn from a	Section 3.3.2, pp. 16-17 Section 3.3.3, p. 17 Section 3.8.1, p. 21

S/N	COMMENTS BY EXTERNAL EXAMINER	ACTION TAKEN	INDICATOR
	validation procedures.	different image; hard negatives are FAISS nearest neighbours in embedding space and are reserved exclusively for evaluation, i.e. they are not used as negatives inside the InfoNCE training objective. The same section states that neither procedure produces human annotation of contradiction type (omission, substitution, negation or attribute mismatch), and that validation was limited to monitoring the contrastive loss and selecting the decision threshold on the validation split before held-out testing. The group-based 70/15/15 split by image identifier, which prevents leakage of an image across splits, is restated in Section 3.3.3.	Section 4.3, p. 24 Appendix F.5, p. 62
3	Evaluate explainability quantitatively, rather than assuming that attention-based visualisations are faithful explanations.	Attention maps and token-importance rankings are now presented as qualitative illustrations rather than verified explanations. The Abstract states explicitly that these visualisations were not evaluated with quantitative faithfulness tests and are not claimed as causal explanations. Section 6.3.3 states that quantitative faithfulness evaluation and causal explanation remain an area for further work, and Section 6.4.3 repeats that formal evaluation of explanation faithfulness is future work. Explainability is declared a delimitation of the study: it was restricted to visual attention maps and token-importance scores inspected qualitatively, with faithfulness tests and structured human rating outside the present scope. Faithfulness testing and human-judged evaluation of explanations are listed under Recommendations for Future	Abstract, pp. iv-v Section 1.8, p. 5 Section 6.3.3, p. 43 Section 6.4.3, p. 44 Section 6.6.2, p. 45 Section 6.7, p. 46

S/N	COMMENTS BY EXTERNAL EXAMINER	ACTION TAKEN	INDICATOR
		Research. No quantitative faithfulness measurement was performed in this study.	
4	Strengthen the analysis statistically, including significance testing and external validation.	A new limitation, Section 6.8.7 "Statistical and External Validation", states that the reported metrics describe only the held-out Flickr30K splits constructed for this study, that formal significance testing and evaluation on an independent external corpus were not completed, and that the comparison with CLIP is therefore descriptive within this benchmark. Section 6.7 records that no external validation corpus was used. Significance testing and external validation are carried forward as future work. No new statistical tests were computed for this revision; the results tables are unchanged.	Section 6.6.2, p. 45 Section 6.7, p. 46 Section 6.8.7, p. 47
5	Sharpen the conceptual distinction between multimodal consistency/contradiction verification and conventional image-text retrieval.	The Abstract now states that, unlike conventional image-text retrieval, which ranks many candidates for a query, consistency verification is a pairwise decision on a single pair. Section 1.8 delimits the task to pairwise verification: given one image and one textual document the model returns a consistency score and a thresholded consistent/inconsistent decision; retrieval-style ranking was not the primary evaluation setting and Recall@K is not used as the main claim. The same statement is repeated in Section 6.7 and in the methodological contribution at Section 6.4.1. The primary metrics remain classification metrics with ROC-AUC and PR-AUC computed on the constructed pairs, and cosine similarity with a decision threshold is described as an explicit decision rule	Abstract, p. iv Section 1.8, p. 5 Section 3.5.5, p. 19 Section 3.9.1, p. 22 Section 6.4.1, p. 44 Section 6.7, p. 46

S/N	COMMENTS BY EXTERNAL EXAMINER	ACTION TAKEN	INDICATOR
		rather than a ranking problem.	
6	Moderate and clearly qualify claims, distinguishing what has been empirically demonstrated from prospective operational applications.	The Abstract no longer motivates the work by reference to an operational domain, and closes with an explicit statement of what the study is limited to. Chapter 6 qualifies its conclusions with "on this proxy benchmark" and "constructed in this study", and the contribution sections now refer to an imbalanced 90/10 class distribution instead of an operationally realistic one. Section 6.6.1 has been retitled from "Recommendations for Practice and Deployment" to "Recommendations for Practice"; the recommendation to fine-tune before deployment is replaced by a recommendation for domain-specific adaptation before any use outside the photograph-and-caption domain evaluated here. Appendix E has been reduced to a single set of alignment statements; the duplicated sections claiming work on real customs data have been deleted and the remaining claims are qualified as prospective ("may, in future work").	Abstract, pp. iv-v Section 6.4, pp. 43-44 Section 6.5, p. 44 Section 6.6.1, p. 45 Section 6.9, p. 47 Appendix E, pp. 58-59

S/N	COMMENTS BY VIVA VOCE PANEL	ACTION TAKEN	INDICATOR
1	The findings cannot show evidence of deployment in customs NII as Flickr30K does not represent the complexities in customs documentation or cargo inspection. This should be reflected in the study.	The sentence that motivated the work by reference to customs Non-Intrusive Inspection has been removed from the Abstract, and the Abstract now states that the findings should not be generalised beyond this benchmark. Section 2.8 now states that although similar consistency requirements can arise in specialised visual and textual settings, the present study does not evaluate those settings, and that	Abstract, pp. iv-v Section 1.8, p. 5 Section 2.8, p. 11 Section 6.7, p. 46 Section 6.8.1, p. 46

S/N	COMMENTS BY VIVA VOCE PANEL	ACTION TAKEN	INDICATOR
		Flickr30K photographs and captions are used as a proxy that does not represent domain-specific imagery or documents. Section 6.8.1 retains and sharpens the proxy-nature limitation: the results demonstrate feasibility on a controlled general-domain benchmark, and performance on purpose-built or domain-specific consistency data should be studied independently. No claim of deployment readiness in customs NII is made anywhere in the revised report.	
2	The change of domain from NII should be changed throughout the report.	References to customs Non-Intrusive Inspection have been removed from the Abstract and from the Problem Statement, the Significance of the Study and the Justification of the Study, where they had previously been listed as a target domain. The duplicated Appendix E sections written around Uganda Revenue Authority cargo risk assessment and "real customs data" have been deleted, and the Agenda 2063 discussion no longer argues the case through customs and border management. Two references remain deliberately and are framed as context rather than as the domain of the study: Section 1.7 (Scope), where customs cargo documentation is named as a potential future application rather than the focus of the work, and Section 2.8 (Literature Review), where the NII literature is reviewed and is now immediately qualified by the statement that this study does not evaluate those settings. The title and the objectives are framed as multimodal consistency verification and required no change.	Title page and Section 1.3 Abstract, pp. iv-v Sections 1.2, 1.5, 1.6, pp. 3-4 Section 1.7, p. 4 Section 2.8, p. 11 Appendix E, pp. 58-59

S/N	COMMENTS BY VIVA VOCE PANEL	ACTION TAKEN	INDICATOR
3	Cross-attention/fusion should have been considered to handle subtle contradictions in hard negatives.	Accepted as a structural limit of the dual-encoder design. The architecture is now declared a delimitation: cross-attention and other fusion mechanisms were not implemented, and the ablations were confined to ResNet-50 and TF-IDF with CLIP as the baseline under the same pair-construction protocol. Section 6.8.4 explains that image-only attention is computed without conditioning on the paired text, so the map cannot highlight regions that support or contradict particular tokens, and that the absence of cross-attention limits fine-grained grounding. Section 6.8.5 links the hard-negative gap (proposed model ROC-AUC approximately 0.70 against approximately 0.85 for CLIP) to global similarity without token-region grounding. Cross-attention and other explicit vision-language interaction mechanisms are carried as priority future work, and the absence of cross-modal attention is named among the remaining challenges in the Final Remarks. No fusion architecture was implemented in this study.	Section 1.8, p. 5 Section 6.6.2, p. 45 Section 6.7, p. 46 Section 6.8.4, p. 47 Section 6.8.5, p. 47 Section 6.9, p. 47
4	The inconsistent labels are generated algorithmically rather than being extensively human-annotated semantic contradictions. Therefore, a randomly mismatched pair may not represent the same type of inconsistency that would occur in a real-world verification system. This should be included in	A named limitation has been added, Section 6.8.2 "Algorithmic Inconsistent Labels": inconsistent labels were generated algorithmically rather than through extensive human annotation of semantic contradiction type; a randomly mismatched pair is not the same phenomenon as a real-world inconsistency arising from omission, substitution, negation or attribute mismatch; and performance on the inconsistency class therefore measures separation from these constructed	Abstract, p. iv Section 3.3.2, pp. 16-17 Section 4.3, p. 24 Section 6.3.1, p. 43 Section 6.8.2, p. 46

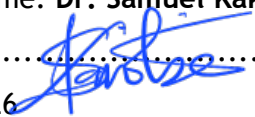
S/N	COMMENTS BY VIVA VOCE PANEL	ACTION TAKEN	INDICATOR
	the limitation.	negatives. The same qualification is carried at the point of construction in Section 3.3.2 and repeated in Section 4.3, which states that a randomly mismatched pair is a controlled proxy for inconsistency and not a labelled real-world contradiction. The Abstract states that the inconsistent pairs were generated algorithmically rather than through extensive human annotation of semantic contradiction type, and Section 6.3.1 repeats this when reporting the benchmark construction finding.	

Candidate's Name: **Mugimba Kakure Jude**

Signature: 

September, 2026

Supervisor's Name: **Dr. Samuel Kakuba**

Signature: 

September, 2026