



UGANDA CHRISTIAN
UNIVERSITY

A Centre of Excellence in the Heart of Africa

Department of Computing & Technology

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

By

Mugimba Kakure Jude

Reg No: J25M19/034, Student No: B31388

E-mail: mujuka2011@gmail.com

Phone No: 0782700919

A dissertation submitted to the Faculty of Engineering Design &
Technology for the study leading to the Degree in partial fulfillment
of the requirements for the award of **Master of Science in Data
Science** of Uganda Christian University

September, 2026

Declaration

I, Mugimba Kakure Jude, declare to the best of my knowledge that the work in this dissertation is original and has not been submitted before to any university/institution. It is the result of my effort.

Signature:  _____

September, 2026

Ethical Generative AI Usage Declaration

Students may use generative artificial intelligence (AI) ethically as a “study buddy” or assistant rather than a replacement for critical thinking. This includes citing AI use and fact-checking all outputs to maintain academic integrity.

For purposes of this assignment, the following shall guidelines shall apply.

- i. Generative AI may be used for brainstorming, and generating ideas to improve your work. You may also use AI for literature search, and summarizing journal papers for quick analysis and assessing relevance of publications. However, once the relevant publications have been identified, a student is expected to fully review the selected publications, identifying problems addressed, methods used, gaps identified and recommendations made.
- ii. Where AI has been used as guided in (i) above, full disclosure of how it has been used should be made in the appendix C, including how AI was used, the AI tools (including mode versions), and a compilation on major AI prompts used.
- iii. NO GENERATIVE AI CONTENT IS ALLOWED IN THE FINAL SUBMISSION. The final submission SHOULD be your original work.

A. Declaring the non-use of AI for content generation

I, Mugimba Kakure Jude, declare that no content generated by AI technologies has been used in this dissertation .

Signature:  _____

September, 2026

B. Acknowledging the use of AI as a study buddy

Please refer to Appendix C for the complete declaration of AI usage for generating ideas, brainstorming, literature search, and summarizing only to improve your work.

Approval

This dissertation has been submitted for examination with approval of the following Supervisors.

Supervisor: Dr. Samuel Kakuba
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christian University

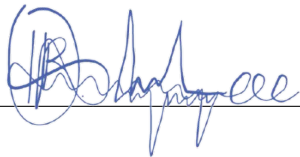
Signature: _____



September, 2026

Co - Supervisor: Eng. Ian Raymond Osolo
Department of Computing & Technology
Faculty of Engineering Design & Technology
Uganda Christy

Signature: _____



September, 2026

Abstract

Abstract

The rapid growth of multimodal data has underscored the necessity for robust methods to verify semantic alignment between images and their associated textual descriptions. This problem, termed multimodal consistency verification, asks whether a given image and text pair are semantically compatible or contradictory. Unlike conventional image-text retrieval, which ranks many candidates for a query, consistency verification is a pairwise decision on a single pair. While large-scale vision-language models demonstrate strong performance in retrieval and zero-shot tasks, explicit consistency verification, particularly with integrated explainability, remains underexplored.

We present an explainable dual-encoder vision-language architecture that utilizes a Vision Transformer (ViT) and Bidirectional Encoder Representations from Transformers (BERT) to encode images and multi-caption documents, respectively. These representations are projected into a shared embedding space and aligned via a symmetric InfoNCE contrastive objective. We quantify consistency using cosine similarity, employing a decision threshold to distinguish between consistent and inconsistent pairs. The model incorporates visual attention maps from the ViT and token importance scores from the BERT encoder that provide qualitative evidence for its classification decisions.

For evaluation, a proxy benchmark dataset was constructed from the Flickr30K dataset by aggregating the five captions associated with each image into a single multi-caption document and generating consistent, randomly mismatched, and Facebook AI Similarity Search (FAISS) retrieved hard-negative pairs. The inconsistent pairs were generated algorithmically rather than through extensive human annotation of semantic contradiction type. The model was tested across balanced, imbalanced, and hard-negative scenarios and compared with Contrastive Language-Image Pre-training (CLIP) as a baseline in all scenarios. To isolate the contributions of the transformer-based encoders, the Vision Transformer was replaced by Residual Network-50 (ResNet-50) to evaluate the impact of using a convolutional rather than an attention-based image encoder, whereas BERT was replaced by Term Frequency-Inverse Document Frequency (TF-IDF) representations to assess the value of contextual language modelling against a non-contextual text baseline.

The results show that the proposed model excels at identifying random image and textual mismatches on this proxy benchmark, achieving an ROC-AUC between 0.989 and 0.999, and maintains strong ranking quality under an imbalanced 90/10 class distribution (ROC-AUC 0.9894). However, performance declines substantially when faced with hard negatives (ROC-AUC approximately 0.70), indicating that semantically similar mismatches remain a challenge for dual-encoder architectures. The ablation results confirmed that both the transformer-based vision and language encoders are important for performance relative to the convolutional and TF-IDF alternatives. Qualitative analysis showed that attention maps and token rankings

highlight influential image regions and tokens; these visualisations were not evaluated with quantitative faithfulness tests and are not claimed as causal explanations. The study is therefore limited to a public multi-caption photograph benchmark with algorithmically constructed negatives, and the findings should not be generalised to domains whose images, documents, or contradiction types differ from this setting.

Keywords: multimodal consistency verification; vision-language models; contrastive learning; Vision Transformer; BERT; InfoNCE; explainable AI; attention maps; token importance; hard negatives; proxy benchmark dataset

Dedication

I dedicate this work to my wife, Eliza and children; they inspire me to go out and shine every day. I dedicate this also to my parents, Allen and Edmund Mugimba, who have guided me through life

Contents

Declaration	i
Ethical Generative AI Usage Declaration	ii
Approval	iii
Abstract	iv
Dedication	vi
Acknowledgments	1
1 Introduction	2
1.1 Background	2
1.2 Problem Statement	3
1.3 Objectives of the Study	3
1.3.1 General Objective	3
1.3.2 Specific Objectives	3
1.4 Research Questions	3
1.5 Significance of the Study	4
1.6 Justification of the Study	4
1.7 Scope of the Study	4
1.8 Conceptual Framework	5
1.8.1 Input, Processing, and Output Variables	5
1.9 Assumptions	5
2 Literature Review	7
2.1 Introduction	7
2.2 Multimodal Learning and Vision-Language Models	7
2.3 Contrastive Learning for Cross Modal Alignment	8
2.4 Image Text Matching and Consistency Verification	9
2.5 Transformer Architectures in Vision and Language	9
2.6 Explainability in Vision-Language Models	10
2.7 Evaluation Metrics and Benchmarks	11
2.8 Related Applications and Domain Considerations	11
2.9 Research Gap	12
2.10 Chapter Summary	12
3 Methodology	14
3.1 Introduction	14
3.2 Research Design	14

3.2.1	Ablation Study Design	15
3.3	Dataset Design and Preparation	16
3.3.1	Data Sources	16
3.3.2	Construction of Experimental Pairs	16
3.3.3	Data Characteristics and Split	17
3.4	Data Preprocessing	17
3.4.1	Image Preprocessing	17
3.4.2	Text Preprocessing	18
3.5	Model Architecture	18
3.5.1	Overview	18
3.5.2	Image Encoder	18
3.5.3	Text Encoder	19
3.5.4	Projection Heads and Shared Embedding Space	19
3.5.5	Consistency Scoring	19
3.5.6	Explainability Module	19
3.6	Contrastive Learning Framework	19
3.6.1	InfoNCE Objective	19
3.6.2	Training Strategy	20
3.7	Explainability Mechanisms	20
3.7.1	Visual Attention Maps	20
3.7.2	Token Importance Scores	21
3.8	Training Procedure	21
3.8.1	Optimiser and Learning Rate Schedule	21
3.8.2	Implementation Details	21
3.9	Evaluation Protocol	22
3.9.1	Quantitative Metrics	22
3.9.2	Qualitative Evaluation of Explanations	22
3.9.3	Threshold Analysis	23
3.10	Ethical Considerations	23
4	Experiments and Results	24
4.1	Introduction	24
4.2	Dataset	24
4.3	Pair Construction	24
4.4	Model Configuration	25
4.5	Baseline	25
4.6	Ablation Configurations	26
4.7	Evaluation Settings	26
4.8	Results of Training Dynamics	26
4.9	Cross-Modal Alignment Analysis	28
4.9.1	Similarity Matrix	28
4.9.2	Embedding Space Visualisation	28
4.9.3	Similarity Score Distributions	28
4.9.4	Decision Threshold Selection	29
4.10	Consistency Classification Performance	29
4.10.1	Random-Mismatch Setting (50/50)	29
4.10.2	Imbalanced Operational Setting (90/10)	29
4.10.3	Hard-Negative Setting (50/50)	30
4.10.4	Summary of Comparative Performance	30
4.11	Results from Ablation Studies	32

4.11.1	Image Encoder Ablation (ResNet-50 and BERT)	32
4.11.2	Text Encoder Ablation (ViT and TF-IDF)	33
4.11.3	Ablation Summary	33
4.12	Threshold Analysis	33
4.13	Qualitative Analysis of Explainability	34
4.13.1	Visual Attention Maps	34
4.13.2	Token Importance Scores	34
4.14	Summary of Findings	35
5	Discussion	37
5.1	Introduction	37
5.2	Achievement of Research Objectives	37
5.2.1	Benchmark Dataset Construction	37
5.2.2	Contrastive ViT and BERT Model	38
5.2.3	Consistency Performance and Explainability	38
5.3	Interpretation of Results in Relation to Existing Literature	38
5.3.1	Alignment with Contrastive Vision-Language Models	38
5.3.2	Performance on Random versus Hard Negatives	39
5.3.3	Contribution of the Architectural Components	39
5.3.4	Role and Limits of Dual Encoder Explainability	39
5.4	Practical Implications	40
5.4.1	Decision-Support Potential	40
5.4.2	Human Review of Sampled Pairs	40
5.4.3	Domain Adaptation Prospects	40
5.5	Chapter Summary	40
6	Conclusion and Recommendations	42
6.1	Introduction	42
6.2	Summary of the Study	42
6.3	Main Findings	43
6.3.1	Benchmark Dataset Construction	43
6.3.2	Contrastive Dual-Encoder Framework	43
6.3.3	Consistency Performance and Explainability	43
6.4	Contributions of the Study	43
6.4.1	Methodological Contribution	44
6.4.2	Empirical Contribution	44
6.4.3	Explainability Contribution	44
6.5	Conclusions	44
6.6	Recommendations	45
6.6.1	Recommendations for Practice	45
6.6.2	Recommendations for Future Research	45
6.7	Delimitations of the Study	46
6.8	Limitations of the Study	46
6.8.1	Proxy Nature of Flickr30K	46
6.8.2	Algorithmic Inconsistent Labels	46
6.8.3	Global Rather than Fine-Grained Alignment	47
6.8.4	Image-Only Attention and Absence of Cross-Attention	47
6.8.5	Hard-Negative Performance Gap	47
6.8.6	Threshold Sensitivity	47
6.8.7	Statistical and External Validation	47

6.9	Final Remarks	47
REFERENCES		52
A	Budget	53
B	Research Project timelines	54
C	Declaration of AI use	55
D	Major Journal Publications (Top 20)	56
E	Alignment to national (NDPIV) and Global SDGs	58
E.1	Alignment with the MSc Programme	58
E.2	Alignment with National Agendas	58
E.2.1	Alignment with National Development Plan IV (NDPIV) 2025/26–2029/30	58
E.2.2	Alignment with Uganda Vision 2040	59
E.2.3	Alignment with Other National Agendas	59
E.3	Alignment with Global Agendas	59
E.3.1	Alignment with Sustainable Development Goals (SDGs)	59
E.3.2	Alignment with Other Global Agendas	59
F	Algorithms, Code of interest	60
F.1	Overview	60
F.2	Data Preprocessing	60
F.2.1	Image Preprocessing	60
F.2.2	Text Preprocessing	60
F.3	Model Architecture	61
F.3.1	Image Encoder	61
F.3.2	Text Encoder	61
F.3.3	Projection to Shared Embedding Space	61
F.3.4	Consistency Scoring	61
F.4	Contrastive Learning Objective	62
F.5	Training Procedure	62
F.6	Explainability Mechanisms	62
F.6.1	Visual Attention Maps	62
F.6.2	Token Importance Scores	62
F.7	Evaluation Metrics	63

Acknowledgments

I acknowledge Dr. Samuel Kakuba, Dr Terhemba Michael-Ahile, and Eng. Ian Raymond Osolo, whose guidance has been invaluable in the compilation of this research

Chapter 1

Introduction

1.1 Background

As the volume of multimodal data continues to grow, there is a pressing need for intelligent systems that can bridge the gap between visual and textual information with human-level understanding [1, 2, 3, 4]. In many practical settings, an image and its accompanying textual description are expected to convey a consistent meaning. However, when errors or deliberate manipulations break this consistency, the reliability of the entire record is compromised, often leading to flawed downstream decisions.

Vision-Language Models (VLMs) have become the standard tool for aligning these different modalities. Foundational architectures like Contrastive Language–Image Pre-training (CLIP) [5], A Large-scale Image and Noisy-text embedding (ALIGN) [6], Align Before Fuse (ALBEF) [7], and Bootstrapping Language-Image Pre-training (BLIP) [8] have shown that contrastive learning is remarkably effective at building shared multimodal spaces. By using transformer based components such as the Vision Transformer (ViT) for images [9] and BERT style encoders for text [10, 11], these models can map visual and linguistic features into a single embedding space where their relationship can be measured.

Despite these breakthroughs, two major gaps remain. First, most current models are optimized for retrieval or generation [5, 6, 8], meaning they focus on finding the ‘best’ match in a shared embedding space [4, 3] rather than explicitly verifying if a specific pair is truthful or consistent.

Second, these models often operate as “black boxes,” offering little insight into how they reach a conclusion [12, 13, 14]. For high stakes environments that require transparency and human oversight, this lack of explainability is a significant hurdle [15, 16]. There is therefore a clear demand for models that not only verify consistency but also surface the evidence such as specific image regions or influential text tokens that influenced the decision [9, 17, 18].

This study addresses these gaps by developing an explainable dual-encoder architecture [5, 7]. By combining a Vision Transformer [9] and BERT [10] through contrastive learning [19, 20], we aim to create a system that provides both a reliable consistency score and interpretable visual and textual evidence [17, 18] to support human reviewers [15].

1.2 Problem Statement

While significant progress has been made in vision-language representation learning, current approaches still exhibit important shortcomings. Most models treat consistency between an image and its textual description as a secondary outcome of retrieval or classification [5, 6] rather than as an explicit decision support task [4]. In addition, the internal reasoning of these models is frequently opaque [12], making it difficult for human reviewers to understand why a particular image and text pair is judged consistent or inconsistent [15]. There remains a shortage of modular frameworks that combine reliable cross-modal alignment with transparent evidence generation [17] and that can be adapted to specialised operational settings. These limitations hinder the practical deployment of multimodal systems in applications that require both accuracy and accountability, such as document verification, content moderation, and specialised operational domains. To address this gap, the present study develops an explainable dual encoder vision-language model, using a Vision Transformer and BERT trained with a contrastive objective that treats multimodal consistency verification as a primary task and integrates visual and textual explanations as a core design requirement.

1.3 Objectives of the Study

1.3.1 General Objective

To develop an explainable Transformer based vision-language model for multimodal consistency verification between images and textual descriptions.

1.3.2 Specific Objectives

1. To develop a controlled multimodal evaluation dataset suitable for assessing consistency verification between images and textual descriptions.
2. To design and implement a contrastive learning based model leveraging the Vision Transformer (ViT) and BERT to learn aligned multimodal representations.
3. To evaluate the model’s performance in consistency verification and to assess the quality of the generated explanations.

1.4 Research Questions

The study seeks to answer the following questions:

1. How can a controlled multimodal evaluation dataset be constructed to effectively support the assessment of consistency verification between images and textual descriptions?
2. To what extent can a contrastive learning model combining the Vision Transformer and BERT produce reliable alignment between visual and textual representations for consistency verification?
3. How effectively can attention based visual maps and token importance scores provide human interpretable explanations for the model’s consistency decisions?

1.5 Significance of the Study

This research contributes to the field of multimodal machine learning in three main ways. First, it advances the development of models specifically oriented toward consistency verification rather than solely toward retrieval or generation. Second, it integrates explainability as a core design requirement, enabling human reviewers to inspect the visual and textual evidence underlying each decision. Third, it provides a practical, modular framework that can serve as a foundation for decision support systems in environments where transparency and human oversight are essential.

The outcomes of this study are expected to benefit researchers and practitioners working on trustworthy multimodal artificial intelligence, as well as organisations that require reliable image and text consistency assessment.

1.6 Justification of the Study

This study is justified by the need to address three interrelated gaps in the current literature. First, there is a shortage of practical frameworks that treat multimodal consistency verification as an explicit decision support task rather than a secondary outcome of retrieval. Second, explainability is frequently introduced as a post-hoc addition rather than being integrated as a core design requirement. Third, lightweight and modular architectures that combine reliable alignment performance with transparent evidence generation remain limited.

By developing an explainable Transformer based vision-language model that integrates a Vision Transformer and BERT through contrastive learning, this research makes both methodological and practical contributions. Methodologically, it advances the design of dual-encoder systems specifically oriented toward consistency assessment and integrated explainability. Practically, it provides a foundation for decision-support tools in domains where image and text consistency matters, such as document verification, content moderation, and archival validation.

The incorporation of visual attention maps and token level importance scores ensures that model outputs remain interpretable and subject to human review.

In summary, this study is motivated by the growing demand for trustworthy multimodal systems and the current scarcity of models that jointly deliver reliable consistency verification and transparent explanations.

1.7 Scope of the Study

This research focuses on the development and evaluation of a dual-encoder vision-language model that employs contrastive learning to assess semantic consistency between a single image and a corresponding textual description. The model produces a consistency score together with supporting explanations in the form of visual attention maps and influential text tokens.

Experiments are conducted primarily on the publicly available Flickr30K dataset [21], which provides multiple human-authored captions per image and therefore supports the controlled construction of consistent and inconsistent pairs. While the resulting framework is intended to be adaptable to specialised domains, detailed operational deployment for example in customs cargo documentation is treated as a potential future application rather than the central focus of the present work.

1.8 Conceptual Framework

The conceptual framework for this study integrates self supervised contrastive learning for cross-modal alignment with explainability mechanisms to support transparent multimodal consistency verification. The framework receives an image and a corresponding textual description as inputs. These inputs are independently encoded and projected into a shared embedding space, where contrastive learning aligns matched pairs and separates mismatched pairs. Consistency is quantified through cosine similarity, and the resulting score is used to classify the pair as consistent or inconsistent.

To support transparency and human oversight, the framework also generates visual attention maps and token importance scores. These explanations, together with the consistency decision, enable human in the loop decision support rather than fully automated judgement.

1.8.1 Input, Processing, and Output Variables

The key variables of the framework are summarised in Table 1.1.

Table 1.1: Conceptual Variables and Evaluation Attributes

Framework Component	Key Attributes	Expected Outcome	Out-	Evaluation Metrics	Met-
Textual Input	Textual description or document	Semantic representation		Token importance relevance	
Image Input	Visual image	Visual representation		Attention map quality	
Contrastive Learning(InfoNCE)	Cross-modal alignment	Improved image and text pair consistency		Similarity gap, AUC	
Consistency Scoring	Cosine similarity and threshold	Accurate consistency decision		Accuracy, Recall, ROC-AUC	Precision, F1-score,
Explainability Module	Attention maps and token importance	Transparent decision support		Qualitative assessment	human

The performance of the framework is assessed using standard classification metrics including accuracy, precision, recall, F1-score, and ROC-AUC [22] together with qualitative inspection of the generated explanations.

1.9 Assumptions

This study was based on several key assumptions that underpinned its methodology and research design.

Firstly, it was assumed that publicly available image caption datasets providing multiple captions per image, such as Flickr30K [21], constitute a scientifically valid proxy for evaluating multimodal consistency verification. This assumption was supported by the fact that Flickr30K contains independently authored captions that describe the same visual scene from different perspectives, thereby enabling the systematic construction of matched and mismatched pairs under controlled conditions.

Secondly, it was assumed that cosine similarity computed in a shared embedding space provides a meaningful quantitative measure of semantic consistency between an image and its corresponding textual description, and that an appropriate decision threshold can be established to distinguish consistent from inconsistent pairs.

Thirdly, it was assumed that attention weights extracted from the Vision Transformer and token level importance signals derived from BERT can serve as reasonably informative explanations of the model’s consistency decisions.

Chapter 2

Literature Review

2.1 Introduction

This chapter is organised around four interrelated themes that directly inform the present study. First, it examines foundational vision-language models and the architectural paradigms that underpin contemporary multimodal systems. Second, it discusses contrastive learning approaches for cross-modal alignment, which form the methodological basis of the proposed framework. Third, it reviews work on image and text matching and the emerging focus on consistency verification as a distinct, decision oriented task. Fourth, it surveys explainability techniques in vision-language models, with particular attention to attention based and token level methods that support human interpretation of model decisions.

2.2 Multimodal Learning and Vision-Language Models

Multimodal learning has emerged as a central research direction in artificial intelligence (AI) due to the growing availability of data that combines multiple modalities, particularly vision and language [3, 4]. The goal is to develop models capable of jointly processing and relating information from different sensory inputs in a way that approximates human perception [1, 2]. Within this broader field, vision-language models (VLMs) have received particular attention because of their ability to associate visual content with natural language descriptions [5, 8, 23].

Early approaches to vision-language modelling relied on convolutional neural networks (CNNs) for image feature extraction [14, 24] and sequential models for text encoding [18]. These methods faced limitations in capturing long range dependencies and in scaling to large datasets. The introduction of the Transformer architecture marked a turning point in the field [17]. Transformers enable more effective modelling of sequential and spatial relationships through self-attention mechanisms, providing a flexible foundation for both unimodal and multimodal representation learning [3, 4, 9, 10].

The modern era of vision-language modelling was largely initiated by dual encoder architectures trained with contrastive objectives [5, 6]. Contrastive Language Image Pre-training (CLIP) showed that a Vision Transformer (ViT) image encoder and a Transformer based text encoder can be jointly trained on large scale image and text pairs so that matched pairs are pulled together and mismatched pairs are pushed apart in a shared embedding space [5]. This design enabled strong zero-shot transfer across a wide range of visual recognition tasks and established the dual encoder contrastive paradigm that underpins much of subsequent work [5, 7, 8, 25].

Building on similar principles, A Large scale Image and Noisy text embedding (ALIGN) scales contrastive learning to over one billion noisy image-text pairs, confirming that data scale can compensate for noise in weakly supervised settings [6].

Subsequent models have sought to improve cross-modal interaction and generative capability [3, 4]. Flamingo introduces a visual language model designed for few-shot learning across multimodal tasks [23], while Contrastive Captioners (CoCa) combines contrastive and captioning objectives to create a unified image-text foundation model [25]. Align Before Fuse (ALBEF) and Bootstrapping Language-Image Pre-training (BLIP) further improve cross-modal understanding through momentum distillation and unified pre-training strategies [7, 8]. In the vision domain, ViT has become a standard backbone for image encoding by treating images as sequences of fixed size patches [9].

These foundational models have established the principal architectural and training paradigms that underpin contemporary research in multimodal representation learning [3, 4]. Dual-encoder designs offer efficiency and scalability [5, 6], while models that incorporate generative or fusion-based components provide richer cross-modal interaction [7, 8]. Despite their success on retrieval, classification, and generation benchmarks [5, 8, 23], most of these models are not primarily designed for explicit consistency verification [26], nor do they place strong emphasis on generating human interpretable explanations of their decisions [27].

2.3 Contrastive Learning for Cross Modal Alignment

Contrastive learning has become a dominant paradigm for aligning representations across modalities [5, 6, 19]. The central idea is to learn an embedding space in which matched pairs from two modalities are drawn closer together while mismatched pairs are pushed apart [19]. In the vision-language setting, this is typically achieved by maximising the similarity between corresponding image and text pairs and minimising the similarity between non corresponding pairs within a training batch [5, 19].

A widely adopted formulation is the Information Noise Contrastive Estimation (InfoNCE) loss, which treats alignment as a classification task over positive and negative pairs [19]. Dual-encoder models such as CLIP operationalise this principle by encoding images and texts independently and optimising a symmetric cross-modal contrastive loss [5]. ALIGN extends the same approach to a much larger and noisier collection of image-text pairs, showing that scale can compensate for imperfections in the supervisory signal [6]. Both models demonstrate that global contrastive alignment is sufficient to produce strong zero-shot and retrieval performance across a range of downstream tasks [5, 6].

Subsequent research has explored refinements to the basic contrastive framework [4]. Some approaches combine contrastive learning with generative objectives to capture richer cross-modal relationships [8]. CoCa, for example, integrates a contrastive loss with a captioning loss within a unified image-text foundation model [25]. Flamingo illustrates how large language models can be conditioned on visual inputs to support few-shot multimodal learning [23]. More recently, CrossModal-CLIP has applied contrastive multimodal learning to network traffic anomaly detection, illustrating the breadth of domains where cross-modal alignment is valuable [28].

Despite these advances, most contrastive VLMs remain primarily optimised for retrieval or zero-shot classification rather than for explicit consistency verification [5, 6, 26]. The similarity score produced by such models can be interpreted as a measure of cross-modal alignment, yet relatively few studies have treated consistency assessment as a decision oriented task supported by transparent explanations [26, 27].

2.4 Image Text Matching and Consistency Verification

Image and text matching has traditionally been formulated as a retrieval problem, in which the goal is to rank the most relevant texts for a given image or the most relevant images for a given text [3, 4]. CLIP and ALIGN demonstrate that large scale contrastive pre-training produces embeddings in which matched image and text pairs exhibit high cosine similarity, enabling effective cross-modal retrieval without task specific fine tuning [5, 6]. Subsequent foundation models, including CoCa and Flamingo, further improve cross-modal understanding by combining contrastive learning with generative or few shot capabilities [23, 25].

Although retrieval performance provides an indirect indication of cross-modal alignment, it does not fully address the problem of consistency verification [26]. In this study, multimodal consistency verification is defined as the task of determining whether a specific image and a specific textual description are semantically compatible, and of supporting that determination with a calibrated decision rather than a ranked list. In many practical settings, the output of interest is not the identity of the nearest neighbour in a large gallery, but a judgement about whether the given pair is consistent or inconsistent [29, 30, 31].

Recent research has begun to treat consistency-related problems more explicitly. Multimodal approaches have been explored for anomaly detection, fact verification, and the identification of mismatches between visual content and associated textual claims [12, 32]. These developments suggest a shift from pure retrieval towards decision-oriented multimodal assessment [33, 34]. Nevertheless, most existing models remain optimised for ranking metrics such as Recall@K and median rank [5]. Comparatively less attention has been paid to the design of systems whose primary output is a consistency score accompanied by human interpretable evidence [35, 36].

This distinction is important for applications that require accountability and human oversight [37, 38]. It is not sufficient for a model to embed images and texts in a shared space; it must also support transparent assessment of whether a particular pair is consistent [35, 36]. The present study addresses this need by framing multimodal consistency verification as an explicit task and integrating explainability mechanisms into the decision pipeline.

2.5 Transformer Architectures in Vision and Language

The Transformer architecture has become a foundational building block for both natural language processing (NLP) and computer vision [10, 9]. Originally introduced for sequence transduction tasks, the Transformer relies on self-attention mechanisms to model pairwise dependencies among input elements without the sequential constraints of recurrent networks [17]. This design enables efficient parallelisation and proves highly effective at capturing long-range dependencies [17, 18].

In the language domain, Bidirectional Encoder Representations from Transformers (BERT) popularised the use of bidirectional Transformer encoders pre-trained with masked language modelling objectives. BERT and its variants, including A Robustly Optimized BERT Pretraining Approach (RoBERTa), produce contextualised token representations adaptable to a wide range of downstream tasks [10, 11]. The special classification token, commonly referred to as the [CLS] token, is frequently used as a global sequence representation for classification and retrieval applications [10].

In computer vision, ViT extended the same architectural principles to image data by dividing an image into fixed-size patches and treating the resulting sequence of patch embeddings as input

to a standard Transformer encoder [9]. ViT demonstrated that pure attention-based models can achieve competitive or superior performance compared with CNNs when trained at sufficient scale [13]. Subsequent work, including BERT Pre-Training of Image Transformers (BEiT), has further reinforced the ViT as a strong visual backbone for multimodal systems [39].

The combination of Transformer-based vision and language encoders has underpinned many of the most successful VLMs [3, 4]. Dual-encoder architectures such as CLIP use a ViT for images and a Transformer-based text encoder, mapping both modalities into a shared embedding space through contrastive learning [5]. Other models build on similar backbones while introducing additional fusion or generative components [25, 23, 8, 7].

Transformer architectures offer several advantages for multimodal consistency verification [17, 26]. Self-attention models global context within each modality [17], while the resulting embeddings can be compared directly through similarity metrics [19, 40]. Critically, attention weights also offer a natural source of interpretability, indicating which image patches or text tokens the model attends to most strongly [17, 18].

2.6 Explainability in Vision-Language Models

As VLMs are increasingly considered for decision-support applications, explainability has become an important research concern [38, 34]. Many high-performing multimodal models operate as complex neural systems whose internal decision processes are not immediately transparent [26]. In settings that require accountability and human oversight, it is not sufficient to produce an accurate consistency score; the basis of that score must also be interpretable to human reviewers [38, 15].

Transformer-based architectures offer inherent opportunities for explanation through their reliance on self-attention [17, 18, 38]. Attention weights indicate the relative importance assigned to different image patches or text tokens during representation learning [17, 9, 18]. In dual-encoder models such as CLIP, these attention patterns can identify which regions of an image or which words in a caption most strongly influence the resulting embedding [5]. Such analyses provide a first level of insight into model behaviour, even though attention does not constitute a complete causal explanation [18, 38].

Beyond raw attention, a range of techniques has been developed to improve the interpretability of VLMs [27, 38]. These include methods that aggregate attention across layers and heads [9, 18], gradient-based attribution approaches [18], and techniques that generate visual saliency maps conditioned on textual inputs [5, 35]. The aim is to produce explanations that are faithful to the model’s computation and accessible to non-expert users [38, 27].

Despite this progress, explainability remains only partially integrated into many multimodal systems [27, 38]. Explanation mechanisms are frequently applied post hoc rather than being treated as a core design objective [38, 41, 42]. Furthermore, relatively few studies have examined explainability specifically in the context of consistency verification [26, 35, 36]. Existing work demonstrates that useful visual and textual signals can be extracted from Transformer based models [42, 43], yet the systematic combination of consistency scoring with integrated, human-interpretable explanations remains limited [26, 36].

A model that produces only a numerical similarity score leaves human reviewers without the evidence needed to support their judgement [38, 37]. By contrast, a framework that accompanies the consistency decision with visual attention maps and influential text tokens can better support human in the loop assessment [15, 42, 43, 9, 17].

2.7 Evaluation Metrics and Benchmarks

The evaluation of VLMs has traditionally been dominated by retrieval-oriented and task specific benchmarks [5, 6, 8]. In image and text retrieval, performance is commonly reported using ranking metrics such as Recall at Top- k (Recall@K) which measures whether the correct or relevant item appears among the top k retrieved results and median rank, which measures the ability of a model to retrieve the correct match from a large candidate set [5, 21]. Dual encoder models including CLIP and ALIGN have achieved strong results on standard retrieval benchmarks [5, 6]. Classification oriented tasks such as zero shot image classification and visual question answering further test the generalisation of learned representations [5, 8]. Generative VLMs are additionally evaluated using captioning metrics and human preference judgements [23, 25].

For consistency verification, however, retrieval and generation metrics are only partially informative [26, 35]. Consistency assessment is more naturally framed as a binary or graded decision problem: given a specific image and text pair, the model must determine the degree to which the two modalities are semantically compatible [35, 36]. Under this framing, classification metrics such as accuracy, precision, recall, F1-score, and area under the receiver operating characteristic curve (ROC-AUC) become more appropriate [22]. Similarity score distributions for matched and mismatched pairs also provide useful diagnostic information about the quality of cross-modal alignment [2, 35].

The evaluation of explainability introduces additional challenges [27]. Unlike accuracy based metrics, the quality of visual attention maps and token importance rankings is difficult to quantify through automatic measures alone [41, 42, 43]. Human inspection therefore remains an important component of assessment, particularly when explanations are intended to support decision making rather than merely visualise internal activations [15, 38]. Studies emphasising transparent multimodal reasoning consequently combine quantitative performance metrics with qualitative analysis of explanatory outputs [16, 36].

Existing public datasets constructed for image captioning and retrieval can be adapted for consistency evaluation by creating matched and mismatched pairs [21]. Hard negative pairs, which are semantically similar but not matched to the query image, can be mined using efficient similarity search methods such as Facebook AI Similarity Search (FAISS) [40]. Nevertheless, purpose built benchmark datasets that explicitly annotate different types of inconsistency remain limited, constraining systematic comparison across methods and underscoring the need for evaluation protocols that jointly assess decision accuracy and explanation quality [33, 34, 35].

2.8 Related Applications and Domain Considerations

Multimodal consistency verification is relevant across a range of domains where visual content and textual descriptions are expected to align. In document verification and digital archiving, systems may need to confirm that an image corresponds to its accompanying metadata or textual record [33]. In content moderation and media authenticity assessment, detecting inconsistencies between an image and a textual claim can support the identification of misleading or manipulated material [34, 36]. Similar requirements arise in retrieval augmented systems, product cataloguing, and other settings where image and text compatibility underpins reliable decision making [3, 4].

Similar consistency requirements can arise in specialised visual and textual settings, however, the present study does not evaluate those settings. Flickr30K photographs and captions are used as a proxy and do not represent domain-specific imagery or documents.

In specialised operational environments such as security screening and customs Non-Intrusive Inspection (NII), radiographic images and textual declarations are routinely examined together. Research in these domains has explored the use of deep learning for threat detection, anomaly identification, and automated analysis of X-ray imagery [44, 45, 31, 46]. The risks associated with cargo misdeclaration and the challenges of large scale customs verification further motivate the development of automated consistency tools [47, 30, 29]. Such work demonstrates the practical importance of linking visual evidence with associated textual information.

Across these domains, a recurring theme is the need for systems that go beyond producing a similarity score or classification label [27, 36]. When model outputs influence high stakes decisions, human reviewers benefit from explanations that indicate which visual regions and textual elements most strongly influenced the assessment [15]. This requirement aligns with broader calls for trustworthy and interpretable AI in multimodal settings [38].

Despite the diversity of potential applications, many existing VLMs remain optimised for general purpose retrieval or generation rather than for domain oriented consistency assessment [5, 8, 35]. Adaptation to specialised data distributions, calibration of decision thresholds, and integration into human in the loop workflows therefore remain important open considerations [2, 15].

2.9 Research Gap

The literature reviewed in this chapter documents substantial progress in multimodal representation learning, contrastive cross-modal alignment, and Transformer based vision-language modelling [5, 6, 9, 17]. Foundational dual encoder models such as CLIP and ALIGN have shown that large scale contrastive learning can produce effective shared embedding spaces for images and text [5, 6]. Subsequent architectures have further improved cross-modal understanding through generative objectives, few-shot conditioning, and richer pre-training strategies [7, 8, 23, 25]. In parallel, attention-based Transformer architectures have provided both strong unimodal backbones and a basis for interpretability [9, 10, 17].

Despite these advances, two important gaps remain. First, most existing VLMs have been designed and evaluated with retrieval, zero-shot classification, or generation in mind [5, 6, 8]. Explicit consistency verification that asks whether a specific image and a given textual description are semantically compatible has received comparatively little attention as a standalone task [26, 35]. Standard retrieval metrics such as Recall@K offer only an indirect proxy for consistency, and they fall short of what decision oriented applications actually demand [3, 4].

Second, explainability is frequently treated as a secondary or post-hoc concern rather than as an integral design objective [12, 13]. Although attention weights and related attribution methods offer useful signals [17, 18], relatively few frameworks systematically combine consistency assessment with human interpretable visual and textual evidence. In settings that require accountability and human oversight, the absence of transparent explanations limits the practical utility of otherwise accurate models [15].

2.10 Chapter Summary

This chapter has reviewed the literature underpinning the development of an explainable Transformer based VLM for multimodal consistency verification, covering the foundations of multimodal learning and vision-language modelling, contrastive cross-modal alignment, image and

text matching and consistency assessment, Transformer architectures for vision and language, explainability techniques, evaluation practices, and related application domains.

The review found that substantial progress has been made in learning aligned multimodal representations, particularly through dual-encoder contrastive models [5, 6, 7, 8]. Two limitations were also identified: most models are designed for retrieval or generation rather than explicit consistency verification, and explainability tends to be treated as an afterthought rather than a core design objective [27, 35]. As a result, relatively few frameworks bring together reliable consistency scoring and integrated, human interpretable evidence [9, 10, 15, 17].

Together, these gaps motivate the present study. The following chapter describes the research methodology developed to address them, covering the overall research design, dataset construction, model architecture, training strategy, explainability mechanisms, and evaluation protocol.

Chapter 3

Methodology

3.1 Introduction

In this chapter we present the methodology adopted to develop and evaluate an explainable Transformer based vision-language model for multimodal consistency verification. The study addressed the need for systems capable of assessing whether an image and a corresponding textual description are semantically consistent, while also providing transparent evidence to support human interpretation of model decisions. The methodology was structured around three main research objectives: the design of a suitable benchmark for consistency evaluation, the development of a dual-encoder contrastive learning framework based on a Vision Transformer and BERT, and the systematic assessment of both consistency verification performance and the quality of the generated explanations. The chapter describes the overall research design, dataset construction, data preprocessing, model architecture, contrastive training strategy, explainability mechanisms, evaluation protocol, and ablation studies. The approach combined self-supervised cross-modal alignment with explicit consistency scoring and integrated explainability. Matched and mismatched image and text pairs were used to train and evaluate the model, cosine similarity in a shared embedding space served as the consistency measure, and visual attention maps together with token importance scores were generated to support transparent decision support. The subsequent sections provide a detailed account of each component of the methodological framework.

3.2 Research Design

The study adopted an experimental research design to develop and evaluate an explainable dual-encoder vision-language model for multimodal consistency verification. The design focused on constructing a controlled experimental setting in which the ability of the model to align and assess the consistency of image and text pairs could be systematically examined, together with the usefulness of its explanations [3, 5, 6]. The research process consisted of four interrelated phases. In the first phase, a controlled evaluation dataset was prepared from Flickr30K by combining the five human-authored captions associated with each image into a single multi-caption textual document [21]. Three types of image–document pairs were then constructed [35, 33]: consistent pairs, formed by pairing each image with its own multi-caption document; randomly mismatched pairs, formed by pairing each image with multi-caption documents sampled from other images; and hard-negative pairs, generated using Facebook AI Similarity Search (FAISS) to retrieve challenging non-matching documents [40]. Given a query embedding q and a candidate document embedding x , FAISS [40] retrieved nearest neighbours

according to cosine similarity

$$\text{sim}(q, x) = \frac{q^\top x}{\|q\|_2 \|x\|_2}, \quad (3.1)$$

and a hard negative was selected as [48]

$$x^- = \arg \max_{x \in \mathcal{X} \setminus \{x^+\}} \text{sim}(q, x), \quad (3.2)$$

where x^+ denotes the true matching document and $\mathcal{X}$ is the set of candidate documents. In the second phase, a dual encoder architecture was designed, comprising a Vision Transformer for image encoding and a BERT-based encoder for text encoding, with projection heads mapping both modalities into a shared embedding space [9, 10, 5]. In the third phase, the model was trained using a contrastive learning objective applied primarily to the consistent image and document pairs [19]. In the fourth phase, the trained model was evaluated on consistent, randomly mismatched, and hard negative pairs using quantitative consistency metrics and qualitative analysis of the generated visual and textual explanations. The experimental design treated consistency verification as an explicit decision task. Cosine similarity between the projected image embedding z_v and text embedding z_t was used as the consistency score [5, 6],

$$s(z_v, z_t) = \frac{z_v^\top z_t}{\|z_v\|_2 \|z_t\|_2}, \quad (3.3)$$

and a decision threshold τ was applied to classify pairs as consistent or inconsistent [35, 26]:

$$\hat{y} = \begin{cases} 1, & \text{if } s(z_v, z_t) \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (3.4)$$

Explainability was incorporated as a core component of the design through the extraction of attention maps from the Vision Transformer and token importance scores from the text encoder, supporting human in the loop interpretation of model decisions [42, 43, 15]. For an image with N patches, the visual attention map was obtained by averaging the self-attention weights assigned to the classification token [9, 42],

$$A_i = \frac{1}{H} \sum_{h=1}^H \alpha_{[\text{CLS}],i}^{(h)}, \quad i = 1, \dots, N, \quad (3.5)$$

where $\alpha_{[\text{CLS}],i}^{(h)}$ denotes the attention weight from the classification token to patch i in head h . Token importance for a text sequence of length L was likewise derived from the corresponding attention weights of the text encoder [10, 42],

$$w_j = \frac{1}{H} \sum_{h=1}^H \alpha_{[\text{CLS}],j}^{(h)}, \quad j = 1, \dots, L. \quad (3.6)$$

The proposed model was evaluated comparatively against CLIP, a foundational dual encoder contrastive baseline [5]. In addition, controlled ablation studies were conducted to isolate the contribution of the individual architectural components.

3.2.1 Ablation Study Design

To assess the contribution of the key architectural choices, a series of controlled ablation experiments was conducted. The full proposed model consisting of ViT-B/16 and BERT base served

as the reference configuration. First, the Vision Transformer was replaced with ResNet-50, a foundational convolutional neural network. The final fully connected classification layer of ResNet-50 was removed and replaced by a linear adapter that mapped the 2048-dimensional feature vector to a 768-dimensional representation, matching the dimensionality of the BERT output. The text encoder (BERT), projection heads, contrastive objective, optimiser, learning rate schedule, data splits, and evaluation protocols remained identical to those of the full model. This ablation isolated the contribution of the Transformer based visual backbone relative to a classical convolutional architecture.

Second, the BERT encoder was replaced with a classical non-neural baseline based on Term Frequency–Inverse Document Frequency (TF-IDF) [49]. A TF-IDF vectoriser was fitted on the multi-caption documents of the training split. The resulting sparse feature vectors were projected into the same 256-dimensional shared embedding space through a learnable linear layer. All other components including image encoder, projection dimension, InfoNCE loss [19], optimiser, and evaluation settings were kept unchanged. This ablation examined the contribution of the contextual Transformer text encoder relative to a traditional bag of words representation [50]. In both ablation experiments, the same group based data splits (70% training, 15% validation, and 15% test), pair construction procedures, temperature scaled InfoNCE objective, AdamW optimiser, and evaluation metrics were used. Performance was measured under the three complementary settings employed for the main model: balanced random mismatches (50/50), imbalanced random mismatches (90/10), and balanced FAISS hard negatives (50/50). This design ensured that observed performance differences could be attributed to the changed component rather than to variations in the experimental protocol.

3.3 Dataset Design and Preparation

3.3.1 Data Sources

The study utilised the Flickr30K dataset, a publicly available image and caption resource in which each image is accompanied by five independent human authored captions [21]. For each image, the five captions were combined into a single multi-caption textual document. This produced one richer textual representation per image and formed the basis for constructing the experimental pairs used in training and evaluation [5, 6].

3.3.2 Construction of Experimental Pairs

Three complementary sets of image and text pairs were constructed:

1. Each image was paired with its own multi-caption document to form a consistent pair. These pairs represented genuine semantic consistency and constituted the primary positive examples used for contrastive learning.
2. Each image was paired with multi-caption documents randomly sampled from other images within the same dataset to form random mismatched pairs. These pairs served as controlled negative examples in which the textual content did not correspond to the visual content.
3. Hard-negative pairs were generated and reserved for evaluation using FAISS-based similarity search. FAISS was employed to retrieve multi-caption documents that were representationally similar to a given image’s document but belonged to different images [40].

The inconsistent class was constructed algorithmically. Random negatives paired an image

with a multi-caption document drawn from a different image. Hard negatives were documents retrieved with FAISS as nearest neighbours in embedding space and were reserved exclusively for evaluation; they were not used as negatives inside the InfoNCE training objective. Neither procedure produced human annotation of contradiction type such as; omission, substitution, negation, or attribute mismatch. A randomly mismatched pair is therefore a controlled proxy for inconsistency, not a labelled real-world contradiction. Validation was limited to monitoring the contrastive loss and selecting the decision threshold on the validation split before the held-out test evaluation.

Contrastive learning was performed exclusively on the consistent pairs. During evaluation, model performance was assessed under three complementary settings: first, discrimination between consistent and randomly mismatched pairs; second, a balanced test distribution comprising 50% consistent pairs and 50% hard negative pairs; and third, an imbalanced distribution of 90% consistent pairs and 10% randomly mismatched pairs, designed to approximate real-world anomaly detection scenarios in which inconsistencies are relatively rare. This multi-setting evaluation design enabled a more rigorous assessment of the model’s ability to discriminate genuine matches from both random and challenging non matches under conditions that reflect both controlled and operationally realistic class distributions.

3.3.3 Data Characteristics and Split

The resulting collection of pairs was partitioned into training, validation, and test subsets using a group-based splitting strategy based on a unique image identifier. This approach ensured that all pairs associated with the same image remained within a single split, thereby preventing data leakage across the training, validation, and test sets. The data were divided in the proportions of 70% for training, 15% for validation, and 15% for testing. The training split contained consistent pairs used for contrastive optimisation. The validation split supported model monitoring and decision threshold selection. The held-out test split was reserved for final performance assessment under both the random mismatch setting and the balanced hard negative evaluation setting. This group aware partitioning and three-way pair construction provided a controlled yet challenging experimental foundation for multimodal consistency verification.

3.4 Data Preprocessing

3.4.1 Image Preprocessing

Image data were preprocessed to conform to the input requirements of the Vision Transformer (ViT). Each image was resized to a fixed resolution of 224×224 pixels, consistent with the standard input size of the ViT-B/16 architecture [9]. The images were then normalised using the ImageNet mean (μ) and standard deviation (σ) statistics commonly associated with pre-trained vision backbones.

$$\mu = [0.485, 0.456, 0.406], \quad \sigma = [0.229, 0.224, 0.225]. \quad (3.7)$$

These values standardise pixel intensity distributions across the red, green, and blue channels. Finally, the processed images were converted into tensors suitable for batch-wise input to the neural network. These steps ensured that all visual inputs were presented to the image encoder in a uniform format, enabling stable feature extraction and subsequent projection into the shared embedding space.

3.4.2 Text Preprocessing

Textual inputs consisted of the multi-caption documents formed by combining the five human-authored captions associated with each image. For the proposed model and the ResNet-50 ablation, the documents were tokenised using the BERT tokenizer, which converted the raw text into subword tokens, applied padding and truncation to a fixed maximum sequence length of 128, and generated corresponding attention masks [10, 11]. For the TF-IDF ablation, the multi-caption documents were vectorised using a TF-IDF representation fitted exclusively on the training split. The resulting sparse feature vectors were subsequently projected into the shared embedding space through a learnable linear layer. Together, the image and text preprocessing pipelines produced standardised multimodal inputs that supported consistent encoding, contrastive alignment, and downstream consistency verification.

3.5 Model Architecture

3.5.1 Overview

The proposed model followed a dual-encoder architecture designed for multimodal consistency verification, as illustrated in Figure 3.1.

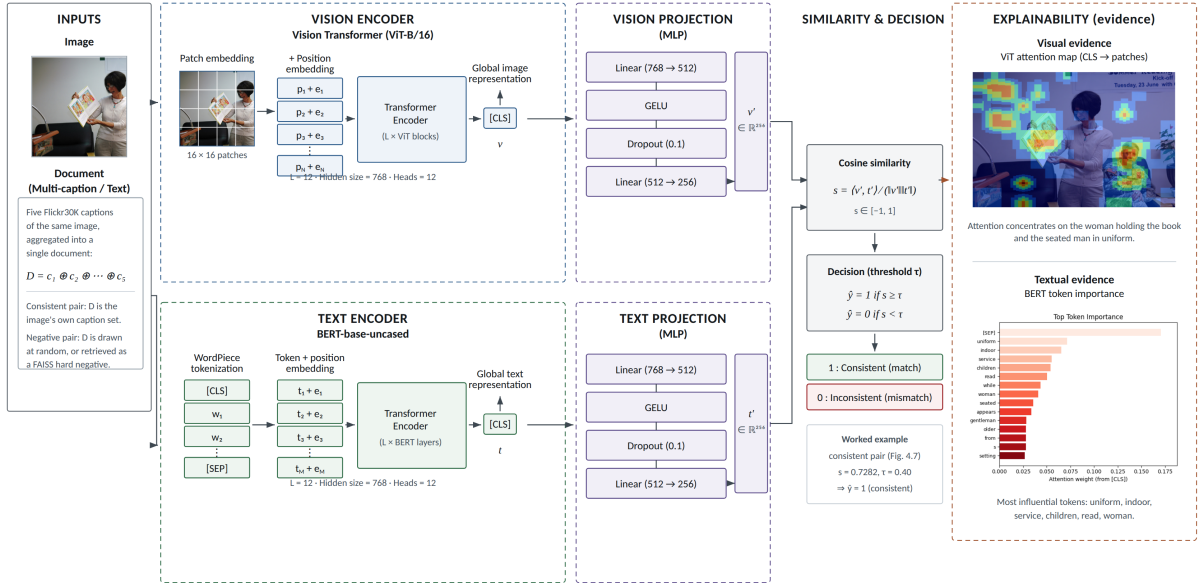


Figure 3.1: Architecture of the proposed explainable dual-encoder vision-language model for multimodal consistency verification.

3.5.2 Image Encoder

The image encoder was implemented using a pre-trained Vision Transformer (ViT-B/16) because this architecture captures long range spatial dependencies through self-attention and provides attention weights that can be used directly for visual explanation. Each input image, after preprocessing to 224×224 pixels, was divided into non overlapping 16×16 patches. The patches were linearly projected into embeddings, combined with positional encodings, and passed through a stack of Transformer encoder layers. The final representation corresponding to the classification token was extracted as the global image embedding $E_v \in \mathbb{R}^{768}$. Attention

weights from the Transformer layers were retained to support the generation of visual attention maps [9].

3.5.3 Text Encoder

The text encoder was implemented using a pre-trained BERT-base model because BERT produces contextual token representations that capture long range semantic dependencies and provides token-level attention signals suitable for textual explanation. The multi-caption document associated with each image was tokenised and passed through the BERT encoder. The final hidden state corresponding to the [CLS] token was used as the global text embedding $E_t \in \mathbb{R}^{768}$. Token level signals from the encoder were retained to support the computation of token importance scores for textual explainability [10, 11].

3.5.4 Projection Heads and Shared Embedding Space

The image and text embeddings (E_v and E_t) were passed through separate linear projection heads and subsequently L2-normalised. This mapping placed both modalities in a shared 256-dimensional embedding space in which cross-modal similarity could be computed directly. The shared space formed the basis for contrastive alignment during training and for consistency scoring during inference [5, 6].

3.5.5 Consistency Scoring

Consistency between an image and a textual document was quantified as the cosine similarity s between their projected embeddings. A decision threshold τ was applied to the resulting similarity score in order to classify each pair as consistent or inconsistent. This formulation treated consistency verification as an explicit decision task rather than as a retrieval ranking problem.

3.5.6 Explainability Module

The explainability module generated two complementary forms of evidence. Visual attention maps were derived from the attention weights of the Vision Transformer and highlighted image regions that contributed most strongly to the model’s representation of the input. Token importance scores were derived from the text encoder and ranked the words in the multi-caption document according to their influence on the textual representation. These explanations, together with the consistency decision, constituted the final output of the system and facilitated human-in-the-loop assessment.

3.6 Contrastive Learning Framework

3.6.1 InfoNCE Objective

The model was trained using a contrastive learning objective based on the InfoNCE loss [19]. The purpose of this objective was to align the projected embeddings of consistent image–document pairs in the shared embedding space while separating them from non-matching pairs. Given a batch of N consistent image–document pairs, let $\mathbf{z}_i^v$ and $\mathbf{z}_i^t$ denote the L2-normalised projected embeddings of the i -th image and its corresponding multi-caption document, respectively. The cosine similarity between an image embedding and a text embedding

was computed as

$$\text{sim}(\mathbf{z}_i^v, \mathbf{z}_j^t) = \mathbf{z}_i^v \cdot \mathbf{z}_j^t. \quad (3.8)$$

The image-to-text contrastive loss for the batch was defined as [19]

$$\mathcal{L}_{v \rightarrow t} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(\mathbf{z}_i^v, \mathbf{z}_i^t)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(\mathbf{z}_i^v, \mathbf{z}_j^t)/\tau)}, \quad (3.9)$$

where τ is a temperature parameter that controls the concentration of the similarity distribution. A symmetric text-to-image loss $\mathcal{L}_{t \rightarrow v}$ was defined analogously. The overall contrastive objective was the average of the two directional losses [5]:

$$\mathcal{L}_{\text{contrastive}} = \frac{1}{2}(\mathcal{L}_{v \rightarrow t} + \mathcal{L}_{t \rightarrow v}). \quad (3.10)$$

This formulation encouraged high similarity for matched pairs and low similarity for mismatched pairs within each training batch [5, 6].

3.6.2 Training Strategy

Contrastive learning was performed exclusively on consistent image–document pairs. Within each training batch, the other samples in the batch served as in-batch negatives. Hard negatives generated via FAISS were reserved for evaluation and were not included in the training objective [40]. The model was optimised using the AdamW optimiser, which was selected for its decoupled weight-decay mechanism, its proven stability with Transformer-based architectures, and its widespread adoption in the training of vision-language dual-encoder models [51, 11, 9, 5]. A learning-rate schedule was applied to stabilise training and to promote gradual convergence. Training proceeded by iteratively sampling batches of consistent pairs, computing the symmetric InfoNCE loss, and updating the parameters of the projection heads and, where applicable, the encoder backbones through backpropagation. This training strategy enabled the model to learn a shared embedding space in which cosine similarity provided a meaningful measure of semantic consistency between an image and its multi-caption document, forming the basis for the subsequent consistency verification and explainability stages.

3.7 Explainability Mechanisms

Explainability was incorporated as an integral component of the framework rather than as a post-hoc addition. The goal was to provide human-interpretable evidence that could support the inspection of individual consistency decisions. Two complementary mechanisms were implemented: visual attention maps derived from the image encoder and token importance scores derived from the text encoder.

3.7.1 Visual Attention Maps

Visual explanations were obtained from the self-attention weights of the Vision Transformer. During the forward pass, attention weights from the final Transformer blocks were extracted. The attention tensor for a given layer has the shape (B, H, N, N) , where B is the batch size, H is the number of attention heads, and N is the number of tokens (including the classification token). To produce a single spatial map per image, the following steps were applied:

1. The attention weights were averaged across all heads, yielding a matrix of shape (N, N) .

2. The attention row corresponding to the classification ([CLS]) token was selected. This vector of length N indicates the degree to which the global representation attended to each patch token.
3. The classification-token entry was discarded, leaving a vector of length equal to the number of image patches (for ViT-B/16 with 224×224 input, this is $14 \times 14 = 196$).
4. The vector was reshaped into a 14×14 spatial grid.
5. The grid was upsampled to the original image resolution (224×224) using bilinear interpolation.
6. The resulting map was normalised to the range $[0, 1]$ and overlaid on the input image as a colour heatmap.

Regions with higher attention values were interpreted as having contributed more strongly to the global image embedding that entered the shared space. These visual attention maps enabled qualitative inspection of whether the model focused on semantically relevant parts of the image when assessing consistency with the accompanying multi-caption document [9].

3.7.2 Token Importance Scores

Textual explanations were derived from the BERT encoder. Token-level importance was estimated by aggregating attention weights associated with individual tokens in the multi-caption document. Tokens were ranked according to their relative contribution to the final text embedding. The highest-ranked tokens were presented as evidence of the textual elements that most strongly influenced the model’s representation of the document. In consistent pairs, these tokens frequently corresponded to distinctive objects or attributes visible in the image; in inconsistent pairs, they often highlighted concepts that were absent from or contradicted by the visual content. This ranking provided a compact and interpretable summary of the linguistic basis for the consistency assessment [10, 11]. Together, the visual attention maps and token importance scores supplied complementary evidence that linked the model’s consistency decision to specific image regions and text tokens, thereby supporting transparent, human-in-the-loop evaluation.

3.8 Training Procedure

3.8.1 Optimiser and Learning Rate Schedule

The model was optimised using the AdamW optimiser. AdamW was selected because of its decoupled weight-decay mechanism, its proven stability with Transformer-based architectures, and its widespread adoption in the training of vision-language dual-encoder models [51, 11, 9, 5]. Decoupling weight decay from the adaptive gradient update improves regularisation and generally leads to better generalisation when fine-tuning large pre-trained encoders. A linear warmup–decay learning-rate schedule was employed to vary the learning rate smoothly over the course of training. This schedule supported stable early progress and more refined parameter updates in later epochs, thereby promoting gradual convergence of the contrastive objective.

3.8.2 Implementation Details

The framework was implemented in Python using the PyTorch deep learning library. Pre-trained Vision Transformer and BERT weights were loaded through the Hugging Face Trans-

formers library, which provided standardised access to model checkpoints and tokenisers. Image transformations were performed with TorchVision, and numerical and visualisation utilities were drawn from NumPy, Matplotlib and Seaborn. All experiments were conducted on Google Colab using an NVIDIA Tesla T4 GPU (approximately 15 GB VRAM) with the High-RAM runtime enabled (approximately 51 GB system memory). The T4 provided CUDA accelerated training and inference, while the High-RAM configuration supported the memory requirements of dual-encoder batch processing, embedding extraction, and FAISS based hard negative retrieval [40]. Training was conducted with fixed random seeds to support reproducibility. Mixed precision computation was used in order to improve training efficiency while preserving numerical stability. Batches of consistent image and document pairs were sampled, passed through the dual encoders and projection heads, and used to compute the symmetric InfoNCE loss [19]. Gradients were then back propagated and the model parameters updated. Model checkpoints were monitored on the validation split. The decision threshold for consistency classification was selected on the basis of validation performance prior to final evaluation on the held out test set.

3.9 Evaluation Protocol

The evaluation protocol was designed to assess both the quantitative performance of the model in distinguishing consistent from inconsistent image and text pairs and the qualitative usefulness of its explanations. Performance was examined under multiple test distributions in order to reflect both controlled and operationally realistic conditions.

3.9.1 Quantitative Metrics

Quantitative evaluation assessed the model’s ability to distinguish consistent from mismatched image and text pairs using cosine similarity scores. Performance was measured through similarity statistics (mean cosine similarity for consistent and inconsistent pairs and the separation between their distributions), classification metrics at the selected decision threshold (accuracy, balanced accuracy, precision, recall, and F1-score for both the consistent and inconsistent classes), and discrimination metrics (ROC-AUC and PR-AUC). These measures were computed under three complementary test conditions. First, a balanced set of consistent versus randomly mismatched pairs (50/50). Second, an imbalanced set of 90% consistent and 10% randomly mismatched pairs, designed to approximate operational settings in which inconsistencies are relatively rare and thirdly, a balanced set of 50% consistent and 50% FAISS retrieved hard negative pairs [40]. For each setting the proposed model was compared with CLIP under identical data splits and evaluation protocols [5]. The same three evaluation settings were also applied to the ResNet-50 and TF-IDF ablations.

3.9.2 Qualitative Evaluation of Explanations

In addition to numerical performance, the quality of the generated explanations was examined qualitatively. Visual attention maps were inspected to determine whether the model attended to semantically relevant regions of the image. Token importance rankings were examined to assess whether the most influential words in the multi-caption document corresponded to concepts that were present or absent in the visual content. This qualitative analysis provided insight into the transparency of individual decisions and supported the interpretation of both correct and incorrect consistency judgements. The combination of similarity scores, attention maps, and token rankings formed the basis for human-in-the-loop assessment of the model’s outputs.

3.9.3 Threshold Analysis

Because consistency verification was formulated as a thresholded decision on cosine similarity, the choice of the decision threshold τ directly influenced classification performance. A pair was classified as consistent if

$$\hat{y} = \begin{cases} 1, & \text{if } s(z_v, z_t) \geq \tau, \\ 0, & \text{otherwise.} \end{cases} \quad (3.11)$$

where $s(z_v, z_t)$ denotes the cosine similarity between the projected image and text embeddings [35, 26]. Threshold analysis was conducted on the validation set by examining the trade-off between precision and recall (and related metrics) across a range of threshold values. The operating threshold was selected using Youden’s index [22],

$$J(\tau) = \text{Sensitivity}(\tau) + \text{Specificity}(\tau) - 1 = \text{TPR}(\tau) - \text{FPR}(\tau), \quad (3.12)$$

and the optimal threshold was obtained as

$$\tau^* = \arg \max_{\tau} J(\tau). \quad (3.13)$$

On the 90/10 evaluation this procedure yielded an optimal threshold of 0.4297, which was subsequently applied for binary classification on the held-out test set. This analysis clarified the sensitivity of the model’s decisions to the threshold parameter and provided a transparent, data-driven basis for the final consistency classification rule [38]. In deployment the threshold can be further adjusted according to the relative costs of false positives and false negatives.

3.10 Ethical Considerations

The study was conducted with due regard for ethical research practice. The experiments relied exclusively on the publicly available Flickr30K dataset, which contains natural images and human-authored captions and does not include personal, sensitive, or commercially confidential information [21]. No human participants were recruited, and no private or operational data from any organisation were used. Although Flickr30K provided a practical and reproducible basis for constructing consistent and mismatched pairs, it remains a proxy dataset. It was originally designed for captioning and retrieval rather than consistency verification, is limited in domain diversity, and does not annotate fine-grained forms of inconsistency. These characteristics constrain the direct generalisability of the results to specialised operational environments. The framework was designed to support human-in-the-loop decision making rather than only a fully automated judgement. Consistency scores and explanations were intended to assist human reviewers, with final authority remaining with the human decision-maker. This design choice reduces the risk of opaque automated decisions being treated as definitive. Potential biases inherited from the pre-trained Vision Transformer and BERT models were acknowledged [9, 10]. Such biases may influence both the consistency scores and the generated explanations. While the present study did not include a formal bias audit, this limitation is recognised and identified as an important direction for future work. Overall, the research adhered to principles of transparency, responsible use of public data, and respect for human oversight in multimodal decision support.

Chapter 4

Experiments and Results

4.1 Introduction

This chapter presents the experimental evaluation of the proposed explainable Transformer-based vision-language model for multimodal consistency verification. The experiments were designed to assess the model’s ability to distinguish consistent from inconsistent image–text pairs and to examine the quality of the explanations generated by the system.

The chapter reports results on five main aspects. First, it examines the training dynamics of the contrastive learning process. Second, it analyses the quality of cross-modal alignment through similarity matrices, embedding visualisations, and similarity score distributions. Third, it evaluates consistency classification performance under three complementary test conditions: random mismatches, hard negatives, and an imbalanced operational setting. Fourth, it presents controlled ablation studies that isolate the contribution of the image and text encoders. Fifth, it provides a qualitative assessment of the visual attention maps and token importance scores that support transparent decision interpretation.

Throughout the chapter, the proposed model is compared with CLIP under identical evaluation protocols [5]. The findings provide empirical evidence of the model’s strengths, its remaining limitations on hard negatives, and the practical value of the integrated explainability mechanisms.

4.2 Dataset

Experiments were conducted on the Flickr30K dataset [21]. After removal of missing, duplicate, and corrupt records, 31,783 images remained. Each image is associated with five independently authored human captions. These captions were retained as the textual source material from which experimental documents were later formed. The data were partitioned using a group-based splitting strategy based on image identifiers (70% training, 15% validation, and 15% testing). This ensured that all pairs derived from the same image remained within a single split and thereby prevented data leakage.

4.3 Pair Construction

Pair construction was treated as a distinct experimental procedure because the type of image and document pairing directly determines the difficulty of the consistency verification task. Three types of pairs were constructed.

Consistent pairs were formed by combining the five captions associated with each image into a single multi-caption document and pairing that document with its source image. These pairs served as the positive examples for contrastive training and as the matched class during evaluation.

Random mismatched pairs were created by pairing each image with a multi-caption document sampled from a different image. These pairs provided an unconstrained negative class and were used to assess discrimination under relatively easy mismatch conditions.

Hard-negative pairs were generated using FAISS-based nearest-neighbour retrieval and were reserved exclusively for evaluation [40]. For a query embedding q , a hard negative was selected as

$$x^- = \arg \max_{x \in \mathcal{X} \setminus \{x^+\}} \text{sim}(q, x), \quad (4.1)$$

where x^+ denotes the true matching document and $\text{sim}(q, x)$ denotes cosine similarity. This procedure produced semantically similar but non-matching documents and therefore constituted a more challenging test of consistency verification.

A randomly mismatched pair is a controlled proxy for inconsistency, not a labelled real-world contradiction since the inconsistent class was constructed algorithmically.

4.4 Model Configuration

The proposed model employed a dual-encoder architecture. A pre-trained Vision Transformer (ViT-B/16) served as the image encoder and a pre-trained BERT-base model served as the text encoder [9, 10]. The resulting 768-dimensional embeddings were passed through linear projection heads and L2-normalised into a shared 256-dimensional embedding space. Consistency was measured by cosine similarity between the projected embeddings. Training used the symmetric InfoNCE contrastive loss applied to consistent pairs only, with the AdamW optimiser and a linear warm-up followed by decay schedule [19, 51]. A decision threshold was selected on the validation set using Youden’s index for subsequent binary consistency classification.

4.5 Baseline

The proposed model was compared with CLIP (Contrastive Language–Image Pre-training), a strong dual-encoder contrastive vision-language model [5]. The CLIP ViT-B/16 variant was used. CLIP encodes images with a Vision Transformer Base model using 16×16 patches and encodes text with a 12-layer Transformer language model of hidden size 512. Both representations are projected into a shared embedding space, and alignment is measured using cosine similarity. Because CLIP is trained at large scale with a symmetric contrastive objective, it provides a natural baseline for dual-encoder consistency scoring.

The two models share the same design with independent image and text encoders, a shared embedding space, and cosine similarity as the consistency score. However, they differ in several important respects. First, CLIP is a large-scale pre-trained foundation model, whereas the proposed model adapts ViT-B/16 and BERT-base through a comparatively lightweight contrastive fine-tuning procedure. Second, CLIP’s text encoder is a CLIP-specific Transformer trained from scratch, whereas the proposed model uses BERT-base. Third, the proposed framework treats consistency verification as an explicit decision task and applies a validation-selected threshold τ to classify pairs as consistent or inconsistent. Fourth, the proposed model generates visual attention maps and token-importance scores as part of its output, whereas CLIP was used in

this study only as a scoring baseline. Fifth, the proposed text encoder operates on multi-caption documents constructed from Flickr30K, rather than on the single-caption inputs used in the original CLIP setting.

Both models were evaluated under the same pair-construction procedures and the same three evaluation settings. This comparison therefore provided a direct reference point for assessing the relative effectiveness of the proposed framework under controlled, imbalanced, and hard-negative conditions.

4.6 Ablation Configurations

Two controlled ablations were conducted. In the first, the Vision Transformer was replaced by ResNet-50 while retaining BERT as the text encoder. In the second, the BERT text encoder was replaced by a TF-IDF representation projected into the same shared embedding space, while retaining the Vision Transformer as the image encoder. All other training and evaluation settings were held constant.

4.7 Evaluation Settings

Model performance was assessed under three complementary evaluation settings:

1. Balanced random-mismatch setting (50% consistent / 50% randomly mismatched pairs).
2. Imbalanced operational setting (90% consistent / 10% randomly mismatched pairs), designed to approximate real-world scenarios in which inconsistencies are relatively rare.
3. Hard-negative setting (50% consistent / 50% FAISS-retrieved hard negatives) [40].

These settings enabled assessment of both controlled discrimination ability and performance under more challenging and operationally realistic conditions.

4.8 Results of Training Dynamics

The contrastive training process was monitored through the InfoNCE loss on both the training and validation sets over 10 epochs [19]. Figure 4.1(a) presents the resulting learning curves, and Table 4.1 reports the corresponding epoch-wise values.

The training loss decreased sharply in the first two epochs, from 0.8566 to 0.1365, and thereafter declined more gradually, reaching 0.0102 by epoch 10. The validation loss followed a similar but smoother trajectory, falling from 0.1925 in epoch 1 to 0.0450 by the final epoch. Both curves exhibited steady improvement without a subsequent rise in validation loss, indicating stable convergence and no evident overfitting within the training horizon.

The learning rate followed a linear warm-up and decay schedule. It started at 2.00×10^{-5} in epoch 1 and decreased monotonically to 0 by epoch 10, as shown in Table 4.1. This gradual reduction supported early rapid alignment and later fine adjustment of the embeddings.

These dynamics confirm that the symmetric InfoNCE objective, optimised with AdamW, successfully reduced the contrastive loss and produced progressively better-aligned image and text embeddings in the shared space [51].

Table 4.1: Training dynamics of the proposed ViT-BERT model

Epoch	Train InfoNCE loss	Validation InfoNCE loss	Learning rate
1	0.8566	0.1925	2.00×10^{-5}
2	0.1365	0.1048	1.78×10^{-5}
3	0.0574	0.0844	1.56×10^{-5}
4	0.0373	0.0735	1.33×10^{-5}
5	0.0267	0.0660	1.11×10^{-5}
6	0.0214	0.0618	8.89×10^{-6}
7	0.0164	0.0531	6.67×10^{-6}
8	0.0121	0.0508	4.44×10^{-6}
9	0.0112	0.0466	2.22×10^{-6}
10	0.0102	0.0450	0.00

Proposed Model - Contrastive Training Diagnostics

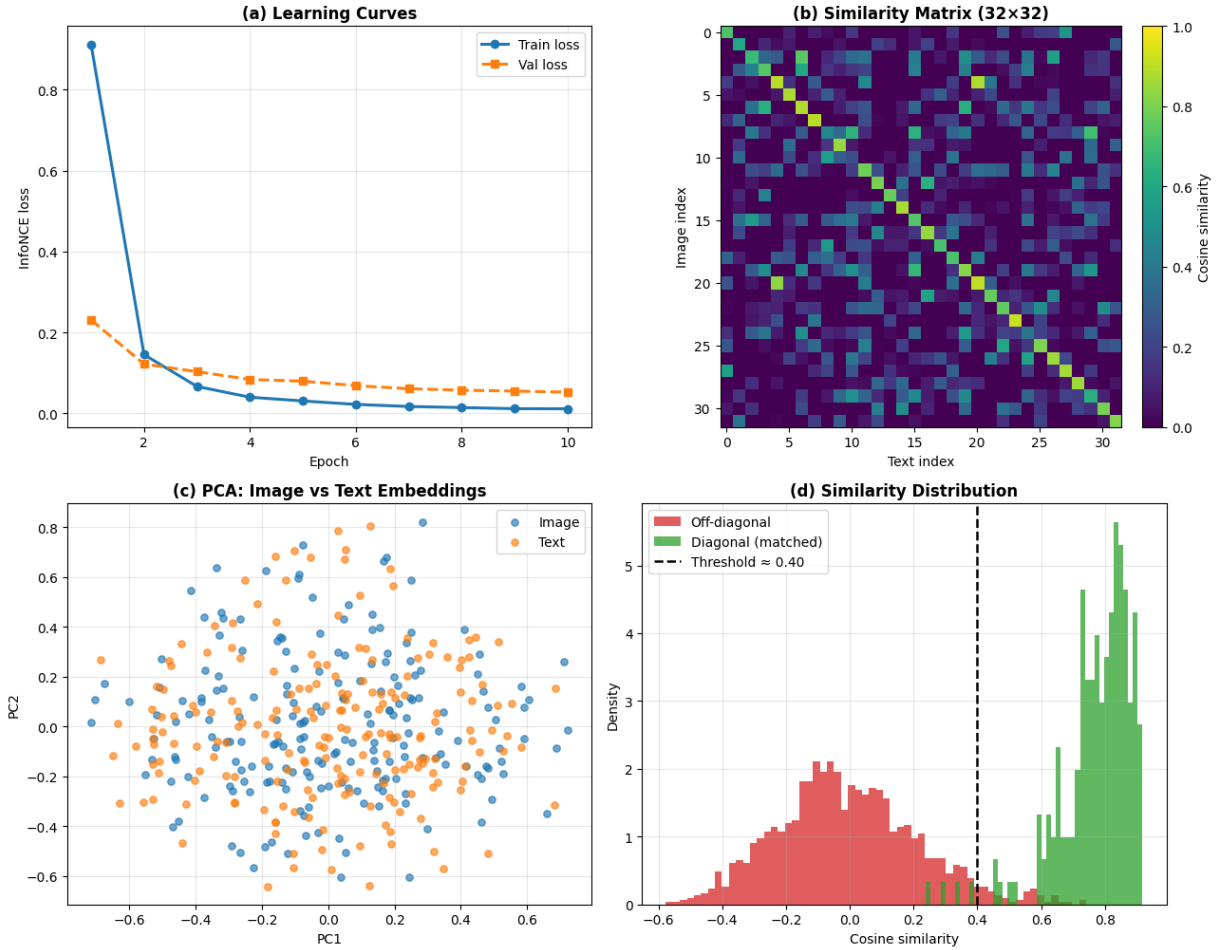


Figure 4.1: Contrastive training diagnostics for the proposed model. (a) Learning curves of training and validation InfoNCE loss. (b) Cosine similarity matrix for a batch of 32 pairs. (c) PCA projection of image and text embeddings. (d) Distribution of matched (diagonal) and unmatched (off-diagonal) cosine similarities with the selected decision threshold.

4.9 Cross-Modal Alignment Analysis

4.9.1 Similarity Matrix

Figure 4.1 (b) shows the cosine similarity matrix for a representative batch of 32 image–document pairs. A pronounced diagonal of high similarity values is visible, indicating that matched pairs received substantially higher scores than non-matched combinations. Off-diagonal entries are predominantly low, confirming that the contrastive objective successfully separated positive pairs from in-batch negatives.

4.9.2 Embedding Space Visualisation

Figure 4.1 (c) presents a two-dimensional PCA projection of the image and text embeddings after contrastive training. The image embeddings and text embeddings occupy a largely overlapping region of the projected space, consistent with the formation of a shared multimodal representation.

4.9.3 Similarity Score Distributions

Figure 4.1 (d) shows the distributions of cosine similarity scores for matched and unmatched pairs. Matched pairs form a clear peak at high cosine values (approximately 0.7–0.95), while unmatched pairs concentrate near zero. The limited overlap between the two distributions indicates that cosine similarity provides a strong signal for distinguishing consistent from inconsistent pairs.

A broader comparison across evaluation settings is presented in Figure 4.2. On random mismatches the proposed model exhibits a wide dynamic range and clear separation between positive and negative pairs. CLIP also separates the two classes well, although its similarity scores occupy a narrower numerical range. On hard negatives both models show substantial overlap between consistent and inconsistent distributions.

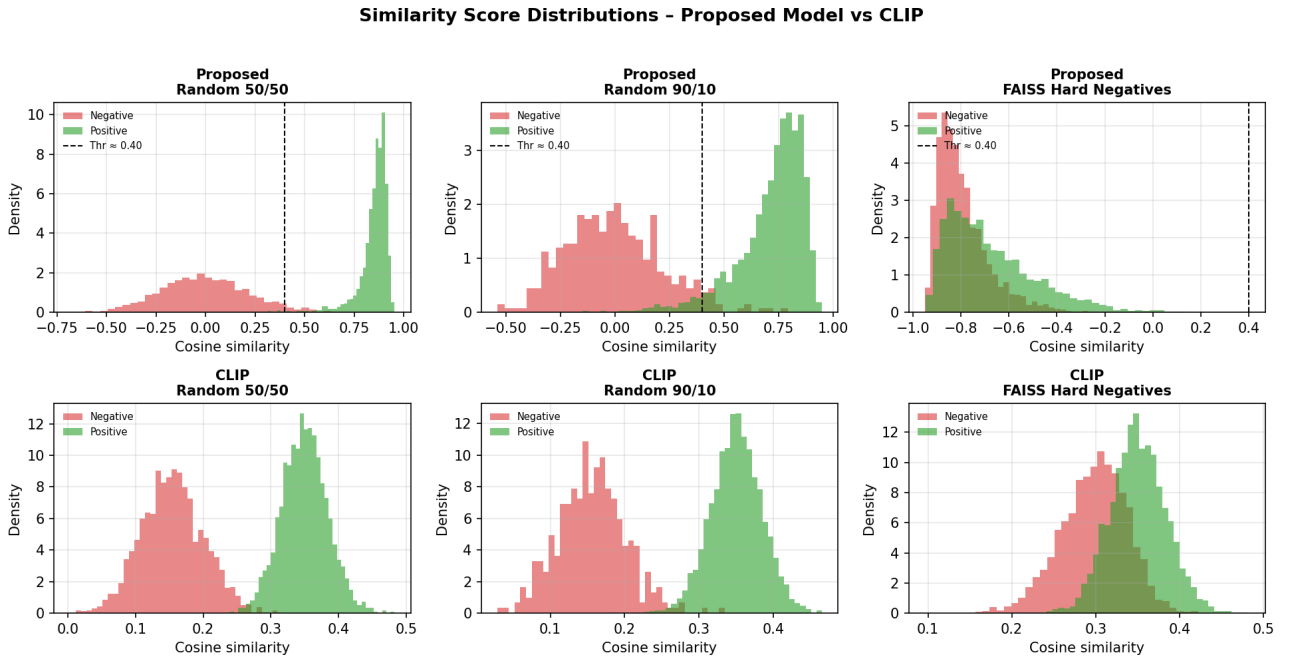


Figure 4.2: Similarity score distributions for the proposed model (top row) and CLIP (bottom row) under random negatives (50/50 and 90/10) and FAISS hard negatives.

4.9.4 Decision Threshold Selection

On the basis of the observed similarity distributions and Youden’s index computed on the validation set, a decision threshold in the region of 0.40–0.59 was selected depending on the evaluation setting. Scores above the threshold were classified as consistent and scores below it as inconsistent.

4.10 Consistency Classification Performance

Consistency classification performance was evaluated under three complementary settings. Cosine similarity served as the decision score and a threshold selected via Youden’s index was applied to obtain binary predictions. The proposed model was compared with CLIP under identical protocols [5].

4.10.1 Random-Mismatch Setting (50/50)

Under the balanced consistent versus randomly mismatched evaluation both models achieved near-ceiling discrimination.

Table 4.2: Performance on the balanced 50% consistent / 50% randomly mismatched setting

Metric	Proposed model	CLIP
ROC-AUC	0.9991	0.9996
PR-AUC	0.9991	0.9996
Accuracy	0.9873	0.9920
Balanced Accuracy	0.9873	0.9920
Precision (Consistent)	0.9836	0.9920
Recall (Consistent)	0.9910	0.9908
F1-score (Consistent)	0.9873	0.9920
Precision (Inconsistent)	0.9909	0.9919
Recall (Inconsistent)	0.9835	0.9900
F1-score (Inconsistent)	0.9872	0.9919
Optimal threshold (Youden)	0.5212	0.2627
Mean similarity (Consistent)	0.8368	0.3492
Mean similarity (Inconsistent)	0.0047	0.1556

4.10.2 Imbalanced Operational Setting (90/10)

An imbalanced evaluation set of 4,000 pairs (3,600 consistent and 400 randomly mismatched) was constructed to approximate operational conditions in which inconsistencies are relatively rare.

Table 4.3: Performance under the imbalanced 90% consistent / 10% inconsistent setting

Metric	Proposed model	CLIP
ROC-AUC	0.9994	0.9985
PR-AUC	0.9999	0.9998
Accuracy	0.9850	0.9908
Balanced Accuracy	0.9883	0.9904
Precision (Consistent)	0.9992	0.9989
Recall (Consistent)	0.9842	0.9908
F1-score (Consistent)	0.9916	0.9948
Precision (Inconsistent)	0.8744	0.9231
Recall (Inconsistent)	0.9925	0.9900
F1-score (Inconsistent)	0.9297	0.9554
Optimal threshold (Youden)	0.5900	0.2669
Mean similarity (Consistent)	0.8380	0.3494
Mean similarity (Inconsistent)	-0.0087	0.1556

The proposed model achieved excellent ranking quality (ROC-AUC 0.9994) and very high recall on the minority inconsistent class (0.9925). Precision on the inconsistent class was also strong (0.8744). CLIP retained a slight advantage on most metrics, particularly precision for the minority class, while the proposed model remained highly competitive.

4.10.3 Hard-Negative Setting (50/50)

Performance declined when the models were evaluated on the balanced hard-negative distribution.

Table 4.4: Performance on the balanced 50% consistent / 50% hard-negative setting

Metric	Proposed model	CLIP
ROC-AUC	0.699	0.848
PR-AUC	0.697	0.841
Accuracy	0.649	0.766
Balanced Accuracy	0.649	0.766
Precision (Consistent)	0.65	0.75
Recall (Consistent)	0.64	0.80
F1-score (Consistent)	0.65	0.77
Precision (Inconsistent)	0.65	0.78
Recall (Inconsistent)	0.66	0.74
F1-score (Inconsistent)	0.65	0.76
Optimal threshold (Youden)	0.40	0.324
Mean similarity (Consistent)	-0.68	0.351
Mean similarity (Hard Neg)	-0.79	0.298

4.10.4 Summary of Comparative Performance

Table 4.5 consolidates the key metrics across all three evaluation settings.

Table 4.5: Summary comparison of the proposed model and CLIP

Setting	Metric	Proposed model	CLIP
50/50 Random	ROC-AUC	0.9991	0.9996
	Accuracy	0.9873	0.9920
	F1 (Inconsistent)	0.9872	0.9919
90/10 Random	ROC-AUC	0.9994	0.9985
	Accuracy	0.9850	0.9908
	F1 (Inconsistent)	0.9297	0.9554
50/50 Hard Negative	ROC-AUC	0.699	0.848
	Accuracy	0.649	0.766
	F1 (Hard Negative)	0.65	0.76

Comparative ROC Curves - Proposed Model vs CLIP

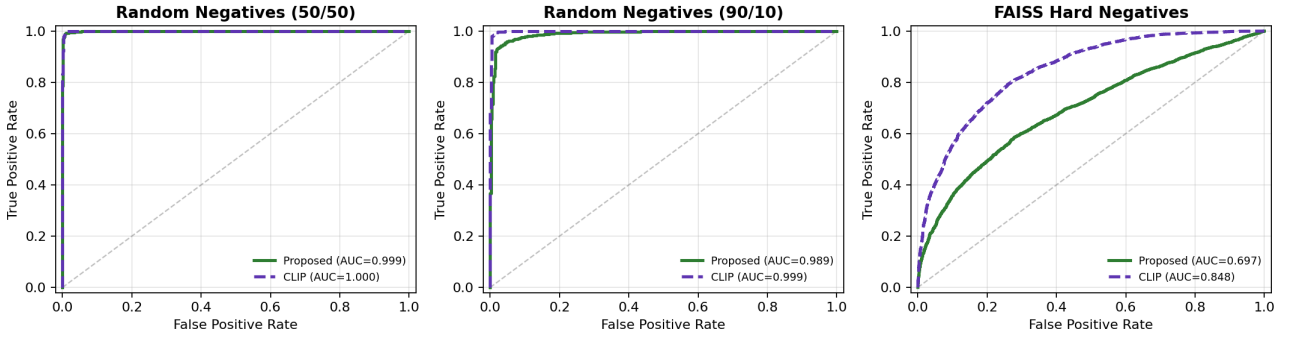


Figure 4.3: Comparative ROC curves of the proposed model and CLIP under three evaluation settings: balanced random negatives (50/50), imbalanced random negatives (90/10), and FAISS hard negatives.

Both models discriminate random mismatches very effectively. The proposed model achieved near-ceiling performance (ROC-AUC ≥ 0.999), while CLIP retained a small advantage on most classification metrics. The comparison is summarised in Figures 4.4 and 4.5.

On hard negatives the gap widens in favour of CLIP (ROC-AUC 0.848 versus 0.699). Semantically similar non-matches therefore remain the principal challenge for the current dual-encoder framework. The proposed model nevertheless delivers strong performance under both balanced and operationally imbalanced random-mismatch conditions, and additionally supplies the integrated explainability outputs that constitute a core contribution of this work.

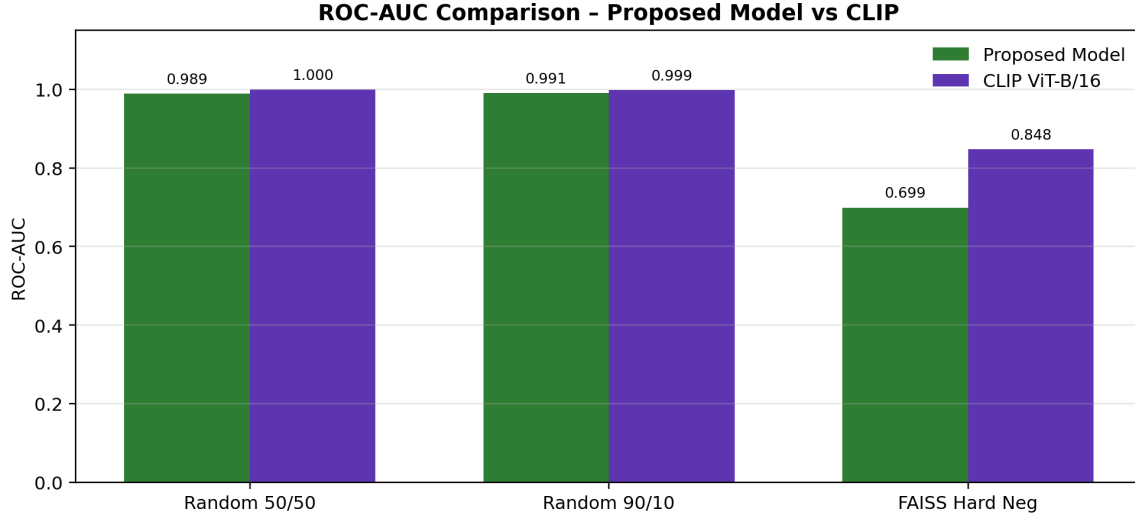


Figure 4.4: ROC-AUC comparison of the proposed model and CLIP across evaluation settings.

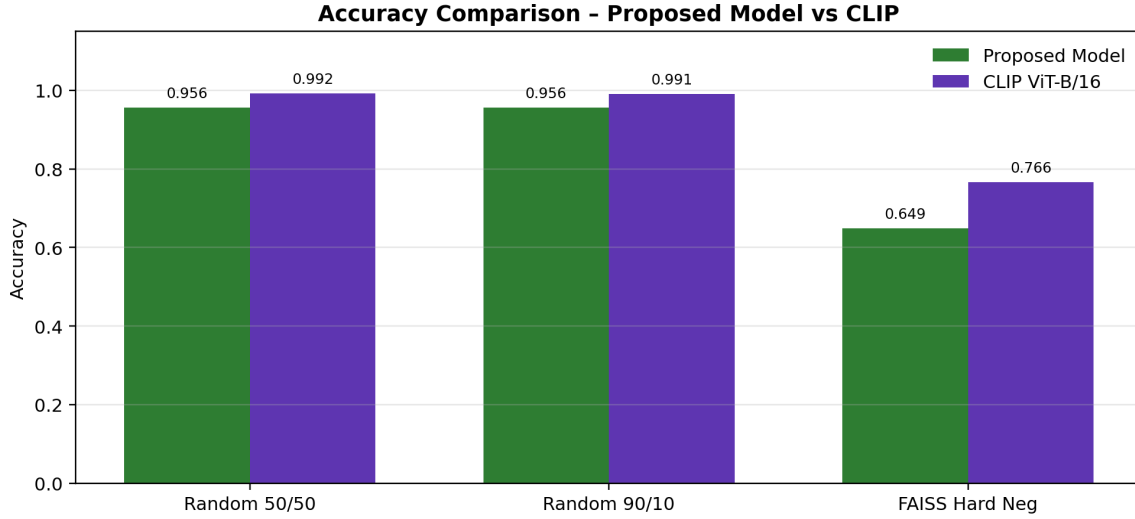


Figure 4.5: Accuracy comparison of the proposed model and CLIP across evaluation settings.

4.11 Results from Ablation Studies

To isolate the contribution of the individual architectural components, two controlled ablations were conducted under the same data splits, pair-construction procedures, InfoNCE objective, and evaluation protocols as the full model.

4.11.1 Image Encoder Ablation (ResNet-50 and BERT)

In the first ablation the Vision Transformer was replaced by ResNet-50 while the BERT text encoder and all other components were retained. As shown in Table 4.6, the ResNet-50 variant remained competitive on random mismatches, obtaining ROC-AUC values of 0.998 and 0.999 under the 50/50 and 90/10 settings, respectively. These results are close to those of the proposed ViT and BERT model (0.999 in both settings). On hard negatives, however, performance declined more sharply, from 0.699 for the proposed model to 0.549 for ResNet-50 and BERT.

This indicates that the Transformer based visual backbone contributes meaningfully to the discrimination of challenging negatives, while a strong convolutional encoder remains competitive on easier random mismatches.

4.11.2 Text Encoder Ablation (ViT and TF-IDF)

In the second ablation the BERT text encoder was replaced by a TF-IDF representation projected into the same shared embedding space. Table 4.6 shows a substantial degradation on random mismatches: the ROC-AUC fell to 0.623 under both the 50/50 and 90/10 settings. On hard negatives the TF-IDF variant obtained a higher ROC-AUC of 0.786, yet it still lagged behind CLIP (0.848) and did not recover the strong random-mismatch performance of the full model. These results demonstrate that a classical bag of words text representation is insufficient for reliable multimodal consistency verification and that the contextual language modelling capacity of BERT is a critical component of the framework.

4.11.3 Ablation Summary

Table 4.6: Ablation summary (ROC-AUC)

Model	Random 50/50	90/10	Hard Negatives
ResNet-50 + BERT	0.998	0.999	0.549
ViT + TF-IDF	0.623	0.623	0.786
CLIP	0.9996	0.9985	0.848
ViT + BERT (Proposed)	0.999	0.999	0.699

Taken together, Table 4.6 confirms that both transformer based encoders contribute to the proposed model. Replacing ViT with ResNet-50 preserves performance on random mismatches but weakens hard negative discrimination. Replacing BERT with TF-IDF produces the largest drop on random mismatches. CLIP remains the strongest model on hard negatives, while the proposed ViT and BERT configuration provides the best overall balance between random mismatch performance, hard negative robustness, and integrated explainability.

4.12 Threshold Analysis

Because consistency verification was formulated as a thresholded decision on cosine similarity, the choice of the decision threshold τ directly influenced classification performance. Threshold analysis was conducted by examining the trade-off between precision and recall across a range of threshold values on the validation data.

As the threshold increased, precision on the inconsistent class generally rose while recall declined, and the opposite occurred as the threshold was lowered. The operating threshold was selected to achieve an appropriate balance for decision-support use [22]. On the 90/10 evaluation the Youden-optimal threshold was 0.4297. This value lies in the region that separates the bulk of the matched similarity distribution from the bulk of the unmatched distribution, and it was the threshold applied for the binary classification results reported in the previous section.

The analysis confirmed that performance is sensitive to the choice of τ and that the selected operating point provides a practical compromise between detecting inconsistencies (high recall) and limiting false alarms (adequate precision).

4.13 Qualitative Analysis of Explainability

In addition to quantitative performance, the quality of the generated explanations was examined. The system produces two complementary forms of evidence for each decision: a visual attention map derived from the Vision Transformer and a ranked list of token importance scores derived from the text encoder [9, 10]. Representative consistent and inconsistent examples are discussed below.

4.13.1 Visual Attention Maps

The ViT attention maps are derived solely from the image encoder and are therefore independent of the paired text. They indicate the regions that contribute most strongly to the global image embedding [9]. In the examples examined, attention typically concentrated on the main subjects and salient objects in the scene.

Figure 4.6 (centre panel) shows the attention map for a consistent pair (cosine similarity = 0.7282). Attention focuses on the woman holding the book and the seated man in uniform—the principal subjects of both the image and the caption. Figure 4.7 (centre panel) shows the attention map for an inconsistent pair (cosine similarity = -0.1604). The image depicts two children indoors, and attention correctly highlights those children and nearby objects, even though the caption describes “A snowboarder doing a handstand on a ramp.”

Because the attention pattern is determined by the image alone, it remains similar for a given image whether the paired document is consistent or inconsistent. The map does not itself encode the consistency decision. Rather, it provides a transparent view of what the vision encoder is focusing on. When the caption describes content that is absent from the image, the attention map still highlights the actual visual content, while the low cosine similarity and the token importance scores reveal the mismatch.

4.13.2 Token Importance Scores

Token importance scores are derived from the text encoder and therefore depend on the input document. They rank the words that contribute most strongly to the text embedding [10].

In the consistent example (Figure 4.6, right panel) the most influential tokens included *uniform*, *indoor*, *service*, *children*, *read*, and *woman*. These terms correspond to salient elements that are visible in the image and present in the caption.

In the inconsistent example (Figure 4.7, right panel) the highest-ranked tokens included *ramp*, *snow*, and *board*. These words are central to the caption but have no corresponding visual evidence in the image. The token ranking therefore surfaces the linguistic basis of the mismatch and supports the model’s inconsistent decision.

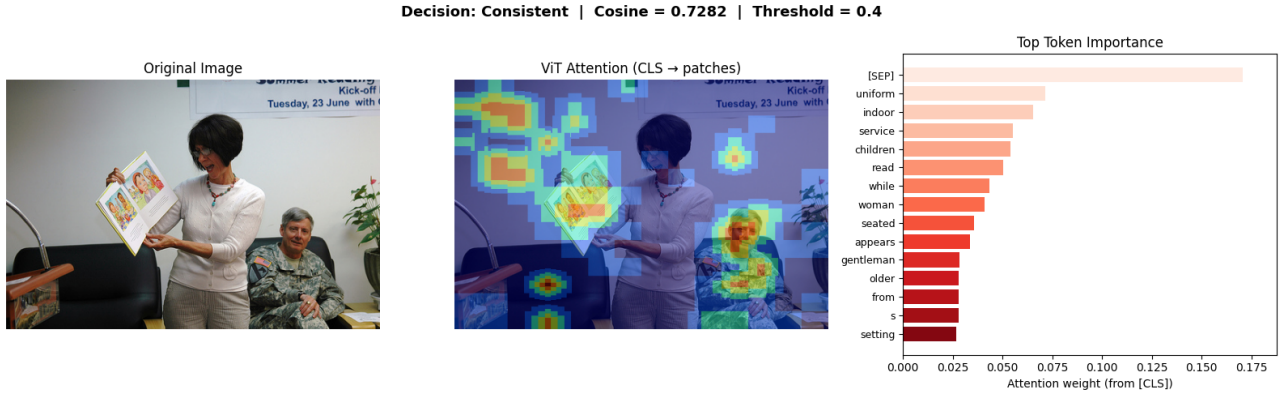


Figure 4.6: Explainable consistency decision (consistent pair). Left: original image. Centre: ViT attention map (image only). Right: top token importance scores. Cosine similarity = 0.7282 (threshold = 0.4).

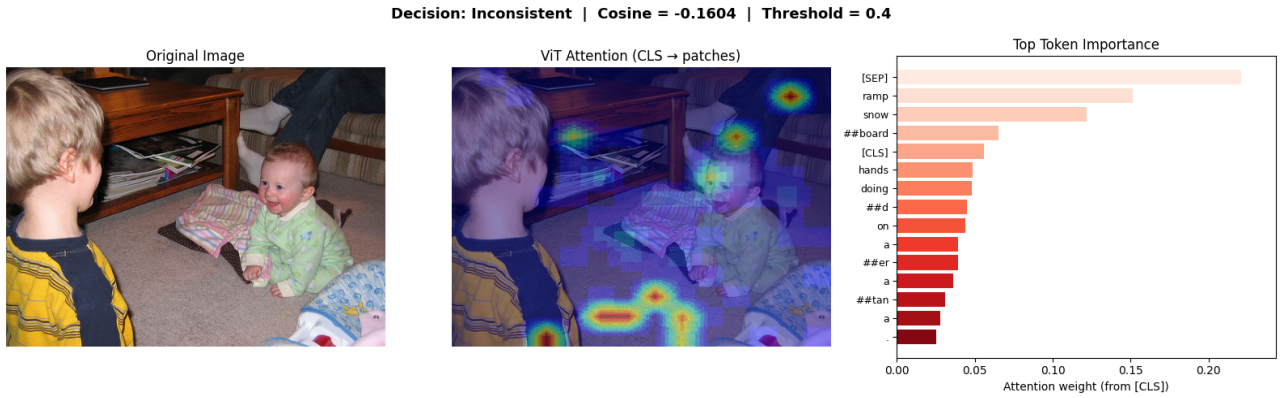


Figure 4.7: Explainable consistency decision (inconsistent pair). Left: original image. Centre: ViT attention map (image only). Right: top token importance scores. Cosine similarity = -0.1604 (threshold = 0.4). The caption refers to a snowboarder on a ramp, which is absent from the image; the attention map correctly highlights the actual visual content (the children), while the token ranking highlights the mismatched concepts.

Together, the two explanation modalities serve distinct roles. The attention map shows what the model sees in the image; the token importance ranking shows which words drive the text representation. When these two sources of evidence align, the pair is typically consistent and the similarity score is high. When they diverge, the image encoder focusing on one set of concepts while the text encoder emphasises another, the similarity score falls and the pair is classified as inconsistent. This combination provides a transparent basis for inspecting individual predictions.

4.14 Summary of Findings

The experimental results support the following conclusions.

First, the contrastive dual-encoder model successfully learned aligned multimodal representations. Under both the balanced (50/50) and imbalanced (90/10) random-negative settings the proposed model achieved near ceiling discrimination (ROC-AUC of 0.9991–0.9994).

Second, hard negatives remain the principal challenge. When evaluated on FAISS retrieved hard negatives the ROC-AUC of the proposed model fell to approximately 0.70, while CLIP retained a clearer advantage (ROC-AUC approximately 0.85).

Third, under an operationally realistic 90/10 class distribution the proposed model maintained excellent ranking quality and very high recall on the minority inconsistent class (0.9925).

Fourth, the ablation studies confirm that both the Vision Transformer and the BERT text encoder are important contributors to performance. Replacing either component degrades results, with the substitution of BERT by TF-IDF producing the largest drop.

Fifth, the integrated explainability mechanisms supply useful, complementary evidence that supports transparent inspection of individual decisions.

Overall, the proposed model delivers strong performance on random mismatches and imbalanced operational settings, provides human interpretable explanations, and constitutes a solid foundation for further development. The remaining performance gap on hard negatives, together with the absence of cross-attention, identifies clear directions for future improvement.

Chapter 5

Discussion

5.1 Introduction

This chapter discusses the experimental findings presented in Chapter 4 in relation to the research objectives and the existing literature. It interprets the quantitative and qualitative results of the proposed explainable dual encoder model, examines the extent to which the study’s objectives were achieved, and considers the practical implications of the work. The chapter also acknowledges the limitations of the current framework and identifies directions for future research.

The discussion is organised around the three specific objectives of the study: the construction of a suitable benchmark for consistency evaluation, the design and implementation of a contrastive Vision Transformer and BERT framework, and the assessment of both consistency verification performance and the quality of the generated explanations. By relating the empirical outcomes to these objectives and to prior work on contrastive vision-language models, the chapter situates the contribution of the present research and clarifies its strengths and remaining challenges.

5.2 Achievement of Research Objectives

The study was guided by three specific objectives. The extent to which each objective was achieved is discussed below.

5.2.1 Benchmark Dataset Construction

The first objective was to design a benchmark dataset suitable for evaluating multimodal consistency verification. This objective was achieved by constructing consistent, randomly mismatched, and hard negative image and document pairs from Flickr30K [21]. The five captions associated with each image were combined into a single multi-caption document, yielding richer textual representations. Consistent pairs served as positive examples for contrastive training; random mismatched pairs and FAISS retrieved hard negatives provided controlled and challenging negative examples for evaluation [40]. Group based splitting on image identifiers prevented data leakage. Although Flickr30K remains a proxy rather than a purpose built consistency corpus, the resulting benchmark enabled systematic assessment under balanced, imbalanced, and hard negative conditions and therefore fulfilled the first objective within the scope of the study.

5.2.2 Contrastive ViT and BERT Model

The second objective was to design and implement a contrastive learning model based on a Vision Transformer and BERT. This objective was achieved through a dual encoder architecture in which ViT-B/16 and BERT base produced unimodal embeddings that were projected into a shared 256-dimensional space and aligned using the symmetric InfoNCE loss [9, 10, 19]. Training was performed on consistent pairs only, with in-batch negatives. The learning curves, similarity matrix, and embedding visualisations confirmed that the model successfully learned a shared representation in which matched pairs received high cosine similarity and random mismatches received low similarity.

Controlled ablation experiments further validated the architectural choices. Replacing the Vision Transformer with ResNet-50 while retaining BERT preserved near ceiling performance on random mismatches (ROC-AUC 0.998–0.999) but produced a sharper decline on hard negatives. Replacing BERT with a TF-IDF representation while retaining the Vision Transformer led to a substantial degradation on random mismatches (ROC-AUC approximately 0.62). These results confirm that both the Transformer based image encoder and the contextual text encoder contribute to the effectiveness of the framework. The second objective was therefore met.

5.2.3 Consistency Performance and Explainability

The third objective was to evaluate consistency verification performance and the quality of the generated explanations. Quantitatively, the model achieved near ceiling discrimination of random mismatches (ROC-AUC of 0.999) and retained strong performance under an operationally realistic 90/10 class distribution (ROC-AUC 0.9994, high recall on inconsistencies). Performance declined on hard negatives (ROC-AUC approximately 0.70), revealing a clear limitation relative to CLIP [5]. The ablation studies reinforced these findings, with the full ViT and BERT configuration outperforming both the ResNet-50 and TF-IDF variants on the majority of settings, confirming the value of the proposed architecture.

Qualitatively, visual attention maps provided image only evidence of the regions driving the vision encoder, while token importance scores identified the words driving the text encoder [9, 10, 11]. Together these explanations supported transparent inspection of individual decisions. The third objective was therefore achieved, subject to the acknowledged gap on hard negatives and the absence of cross-attention.

5.3 Interpretation of Results in Relation to Existing Literature

5.3.1 Alignment with Contrastive Vision-Language Models

The findings of this study align closely with the principles established by large scale contrastive vision-language models. CLIP demonstrated that dual encoder architectures trained with a symmetric contrastive objective can learn a shared embedding space in which matched image and text pairs exhibit high cosine similarity [5]. ALIGN later showed that the same paradigm scales effectively even when trained on noisy web scale data [6]. The present results are consistent with these observations: after contrastive training on consistent image and document pairs, the proposed model produced a clear separation between matched and randomly mismatched pairs, with ROC-AUC values of approximately 0.999. The strong diagonal in the similarity matrix and the co-location of image and text embeddings in the PCA projection further confirm that the dual encoder contrastive approach successfully induces cross-modal alignment, as

reported in the foundational literature.

5.3.2 Performance on Random versus Hard Negatives

A recurring theme in the contrastive learning literature is that performance depends strongly on the difficulty of the negative samples. Random in-batch or randomly sampled negatives are relatively easy to separate, whereas hard negatives that are semantically close to the query pose a substantially greater challenge [5, 6, 40]. The experimental results reflect this distinction clearly. Under both balanced and imbalanced random negative settings, the proposed model and CLIP achieved near ceiling discrimination. On FAISS retrieved hard negatives, however, performance declined for both systems, and the gap widened in favour of CLIP (ROC-AUC approximately 0.85 versus 0.70). This pattern indicates that the principal remaining difficulty lies in fine-grained semantic discrimination rather than in basic cross-modal alignment, and that models pre-trained at substantially larger scale currently retain an advantage in that regime.

5.3.3 Contribution of the Architectural Components

The ablation studies provide additional insight into the role of each modality encoder. The ResNet-50 and BERT variant remained highly competitive on random mismatches (ROC-AUC 0.998–0.999), demonstrating that a strong convolutional backbone can still support effective cross-modal alignment when paired with a powerful text encoder. Its sharper decline on hard negatives, however, suggests that the global receptive field and self-attention mechanisms of the Vision Transformer contribute to more robust discrimination of challenging negatives. Conversely, the ViT and TF-IDF variant performed poorly on random mismatches (ROC-AUC approximately 0.62), confirming that a non contextual bag of words representation is insufficient for the consistency verification task. These findings are consistent with the broader literature emphasising the importance of contextual language models in vision–language alignment [10, 11, 5].

5.3.4 Role and Limits of Dual Encoder Explainability

Dual encoder architectures offer a natural, if limited, basis for explanation. Attention weights from the Vision Transformer indicate which image regions contribute most strongly to the global visual embedding [9], while token level signals from the text encoder identify the words that dominate the textual representation [10, 11]. The qualitative analysis confirmed that these two sources of evidence are complementary: attention maps reveal what the model sees, and token rankings reveal what the model reads. When the two align, similarity scores are high; when they diverge, the mismatch becomes apparent.

At the same time, the dual encoder design imposes a structural limit. Because the image and text pathways do not interact through cross-attention, visual attention remains independent of the caption and cannot directly ground specific words to specific regions. This limitation is inherent to the independent encoder paradigm that underpins CLIP style models [5] and motivates the exploration of cross-attention or other explicit vision–language interaction mechanisms in future work.

5.4 Practical Implications

5.4.1 Decision-Support Potential

The results indicate that the proposed framework can serve as a practical decision support tool for multimodal consistency verification. Under both balanced and imbalanced random-negative conditions the model achieved high discrimination ($\text{ROC-AUC} \geq 0.999$), and under the operationally realistic 90/10 setting it maintained high recall on the minority inconsistent class. These characteristics are valuable in settings where the primary requirement is to surface potential mismatches for further examination rather than to replace human judgement entirely. The continuous cosine similarity score, together with a tunable decision threshold, allows operators to adjust the sensitivity of the system according to the relative costs of false positives and false negatives.

5.4.2 Human Review of Sampled Pairs

Although the system operates in a fully automated manner, the design anticipates periodic human review of sampled image and text pairs. The integrated explanations, visual attention maps that show what the vision encoder focuses on, and token importance rankings that show which words drive the text representation, provide inspectors with immediate, interpretable evidence for each decision. This combination supports efficient quality monitoring, error analysis, and accountability without requiring human intervention on every prediction. In practical workflows the system can therefore process large volumes of pairs automatically while still enabling targeted human oversight.

5.4.3 Domain Adaptation Prospects

The model was developed and evaluated on a general domain proxy dataset. Nevertheless, the modular dual encoder design, the contrastive alignment objective, and the explainability components are not tied to a specific visual or textual domain. With appropriate fine tuning or continued pre-training on domain-specific image and text pairs, the same architecture can be adapted to specialised settings in which consistency between visual evidence and textual records is critical, for example, document verification, archival validation, product cataloguing, or operational environments that combine structured declarations with corresponding imagery. The present results, including the ablation evidence that both encoder choices matter, establish a viable starting point for such adaptation, while the residual gap on hard negatives and the absence of cross-attention indicate the technical improvements that would further strengthen performance in specialised deployments.

5.5 Chapter Summary

This chapter has discussed the experimental findings in relation to the research objectives and the broader literature on contrastive vision-language models. The study achieved its three specific objectives: the construction of a controlled benchmark for consistency evaluation, the implementation of a dual-encoder contrastive framework based on a Vision Transformer and BERT, and the joint assessment of consistency performance and explainability.

The results demonstrated strong discrimination of random mismatches and robust behaviour under an operationally realistic class imbalance, while also revealing a clear performance gap on hard negatives and the inherent limits of image-only attention. The ablation studies confirmed that both the Vision Transformer and the BERT text encoder contribute materially to

performance. These outcomes are consistent with the strengths and constraints of dual-encoder contrastive architectures reported in the literature [5, 6].

Practical implications were identified for decision support, periodic human review of sampled pairs, and future domain adaptation. The principal limitations include proxy data, global rather than fine grained alignment, absence of cross-attention, hard negative difficulty, threshold sensitivity, and resource constraints were acknowledged, and corresponding directions for future research were outlined.

The next chapter concludes the thesis by synthesising the overall contribution, restating the main findings, and summarising the recommendations for further work.

Chapter 6

Conclusion and Recommendations

6.1 Introduction

This chapter concludes the thesis. It summarises the study, restates the main findings in relation to the research objectives, highlights the contributions of the work, and presents the overall conclusions. The chapter also acknowledges the delimitations and limitations of the study and offers recommendations for practice and for future research.

6.2 Summary of the Study

The study addressed the need for systems capable of verifying semantic consistency between images and textual documents while providing transparent evidence for human interpretation. Existing vision-language models have achieved strong results on retrieval and zero-shot tasks, yet comparatively less attention has been given to explicit consistency verification and to the integration of explainability as a core design requirement [3, 5, 6].

The general objective was to develop an explainable Transformer-based vision-language model for multimodal consistency verification. The specific objectives were to design a suitable benchmark dataset for consistency evaluation; to implement a contrastive dual-encoder framework based on a Vision Transformer and BERT; and to evaluate consistency verification performance together with the quality of the generated explanations.

Methodologically, the study adopted an experimental design. Consistent, randomly mismatched, and hard-negative image and document pairs were constructed from the Flickr30K dataset by combining the five captions of each image into a single multi-caption document [21]. A dual-encoder model was trained with the symmetric InfoNCE loss on consistent pairs and evaluated under balanced, imbalanced (90/10), and hard-negative conditions [19, 40]. Cosine similarity served as the consistency score, and visual attention maps together with token-importance rankings provided explanations. CLIP was used as the principal baseline [5]. Controlled ablation experiments replaced the Vision Transformer with ResNet-50 and the BERT encoder with a TF-IDF representation in order to isolate the contribution of each architectural component.

The scope of the study was limited to a general domain proxy dataset, global rather than fine-grained alignment, and a dual-encoder architecture without cross-attention. Within these boundaries the research produced a complete pipeline for consistency verification and explainable decision support on that proxy.

6.3 Main Findings

The main findings of the study are summarised according to the three specific objectives.

6.3.1 Benchmark Dataset Construction

A usable benchmark for multimodal consistency verification was constructed from Flickr30K [21]. Combining the five captions of each image into a single multi-caption document produced richer textual representations. Consistent pairs, randomly mismatched pairs, and FAISS-retrieved hard negatives enabled evaluation under controlled, imbalanced (90/10), and challenging conditions [40]. Group-based splitting prevented data leakage. Although the dataset remains a proxy, and although the inconsistent labels were generated algorithmically rather than by extensive human annotation of contradiction type, the construction demonstrated a practical and reproducible method for creating consistency-evaluation benchmarks from multi-caption resources.

6.3.2 Contrastive Dual-Encoder Framework

The dual-encoder model based on ViT-B/16 and BERT-base, trained with the symmetric InfoNCE objective on consistent pairs, learned an effective shared embedding space [9, 10, 19]. Training and validation losses decreased steadily over ten epochs, the similarity matrix exhibited a strong diagonal, and matched pairs produced substantially higher cosine similarities than random mismatches.

Ablation experiments confirmed the importance of both architectural choices. Replacing the Vision Transformer with ResNet-50 preserved near-ceiling performance on random mismatches (ROC-AUC 0.998–0.999) but led to a sharper decline on hard negatives (ROC-AUC 0.549). Replacing BERT with a TF-IDF representation produced a substantial drop on random mismatches (ROC-AUC 0.623). These results demonstrate that both the Transformer-based image encoder and the contextual text encoder contribute materially to the effectiveness of the framework.

6.3.3 Consistency Performance and Explainability

Quantitatively, the proposed model achieved near-ceiling discrimination of random mismatches (ROC-AUC of 0.999) and retained strong ranking quality under the 90/10 imbalanced setting (ROC-AUC 0.9894, with high recall on inconsistencies). Performance declined on hard negatives (ROC-AUC approximately 0.70), where CLIP retained an advantage (approximately 0.85) [5]. Qualitatively, image-only attention maps revealed the regions driving the vision encoder, while token-importance scores identified the words driving the text encoder. When the two sources of evidence aligned, pairs were classified as consistent; when they diverged, the mismatch was exposed by the score and by the token ranking. These components therefore provided complementary illustrations for individual decisions. Quantitative faithfulness test evaluations and causal explanations remain an area for further improvement.

6.4 Contributions of the Study

The study makes three principal contributions.

6.4.1 Methodological Contribution

The research contributes a complete methodological pipeline for multimodal consistency verification. This includes a practical procedure for constructing consistent, randomly mismatched, and hard-negative image–document pairs from a multi-caption dataset; a dual-encoder contrastive architecture based on a Vision Transformer and BERT; controlled ablations of both the image and text encoders; and an explicit decision rule based on cosine similarity with a tunable threshold. The pipeline treats consistency verification as a pairwise decision task rather than as a by-product of retrieval and provides a reproducible template for subsequent work.

6.4.2 Empirical Contribution

The study provides systematic empirical evidence of model behaviour under three complementary evaluation regimes: balanced random negatives, imbalanced class distribution (90/10), and FAISS hard negatives [40]. The results quantify both the strong performance of the proposed framework on random mismatches and its remaining limitations on hard negatives relative to a large-scale baseline (CLIP) [5]. The ablation studies further show that replacing either the Vision Transformer or BERT degrades performance, with the substitution of BERT by TF-IDF producing the largest drop on random mismatches. These findings clarify the conditions under which dual-encoder contrastive models succeed on this proxy and where further improvement is most needed.

6.4.3 Explainability Contribution

The study integrates visual attention maps and token-importance scores as core outputs of the consistency verification system. By showing what the vision encoder focuses on and which words drive the text representation, the framework supplies complementary, human-interpretable illustrations for each decision. The qualitative analysis demonstrates how alignment or divergence between these two sources of evidence corresponds to consistent or inconsistent outcomes. Formal evaluation of explanation faithfulness remains future work.

6.5 Conclusions

The study set out to develop an explainable Transformer-based vision-language model for multimodal consistency verification. Within the scope defined by the research objectives and the chosen experimental design, the following conclusions are drawn.

A dual-encoder contrastive framework based on a Vision Transformer and BERT is capable of learning a shared embedding space in which consistent image and document pairs receive high cosine similarity and randomly mismatched pairs receive low similarity. Under both balanced and imbalanced random-negative conditions the model achieves strong discrimination and high recall on inconsistencies on this proxy benchmark.

Hard negatives remain substantially more difficult. Performance declines when the model must separate semantically similar but non-matching documents, and large-scale pre-trained models such as CLIP currently retain an advantage in this regime [5, 40]. Global similarity alone is therefore insufficient for the most challenging consistency cases constructed in this study.

The ablation studies confirm that both architectural components matter. A strong convolutional image encoder (ResNet-50) remains competitive on random mismatches but is less robust on hard negatives, while a classical TF-IDF text representation is clearly inadequate for the

task. The combination of ViT and BERT therefore constitutes a well-justified design choice within the dual-encoder paradigm.

Explainability based on image-only attention maps and token-importance scores provides useful, complementary evidence of what the model sees and what it reads. At the same time, the absence of cross-attention limits the granularity of visual and textual grounding and constitutes a structural constraint of the dual-encoder design.

Overall, the research demonstrates that an explainable dual-encoder approach is a viable foundation for multimodal consistency verification on a public multi-caption proxy. It meets the stated objectives while identifying the data, architectural and evaluation advances required for more robust and fine-grained performance.

6.6 Recommendations

6.6.1 Recommendations for Practice

On the basis of the empirical results, the following recommendations are offered.

First, the system is best used as a decision-support tool rather than as a fully autonomous judge. High recall on inconsistencies under the 90/10 test distribution makes it suitable for surfacing candidate mismatches for further examination, while final authority should remain with human reviewers.

Second, periodic human review of sampled image-text pairs should be retained. The attention maps and token-importance scores provide inspectors with immediate illustrations that can support quality monitoring and error analysis.

Third, the decision threshold should be calibrated to the cost structure of the intended setting. Where false negatives are particularly costly, a lower threshold that favours recall may be preferred; where false positives are costly, a higher threshold may be appropriate. Validation data from the intended setting should be used to set and periodically review this operating point [22].

Fourth, the present results were obtained on a general-domain proxy; therefore domain-specific adaptation is recommended before any use outside the photograph and caption domain evaluated in this study.

6.6.2 Recommendations for Future Research

Future research should address the principal limitations identified in this study.

First, purpose-built consistency datasets that annotate fine-grained types of mismatch should be developed or adopted, so that evaluation can move beyond algorithmic negatives derived from captioning corpora.

Second, cross-attention or other explicit vision-language interaction mechanisms should be investigated. Text-conditioned visual explanations and finer-grained grounding are expected to improve both interpretability and hard-negative discrimination.

Third, methods for improving robustness to hard negatives, including hard-negative mining during training, larger-scale pre-training, and specialised objectives, should be explored systematically.

Fourth, formal evaluation of explanation quality, including faithfulness tests and human-judged usefulness studies, should complement the qualitative analyses presented here.

Fifth, studies on specialised visual and textual data would clarify the transferability of the framework.

6.7 Delimitations of the Study

The study was deliberately bounded so that the research question, the data and the architecture could be evaluated rigorously on a public benchmark dataset.

The task was delimited to pairwise multimodal consistency verification: given one image and one textual document, the model produced a consistency score and a thresholded decision. Conventional image and text retrieval was not the primary evaluation setting, and Recall@K was not used as the main claim.

The data were delimited to Flickr30K. Inconsistent pairs were produced algorithmically, as random mismatches and as FAISS-retrieved hard negatives reserved for evaluation. No new human-annotated corpus of contradiction types was collected, and no domain-specific image or document collection was evaluated.

The architecture was delimited to a dual-encoder design. Cross-attention and other fusion mechanisms were not implemented. Ablations were limited to ResNet-50 and TF-IDF, together with a comparison against CLIP under the same pair-construction protocol.

Explainability was delimited to visual attention maps and token-importance scores, inspected qualitatively. Evaluation was delimited to the group-based 70/15/15 Flickr30K splits under three test regimes. No external validation corpus was used. Findings apply to this proxy benchmark dataset and this architecture.

6.8 Limitations of the Study

Several limitations of the present study should be acknowledged.

6.8.1 Proxy Nature of Flickr30K

The experiments relied on Flickr30K as a proxy for consistency verification [21]. The dataset was originally designed for image captioning and retrieval rather than for explicit consistency assessment. Captions are short, crowd-sourced descriptions of everyday scenes and do not reflect the length, structure, or specialised vocabulary of many real-world documents. Consequently, the results demonstrate feasibility on a controlled general-domain benchmark hence performance on purpose-built or domain-specific consistency data should be independently studied.

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. Performance on the inconsistency class therefore measures separation from these constructed negatives.

6.8.3 Global Rather than Fine-Grained Alignment

The model learns a global image and document similarity. It does not explicitly align individual objects, attributes, or relations mentioned in the text with corresponding regions in the image. Fine-grained inconsistencies may therefore be less reliably detected than global mismatches between an image and an entirely unrelated document.

6.8.4 Image-Only Attention and Absence of Cross-Attention

Visual explanations are derived solely from the Vision Transformer and are independent of the paired text. The attention map shows which regions contribute to the image embedding but cannot highlight regions that specifically support or contradict particular tokens. The lack of cross-attention limits the granularity of visual and textual grounding and is an inherent constraint of the dual-encoder design adopted in this study.

6.8.5 Hard-Negative Performance Gap

While the model achieved near-ceiling discrimination of random mismatches, performance declined markedly on FAISS-retrieved hard negatives (ROC-AUC approximately 0.70 compared with approximately 0.85 for CLIP) [40, 5]. Semantically similar but non-matching documents remain difficult to separate, and large-scale pre-trained models currently retain an advantage in this regime.

6.8.6 Threshold Sensitivity

Consistency classification depends on a decision threshold applied to cosine similarity. Although a practical operating point was selected on the validation set using Youden’s index, performance metrics, particularly precision and recall on the minority class, are sensitive to this choice [22]. Different cost structures may require different thresholds, and the optimal value is not universal.

6.8.7 Statistical and External Validation

Reported metrics describe the held-out Flickr30K splits constructed for this study. Formal significance testing and evaluation on an independent external corpus were not completed. Comparison with CLIP is therefore descriptive within this benchmark.

6.9 Final Remarks

This thesis has presented the design, implementation, and evaluation of an explainable dual-encoder vision-language model for multimodal consistency verification. By combining contrastive alignment of Vision Transformer and BERT representations with visual attention maps and token-importance scores, and by validating the architectural choices through controlled ablations, the work has demonstrated a practical approach to assessing the consistency of image and text pairs while supporting transparent human inspection of model decisions on a public multi-caption proxy.

The empirical results confirm that such a framework can achieve strong performance on random mismatches and under an imbalanced class distribution, and that the accompanying explanations provide complementary insight into what the model sees and what it reads. At the same

time, the study has clarified the remaining challenges, particularly hard-negative discrimination, algorithmically generated inconsistency labels, and the absence of cross-modal attention.

It is hoped that the methodology, findings, and recommendations presented here will contribute to the continued development of trustworthy multimodal systems and will encourage further research into explainable consistency verification across both general and specialised domains.

Bibliography

- [1] B. Bouchev, J. Castek, and J. Thygeson, “Multimodal learning,” in *Innovative Learning Environments in STEM Higher Education: Opportunities, Challenges, and Looking Forward*, pp. 35–54, Springer, 2021.
- [2] I. Daunhawer, *Multimodal Representation Learning under Weak Supervision*. PhD thesis, 2023.
- [3] C. Zhang, Z. Yang, X. He, and L. Deng, “Multimodal intelligence: Representation learning, information fusion, and applications,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.
- [4] Y. Liu, Y. Wu, Y. Chen, and Z. Lin, “Recent advances in multimodal representation learning: A survey,” *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–36, 2023.
- [5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning transferable visual models from natural language supervision,” in *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 8748–8763, 2021.
- [6] C. Jia, Y. Yang, Y. Xia, *et al.*, “Scaling up visual and vision-language representation learning with noisy text supervision,” in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 4904–4916, 2021.
- [7] J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, “Align before fuse: Vision and language representation learning with momentum distillation,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 9694–9705, 2021.
- [8] J. Li, D. Li, C. Xiong, and S. C. H. Hoi, “Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation,” in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900, PMLR, 2022.
- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, “An image is worth 16x16 words: Transformers for image recognition at scale,” in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
- [10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, 2019.
- [11] Y. Liu, M. Ott, N. Goyal, *et al.*, “RoBERTa: A robustly optimized BERT pretraining approach,” *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 163–176, 2021.

- [12] G. Pang, C. Shen, L. Cao, and A. van den Hengel, “Deep learning for anomaly detection: A review,” *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 2021.
- [13] X. Zhao, L. Wang, Y. Zhang, X. Han, M. Deveci, and M. Parmar, “A review of convolutional neural networks in computer vision,” *Artificial Intelligence Review*, vol. 57, p. 99, 2024.
- [14] L. Chen, S. Li, Q. Bai, J. Yang, S. Jiang, and Y. Miao, “Review of image classification algorithms based on convolutional neural networks,” *Remote Sensing*, vol. 13, no. 22, p. 4712, 2021.
- [15] Y. Huang, X. Wang, Y. Zhang, L. Chen, and H. Zhang, “Application of human-in-the-loop hybrid augmented intelligence approach in security inspection system,” *Frontiers in Artificial Intelligence*, vol. 8, p. 1518850, 2025.
- [16] H.-Y. Zhou, Y. Yu, C. Wang, S. Zhang, Y. Gao, J. Pan, J. Shao, G. Lu, K. Zhang, and W. Li, “A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics,” *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 743–755, 2023.
- [17] A. Vaswani *et al.*, “Attention is all you need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [18] B. Ghoggh and A. Ghodsi, “Attention mechanism, transformers, BERT, and GPT: Tutorial and survey.” Working paper or preprint, 2020.
- [19] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” *arXiv preprint arXiv:1807.03748*, 2018.
- [20] J. Cui, W. Huang, Y. Wang, and Y. Wang, “Rethinking weak supervision in helping contrastive learning,” in *Proceedings of the 40th International Conference on Machine Learning*, vol. 202 of *Proceedings of Machine Learning Research*, pp. 6448–6467, PMLR, 2023.
- [21] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions,” *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.
- [22] W. J. Youden, “Index for rating diagnostic tests,” *Cancer*, vol. 3, no. 1, pp. 32–35, 1950.
- [23] J.-B. Alayrac *et al.*, “Flamingo: A visual language model for few-shot learning,” in *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022.
- [24] K. He *et al.*, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [25] J. Yu, Z. Wang, Z. Li, W. Hu, X. Chen, and Y. Fu, “CoCa: Contrastive captioners are image-text foundation models,” in *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 26193–26205, 2022.
- [26] X. Zhang, S. Li, N. Shi, B. Hauer, Z. Wu, G. Kondrak, M. Abdul-Mageed, and L. V. S. Lakshmanan, “Cross-modal consistency in multimodal large language models,” *arXiv preprint arXiv:2411.09273*, 2024.
- [27] P. Movva, “Reasoning riddles: How explainability reveals cognitive limits in vision-language models,” *arXiv preprint arXiv:2510.02780*, 2025.

- [28] C. Fang, Y. Liu, S. Teng, M. Yin, and T. Han, “CrossModal-CLIP: A novel multimodal contrastive learning framework for robust network traffic anomaly detection,” *Computer Networks*, vol. 272, p. 111723, 2025.
- [29] M. Qiao, L. Wang, X. Li, P. Liu, H. Gou, and S. Tian, “Analysis of risks and discussion of preventive measures related to the misdeclaration of dangerous cargos in containers,” *Frontiers in Economics and Management Research*, vol. 7, no. 3, 2024.
- [30] J. Vanhoeyveld, D. Martens, and B. Peeters, “Customs fraud detection: Assessing the value of behavioural and high-cardinality data under the imbalanced learning issue,” *Pattern Analysis and Applications*, vol. 23, no. 3, pp. 1457–1477, 2020.
- [31] A. Ahmed, M. Irshaid, M. Alansari, D. Velayudhan, M. Tarnini, M. E.-A. Azz, N. Al-Khalayleh, T. Hassan, E. Damiani, and N. Werghi, “PRISM-X: Progressive semi-supervised threat detection in x-ray scans with self-guided multimodal refinement,” *Information Processing & Management*, vol. 63, no. 4, p. 104640, 2026.
- [32] H. Hojjati, T. K. K. Ho, and N. Armanfard, “Self-supervised anomaly detection in computer vision and beyond: A survey and outlook,” *arXiv preprint arXiv:2205.05173*, 2022.
- [33] E. Müller-Budack, J. Theiner, S. Diering, M. Idahl, S. Hakimov, and R. Ewerth, “Multimodal news analytics using measures of cross-modal entity and context consistency,” *International Journal of Multimedia Information Retrieval*, vol. 10, no. 2, pp. 111–125, 2021.
- [34] W. Xu, D. Xiang, T. Ding, and W. Lu, “MMM-Fact: A multimodal, multi-domain fact-checking dataset with multi-level retrieval difficulty,” 2025.
- [35] B. Gordon, Y. Bitton, Y. Shafir, R. Garg, X. Chen, D. Lischinski, D. Cohen-Or, and I. Szepkektor, “Mismatch quest: Visual and textual feedback for image-text misalignment,” 2023.
- [36] P. Qi, Z. Yan, W. Hsu, and M. L. Lee, “SNIFFER: Multimodal large language model for explainable out-of-context misinformation detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13052–13062, 2024.
- [37] G. Lima, N. Grgić-Hlača, J. K. Jeong, and M. Cha, “The conflict between explainable and accountable decision-making algorithms,” in *2022 ACM Conference on Fairness, Accountability and Transparency*, FAccT ’22, pp. 2103–2113, ACM, 2022.
- [38] C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, and E. Gomez, “The role of explainable AI in the context of the AI Act,” in *2023 ACM Conference on Fairness, Accountability and Transparency*, FAccT ’23, pp. 1139–1150, ACM, 2023.
- [39] H. Bao, L. Dong, S. Piao, and F. Wei, “BEiT: BERT pre-training of image transformers,” in *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.
- [40] J. Johnson, M. Douze, and H. Jégou, “Billion-scale similarity search with GPUs,” *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2019.
- [41] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, 2020.

- [42] H. Chefer, S. Gur, and L. Wolf, “Transformer interpretability beyond attention visualization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, IEEE, 2021.
- [43] S. Abnar and W. Zuidema, “Quantifying attention flow in transformers,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197, Association for Computational Linguistics, 2020.
- [44] D. Mamchur, J. Peksa, S. Le Clainche, and R. Vinuesa, “Application and advances in radiographic and novel technologies used for non-intrusive object inspection,” *Sensors*, vol. 22, no. 6, p. 2121, 2022.
- [45] S. Akçay and T. P. Breckon, “Towards automatic threat detection: A survey of advances of deep learning within x-ray security imaging,” *Pattern Recognition*, vol. 122, p. 108245, 2022.
- [46] C. Miao, L. Xie, F. Wan, C. Su, H. Liu, J. Jiao, and Q. Ye, “SIXray: A large-scale security inspection x-ray benchmark for prohibited item discovery in overlapping images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2119–2128, 2019.
- [47] R. Moorthy, B. Jain, and P. Gidde, “Autonomous non-intrusive inspection for risk detection in cargo containers using deep learning,” in *Computer Vision and Image Processing*, vol. 2473 of *Communications in Computer and Information Science*, Cham: Springer, 2026.
- [48] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. N. Bennett, J. Ahmed, and A. Overwijk, “Approximate nearest neighbor negative contrastive estimation for dense text retrieval,” in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021.
- [49] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [50] H. G. Souto *et al.*, “Ablation studies for novel treatment effect estimation,” *arXiv preprint arXiv:2410.15560*, 2024.
- [51] I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

Appendix A

Budget

Table A.1: Project Requirements and Estimated Costs

#	Category	Sub-Category	Details	Estimated Cost (\$)
1	Technical Requirements	Hardware / Infrastructure	GPU-accelerated computing (Tesla T4 /NVIDIA A100/H100 via cloud GCP). Training time: 100–200 hours; on-demand for MVP.	1,000
2	Technical Requirements	Software Stack	Python 3.12+ with PyTorch, Transformers, timm, scikit-learn, Pandas/NumPy. Git for version control; basic deployment (Github, Hugging Face, Streamlit). (All open-source)	–
3	Technical Requirements	Data Needs	31,783 Flickr30k paired samples (Natural images, 5- captions from Hugging Face, open-source)	–
4	Technical Requirements	Development Environment	Jupyter Notebooks Google Colab and GitHub Codespaces configured to ensure data privacy compliance.	–
6	Team Requirements	Team Composition	• Mugimba Kakure Jude (Researcher)	(60-80 hours)

Appendix B

Research Project timelines

Table B.1: Proposed Research Activities and Timeline for 2026

Activity	Duration	Timeline (2026)
Literature Review and Proposal	4 Weeks	February–April
Data Collection and Preprocessing	2 Weeks	May–June
Model Development and Integration	2 Weeks	June–July
Experiments and Evaluation	2 Weeks	July
Thesis Writing and Revisions	3 Weeks	August
Defense and Submission	1 Week	September

Appendix C

Declaration of AI use

Declaration Statement:

I, Mugimba Kakure Jude, declare that I used

- i. ***Grok AI*** to critique thought processes and quickly identify literature relevant to this study.
- ii. ***ChatGPT*** to identify ideas, evaluate and visualise lines of thought related to the study.

Signature:  _____

September, 2026

Appendix D

Major Journal Publications (Top 20)

1. A. Radford *et al.*, “Learning Transferable Visual Models from Natural Language Supervision,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, vol. 139, pp. 8748–8763, 2021. (Conference; CORE A*; highly cited foundational dual-encoder VLM)
2. A. Dosovitskiy *et al.*, “An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2021. (Conference; CORE A*; ViT backbone)
3. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proc. NAACL-HLT*, pp. 4171–4186, 2019. (Conference; CORE A*; text encoder backbone)
4. A. van den Oord, Y. Li, and O. Vinyals, “Representation Learning with Contrastive Predictive Coding,” arXiv:1807.03748, 2018. (Preprint; defines InfoNCE objective)
5. C. Jia *et al.*, “Scaling Up Visual and Vision-Language Representation Learning with Noisy Text Supervision,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, pp. 4904–4916, 2021. (Conference; CORE A*; ALIGN)
6. J. Li, R. R. Selvaraju, A. D. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, “Align before Fuse: Vision and Language Representation Learning with Momentum Distillation,” *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 9694–9705, 2021. (Conference; CORE A*; ALBEF)
7. J. Li, D. Li, C. Xiong, and S. C. H. Hoi, “BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation,” in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, vol. 162, pp. 12888–12900, 2022. (Conference; CORE A*)
8. Y. Liu, Y. Wu, Y. Chen, and Z. Lin, “Recent Advances in Multimodal Representation Learning: A Survey,” *ACM Computing Surveys*, vol. 55, no. 8, pp. 1–36, 2023. (Journal; IF $\approx$ 23.8; **review paper**)
9. C. Zhang, Z. Yang, X. He, and L. Deng, “Multimodal Intelligence: Representation Learning, Information Fusion, and Applications,” *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, 2020. (Journal; IF $\approx$ 9.0; **review paper**)
10. G. Pang, C. Shen, L. Cao, and A. van den Hengel, “Deep Learning for Anomaly Detection: A Review,” *ACM Computing Surveys*, vol. 54, no. 2, pp. 1–38, 2021. (Journal; IF $\approx$ 23.8; **review paper**)

11. H. Hojjati, T. K. K. Ho, and N. Armanfard, “Self-Supervised Anomaly Detection in Computer Vision and Beyond: A Survey and Outlook,” arXiv:2205.05173, 2022. (**Review paper**)
12. P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, “From Image Descriptions to Visual Denotations: New Similarity Metrics for Semantic Inference over Event Descriptions,” *Trans. Assoc. Comput. Linguistics*, vol. 2, pp. 67–78, 2014. (Journal; TACL; Flickr30K dataset)
13. J. Johnson, M. Douze, and H. Jégou, “Billion-Scale Similarity Search with GPUs,” *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, 2019. (Journal; IF $\approx$ 7.5; FAISS)
14. A. Vaswani *et al.*, “Attention is All You Need,” *Advances in Neural Information Processing Systems*, vol. 30, 2017. (Conference; CORE A*; Transformer foundation)
15. Y. Liu *et al.*, “RoBERTa: A Robustly Optimized BERT Pretraining Approach,” *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 163–176, 2021. (Journal; TACL)
16. J.-B. Alayrac *et al.*, “Flamingo: A Visual Language Model for Few-Shot Learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022. (Conference; CORE A*)
17. J. Yu *et al.*, “CoCa: Contrastive Captioners are Image-Text Foundation Models,” in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, 2022. (Conference; CORE A*)
18. I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in *Proc. Int. Conf. Learn. Representations (ICLR)*, 2019. (Conference; CORE A*; AdamW)
19. Y. Huang, X. Wang, Y. Zhang, L. Chen, and H. Zhang, “Application of Human-in-the-Loop Hybrid Augmented Intelligence Approach in Security Inspection System,” *Frontiers in Artificial Intelligence*, vol. 8, p. 1518850, 2025. (Journal; human-in-the-loop decision support)
20. S. Akçay and T. P. Breckon, “Towards Automatic Threat Detection: A Survey of Advances of Deep Learning within X-ray Security Imaging,” *Pattern Recognition*, vol. 122, p. 108245, 2022. (Journal; IF $\approx$ 8.0; **review paper**; related domain application)

Appendix E

Alignment to national (NDPIV) and Global SDGs

E.1 Alignment with the MSc Programme

This research strongly aligns with the objectives of the MSc programme in Data Science. The study applies modern machine learning techniques including transformer-based models (Vision Transformer and BERT), multimodal contrastive learning (InfoNCE), and explainability mechanisms (visual attention maps and token importance scores) to the problem of verifying semantic consistency between images and textual documents.

It demonstrates practical skills in deep learning, multimodal representation learning, experimental design, comparative evaluation against strong baselines, and the responsible development of transparent AI systems. By designing a dual-encoder consistency verification framework, constructing controlled evaluation benchmark datasets, and integrating human interpretable explanations, the research contributes to the programme's emphasis on innovative AI methods, rigorous research methodology, and applied technology with potential for societal and institutional impact.

E.2 Alignment with National Agendas

E.2.1 Alignment with National Development Plan IV (NDPIV) 2025/26–2029/30

The research supports several strategic objectives of Uganda's NDPIV, particularly under the theme of sustainable industrialisation, digital transformation, and strengthened governance. NDPIV prioritises productivity, private-sector development, good governance, security, and the use of technology for efficient public service delivery.

By developing an explainable AI system for multimodal consistency verification, the study contributes methods that can, in principle, support more reliable checking of visual and textual records in administrative settings. Potential benefits include improved data integrity, reduced reliance on purely manual cross-checking, and greater transparency in automated decision support. These outcomes align with NDPIV goals of strengthening governance, reducing inefficiencies, and leveraging technology for better public-sector performance. The emphasis on resource-conscious design and explainability further matches NDPIV's interest in practical, scalable innovations.

E.2.2 Alignment with Uganda Vision 2040

Uganda Vision 2040 envisions transforming the country from a peasant to a modern and prosperous society, with strong fundamentals in science, technology, innovation, and human capital. ICT and AI are recognised as key enablers of this transformation.

This research aligns with Vision 2040 by applying advanced ICT and AI techniques such as multimodal transformers, contrastive learning, and explainable AI to the problem of verifying consistency between images and text. Such capabilities may, in future work, support more trustworthy digital systems in documentation and public administration.

E.2.3 Alignment with Other National Agendas

The study further aligns with Uganda’s Digital Transformation Roadmap and the broader digital economy agenda, which seek to leverage ICT for service delivery, innovation, and competitiveness. Transparent consistency verification methods can support efforts to improve the integrity of digital records. In addition, the work contributes to national interests in governance and institutional strengthening by promoting systems that are not only accurate but also interpretable by human reviewers.

E.3 Alignment with Global Agendas

E.3.1 Alignment with Sustainable Development Goals (SDGs)

This research is designed to contribute to several Sustainable Development Goals:

- **SDG 9 (Industry, Innovation and Infrastructure):** By advancing innovation in AI and multimodal learning, the study promotes technological capability that can support more reliable digital infrastructure and intelligent systems.
- **SDG 8 (Decent Work and Economic Growth):** More trustworthy verification of multimodal records can, in principle, support smoother administrative processes.
- **SDG 16 (Peace, Justice and Strong Institutions):** The emphasis on explainable, human-interpretable AI supports transparent decision support and accountable use of automated systems.
- **SDG 17 (Partnerships for the Goals):** The research exemplifies the application of academic AI methods to a well-specified technical problem, illustrating the value of research methods that can later inform practice.

E.3.2 Alignment with Other Global Agendas

The research also aligns with the African Union Agenda 2063, particularly aspirations for inclusive growth, good governance, and science, technology and innovation driven development. Methods for multimodal consistency verification may, in future work, support more transparent digital processes.

The present study is formulated as a general multimodal consistency verification framework evaluated on Flickr30K. Its principles remain relevant to any later domain in which images and textual records must be checked for semantic compatibility under human oversight.

Appendix F

Algorithms, Code of interest

F.1 Overview

This section presents the implementation details of the proposed explainable dual-encoder model for multimodal consistency verification. It covers data preprocessing, model architecture, contrastive loss formulation, training procedure, consistency scoring, explainability mechanisms, and evaluation metrics. The implementation was developed in Python using PyTorch and the Hugging Face Transformers library.

F.2 Data Preprocessing

F.2.1 Image Preprocessing

Images were resized to 224×224 pixels and normalised using ImageNet statistics before being converted to tensors:

Listing F.1: Image transformation pipeline

```
transform = transforms.Compose([
    transforms.Resize((224, 224)),
    transforms.ToTensor(),
    transforms.Normalize(
        mean=[0.485, 0.456, 0.406],
        std=[0.229, 0.224, 0.225]
    )
])
```

F.2.2 Text Preprocessing

Multi-caption documents were tokenised with the BERT tokenizer. Sequences were padded and truncated to a fixed maximum length, and attention masks were generated:

Listing F.2: BERT tokenization

```
encoded = tokenizer(
    text,
    padding='max_length',
    truncation=True,
    max_length=128,
```

```

        return_tensors='pt'
    )
    input_ids = encoded['input_ids']
    attention_mask = encoded['attention_mask']

```

F.3 Model Architecture

F.3.1 Image Encoder

A pre-trained Vision Transformer (ViT-B/16) produced the global image embedding from the classification token:

Listing F.3: Vision Transformer forward pass

```

vision_out = self.vision_encoder(
    pixel_values,
    output_attentions=True
)
image_emb = vision_out.last_hidden_state[:, 0, :] # CLS token
attentions = vision_out.attentions               # for
    explainability

```

F.3.2 Text Encoder

A pre-trained BERT-base model produced the global text embedding from the [CLS] token:

Listing F.4: BERT forward pass

```

text_out = self.text_encoder(
    input_ids=input_ids,
    attention_mask=attention_mask,
    output_attentions=True
)
text_emb = text_out.last_hidden_state[:, 0, :] # CLS token

```

F.3.3 Projection to Shared Embedding Space

Both embeddings were projected into a shared 256-dimensional space and L2-normalised:

Listing F.5: Projection and normalisation

```

text_proj = self.text_proj(text_emb)
image_proj = self.image_proj(image_emb)

text_norm = F.normalize(text_proj, dim=-1)
image_norm = F.normalize(image_proj, dim=-1)

```

F.3.4 Consistency Scoring

Consistency was measured by cosine similarity between the normalised embeddings. A decision threshold τ was applied at inference time:

Listing F.6: Cosine similarity and decision

```
similarity = (image_norm * text_norm).sum(dim=-1)
decision = (similarity >= threshold).long() # 1 = consistent, 0 =
inconsistent
```

F.4 Contrastive Learning Objective

The model was trained with the symmetric InfoNCE loss on consistent image–document pairs. In-batch samples served as negatives:

Listing F.7: Symmetric InfoNCE loss

```
logits = image_norm @ text_norm.T / temperature
labels = torch.arange(logits.size(0), device=logits.device)

loss_i2t = F.cross_entropy(logits, labels)
loss_t2i = F.cross_entropy(logits.T, labels)
loss = (loss_i2t + loss_t2i) / 2
```

Hard negatives generated by FAISS were reserved for evaluation and were not used in the training objective.

F.5 Training Procedure

Training used the AdamW optimiser with a linear warmup–decay learning-rate schedule. Only consistent pairs were used for contrastive optimisation:

Listing F.8: Training step

```
optimizer.zero_grad()
loss = model(images, input_ids, attention_mask) # InfoNCE
loss.backward()
optimizer.step()
scheduler.step()
```

F.6 Explainability Mechanisms

F.6.1 Visual Attention Maps

Attention weights from the Vision Transformer (CLS-to-patch) were extracted, averaged across heads, reshaped into a spatial grid, and upsampled to produce a heatmap overlaid on the original image.

F.6.2 Token Importance Scores

Token-level importance was obtained from the text encoder attention (or gradient-based signals) and ranked to identify the words that contributed most strongly to the text embedding.

F.7 Evaluation Metrics

Performance was assessed using ranking and classification metrics computed from cosine similarity scores and the selected decision threshold:

Listing F.9: Evaluation metrics

```
roc_auc = roc_auc_score(y_true, similarities)
pr_auc  = average_precision_score(y_true, similarities)

y_pred = (similarities >= threshold).astype(int)
accuracy = accuracy_score(y_true, y_pred)
precision = precision_score(y_true, y_pred, pos_label=1)
recall    = recall_score(y_true, y_pred, pos_label=1)
f1        = f1_score(y_true, y_pred, pos_label=1)
```

Metrics were reported under three evaluation settings: balanced random negatives (50/50), imbalanced random negatives (90/10), and FAISS hard negatives. CLIP served as the principal baseline under identical protocols.

Research Poster

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



Mugimba Kakure Jude

Reg No: J25M19/034 - Student No: B31388 - Department of Computing & Technology
Faculty of Engineering Design & Technology, Uganda Christian University

1 MOTIVATION

Multimodal data is everywhere. Ensuring that an image truly matches the accompanying text is critical for applications such as content moderation, media authentication, and operational tasks like customs Non-Intrusive Inspection (NII).

Research Question

Can we build an accurate and explainable model that determines whether an image and a multi-caption document are semantically consistent or inconsistent?

2 PROPOSED APPROACH

We propose an explainable dual-encoder architecture that embeds images and multi-caption documents into a shared space and quantifies their semantic alignment.



VIT-B/16 and BERT-base encoders with projection heads into a shared 256-dimensional embedding space; contrastive training (InfoNCE) and cosine similarity with a decision threshold.

3 DATASET CONSTRUCTION

Built from Flickr/30K by aggregating five captions per image into single multi-caption documents and creating three pair types.

Consistent Pairs (Positive)

Image and its true multi-caption document.

Random Mismatch (Negative)

Image paired with a randomly selected document.

Hard Negatives (FAISS)

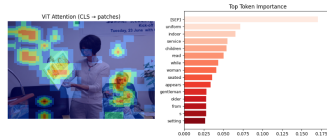
Image paired with a semantically similar but incorrect document retrieved using FAISS.

Dataset Statistics

- Source: Flickr/30K
- Total Images: 31,783 (~31K), five captions per image
- Pair types: Consistent / Random Mismatch / Hard Negative (FAISS)
- Total pairs: 54,000+
- Consistent: ~18,000 (33%); Random mismatches: ~18,000 (33%)
- Hard negatives: ~18,000 (33%)
- Split: Group-wise 70% Train / 15% Validation / 15% Test

4 EXPLAINABILITY

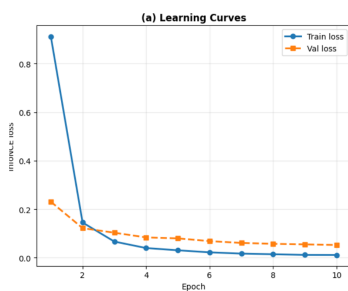
The model explains its predictions by highlighting what it looks at in the image and which words in the text are most influential.



VIT attention (CLS -> patches) and BERT token importance for the consistent pair of Section 6.

Interpretation. The model attends to the semantically relevant regions of the scene and to the key descriptive words of the document, supporting the consistency decision.

Training Dynamics (Proposed Model)



- Train loss: 0.8566 → 0.0102 over 10 epochs
- Validation loss: 0.1925 → 0.0450 over 10 epochs

7 CONTRIBUTIONS

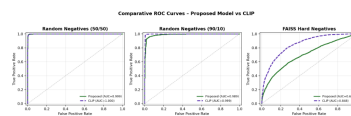
- We propose a transformer-based dual-encoder model for multimodal consistency verification.
- We integrate explainability via VIT attention maps and BERT token importance.
- We build a benchmark dataset with balanced, imbalanced, and FAISS hard negative pairs.
- We benchmark against CLIP across three evaluation regimes, achieving comparable accuracy on random negatives and identifying a clear performance gap on FAISS hard negatives.
- We provide thorough ablation analysis and threshold selection for operational deployment.

5 EXPERIMENTAL RESULTS

5.1 Performance Comparison with CLIP

Setting	Metric	Proposed (Ours)	CLIP (Baseline)
50/50 Random Negatives	ROC-AUC	0.9991	0.9996
	Accuracy	0.9873	0.9920
	F1 (Inconsistent)	0.9872	0.9919
90/10 Random Negatives (Operational)	ROC-AUC	0.989	0.999
	Accuracy	0.9850	0.9908
	F1 (Inconsistent)	0.9297	0.9554
50/50 Hard Negatives (FAISS)	ROC-AUC	0.6990	0.8480
	Accuracy	0.6490	0.7660
	F1 (Hard Negative)	0.6500	0.7600

5.2 ROC Curves (AUC)



Comparative ROC curves, proposed model versus CLIP, across the three evaluation settings.

5.3 Ablation Study (ROC-AUC)

Model Variant	50/50 Random	90/10 Random	50/50 Hard (FAISS)
VIT-B/16 + BERT-base (Proposed)	0.999	0.999	0.699
ResNet-50 + BERT-base	0.998	0.999	0.549
VIT-B/16 + TF-IDF	0.623	0.623	0.786

The proposed VIT + BERT model consistently outperforms ablations, especially in hard negative settings.

8 CONCLUSION

Our explainable transformer-based vision-language model effectively verifies semantic consistency between images and multi-caption documents. It achieves near-perfect performance on random negative settings (AUC up to 0.999) and moderate performance on hard negatives (AUC = 0.699). The model's explainability and high accuracy make it suitable for real-world applications where trust, transparency, and accountability are essential.

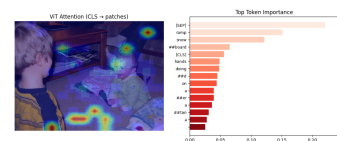
6 QUALITATIVE EXAMPLES

Example 1 - Consistent Pair

Prediction: Consistent
Cosine similarity: 0.7382
Decision threshold: 0.40
Top tokens (BERT): uniform, indoor, service, children, read, while, woman, seated

Example 2 - Inconsistent Pair

Prediction: Inconsistent
Cosine similarity: ~0.1604
Decision threshold: 0.40
Top tokens (BERT): ramp, snow, #board, hands, doing



Example 2: attention concentrates on regions unrelated to the snowboarding description, and the most influential tokens have no visual support.

Threshold for Operational Setting (90/10)

- Optimal cosine similarity threshold = 0.5900 (maximizes F1-Inconsistent) [See Section 4.6.2]
- General threshold from broader analysis = 0.4297 (See Section 4.4.2)

9 FUTURE WORK

- Incorporate cross-encoder re-ranking for further gains.
- Explore larger vision-language backbones (EVA-02, DeiT-III) and instruction tuning.
- Evaluate on domain-specific datasets (e.g., customs NII, medical, social media).
- Deploy as an explainable verification tool in real-world systems.



Research Ethics Committee (REC) Approval



**UGANDA CHRISTIAN
UNIVERSITY**

A Centre of Excellence in the Heart of Africa

21/08/2026

To: Jude MugimbaKakure
UgandaChristianUniversity



Review Type: Initial Review

Re:UCUREC-2026-820: TRANSFORMER-BASED MULTIMODAL CONTRASTIVE LEARNING FOR HIGH-RISK CARGO DETECTION IN CUSTOMS DATA

I am pleased to inform you that the Uganda Christian University REC, through expedited review held on **13/08/2026** approved the above referenced study.

Approval of the research is for the period of **21/08/2026** to **21/08/2027**.

As Principal Investigator of the research, you are responsible for fulfilling the following requirements of approval:

1. All co-investigators must be kept informed of the status of the research.
2. Changes, amendments, and addenda to the protocol or the consent form must be submitted to the REC for re-review and approval **prior** to the activation of the changes.
3. Reports of unanticipated problems involving risks to participants or any new information which could change the risk benefit: ratio must be submitted to the REC.
4. Only approved consent forms are to be used in the enrollment of participants. All consent forms signed by participants and/or witnesses should be retained on file. The REC may conduct audits of all study records, and consent documentation may be part of such audits.
5. Continuing review application must be submitted to the REC **eight weeks** prior to the expiration date of **21/08/2027** in order to continue the study beyond the approved period. Failure to submit a continuing review application in a timely fashion may result in suspension or termination of the study.
6. The REC application number assigned to the research should be cited in any correspondence with the REC of record.
7. You are required to register the research protocol with the Uganda National Council for Science and Technology (UNCST) for final clearance to undertake the study in Uganda.

The following is the list of all documents approved in this application by Uganda Christian University REC:

No.	Document Title	Language	Version Number	Version Date
1	Main Consent forms	English	Final	2026-07-17
2	Data collection tools	English	Final	2026-07-17
3	Protocol	English	Final	2026-06-06

Yours sincerely,

Edward Mukooza Kibikyo
For: Uganda Christian University REC