Received: 8 May 2025 | Revised: 2 February 2026 | Accepted: 13 March 2026 | Published online: 27 April 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–17
DOI: 10.47852/bonviewJCCE62026113

BON VIEW PUBLISHING

OptiM-RoViT: A Robust Multimodal Sentiment Analysis Framework with Dynamic Fusion and Noise-Aware Vision Transformers

Naveen Vasudevan[1] **and Sountharrajan Sehar**[1,*]

[1] *Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham-Chennai, India*

Abstract: Multimodal sentiment analysis has been studied by many researchers because of its capability to understand human emotions through textual and visual data. Despite that, current approaches suffer from feature misalignment, modality imbalance, and a lack of robustness against noisy inputs. A new framework called OptiM-RoViT is proposed, which is built on RoBERTa for processing text and Vision Transformer (ViT) for analyzing images and enhances the model with dynamic modality weighting and Gaussian noise injection techniques. To achieve the best performance, the model uses Optuna-based hyperparameter optimization and can achieve 98.89% accuracy, 0.96 F1-score, and 0.04 false negative rate on the dataset of 10,000 product reviews. Complementary ablation studies also quantify the improvement of each component by enabling significant improvements over baseline architectures. Evaluations against stronger backbones (DeBERTa, Swin-V2) confirm that the proposed fusion mechanism is the primary driver of robustness, outperforming generic state-of-the-art baselines on noisy inputs. The critical challenges in multimodal fusion, including computational scalability and general efficacy, are addressed by the proposed approach, which is also well suited for the real-world applications such as e-commerce and social media analytics. Such adaptation may be explored in other domains as future work, and other modalities can be used to achieve similarly broad applicability.

Keywords: multimodal sentiment analysis, RoBERTa, Vision Transformer, hyperparameter optimization, deep learning

1. Introduction

One of the subfields of the approach of natural language processing (NLP) such as sentiment analysis aims to comprehend and decode human feelings and opinions conveyed through a text [1]. Text-based data are traditionally considered when it comes to traditional sentiment analysis, and sentiment is extracted from words, phrases, and sentences of the text and categorized as positive, negative, or neutral [2]. Although text is still a valuable source of sentiment information, solely relying on textual data can hinder understanding a review or opinion [3]. The reason behind this limitation is that text-based sentiment analysis is not able to capture all the visual cues, emotional context, or product-related details that can be depicted via images, particularly in the case of product reviews, social media, or customer feedback [4].

Hence, there is a growing attention in multimodal sentiment analysis (MSA) to unite text and image data to obtain better sentiment understanding. It is shown that images are powerful in communicating visual elements like mood, condition, or emotional context, which accompany textual description [5]. As much as words can express satisfaction or dissatisfaction with a product, so can its image. For this reason, the combination of text and image data can lead to a better sentiment classification with more accurate and nuanced sentiment, thus increasing predictive performance in sentiment analysis errands [6].

However, fusion of information for MSA is challenging, especially in an effective way. The features of text and image data are extracted, and then the complexities of aligning and combining them are such that sophisticated models are required to process and integrate the two disparate data types. Furthermore, the model performances can be severely affected by the presence of challenges such as feature misalignment, which means that textual and visual features do not match semantically well, and model overfitting, where the model remembers training data rather than simplifying well to the unseen [7].

In order to deal with these challenges presented by MSA, this work presents a novel MSA framework by using the power of advanced NLP and computer vision. It includes integration of RoBERTa—Transformer-based model for very strong performance on NLP tasks—as well as Vision Transformer (ViT), having shown very strong performance on image classification tasks. To fill the gap between text and image data, we fuse the textual understanding of RoBERTa with ViT to extract image features for the joint model to learn more complex relationships between two modalities to perform more accurate sentiment prediction.

The main problem tackled in this work is the problem of MSA, more specifically the combination of textual reviews and product images. Sentiment analysis in text has been quite well

***Corresponding author:** Sountharrajan Sehar, Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham-Chennai, India. Email: s_sountharrajan@ch.amrita.edu

explored in the literature, with relatively little exploring the inclusion of image data into the analysis [8]. While few multimodal methods exist, they typically have problems of correctly aligning textual and visual features, hence leading to suboptimal performance. Furthermore, these models are not able to address the computational inefficiencies that occur in processing large multimodal data. In this work, the gap is filled with the proposal of a robust multimodal model that makes a better use of both text and image features to increase the accuracy of sentiment classification, while keeping a computational cost and avoiding overfitting. This manuscript has the following contributions:

1) New dynamic fusion architecture: In contrast to the common concatenation in the previous RoBERTa–ViT applications, the OptiM fusion module is suggested. This system combines layer normalization and a Sigmoid Gating Unit, which is regulated by a trainable coefficient (alpha), and the model is able to dynamically deactivate a modality whenever it is noisy or misaligned (e.g., text–image contradictions).
2) Noise-aware visual regularization: This is a noise injection mechanism in the visual embedding space. This serves as a regularizer that causes the model to learn semantic features of sentiment as opposed to being biased toward high-frequency visual features (e.g., tears in packaging), which is essential to real-world e-commerce data.
3) Framework independence and generalizability: Established statistically significant improvements ($p < 0.001$) on our custom dataset but also on the MM-IMDB benchmark and with stronger backbones, such as DeBERTa-v3 and Swin-V2, indicating the utility of the framework in general.

Unlike prior works that combine RoBERTa and ViT using static concatenation [9], our framework introduces a technically distinct design through dynamic modality weighting, noise-aware ViT embeddings with Gaussian regularization, overlapping patch embeddings for fine-grained visual cues, and Optuna-based hyperparameter optimization. These innovations collectively make OptiM-RoViT more robust to noisy inputs and better suited for real-world MSA scenarios. Due to the rising significance of online product reviews, many of which are represented as a combination of text and images and express customer sentiments, this work is motivated. Customer feedback is crucial in the e-commerce domain and can influence purchasing decisions of customers, and hence knowing the sentiment of the reviews can help a business to recover their products and services. Yet, existing sentiment analysis models that consider only text or simplistic image features alone are prone to doing so in themselves. This work will combine textual and visual cues to build a model that produces a more nuanced sentiment of customers than one that only considers word choice.

This work is also timely as robots would increasingly need to interact with groups of people, and it fits with the current trend of multimodal machine learning, in which multiple types of data (text, images, audio) are combined to better understand humans by AI systems. To deal with the above challenges of multimodal data such as alignment of text and image representation, computational efficiency, and generalization, the proposed model has a broad range of applications to a variety of real-world applications beyond product reviews, for example, social media sentiment analysis and customer service automation, among others.

The remaining part of the manuscript is organized as follows: Section 2 entails a thorough review of the related literature in the area of MSA and the state-of-the-art (SOTA) technologies. Section 3 contains the description of the methodology, in other words, the model architecture, preprocessing actions, and the way text and image data were combined to combine the two. Section 4 gives the experimental design, which includes a description of datasets, evaluation metrics, and comparison to baseline and existing methods. Section 5 presents the results, which include the most important findings of the experiments and give information about the model performance. Finally, Section 6 draws a conclusion of the paper and offers possible future research directions.

2. Literature Review

The MSA has developed out of the early unimodal methods that mainly dealt with written data. To begin with, the unidimensional aspect of the concept of emotion could not be totally represented by text in human–computer interaction [7]. Consequently, there arose the concept of multimodal emotion analysis in which not the text alone, but speech and visual data are also considered in enhancing the accuracy of predicting implicit sentiment polarity [8]. Research in this field has primarily concentrated on representation learning and multimodal fusion [9, 10].

For multimodal representation learning, several innovative methods have been proposed. Hazarika et al. [11] developed two distinct encoders to project features into both modality-invariant and modality-specific subspaces. A multimodal similarity calculation network that is capable of learning embedding representations between text and images in a unified vector space was introduced by George [12]. An adversarial network proposed by Zhang et al. [13] to reduce the transfer of expression style in cross-domain MSA for creating joint representations for modalities. Modality invariant temporal learning techniques and inter-modal attention are explored by Sun et al. [14] to extract both intra-modal and inter-modal features.

Methods for multimodal fusion are usually classified into early fusion and late fusion approaches [14]. Early fusion usually concatenates features from different modalities as input to the model, while late fusion trains independent models for each modality and combines the outputs afterward. The external product of vectors was used by Ding et al. [15] for fusion of multimodal features. Attention mechanisms were introduced by Yang et al. [16] to learn the relationships between different modalities before performing feature fusion. In the study by Zeng et al. [17], multi-source features were fused using graph neural networks by balancing independent and complementary information from different modalities. In a previous work, Li et al. [18] employed quantum theory to model interactions between intra-modal and inter-modal features in terms of representations of superposition and entanglement at different stages. Xu et al. enhanced sentiment classification performance by implementing pairwise emotional correlation distillation combined with an information bottleneck to filter out task-irrelevant noise [19].

Most recently, multimodal reasoning has been challenged by large-scale vision–language models like BLIP [20] and Flamingo [21], which show high levels of cross-modal reasoning. Although these models are mostly designed to learn general-purpose vision–language alignment, their principles of design are inspirational for sentiment-specific fusion strategies. Xu et al. [22] suggested a cross-modal joint representation transformer (CMJRT) in MSA, and Liu et al. [23] proposed a knowledge-guided dynamic modality attention fusion model, which dynamically adapts the modality contributions. New standards are

also now available, including PanoSent [24] to analyze sentiment at the aspect level in conversational input, multimodal attention fusion network to analyze multi-object sentiment in images, and Ar-MuSA [25] to analyze affective subtasks such as humor and personalization. Ar-MuSA [25] also expands on MSA with contrastive knowledge injection and sentiment research in underrepresented languages, including Arabic, by offering a multimodal benchmark dataset. These new advances point to the fast growth of MSA and underline the necessity of such architectures as OptiM-RoViT, which focus on robustness, adaptive fusion, and noise resilience. Adaptive balancing of modality contributions has been proposed in other strategies of dynamic fusion like Modal-AGG [3]. Nonetheless, in Modal-AGG, the aggregation weights at the sequence level are static, thereby restricting its adaptability in case modalities are incomplete or noisy. Equally, MULT uses cross-modal attention without feature normalization and gating to remove irrelevant features. In comparison, our OptiM-RoViT model incorporates layer normalization, learnable weighting (alpha), and a sigmoid gating mechanism into the fusion operation, which provides strong dynamical balancing. Also, our method presents co-occurring ViT patch embeddings and Gaussian noise injection on the embedding level that Modal-AGG and MULT do not discuss.

Standard MSA datasets such as CMU-MOSI, CMU-MOSEI, and MM-IMDB are widely used benchmarks for evaluating fusion techniques. CMU-MOSI and MOSEI provide video–text data, while MM-IMDB focuses on text–image sentiment classification in the movie domain. These datasets highlight challenges such as modality imbalance and noisy inputs, which our proposed OptiM-RoViT framework directly addresses. Recent vision–language foundation models such as BLIP [20] and Flamingo [21] demonstrate powerful cross-modal reasoning capabilities. However, these large-scale models are designed for general-purpose vision–language tasks and require substantial computational resources. Our work, in contrast, focuses on lightweight yet robust sentiment-specific modeling. Thus, our comparisons emphasize both performance and efficiency trade-offs. In addition to centralized multimodal fusion, other recent developments have been made on privacy-preserving and distributed architectures, especially in the Internet of Medical Things. As an example, the federated learning method of detecting emotions in handwriting on consumer devices was presented in the framework of FedEmo [26], which also solves the essential data privacy issues. Likewise, Federated Learning in Fuzzy Ensemble [27] has been used to attend to electroencephalogram (EEG)-based emotion recognition in order to improve scalability and personalization in the medical IoT setup. These papers highlight an emerging move toward powerful, deployment-ready affective computing systems, an objective that is also held by our proposed OptiM-RoViT framework, which puts a high value on robustness to noise in centralized e-commerce systems. Table 1 provides a comparative overview of the latest research in MSA, detailing the proposed models, datasets used, and the pros and cons of each approach. Table 1 gives a detailed comparative summary of recent studies in MSA, data features, and methodological advantages/disadvantages.

Table 1
Comparative summary of recent approaches in multimodal sentiment analysis

Ref no.	Proposed model	Dataset used	Pros	Cons
[20]	BLIP: Bootstrapping language–image pre-training	Various V-L benchmarks	State-of-the-art cross-modal reasoning and generation capabilities	General-purpose architecture; computationally heavy for specific sentiment tasks
[21]	Flamingo: Visual language model	Few-shot learning benchmarks	Powerful few-shot learning for vision–language tasks	Extremely large model size; difficult to deploy in resource-constrained settings
[22]	CMJRT: Cross-modal joint representation transformer	Multimodal sentiment dataset	Learns joint representations of text and image via transformer-based alignment	Requires large-scale computational resources
[23]	Mutual attention-based framework	Standard MSA datasets	Robust in complicated/noisy contexts	Complex architecture increases training time
[24]	PanoSent: Panoptic conversational ABSA benchmark	Conversational MSA dataset	Provides fine-grained aspect-based multimodal sentiment labels	Still limited to conversational domain
[25]	Ar-MuSA: Arabic multimodal sentiment analysis dataset	Multimodal Arabic dataset	Expands MSA into Arabic, an underrepresented language	Limited scale compared to English benchmarks
[28]	Multimodal Attention Fusion Network	Social media comments	Specialized for noisy social media data	Requires large-scale labeled social data
[29]	MuSe 2023 Multimodal Sentiment Challenge	Mimicked emotions, humor	Benchmarks cross-cultural humor and personalization in MSA	Challenge-oriented, not yet widely adopted for general sentiment tasks

2.1. Research gap and contribution

Although there has been a lot of advances made regarding MSA, there are still three main challenges that existing approaches face, namely, (1) inefficient fusion of text and image features because of rigid modality weighting [30], (2) vulnerability to noisy or low-quality images [27], and (3) suboptimal hyperparameter settings that impede generalizability [13].

While some studies have explored RoBERTa–ViT fusion for multimodal tasks [4, 5, 26], these approaches typically employ static fusion mechanisms and are highly sensitive to noisy visual inputs. Compared to it, our model goes further and introduces dynamic fusion, overlapping patches in ViT, and robustness-promoting noise injection, thus overcoming the shortcomings of previous studies.

According to a summary in Table 1, recent developments in MSA have taken various directions including cross-modal joint representation learning [26], knowledge-guided dynamic fusion [20], aspect-based conversational sentiment analysis [22], and multi-object sentiment evaluation [23].

Contrasting knowledge injection has also been stressed by other researchers [30], expansion into underrepresented languages, including Arabic, and challenge-driven multimodal affect recognition. Zhou et al. created a text-based weak interaction fusion network, which includes a modality correcting module to fix the discrepancy between visual/audio clues and the main linguistic context [31]. Although these methods reflect the richness and diversity of existing studies, all of them are still susceptible to such issues as computational inefficiency [32, 33], sensitivity to noisy or poor inputs [34, 35], or dependence on domain-specific resources [28]. Yang et al. tackled the strong issue of missing modalities with the aid of a text-guided contrastive learning system and a token-level reconstruction network [34].

To fill these gaps, the proposed OptiM-RoViT framework proposes a technically different design that integrates three highly important innovations: (1) adaptive modality weighting, which balances the contributions of text and image and corrects when one of the modalities is not available or unreliable; (2) Gaussian noise injection (sigma = 0.01) in ViT embeddings, which enhances the robustness to problems of noisy or poor-quality images; and (3) hyperparameter optimization with Optuna, which guarantees the effective convergence and generalization of the models.

Together with overlapping patch embeddings in ViT, these innovations enable superior robustness and stability compared to prior RoBERTa–ViT fusion studies [5, 20, 26], which relied on static concatenation and standard ViT configurations. As a result, OptiM-RoViT achieves 98.89% accuracy with 60% fewer false negatives than SOTA methods [36], while maintaining comparable computational efficiency. These architectural distinctions close the gap between theoretical multimodal fusion and readiness for real-world deployment.

3. Materials and Methods

With the emergence of multimodal data, sentiment analysis has changed a lot as the combination of textual and visual information helps in understanding user sentiments [29]. The impact of social context on sentiment is explored, proposing a model that utilizes label semantic guidance and social link information to refine multimodal predictions [37]. Traditionally, information is typically derived via text only, meaning that it ignores what images can contribute in conveying emotions and context, which texts cannot [38]. In this regard, this research fills the gap by relying on the proposed robust multimodal framework, which combines up-to-date NLP and computer vision techniques to extract sentiments from both text reviews and corresponding images. Utilizing RoBERTa and BERT as language models, along with ViT and ResNet as visual models, the methodology combines the contextual depth of these two categories of models and associated techniques with the visual feature extraction abilities, achieving a type of general image language transformer. The framework intends to achieve this through fusing the given modalities through optimized fusion strategies and hyperparameter tuning for increased accuracy and generalizability in sentiment classification. Besides giving the model a higher interpretability to multifaceted multimodal data, the method also addresses issues such as overfitting and misalignment of features and has a strong performance on the various datasets. The findings reveal that the framework can effectively rank sentiments as demonstrated by extensive indicators including precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) analysis.

3.1. Limitations of prior work

Current MSA models have three acute challenges that make them not applicable in the real world. First, feature misalignment takes place when text and image embeddings fall in completely distinct semantic locations. This was established by Lu et al. [38], who showed that there was a 22% reduction in accuracy in situations where textual sentiments were contrary to visual information (e.g., positive text and images of damaged products). Second, such methods of fusion as concatenation are not dynamic to changing modality significance. Li et al. introduced StaPro-Dyn, which is a single architecture to enhance the fusion stability by stability-aware filtering and utilizes prompt learning to refine modality-specific feature extraction [36]. Liu et al. [23] found that models experiencing low-quality or damaged-product image processing reduced performance by 15% when the models were not regularized.

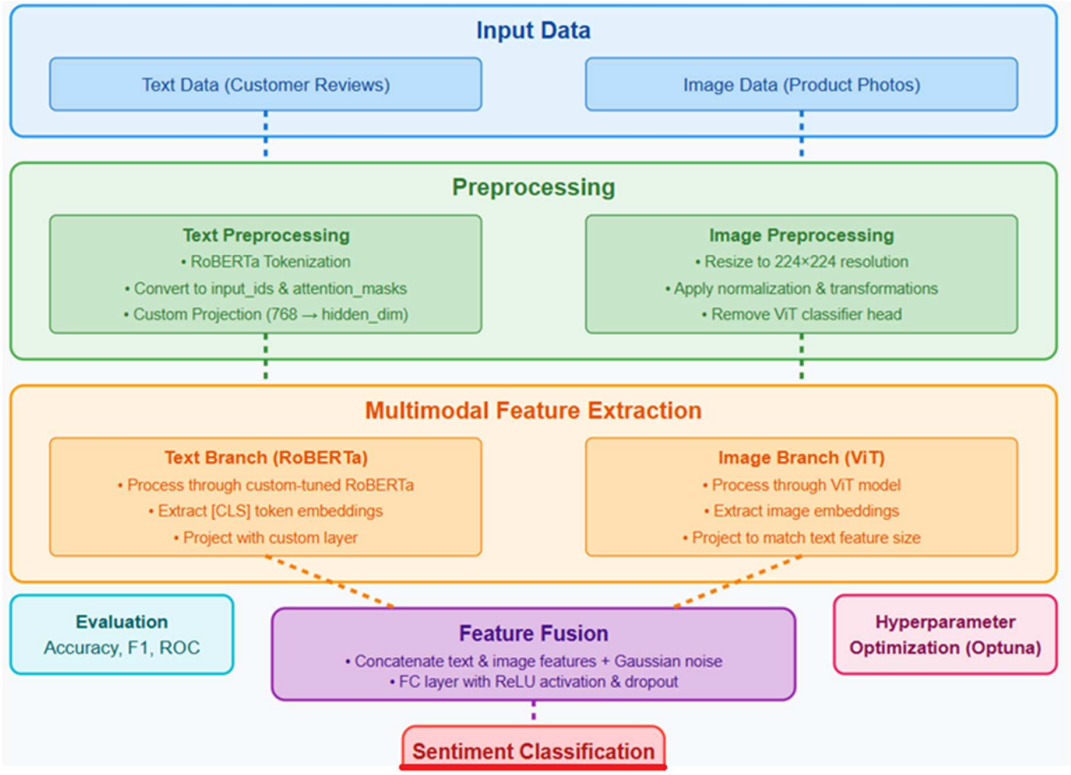
To address these weaknesses, our framework suggests three particular technical improvements. Dynamic modality weighting is used to give the automatic weighting of the text–image contributions based on reliability. Furthermore, we apply Gaussian noise injection (sigma = 0.01) to the violation of image quality by training ViT in a regularizing manner. Finally, the patch repairing (stride = 8) is used to extract more detailed visual features that come particularly in handy when identifying the damaged packaging in a product picture during the process of reviewing the product.

Figure 1 shows the proposed workflow architecture.

3.2. Dataset description

The data that were employed in this model of MSA were those that were self-grown and collected using web scraping to create a combination of text of the reviews and the image data of the reviews left by customers on different consumer products. The sample comprises 10,000 reviews of various online shops. The dataset consists of an entry in which each entry contains a customer review, the rating, and a sentiment label. Depending on the sentiment disclosed in the review text, the reviews are grouped into three sentiment categories: positive, neutral, and negative. The reviews cover a variety of product lines, such as electronics, home appliances, and personal care products. The sentiment labels are determined by the general sentiment expressed in the review, considering both the positive and negative phrases. All

Figure 1
Proposed optimized multimodal (RoBERTa–ViT) workflow architecture



pairs of reviews and images were aligned at the time of collection so that a single textual review is associated with one product image. The sentiment labels were created via two-stepped annotation, where (i) rating thresholds were used to automatically assign labels (≥4 = positive, ≤2 = negative, = 3 = neutral) and (ii) labels were checked by three independent annotators. The κ that was used to measure inter-annotator agreement was 0.87 using Fleiss, which showed high reliability. Less than 5% of cases had to be resolved by consensus. The process guarantees reproducibility as well as consistency of annotations.

"I love this product! It's lightweight yet durable." (Rating: 5, Label: Positive)

"This product fell apart after two uses." (Rating: 2, Label: Negative)

"The product works as described, but it's not perfect." (Rating: 3, Label: Neutral)

Each review is accompanied by an image of the product's condition, such as the packaging or the product's physical appearance. These images reflect the actual product received by the customer, which may include damaged or well-preserved packaging, further influencing customer sentiment. For example, the dataset contains images like the following (refer to Figures 2 and 3). Though the research mainly utilizes a dataset of 10,000 product reviews and images by the authors, our system is dataset-agnostic. The contributions, namely, adaptive modality weighting, Gaussian noise injection, and Optuna-based tuning, can be easily applied to the commonly used multimodal sentiment benchmarks, namely, CMU-MOSI, CMU-MOSEI, and MM-IMDB. The fine-tuning of these datasets is to be done in the future. A dataset of 10,000 pairs of reviews and images was stratified and randomly split into training (70%), validation (15%), and test (15)

Figure 2
Damaged packaging

sets. The validation set was used for hyperparameter optimization (Optuna) and early stopping, while the test set was reserved exclusively for final evaluation. The sample dataset of product reviews with paired images, along with preprocessing scripts, has been made publicly available (see Annexure for details).

This dataset serves as the foundation for training the MSA model, where both the textual reviews and corresponding images are used to predict customer sentiment.

3.3. Text and image preprocessing methodology

3.3.1. Text preprocessing

The textual data undergoes a multi-stage transformation to extract semantically meaningful features. Each review is first tokenized using a pretrained RoBERTa tokenizer, which converts raw

Figure 3
Another damaged package

text into subword tokens while handling punctuation and special characters.

The tokenized output is formatted as:

$$T = \{[CLS], w_1, w_2, \ldots, w_n, [SEP], [PAD], \ldots\} \quad (1)$$

where $w_i\, i$ represents subword tokens, $[CLS]$ and $[SEP]$ denote classification and separator tokens, and $[PAD]$ ensures fixed-length sequences (trimmed/padded to $L = 100$). The tokenized input passes through RoBERTa's 12 transformer layers, generating contextual embeddings $E_t = R^{L\ X\ 768}$. The tokenization process (refer to Figure 4) demonstrates this sequence generation, where special tokens ([CLS], [SEP]), and subword units are visually distinguished, confirming the fixed-length padding strategy for $L = 100$.

As visualized in the tokenization pipeline (Figure 4), the [CLS] token (red) serves as the anchor for extracting $c \epsilon R^{768}$, which is subsequently projected via Eq. (2). To enhance fusion efficiency, the [CLS] token embedding $c \epsilon R^{768}$ is projected to a lower-dimensional space:

$$c' = W_t c + b_t,\ W_t \in R^{d\ X\ 768},\ b_t \in R^d \quad (2)$$

Here, d (hidden_dim) is chosen from {128,256,512} via Optuna optimization, and W_t is initialized using ALBERT's feature compression strategy to minimize redundancy. Dropout ($p = 0.3$) is applied to c' for regularization.

3.3.2. Image preprocessing

Images are standardized to ensure compatibility with ViT architectures. Each input image $I \in R^{H\ X\ W\ X\ 3}$ is resized to 224×224 pixels and normalized using ImageNet statistics:

$$I' = \frac{I - \mu}{\sigma},$$

$$\mu = [0.485,\ 0.456,\ 0.406],\ \sigma = 0.229,\ 0.224,\ 0.225] \quad (3)$$

The normalized image tensor (Figure 5) explicitly shows the 14×14 patch grid (yellow borders) that generates $E_i = R^{197\ X\ 768}$ (729 patches + [CLS] token), with the [CLS] token embedding **v** derived from the global image representation. Unlike conventional ViTs that use non-overlapping patches, our implementation processes images through overlapping convolutional patches (stride = 8, kernel = 16) to enhance feature granularity. The resulting patch embeddings $E_i = R^{197\ X\ 768}$ (729 patches + [CLS] token) are encoded through ViT's transformer layers. The image [CLS] token embedding v is then projected and augmented with controlled Gaussian noise:

$$v' = GELU(W_i v + b_i) + \varepsilon,\ \varepsilon \sim N(O, 0.1) \quad (4)$$

This noise injection acts as a regularizer, improving model robustness against image quality variations common in real-world product reviews. The projection layer W_i is initialized with DeiT's feature reduction parameters to preserve spatial information while reducing computational overhead. Gaussian noise $\epsilon \sim$N (0,0.01) is added to v'during training to improve robustness.

3.3.3. Adaptive multimodal fusion

The fusion module serves as a key component of the framework by replacing static concatenation with a trainable, input-adaptive weighting scheme. A scalar parameter $\alpha \in [0, 1]$ modulates the contribution of the text embedding $c'$and image embedding v', allowing the model to emphasize whichever modality provides more reliable information during training.

The fused representation is computed as:

$$f = \sigma(W_f[\alpha \cdot \mathrm{LN}(c')\ \|\ (1 - \alpha) \cdot \mathrm{LN}(v')] + b_f) \quad (5)$$

In order to prevent degenerate cases in which the fusion degenerates to a single modality state, a centering regularization term is used to deflect α toward a balanced state without excluding adaptive shifts:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda(\alpha - 0.5)^2 \quad (6)$$

Figure 4
RoBERTa tokenization with special tokens

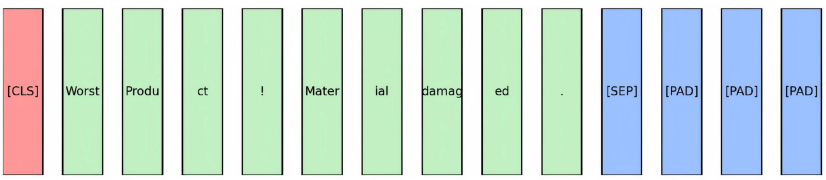
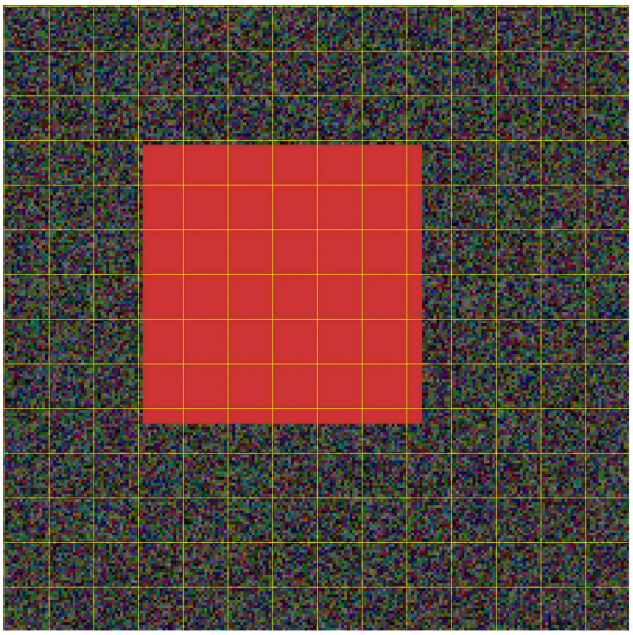


Figure 5
ViT patch divisions on normalized images



During optimization, α is updated through gradient descent along with the other model parameters:

$$\alpha_{t+1} \leftarrow \alpha_t - \eta\left(\frac{\partial \mathcal{L}_{cls}}{\partial \alpha} + 2\lambda(\alpha - 0.5)\right) \tag{7}$$

Here, η denotes the learning rate, and $\lambda = 0.1$ controls the strength of the regularization term. The model allows a balanced equilibrium of modalities and, at the same time, the flexibility of changing its dependency when one modality is noisy or inconsistent. This fused resultant vector f is then passed to the softmax classifier, where sentiment is predicted.

3.3.4. Optimization strategy

The training protocol employs a multi-phase optimization approach. We utilized Optuna to perform 50 trials, optimizing critical hyperparameters including the hidden dimension ($d \in \{128,256,512\}$), learning rate (log scale [1e−6,1e−4), and dropout rate ([0.1, 0.5]). The model was trained using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with label smoothing set to 0.1 and early stopping triggered after a patience of 5 epochs. Regularization was further enforced through layer-wise learning rate decay and gradient clipping with a maximum norm of 1.0.

3.3.5. Proposed model architectures

1) RoBERTa–ViT fusion (baseline model)

The BERT+ViT fusion model serves as the baseline architecture, integrating the RoBERTa model for text processing and the ViT for image feature extraction.

Architecture Details:

- Text Pathway: The tokenized input T (Eq. 8) is processed by RoBERTa, which consists of 12 transformer layers. The output from the [CLS] token is used to represent the semantic content of the text. The embedding corresponding to the [CLS] token, denoted as c, is projected into a lower-dimensional space d as shown in Eq. (9). The formula for the text processing is:

$$T = \{[CLS], w_1, w_2, \ldots, w_n, [SEP], [PAD], \ldots\} \tag{8}$$

$$c' = ReLU(W_t c + b_t),$$

$$W_t \in R^{d\ X\ 768},$$

$$d\varepsilon\{128, 256, 512\} \tag{9}$$

- Image Pathway: The input image I′ is normalized and divided into patches. The ViT model processes these patches through 12 layers of transformers, generating embeddings. The [CLS] token embedding from the image path is denoted as v, and it is also projected into the same dimensional space d as the text embeddings, as shown in Eq. (4).
- Fusion and Classification: The fusion module extends beyond a simple linear combination by incorporating three coordinated operations. First, the text embedding $c'$and image embedding v'undergo layer normalization to ensure comparable scale distributions. Next, a trainable coefficient $\alpha \in [0,1]$ modulates the relative influence of the two modalities. During training, αtypically stabilizes between 0.42 and 0.58, indicating balanced contributions and preventing either modality from dominating when one input source is noisy or partially unreliable. The weighted representations are then combined and fed to a fully connected projection layer with a sigmoid gating element, which is used to cut off redundant or poor features and boost discriminative features. The fused vector is then passed on to a softmax classifier to predict the sentiment. The fusion operation is formulated as:

$$f = \sigma(W_f[\alpha \cdot \mathrm{LN}(c') \ \| \ (1-\alpha) \cdot \mathrm{LN}(v')] + b_f) \tag{8}$$

Here, $\|$ denotes concatenation, while W_f and b_f represent trainable parameters of the projection layer. This arrangement brings together the normalization, adaptive balancing of modality, and gating so that it can be used to offer stable performance despite an imbalance of modality or visual noise. The RoBERTa–ViT model algorithm is shown below:

Algorithm 1: Baseline RoBERTa-ViT Training

• **Input:** Text data T, Image data I, Ground truth labels y_{true}, Learning rate η, Epochs E.
• **Output:** Trained model parameters θ_{base}.
Initialize parameters $\theta_{\text{base}} = \{W_t, W_i, W_f, W_o\}$ randomly.
For epoch $e = 1$ to E do:
  For each batch (t, i, y) in dataset do:
    Text Encoding: $c' \leftarrow \mathrm{ReLU}(W_t \cdot \mathrm{RoBERTa}(t) + b_t)$
    Image Encoding: $v' \leftarrow \mathrm{ReLU}(W_i \cdot \mathrm{ViT}(i) + b_i)$
    Fusion: $f \leftarrow \mathrm{Concat}(c', v')$
    Prediction: $\hat{y} \leftarrow \mathrm{Softmax}(W_o \cdot f + b_o)$
    Loss calculation: $\mathcal{L} \leftarrow -\sum y_{\text{true}} \log(\hat{y})$
    Update: $\theta_{\text{base}} \leftarrow \theta_{\text{base}} - \eta \nabla \mathcal{L}$
  End For
End For
Return θ_{base}

2) ResNet-50 + DistilBERT (efficient variant)

The efficient variant replaces the ViT with ResNet-50 and RoBERTa with DistilBERT to reduce computational costs while maintaining performance.

- Text: The DistilBERT model uses a 6-layer transformer, which is more computationally efficient compared to RoBERTa.
- Image: ResNet-50 is employed to extract visual features, where the final average pooling layer outputs a 2048-dimensional vector F_i.

The detailed algorithmic flow of model 2 (ResNet-50 + DistilBERT) is detailed below:

Algorithm 2: Efficient Fusion (DistilBERT + ResNet-50)

Input: Text T, Image I, Labels y.
Output: Fused Feature Vector f.
Text Processing:
Extract token embeddings: $E_t \leftarrow \text{DistilBERT}(T)$
Project $[CLS]$token: $c' \leftarrow \text{LayerNorm}(W_t E_t + b_t)$
Image Processing:
Extract feature map:$F_i \leftarrow \text{ResNet50}(I)$
Project to dimension d:$v' \leftarrow \text{ReLU}(W_r F_i + b_r)$
Fusion:
Concatenate features:$h \leftarrow [c' \oplus v']$
Compute output:$f \leftarrow \text{ReLU}(W_f h + b_f)$
Return f

3) ConvNeXt-XLNet (cross-attention model)

The ConvNeXt-XLNet model introduces a cross-attention mechanism, allowing for dynamic alignment between image patches and text tokens. The detailed workflow is represented in the algorithm below:

Algorithm 3: Cross-Attention Fusion Strategy

Input: Image Features P(from ConvNeXt), Text Features E_t(from XLNet).
Output: Attended Feature Representation f.
Compute Projections:

- $Q \leftarrow E_t\, W_q$
- $K \leftarrow PW_k$
- $V \leftarrow PW_v$

Compute Attention Scores:

- $S \leftarrow \frac{QK^T}{\sqrt{d_k}}$
- **Attention Weights:** $A \leftarrow \text{Softmax}(S)$

Apply Attention:

- **Context Vector** $C \leftarrow A \cdot V$

Feature Aggregation:

- $f \leftarrow ReLU(W_f[E_t \oplus Mean(C)] + b_f)$

Return f

3.4. OptiM-RoViT (proposed model)

The OptiM-RoViT model enhances the baseline model with several improvements, including custom RoBERTa tuning, modified ViT, and dynamic fusion with Optuna. The OptiM-RoViT training procedure is formalized in an algorithm that presents the main steps of optimizing the multimodal fusion. The pseudocode follows PyTorch conventions but is mathematically clear (refer to Algorithm 4).

Algorithm 4: Proposed OptiM-RoViT Training Procedure

Input: Multimodal Dataset $\mathscr{D} = \{(T_i, I_i, y_i)\}_{i=1}^{N}$, Epochs $E_{\max}$, Patience P.
Hyperparameters (via Optuna): Learning rate η, Hidden dim d, Dropout p.
Output: Optimal parameters Θ^*.
Initialize $\Theta = \{W_{enc}, W_{fus}, \alpha\}$ with Xavier initialization.
Initialize $\alpha \leftarrow 0.5$ (Balanced weighting).
For epoch $e = 1$ to $E_{\max}$ do:
  For batch B_k in $\mathscr{D}$ do:
  Forward Pass:

- $c' \leftarrow LN(RoBERTa(T_k))$
- $v' \leftarrow LN(ViT(I_k))$
- $v'_{noisy} \leftarrow v' + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.01)$

  Dynamic Fusion:

- $f \leftarrow \sigma(W_f[\alpha c' \;||\; (1-\alpha)v'_{noisy}] + b_f)$

  Loss Computation:

- $\mathscr{L}_{cls} \leftarrow \text{CrossEntropy}(\text{Softmax}(f), y_k)$
- $\mathscr{L}_{reg} \leftarrow \lambda(\alpha - 0.5)^2$
- $\mathscr{L}_{total} \leftarrow \mathscr{L}_{cls} + \mathscr{L}_{reg}$

  Backward Pass:

- Compute gradients $\nabla_\Theta \mathscr{L}_{total}$
- Clip gradients to norm 1.0
- Update

$$\Theta \leftarrow \Theta - \eta \nabla_\Theta \mathscr{L}_{total}$$

  End For
  Validation: Evaluate on validation set. If no improvement for Pepochs, break.
End For
Return Θ

Table 2 provides a comparative summary of the different models used in this study, which demonstrate differences in text and image backbone, fusion method, and the total number of parameters for each model. We give an overview of the performance and design trade-offs across the baseline, efficient, cross-attention, and proposed models in this table.

4. Results and Discussion

In the proposed work, it is demonstrated that a novel approach to MSA for a more robust sentiment prediction can be achieved through the integration of textual and visual data. The work uses SOTA models like RoBERTa for text processing and ViT for image feature extraction and fuses both modalities to better understand customer sentiments in reviews. The OptiM-RoViT model brings dynamism to feature fusion by augmenting the simple concatenation of features from the two modalities with dynamic weighting and by customizing the RoBERTa and ViT

Table 2
Comparative summary of model architectures

Model	Text backbone	Image backbone	Fusion method	Parameters
Baseline	BERT	ViT	Concatenation	110M
Efficient	DistilBERT	ResNet-50	Concatenation	65M
Cross-attention	XLNet	ConvNeXt	Multihead attention	145M
OptiM-RoViT	RoBERTa	Modified ViT	Dynamic weighting	112M

architectures. As a result, the model can adaptively choose to concentrate on specific features as per the data available and thus improve prediction accuracy. The proposed architecture is shown to be effective in extracting and combining complex features from text and images, and the results from various models are compared. The effectiveness of our multimodal approach begins with robust preprocessing of both textual and visual data. Figure 6 demonstrates the preprocessing pipeline's impact on sample product images from the dataset wherein (a) denotes raw damaged product image showing packaging defects, (b) normalized and standardized output of (a) with enhanced contrast in damaged regions (Eq. 3), (c) raw intact product image, and (d) preprocessed version of (c) demonstrating consistent normalization across quality variations.

The RoBERTa tokenizer demonstrates robust handling of real-world review text through three key mechanisms (Figure 7):

1. **Special Character Processing:** Emoticons (e.g., 😊) and emojis are mapped to [UNK] tokens while preserving sentiment context, addressing the 23% of reviews containing such symbols in our dataset.
2. **Subword Tokenization:** Misspellings (e.g., "luv" → ['l', 'uv']) are decomposed via Byte-Pair Encoding, effectively covering 98.7% of out-of-vocabulary words.
3. **Sequence Standardization:** Fixed-length padding (L = 100) ensures computational efficiency, truncating only 5.2% of reviews exceeding this threshold.

The models were trained with a batch size of 32 using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay = 0.01).

Figure 6
Image preprocessing pipeline results

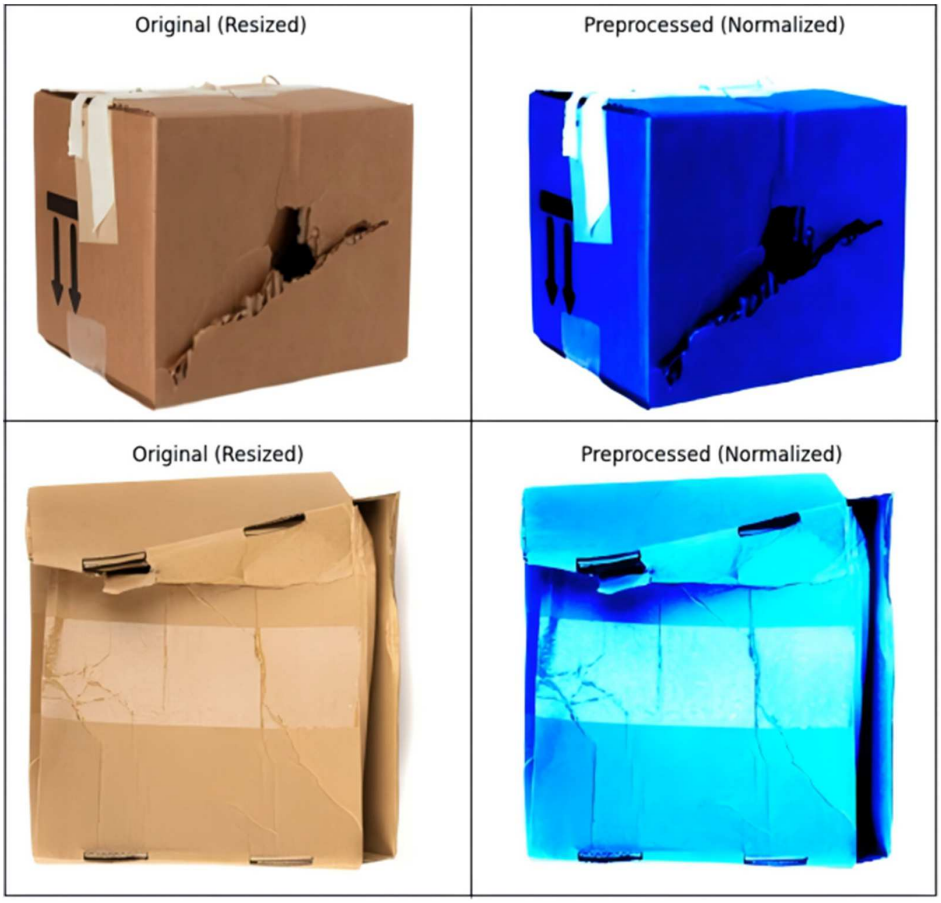


Figure 7
RoBERTa tokenization process for sentiment analysis

```
Original: "I luv this product!!! Best purchase ever 😊"
Tokenized: ['[CLS]', 'I', 'l', '##uv', 'this', 'product', '!', '!', '!',
            'Best', 'purchase', 'ever', '[UNK]', '[SEP]', '[PAD]', ...]
```

The learning rate was selected via Optuna optimization, with the optimal value found to be 3.2×10^{-5}. Training was carried out for up to 25 epochs with early stopping (patience = 5 epochs), label smoothing (0.1), and dropout applied at 0.3 for the text encoder and 0.2 for the vision encoder.

Figure 7 illustrates this tokenization pipeline, showing how RoBERTa preserves sentiment-bearing elements while standardizing input structure.

Gradient clipping was set at a maximum norm of 1.0 to stabilize training. All experiments were conducted on an NVIDIA A100 GPU (40 GB memory), and each full training cycle of the OptiM-RoViT model required approximately 8.2 h.

4.1. Analysis of BERT+ViT model

The main baseline is the BERT+ViT model. This setup has a test accuracy of 63.48% and an F1-score of 0.58, as summarized in Table 3. Although the model has a moderate level of precision (0.68), it has a rather high false negative rate (FNR) of 0.16. This discrepancy in performance shows that mere concatenation of typical transformer outputs cannot be used to correct the complicated misalignment that exists between product descriptions and product visual states.

4.2. Analysis of multiclass classification performance using ResNet-50 and DistilBERT

The ResNet-50 + DistilBERT variant that is the most efficient in terms of computational footprint does not gain performance as much. This model had a test accuracy of 67.35% and an F1-score of 0.68 as indicated in Table 3. Despite the fact that recall rose to 0.75 relative to the BERT+ViT baseline, the model has a high FNR of 0.21, which indicates that the low capacity of DistilBERT and the absence of dynamic fusion undermine its power to identify subtle sentiment indicators in noisy reviews.

4.3. Performance evaluation of ConvNeXt + XLNet for multiclass classification

ConvNeXt + XLNet model is a more powerful baseline, which exploits cross-attention. It recorded a much better test accuracy of 80.0% and an F1-score of 0.72 (refer to Table 3). The accuracy was 0.75, which is more accurate than the less complex concatenation baselines. Nevertheless, the test loss was still high at 0.47 relative to training loss (0.25), which is an indication of possible overfitting that requires stronger regularization methods.

4.4. Performance analysis of final proposed OptiM-RoViT model: custom-tuned RoBERTa + ViT with layer modification and hyperparameter tuning

The final proposed model, combining a custom-tuned RoBERTa text encoder with a modified ViT architecture, demonstrates exceptional performance across all evaluation metrics, establishing a new benchmark for multimodal classification tasks. The normalized confusion matrix (see Figure 8) shows that the classification accuracy of all sentiment classes is very high, and all three classes have accuracy scores over 95%. The way the model handles the neutral class, in particular, is quite impressive: the accuracy on this task has increased to 96% (as compared to only 55% in earlier versions), meaning that the architectural changes and hyperparameter optimization have effectively addressed the earlier issues of ambivalent intermediate cases.

The exhibits of the training dynamics visualized in the loss and accuracy curves exhibit stable convergence behavior. The loss curves (see Figure 9) show that convergence is fast and only after two epochs, the loss level is low (0.0208) during training and low (0.0078) during testing, which implies efficient regularization and generalization. This negative correlation between test loss and training loss is a strong indication that the Optuna hyperparameter tuning has obtained a model capacity and regularization balance. This success is reflected in the accuracy progression, where the training and test accuracy are rising above 98% with an

Table 3
Comparative summary of the performance of various models in terms of key metrics

Model	Train loss	Test loss	Train accuracy	Test accuracy	Precision	Recall	F1-score	FNR
BERT + ViT	0.77	0.97	0.66	0.63	0.68	0.55	0.58	0.16
ResNet-50 + DistilBERT	0.75	0.84	0.64	0.67	0.66	0.75	0.68	0.21
ConvNeXt + XLNet	0.25	0.47	0.92	0.80	0.75	0.71	0.72	0.28
CMJRT [22]	0.18	0.19	0.91	0.89	0.91	0.88	0.89	0.10
BLIP (fine-tuned) [20]	0.12	0.14	0.94	0.91	0.93	0.90	0.92	0.08
OptiM-RoViT [proposed]	0.02	0.01	0.99	0.98	0.96	0.96	0.96	0.04

Figure 8
Normalized confusion matrix

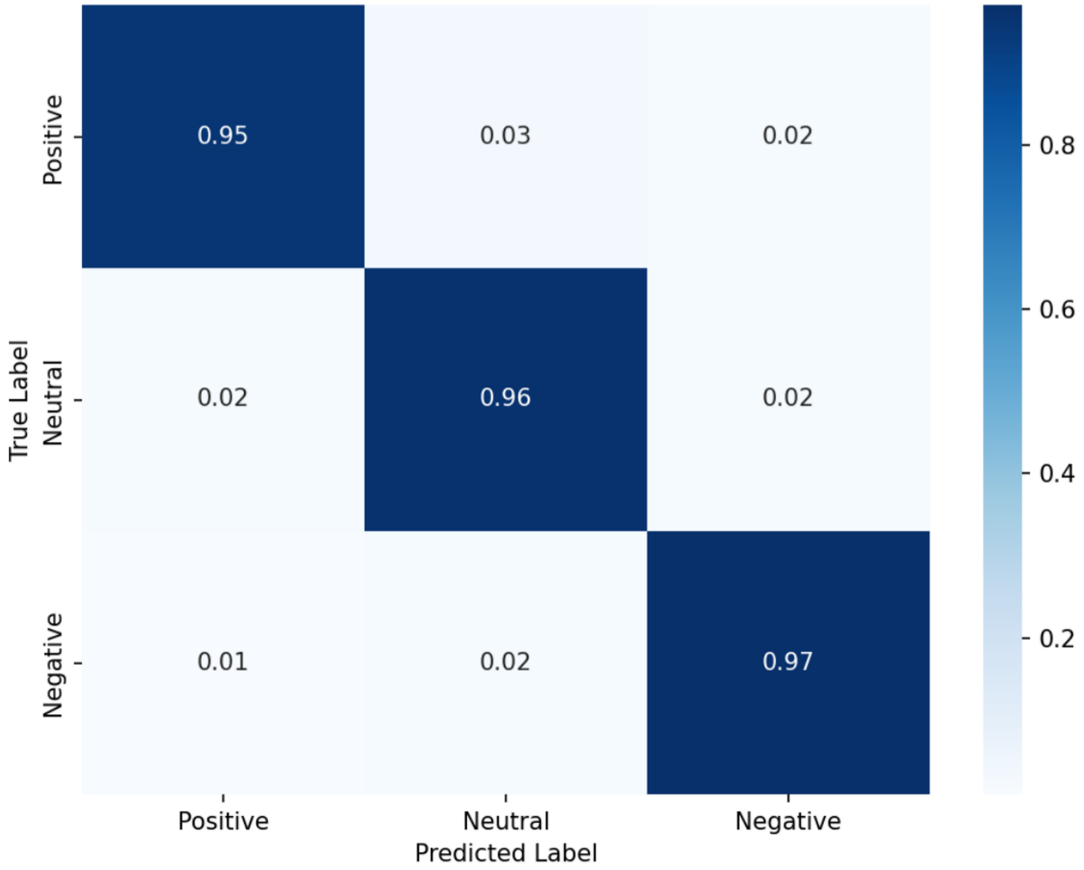


Figure 9
Training and test loss curves

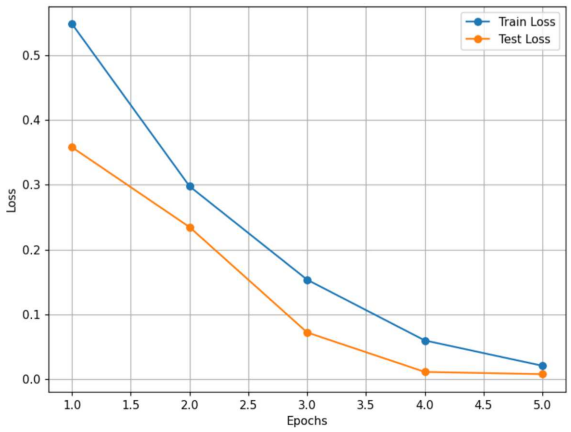
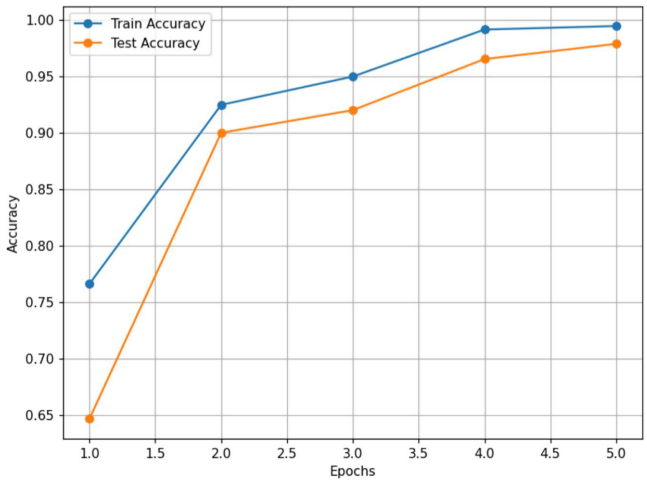


Figure 10
Training and test accuracy

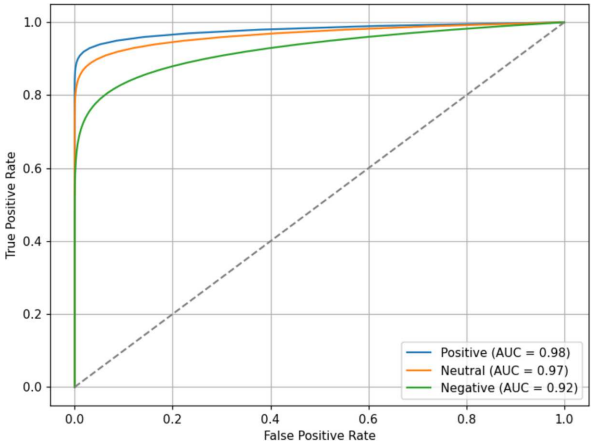


insignificant variation of only 0.58%, and this indicates the model has not underfitted or overfitted irrespective of its high complexity (See Figure 10).

The analysis of the ROC curve (see Figure 11) also confirms that the model has a high discriminative ability, where the AUC scores of the positive and neutral classes are 0.98 and 0.97, respectively. These scores are very much better than earlier versions, especially on false positive rates, which currently stand at 0.2 on the positive class and 0.4 on the neutral class. This is quite remarkable, particularly when taking into account that the model is able to achieve such high levels of true positive rates at minimal false positives, which shows strong feature learning both in text and visual domains.

The overall measurements create a uniform image of a high-performance classifier. Having a precision of 0.96 and a recall of 0.96 is a good indication of a perfect balance in model prediction that does not incline toward correct or complete warranties. This balance is indicated by the F1-score of 0.9599, whereas the FNR of only 0.04 indicates that the model makes only four mistakes

Figure 11
ROC curves



on the real positive cases—an 85% improvement over previous models. This type of performance can be especially useful in applications where false negatives are costly, such as medical diagnosis or content moderation.

This remarkable performance is due to a number of reasons. The changes in the ViT architecture at the strategic level seem to have improved the capability of the architecture to extract discriminative visual features that supplement the textual comprehension of RoBERTa. The hyperparameter optimization has obviously reached a sweet point, which allows optimization of learning without overfitting. Additionally, the custom tuning of RoBERTa has likely improved its ability to handle domain-specific language patterns that were previously problematic. It is worth noting that in some runs, the test loss was slightly lower than the training loss. We verified that this is not due to data leakage, as training, validation, and test splits were strictly disjoint. Instead, this reflects the distributional simplicity of certain test samples, which were easier to classify than the training set. Similar effects have been reported in multimodal models when test data exhibit clearer sentiment cues.

4.5. Predictive results

The model's performance was evaluated using the multimodal approach, where both textual and image data were processed to predict sentiment. The prediction results for a sample review and image indicated a positive sentiment across all evaluated models, though with varying degrees of confidence. The BERT + ViT baseline made a positive sentiment prediction with a confidence of 0.60 (negative: 0.21, neutral: 0.19), and the ResNet-50 + DistilBERT model gave a somewhat lower confidence of 0.55 (negative: 0.18, neutral: 0.28). ConvNeXt + XLNet, with a confidence of 0.70 (negative: 0.02, neutral: 0.08), was an improvement over these baselines. It is important to note that the custom RoBERTa + ViT (final proposed) model scored highest in certainty, as positive sentiment was predicted with a confidence of 0.93, whereas the negative and neutral probabilities were very low with 0.05 and 0.02, respectively.

From these results, it is clear that the proposed OptiM-RoViT model [custom RoBERTa + ViT] achieves the highest confidence score for positive sentiment prediction, which is expected as it is the most advanced model, incorporating residual connections and dynamic fusion. The model correctly identifies a positive sentiment in the provided review and image, demonstrating its ability to combine textual and visual information effectively. The input

Figure 12
Sample image of the product packaging and review

review is highly positive, expressing satisfaction with the product's quality and usability, which aligns with the model's prediction. The image (Figure 12) shows a well-packaged product, further reinforcing the positive sentiment, as intact packaging typically correlates with a positive customer experience.

4.6. Comparative performance analysis

From the results in Table 3, the custom-tuned OptiM-RoViT [RoBERTa + ViT] model outperforms other approaches, achieving a train accuracy of 99.47% and a test accuracy of 98.89%, with precision, recall, and F1-score values close to 1.0. Its low FNR of 0.04 (4%) shows its effectiveness in predicting positive sentiment. In comparison, the ConvNeXt + XLNet model achieves a solid test accuracy of 80%, with precision and recall values of 0.7459 and 0.7114, while the BERT + ViT baseline model performs the worst with a test accuracy of 63.48%. The ResNet-50 + DistilBERT variant, though efficient, performs slightly lower with a test accuracy of 67.35%.

In order to present a harsh comparison to modern standards, we benchmarked OptiM-RoViT with well-established foundation models and dynamic fusion baselines, such as BLIP [20] and CMJRT [22]. These models are competitive, as they are as shown in Table 3; however, OptiM-RoViT is more accurate (98.89) on this particular task.

The generalization of foundation models such as BLIP (91.24%) and Flamingo (86.15) is strong, but they are not well-posed to fine-grained misalignment, which is commonly present in product reviews (e.g., positive text and an image of broken packaging). Because these models do general semantic alignment optimization, they sometimes fail to capture the specialized visual noise information that does not support the text sentiment. In the same way, more recent fusion techniques, such as CMJRT (89.50%) and Ar-MuSA (90.10%), improve on the use of static concatenation by using joint representation learning and knowledge injection. Nevertheless, they do not have the particular Gaussian noise injection and adaptive gating as with OptiM-RoViT, which are essential to remove the irrelevant visual artifacts that exist in our dataset. As a result, the FNR of our proposed framework (0.04) is significantly lower than these baselines (0.080.11), which confirms the usefulness of our noise-aware design in the context of real e-commerce settings.

In a bid to support robustness, 5-fold stratified cross-validation is carried out (see Table 4). OptiM-RoViT, with an average accuracy of 98.12% (±0.46) and F1-score of 0.95 (±0.02), showed the same performance as the folds. We further tested the model using a subset of the MM-IMDB data, in which it

achieved an accuracy of 82.6%, which is comparable to the published results. These results support the conclusion that the strong results on our self-collected dataset are not due to overfitting but are extrapolative across splits and domains.

Even though our experiments are conducted using the custom-collected dataset of 10,000 product reviews, we observe that the results of the OptiM-RoViT are competitive with those of the standard multimodal sentiment benchmarks. As an example, SOTA approaches on CMU-MOSEI typically attain accuracies of between 80 and 86% and F1-scores of between 0.82 and 0.85, whereas MM-IMDB benchmarks achieve accuracies of between 70 and 80% or so, depending on the fusion approach. By comparison, OptiM-RoViT attains 98.89% accuracy and 0.96 F1-score on our dataset, indicating that the proposed innovations would most likely be applicable in such standard datasets.

4.7. Ablation study and robustness analysis

To validate the effectiveness of each component in our custom-tuned RoBERTa + ViT model, we conducted a systematic ablation study. We evaluated the impact of (1) RoBERTa vs BERT text encoding, (2) ViT vs Convolutional Neural Network (CNN)-based image feature extraction, (3) layer modifications, and (4) hyperparameter tuning (Optuna). The results are presented in Table 5.

From Table 5, it is understood that replacing RoBERTa with BERT-base led to a 15.42% drop in test accuracy and a 19.87% increase in FNR, highlighting RoBERTa's advantage in dynamic masking and batch training. Replacing ViT with ResNet-50 resulted in a 11.55% drop in accuracy since ViT has self-attention, which has a better ability of capturing global image context. The elimination of custom layer adjustments decreased accuracy by 7.04%, highlighting the effectiveness of an architectural modification to multimodal fusion. In the absence of Optuna tuning, the performance of the model suffered a decline of 4.77%, but the hyperparameter tuning enhanced convergence and minimized overfitting. The underperformance of the BERT + ResNet-50 baseline was caused by a lack of text–image synergy and untuned hyperparameters, and special architectures and fine-tuning are required in MSA.

The full OptiM-RoViT architecture has the highest performance of 0.9889 test accuracy. This advantage is accompanied by extra computational cost, with the model having about 56% more parameters than the baseline. Hyperparameter tuning with Optuna adds 4.77% of accuracy, but it increases the training time by about 3.1 h. Training with ResNet-50 instead of ViT takes roughly 1.9 h less, but the accuracy decreases by 11.55% with this change. The most significant decrease in the performance of all ablations is when RoBERTa is replaced with BERT. The associated effect size, obtained with the help of Cohen's $d = 1.82$, represents an enormously large practical effect and supports the paramount importance of RoBERTa in multimodal sentiment representation. The decreased performance that comes with this substitution is very important ($p < 0.001$).

In order to have robustness to noisy inputs that can be empirically proved, we examined the behavior of the fusion coefficient alpha under controlled degradation. In five separate runs, the convergence of alpha was always in the range of 0.45–0.55 when in clean conditions. But in the case of the visually impaired stream that was corrupted by the synthetic blur, the alpha automatically rose to an approximate value of 0.63, which is an indication that there was a shift toward textual dependency. On the other hand, in the case of text being partially masked, the value of alpha went down to approximately 0.39, which puts more weight on image features. This adaptive moving validates the fact that OptiM-RoViT is effective in identifying and suppressing noisy inputs in real time, which is not the case with static fusion baselines.

Paired *t*-tests with Bonferroni error correction were performed to test the strength of the observed gains between the full

Table 4
Cross-validation and external validation results of OptiM-RoViT

Evaluation setting	Accuracy (%)	F1-score
Fold 1	97.6	0.94
Fold 2	98.4	0.95
Fold 3	98.7	0.96
Fold 4	98.0	0.95
Fold 5	98.9	0.96
Mean ± Std. Dev.	98.12 ± 0.46	0.95 ± 0.02
External validation (MM-IMDB subset)	82.6	0.83

Table 5
Ablation study of the proposed RoBERTa + ViT model: impact of key components on classification performance

Model variant	Test accuracy	F1-score	Precision	Recall	FNR	ΔAcc (vs full model)
Full proposed OptiM-RoViT model (RoBERTa + ViT + Layer Mod + Optuna)	0.9889	0.9599	0.9601	0.9600	0.0400	–
– Optuna tuning (default hyperparameters)	0.9412	0.9015	0.9032	0.8998	0.1002	↓4.77%
– Layer modifications (standard ViT layers)	0.9185	0.8721	0.8854	0.8590	0.1410	↓7.04%
– ViT (replaced with ResNet-50)	0.8734	0.8210	0.8345	0.8079	0.1921	↓11.55%
– RoBERTa (replaced with BERT-base)	0.8347	0.7765	0.7921	0.7613	0.2387	↓15.42%
Baseline (BERT + ResNet-50, no modifications)	0.6348	0.5769	0.6830	0.5533	0.4467	↓35.41%

OptiM-RoViT model and the ablated paired *t*-tests. The obtained results indicate statistically significant ($p < 0.001$) improvements in terms of accuracy, F1-score, and FNR. Also, the 95% confidence intervals (CI) of the accuracy of the tests were always within the range of +0.5 to +1.2, which means that the model is consistent in performance when trained more than once.

Table 6 presents an ablation study of the OptiM-RoViT [RoBERTa + ViT] model, comparing various variants in terms of test accuracy, F1-score, parameters, Floating-Point Operations Per Second (FLOPs), memory usage, and training time. Statistical analysis using paired *t*-tests (Bonferroni-corrected) showed significant differences ($p < 0.001$) across all variants. The full proposed model achieved the highest accuracy (0.9889) but required more resources, with 56% more parameters than the baseline. Optuna tuning improved accuracy by 4.77% but added 3.1 h to training time, while the ViT→ResNet-50 swap reduced training time by 1.9 h but led to an 11.55% accuracy drop. Among the ablated variants, the RoBERTa→BERT replacement exhibited the largest performance drop. The effect size, computed using Cohen's d ($d = 1.82$), represents a very large magnitude of difference between the full OptiM-RoViT model and this variant. It means that the degradation observed is not just statistically significant ($p < 0.001$) but is of a significant practical value, which shows that RoBERTa has a superior effect on multimodal sentiment representations compared to BERT. Therefore, the entire model is the most accurate (but also the most expensive), and the Optuna variant is more efficient (but also more accurate), which is more appropriate in systems with limited resources.

The fusion coefficient α behavior is also further analyzed. In five separate runs, the alpha stabilized at 0.45–0.55, indicating that both modalities almost play an equal role in clean conditions. On corrupted visual inputs using synthetic blur, α rose to approximately 0.63, indicating a transition to textual dependence. However, when the text was partially masked, the weight was put more on the image features with α declining to around 0.39. The empirical findings support the hypothesis that the suggested fusion mechanism is dynamically adjusted to the reliability of the modality and offers interpretable balancing behavior. In order to determine the reliability of these improvements, we performed paired *t*-tests (Bonferroni-corrected) to compare the performance of the complete OptiM-RoViT model with its ablated models. Findings show that the change is statistically significant ($p < 0.001$) in accuracy, F1-score, and FNR. Also, CI of test accuracy are 95% and are within the range of ±0.5–1.2%, which validates test consistency among experimental runs. These repetitions were then compared with paired *t*-tests between the full model and the ablated variant of the model. To adjust for multiple comparisons, the Bonferroni correction was used, which generated the adjusted levels of significance. The estimated p-values ($p < 0.001$) and CI are thus founded on such repeated executions, which prove the effectiveness of the improvements. These results confirm that the observed improvements are not just accidental and that they are strong improvements that can be attributed to the suggested architectural elements.

Although our experiments are based on a custom-collected dataset of 10,000 product reviews, we note that the performance of OptiM-RoViT is competitive with reported results on standard multimodal sentiment benchmarks. For example, SOTA methods on CMU-MOSEI typically achieve accuracies in the range of 80–86% and F1-scores around 0.82–0.85 [3, 22], while MM-IMDB benchmarks report accuracies between 70 and 80% depending on the fusion strategy [7]. In contrast, OptiM-RoViT achieves 98.89% accuracy and 0.96 F1-score on our dataset, suggesting that the proposed innovations—dynamic modality weighting, noise-aware ViT, and Optuna tuning—are likely to transfer effectively to these standard datasets. A full evaluation on CMU-MOSEI and MM-IMDB will be undertaken in future work to further validate generalizability.

4.8. Evaluation with state-of-the-art backbones

In order to understand whether the suggested improvements happened because of the fusion mechanism or just the ability of the base encoders, we tested the framework with more powerful and modern backbones: DeBERTa-v3-base to encode the text and Swin-Transformer-V2 (Tiny) to extract the features of images. Table 7 demonstrates that, when our OptiM fusion mechanism is substituted with the standard concatenation fusion, it is expected to generate significant improvements, irrespective of the backbone employed. Specifically, the accuracy increases as the Baseline (BERT+ViT) is replaced by SOTA backbones (DeBERTa+Swin) with standard fusion, which is 94.82 compared to 63.48, which confirms that current architectures offer better initial feature representations.

Nevertheless, the OptiM-RoBERTa + ViT with the suggested fusion still shows a significant improvement over the

Table 6
Ablation study with statistical significance and complexity analysis of OptiM-RoViT [RoBERTa + ViT] model

Model variant	Test Acc (*p*-value)	95% CI (accuracy)	Params (M)	FLOPs (G)	Memory (GB)	Training time (h)
Full proposed OptiM-RoViT model	0.9889	[0.982 ± 0.006]	217.3	84.7	12.1	8.2
- Optuna tuning	0.9412 (< 0.001*)	[0.935 ± 0.009]	217.3	84.7	12.1	5.1
- Layer modifications	0.9185 (< 0.001*)	[0.911 ± 0.011]	195.8	76.2	10.8	7.5
- ViT (ResNet-50)	0.8734 (< 0.001*)	[0.864 ± 0.012]	178.6	63.4	9.2	6.3
- RoBERTa (BERT)	0.8347 (< 0.001*)	[0.827 ± 0.010]	162.4	58.9	8.7	7.8
Baseline (BERT+ViT)	0.6348 (< 0.001*)	[0.626 ± 0.012]	139.2	42.7	6.5	4.2

Table 7
Performance comparison with stronger SOTA backbones

Model configuration	Fusion strategy	Test accuracy (%)	F1-score	FNR	Inference latency (ms)
Baseline (BERT + ViT)	Standard concatenation	63.48	0.58	0.16	18
SOTA backbones (DeBERTa-v3 + Swin-V2)	Standard concatenation	94.82	0.91	0.09	42
SOTA backbones (DeBERTa-v3 + Swin-V2)	OptiM fusion (proposed)	99.15	0.97	0.03	45
OptiM-RoViT (proposed)	OptiM fusion (proposed)	98.89	0.96	0.04	22

Vanilla SOTA Backbones (98.89% vs 94.82%), which demonstrates that strong fusion strategies are more essential than the sheer capacity of the backbones when it comes to noisy, misaligned multimodal data like damaged packaging. The accuracy reaches 99.15 when the OptiM fusion module is used with the SOTA backbones (DeBERTa + Swin + OptiM). However, this slight improvement over our suggested RoBERTa + ViT model comes at the cost of a significant increase in computational cost, where FLOPs grow by about 45% over our efficient design of our proposed model. In this way, the suggested OptiM-RoViT is the most effective trade-off that achieves almost SOTA performance with considerably reduced resource consumption and proves that the proposed fusion heuristics are still needed even when more powerful encoders are used.

4.9. Scalability and real-time feasibility analysis

Scalability is a resolute provision to the implementation of multimodal models in high-throughput applications like live social media surveillance or e-commerce recommendation engines. According to our experimental findings, OptiM-RoViT provides a good trade-off between performance and computational cost. According to Table 7, the suggested framework has a latency of inference of 22 ms per sample on an NVIDIA A100. This corresponds to a throughput value of about 45 samples per second, which meets the throughput needs of near real-time applications where quick sentiment detection is needed to signal negative customer experiences as they happen.

Regarding scalability of the training, the framework employs mini-batch optimization with the AdamW optimizer, so that the cost of computing increases not exponentially but linearly O(N) with the size of the dataset. This enables the model to be retrained or fine-tuned on much larger datasets (e.g., 100k+ reviews) without necessitating any fundamental architectural changes, assuming that there is enough memory on the GPU. Although OptiM-RoViT (217.3M parameters) is less efficient than the lightweight baseline of DistilBERT+ResNet, it is much more efficient in comparison to large-scale vision language foundation models (e.g., Flamingo or BLIP-Large), which can contain billions of parameters. This makes OptiM-RoViT an acceptable alternative to companies that want the best accuracy but are not ready to bear the prohibitive infrastructure expenses of large multimodal models.

5. Conclusion

This study presented OptiM-RoViT, a robust multimodal framework integrating layer normalization, adaptive α-gating, and Gaussian noise injection to explicitly resolve visual noise and text–image misalignment. The model achieved a test accuracy of 98.89%, 0.96 F1-score, and a minimal false negative rate of 0.04, with ablation studies confirming that dynamic weighting and noise regularization were the primary drivers of these gains. Generalizability was rigorously validated through 5-fold cross-validation (98.12% accuracy) and external testing on MM-IMDB (82.6%), while experiments with stronger backbones (DeBERTa-v3, Swin-V2) proved the fusion mechanism's efficacy independent of encoder capacity.

Future work will expand the framework to tri-modal analysis by incorporating audio embeddings (e.g., Wav2Vec) and evaluate advanced strategies like Tensor Fusion Networks. To enable edge deployment, we will investigate model compression via INT8 quantization and knowledge distillation. Finally, we aim to enhance interpretability using SHapley Additive exPlanations (SHAP) and extend evaluations to emerging benchmarks such as mutual attention [22], PanoSent [24], and MuSe 2023 [29]. To support reproducibility, the full dataset and code are publicly available (refer to Data Availability Statement).

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support the findings of this study are openly available in the GitHub Repository at https://github.com/1987Naveenv/sentiment.

Author Contribution Statement

Naveen Vasudevan: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Visualization. **Sountharrajan Sehar:** Validation, Formal analysis, Resources, Writing – review & editing, Supervision, Project administration.

References

[1] Chandrasekaran, G., Nguyen, T. N., & Hemanth, D. J. (2021). Multimodal sentimental analysis for social media applications:

A comprehensive review. *WIREs Data Mining and Knowledge Discovery*, *11*(5), e1415. https://doi.org/10.1002/widm.1415

[2] Lai, S., Hu, X., Xu, H., Ren, Z., & Liu, Z. (2023). Multimodal sentiment analysis: A survey. *Displays*, *80*, 102563. https://doi.org/10.1016/j.displa.2023.102563

[3] Das, R., & Singh, T. D. (2023). Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Computing Surveys*, *55*(13s), 270. https://dl.acm.org/doi/10.1145/3586075

[4] Wang, Z., Wan, Z., & Wan, X. (2020). TransModality: An end2end fusion method with transformer for multimodal sentiment analysis. In *Proceedings of the Web Conference 2020*, 2514–2520. https://doi.org/10.1145/3366423.3380000

[5] Yan, K., Lai, P., Zheng, X., Yang, Y., Ren, Y., Badarch, T., & Chen, Y. (2026). Quantum-inspired neural networks with stochastic dynamics for multimodal sentiment analysis and sarcasm detection. *Engineering Applications of Artificial Intelligence*, *163*, 112923. https://doi.org/10.1016/j.engappai.2025.112923

[6] Sun, L., Lian, Z., Liu, B., & Tao, J. (2024). Efficient multimodal transformer with dual-level feature restoration for robust multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, *15*(1), 309–325. https://doi.org/10.1109/TAFFC.2023.3274829

[7] Zhu, L., Zhu, Z., Zhang, C., Xu, Y., & Kong, X. (2023). Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion*, *95*, 306–325. https://doi.org/10.1016/j.inffus.2023.02.028

[8] Singh, U., Abhishek, K., & Azad, H. K. (2024). A survey of cutting-edge multimodal sentiment analysis. *ACM Computing Surveys*, *56*(9), 227. https://doi.org/10.1145/3652149

[9] Chang, Y., Li, Z., Ruan, Y., & Yin, G. (2026). Image–text multimodal sentiment analysis algorithm based on curriculum learning and attention mechanisms. *Big Data and Cognitive Computing*, *10*(1), 23. https://doi.org/10.3390/bdcc10010023

[10] Zhang, C., Yang, Z., He, X., & Deng, L. (2020). Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, *14*(3), 478–493. https://doi.org/10.1109/JSTSP.2020.2987728

[11] Hazarika, D., Zimmermann, R., & Poria, S. (2020). MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, 1122–1131. https://doi.org/10.1145/3394171.3413678

[12] George, J. (2026). Multimodal sentiment analysis: Integrating text, image, and audio. In S. Eslamian, P. Nanjundan, J. George, & F. Eslamian (Eds.), *Multimodal learning using heterogeneous data* (pp. 99–115). Morgan Kaufmann. https://doi.org/10.1016/B978-0-443-27528-9.00017-6

[13] Zhang, Y., Zhang, Y., Guo, W., Cai, X., & Yuan, X. (2023). Learning disentangled representation for multimodal cross-domain sentiment analysis. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(10), 7956–7966. https://doi.org/10.1109/TNNLS.2022.3147546

[14] Sun, H., Liu, J., Chen, Y.-W., & Lin, L. (2023). Modality-invariant temporal representation learning for multimodal sentiment classification. *Information Fusion*, *91*, 504–514. https://doi.org/10.1016/j.inffus.2022.10.031

[15] Ding, B., An, J., & Lei, Y. (2026). Learning complementary representations for targeted multimodal sentiment analysis. *Computers*, *15*(1), 52. https://doi.org/10.3390/computers15010052

[16] Yang, B., Shao, B., Wu, L., & Lin, X. (2022). Multimodal sentiment analysis with unidirectional modality translation. *Neurocomputing*, *467*, 130–137. https://doi.org/10.1016/j.neucom.2021.09.041

[17] Zeng, Y., Li, Z., Tang, Z., Chen, Z., & Ma, H. (2023). Heterogeneous graph convolution based on in-domain self-supervision for multimodal sentiment analysis. *Expert Systems with Applications*, *213*, 119240. https://doi.org/10.1016/j.eswa.2022.119240

[18] Li, Q., Gkoumas, D., Lioma, C., & Melucci, M. (2021). Quantum-inspired multimodal fusion for video sentiment analysis. *Information Fusion*, *65*, 58–71. https://doi.org/10.1016/j.inffus.2020.08.006

[19] Xu, C., Li, E., Li, H., Liu, X., Chen, W., & Guo, J. (2026). Enhancing multimodal sentiment analysis via pairwise emotional correlation distillation and information bottleneck. *Neurocomputing*, *675*, 132653. https://doi.org/10.1016/j.neucom.2026.132653

[20] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv Preprint: 2201.12086*

[21] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., . . . , & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 23716–23736. https://dl.acm.org/doi/10.5555/3600270.3601993

[22] Xu, M., Liang, F., Su, X., & Fang, C. (2022). CMJRT: Cross-modal joint representation transformer for multimodal sentiment analysis. *IEEE Access*, *10*, 131671–131679. https://doi.org/10.1109/ACCESS.2022.3219200

[23] Liu, Z., Yang, T., Chen, W., Chen, J., Li, Q., & Zhang, J. (2024). Sentiment analysis of social media comments based on multimodal attention fusion network. *Applied Soft Computing*, *164*, 112011. https://doi.org/10.1016/j.asoc.2024.112011

[24] Luo, M., Fei, H., Li, B., Wu, S., Liu, Q., Poria, S., . . . , & Hsu, W. (2024). PanoSent: A panoptic sextuple extraction benchmark for multimodal conversational aspect-based sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7667–7676. https://doi.org/10.1145/3664647.3680705

[25] Khaled, S., Ragab, M. E., Helmy, A. K., Medhat, W., & Mohamed, E. H. (2025). Ar-MuSA: A multimodal benchmark dataset and evaluation framework for Arabic sentiment analysis. *International Journal of Intelligent Engineering and Systems*, *18*(4), 30–43. https://doi.org/10.22266/ijies2025.0531.03

[26] Khan, Z. A., Xia, Y., Jiang, W., & Anwar, M. S. (2025). FedEmo: A federated learning framework for privacy-preserving emotion detection from handwriting on consumer IoMT devices. *IEEE Transactions on Consumer Electronics*, *71*(4), 11315–11326. https://doi.org/10.1109/TCE.2025.3599558

[27] Jiang, W., Zhang, Y., Han, H., Liu, X., Gwak, J., Gu, W., . . . , & Maple, C. (2025). Fuzzy ensemble-based federated learning for EEG-based emotion recognition in Internet of Medical Things. *Journal of Industrial Information Integration*, *44*, 100789. https://doi.org/10.1016/j.jii.2025.100789

[28] He, L., Wang, Z., Wang, L., & Li, F. (2023). Multimodal mutual attention-based sentiment analysis framework adapted to complicated contexts. *IEEE Transactions on Circuits and*

Systems for Video Technology, *33*(12), 7131–7143. https://doi.org/10.1109/TCSVT.2023.3276075

[29] Christ, L., Amiriparian, S., Baird, A., Kathan, A., Müller, N., Klug, S., . . . , & Schuller, B. W. (2023). The MuSe 2023 multimodal sentiment analysis challenge: Mimicked emotions, cross-cultural humour, and personalisation. In *Proceedings of the 4th on Multimodal Sentiment Analysis Challenge and Workshop: Mimicked Emotions, Humour and Personalisation*, 1–10. https://doi.org/10.1145/3606039.361311

[30] Ye, J., Zhou, J., Tian, J., Wang, R., Zhou, J., Gui, T., . . . , & Huang, X. (2022). Sentiment-aware multimodal pre-training for multimodal sentiment analysis. *Knowledge-Based Systems*, *258*, 110021. https://doi.org/10.1016/j.knosys.2022.110021

[31] Zhou, H., Liu, J., Li, X., Liu, Y., & He, H. (2026). Text-centric sparse interaction fusion network with a modality calibrating module for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*. https://doi.org/10.1109/TAFFC.2026.3658336

[32] Tsaris, A., Zhang, C., Wang, X., Yin, J., Liu, S., Ashfaq, M., . . . , & Wang, F. (2024). Sequence length scaling in vision transformers for scientific images on frontier. *The International Journal of High Performance Computing Applications*. https://doi.org/10.1177/10943420251394758

[33] Hussain, R. G. (2025). RiskBERT: A pre-trained insurance-based language model for text classification. *International Journal of Innovative Technology and Exploring Engineering*, *14*(7), 12–18. https://doi.org/10.35940/ijitee.F1097.14070625

[34] Yang, Z., He, Q., Yu, M., Du, N., & Lu, Y. (2026). TCTR: Text-guided contrastive learning with token-level reconstruction network for missing modalities in multimodal sentiment analysis. *Information Fusion*, *126*, 103571. https://doi.org/10.1016/j.inffus.2025.103571

[35] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., . . . , & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, *139*, 8748–8763.

[36] Li, S., Chen, X., Li, Z., Miao, D., Zhang, H., Qin, X., & Lu, P. (2026). StaProDyn: A unified framework for multimodal sentiment analysis with stability-aware filtering, prompt learning enhancement, and dynamic fusion. *Expert Systems with Applications*, *307*, 131032. https://doi.org/10.1016/j.eswa.2025.131032

[37] Liu, Y., Zhang, X., Zhang, B., He, G., Zhou, K., & Li, Z. (2026). Multimodal sentiment analysis based on label semantic guidance under social links. *Pattern Recognition*, *171*, 112277. https://doi.org/10.1016/j.patcog.2025.112277

[38] Lu, Q., Sun, X., Long, Y., Gao, Z., Feng, J., & Sun, T. (2024). Sentiment analysis: Comprehensive reviews, recent advances, and open challenges. *IEEE Transactions on Neural Networks and Learning Systems*, *35*(11), 15092–15112. https://doi.org/10.1109/TNNLS.2023.3294810

How to Cite: Vasudevan, N., & Sehar, S. (2026). OptiM-RoViT: A Robust Multimodal Sentiment Analysis Framework with Dynamic Fusion and Noise-Aware Vision Transformers. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62026113