MULTIMODAL MISINFORMATION DETECTION ON SOCIAL MEDIA

RAVI RAGHAVAN [RR1133@SEAS], DHRUV VERMA [VDHRUV@SEAS], RAAFAE ZAKI [RZAKI2@SEAS],

ABSTRACT. This paper presents a multimodal misinformation detection system for social media posts containing both text and images. We first establish unimodal baselines using pretrained BERT for text and ResNet-101 for images, and fine-tune them on the Fakeddit dataset. Next, we develop a conventional multimodal baseline that combines BERT and ResNet-101 embeddings for joint learning. Building on this, we implement a CLIP-guided multimodal approach that leverages joint text–image representations for classification. Experiments show that the CLIP-guided model outperforms both unimodal and conventional multimodal baselines, demonstrating the advantage of CLIP-based feature fusion. Overall, our results highlight the effectiveness of integrating textual and visual cues for accurate misinformation detection.

1. INTRODUCTION

The rise of social media has transformed information sharing, enabling rapid global dissemination but also amplifying misinformation that can mislead public opinion, sway elections, fuel public health misconceptions, etc. Social media posts often combine text and images, making misleading content harder to detect through manual review or unimodal models. Text-only classifiers can fail when images alter or contradict the meaning of the text, while image-only models miss nuanced textual semantics. This creates a pressing need for effective multimodal misinformation detection. Multimodal approaches jointly analyze text and visual cues, capturing the interplay between modalities. By learning richer representations, these methods can more reliably identify misleading content.

1.1. **Contributions.**

In this work, we develop a CLIP-guided multimodal model that harnesses joint text–image representations, achieving improved misinformation detection over both unimodal and multimodal baselines. Additionally, we investigate how modality-specific attention mechanisms allocate focus to textual and visual features, and comparing their performance with standard QKV Attention

2. BACKGROUND

We consider social media posts as tuples $(x_{\text{txt}}, x_{\text{img}})$ containing text and an image. The goal of multimodal misinformation detection is to classify each post as real or fake. In real-world terms, a *real* (`label` = 1) post conveys accurate, factual information, while a *fake* (`label` = 0) post contains false or misleading content, which may be intended to deceive or could arise from parody, satire, or other non-literal contexts. Our objective is to optimize the F1 score on the fake class, since performance on detecting misinformation on social media poses a greater importance than performance on news that is in fact real.

3. RELATED WORK

[2] introduced a large-scale multimodal fake news dataset and showed that simple fusion of BERT text and ResNet image features improves performance, whereas our work builds on this foundation by replacing shallow fusion with learned modality-wise attention. [3] combined CLIP, BERT, and ResNet features using CLIP similarity and modality-wise attention. In contrast, our approach uses the same encoders and performs an ablation between modality-wise attention and QKV multi-head attention to study richer cross-modal interactions. Finally, [1] categorize fake news detection methods into content-based, entity-based, and knowledge-based approaches, with our work falling within content-based multimodal models that rely solely on the textual and visual content of a post. Prior content-based approaches often use shallow fusion of text and image embeddings (e.g., concatenation or averaging) or CLIP similarity–based weighting, which computes the similarity between the CLIP text and image embeddings and adjusts each modality's contribution. While capturing basic alignment, these methods model limited cross-modal interactions. Our work differs by employing a QKV-based multi-head attention mechanism and analyzing attention patterns to understand modality contributions.

4. Approach

4.1. **Baselines.** For comparison, we consider three baselines. The text-only baseline uses BERT on social media post text, extracting embeddings from the [CLS] token of the final BERT layer; a fully connected layer is then trained on top of these embeddings, and we compare using the pretrained versus fine-tuned BERT embeddings. The image-only baseline uses ResNet-101 on social media post images, extracting embeddings from the layer before the final fully connected layer; a fully connected layer is then trained on top of these embeddings, and we compare using pretrained versus fine-tuned ResNet embeddings. The multimodal baseline concatenates BERT and ResNet-101 embeddings and passes them through a fully connected (FC) layer for joint learning. We experiment with two setups: one using pretrained embeddings without fine-tuning, and another where the embeddings are fine-tuned end-to-end along with the FC layer.

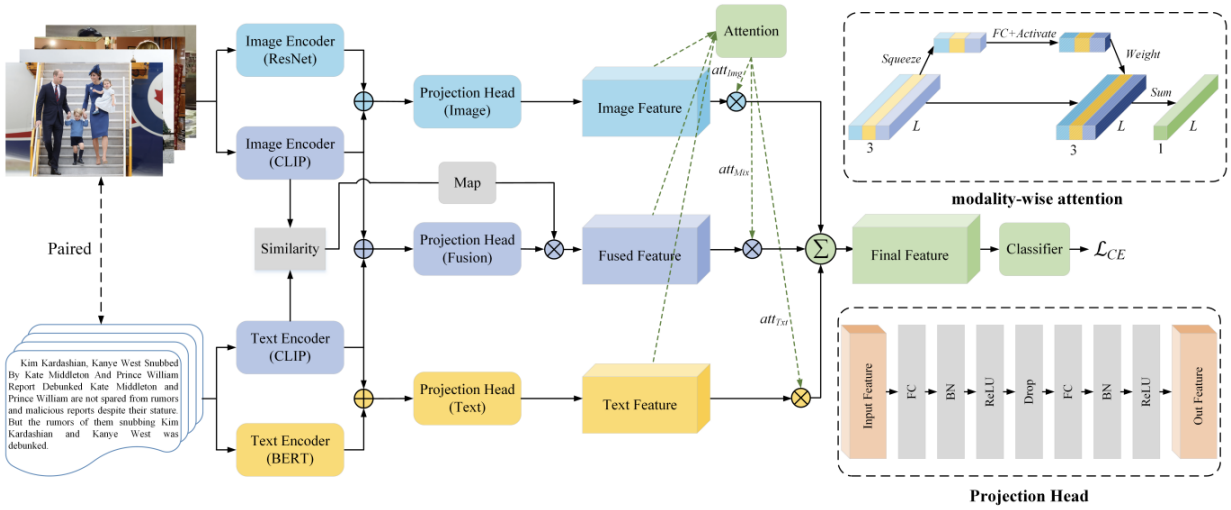


FIGURE 1. CLIP-Guided Multimodal Learning Architecture from [3]

4.2. **CLIP-Guided Multimodal Learning.** To address the challenge of effective multimodal misinformation detection, we adopt a CLIP-guided learning framework, illustrated in Figure 1, that leverages CLIP's cross-modal image–text representations to guide feature fusion. We would like to note that our work is inspired from the methodology presented in [3] on the Fakeddit Dataset.

We denote an input datum as a tuple of text and image, $(x_{\text{txt}}, x_{\text{img}})$. Unimodal features are first extracted using pretrained BERT and ResNet encoders for text and images, respectively, capturing linguistic, emotional, and visual cues. For text, we take the `[CLS]` token embedding from the final BERT layer; for images, we use the feature vector from ResNet immediately before the final fully connected layer. Formally, we compute

$$f_{\text{BERT}} = \text{BERT}_{\text{CLS}}(x_{\text{txt}}) \qquad f_{\text{ResNet}} = \text{ResNet}_{\text{pre-FC}}(x_{\text{img}}) \tag{1}$$

where f_{BERT} and f_{ResNet} denote the unimodal text and image features, respectively. In parallel, we employ CLIP's text and image encoders, denoted by CLIP-T and CLIP-I respectively, to obtain semantically aligned cross-modal representations:

$$f_{\text{CLIP-T}} = \text{CLIP-T}(x_{\text{txt}}) \qquad f_{\text{CLIP-I}} = \text{CLIP-I}(x_{\text{img}}) \tag{2}$$

To enrich the unimodal representations with cross-modal semantic information, we concatenate the CLIP features with their corresponding unimodal features. In addition, we construct a unified multimodal representation by concatenating the CLIP text and image features, where $[\cdot;\cdot]$ denotes feature concatenation.

$$f_{\text{Txt}} = [f_{\text{BERT}}; f_{\text{CLIP-T}}] \qquad f_{\text{Img}} = [f_{\text{ResNet}}; f_{\text{CLIP-I}}] \qquad f_{\text{Mix}} = [f_{\text{CLIP-T}}; f_{\text{CLIP-I}}] \tag{3}$$

Projection: Let $f_{\text{Txt}} \in \mathbb{R}^{d_T}$, $f_{\text{Img}} \in \mathbb{R}^{d_I}$, and $f_{\text{Mix}} \in \mathbb{R}^{d_M}$ denote the concatenated text, image, and multimodal features, respectively. We employ three independent projection heads P_{Txt}, P_{Img}, and P_{Mix} to map the features into a

shared latent space of dimension L. Each projection head is implemented as illustrated in Figure 1, except that we placed Batch Normalization after ReLU, rather than before as shown in the figure.

$$m_{\text{Txt}} = P_{\text{Txt}}(f_{\text{Txt}}), \quad m_{\text{Img}} = P_{\text{Img}}(f_{\text{Img}}), \quad \tilde{m}_{\text{Mix}} = P_{\text{Mix}}(f_{\text{Mix}}) \tag{4}$$

Similarity-Guided Multimodal Modulation: The reliability of multimodal representations in fake news detection critically depends on whether the textual and visual contents convey consistent semantic information. In real-world social media posts, images are frequently reused, loosely related, or even entirely irrelevant to the accompanying text, regardless of the authenticity of the news. As a result, naïvely fusing text and image features may introduce noise and mislead the classifier when cross-modal correlation is weak. To explicitly model this semantic consistency, we compute the cosine similarity between the CLIP text and image embeddings, $f_{\text{CLIP-T}}$ and $f_{\text{CLIP-I}}$.

$$s = \frac{f_{\text{CLIP-T}} \cdot f_{\text{CLIP-I}}}{\|f_{\text{CLIP-T}}\|_2 \, \|f_{\text{CLIP-I}}\|_2} \tag{5}$$

Our goal is to modulate the magnitude of the multimodal feature $\tilde{m}_{\text{Mix}}$ according to text–image semantic consistency, emphasizing aligned pairs while suppressing misaligned ones. Directly using s is problematic as scaling by s changes the magnitude of both aligned and misaligned pairs by the same factor, preventing the model from distinguishing consistent from conflicting pairs. For example, if an aligned pair has cosine similarity $s = 0.8$ and a misaligned pair has $s = -0.8$, directly scaling by s would scale the magnitude of both features by 0.8, failing to properly distinguish them. To overcome this, we map s to a non-negative, bounded modulation coefficient via a sigmoid, establishing an increasing monotonic relationship between cross-modal consistency and multimodal feature contribution.

Although $s \in [-1, 1]$, empirical analysis on the training dataset(Figure 4 in Appendix) shows that CLIP text–image cosine similarities are mostly concentrated in $[0, 0.5]$. Applying a sigmoid directly to values in this narrow range compresses them into a even smaller output interval, causing the modulation coefficients to cluster closely and reducing the model's ability to discriminate between varying degrees of semantic alignment. To mitigate this, we perform the following steps before applying the sigmoid: During training, we compute the mean μ_s and variance σ_s^2 of the cosine similarity scores s across the mini-batch and use them to standardize s, while simultaneously updating running averages of these statistics. During testing, standardization is performed using the stored running averages, analogous to Batch Normalization.

$$\alpha = \text{sigmoid}\left(\frac{s - \mu_s}{\sigma_s}\right) \tag{6}$$

The final multimodal feature is obtained by scaling $\tilde{m}_{\text{Mix}}$ with the normalized similarity score:

$$m_{\text{Mix}} = \alpha \cdot \tilde{m}_{\text{Mix}} \tag{7}$$

Modality-Wise Attention: We perform modality-wise feature aggregation to dynamically weight features from text, image, and multimodal modalities. We assume $m_{\text{Txt}}, m_{\text{Img}}, m_{\text{Mix}} \in \mathbb{R}^{L \times 1}$. The features are concatenated as follows:

$$M = [m_{\text{Txt}}, m_{\text{Img}}, m_{\text{Mix}}] \in \mathbb{R}^{L \times 3}$$

Initial modality weights are computed via global average and max pooling along the feature dimension, summed to form $w_{\text{init}} \in \mathbb{R}^{3 \times 1}$. Specifically, we perform the following operation:

$$w_{\text{init}} = \frac{1}{L} \sum_{i=1}^{L} M_{i,:} + \max_{i=1,\ldots,L} M_{i,:}$$

w_{init} is then passed through two fully connected layers with GELU activation and sigmoid normalization to obtain attention weights for the text, image, and multimodal modalities: $\text{att} = \{\text{att}_{\text{Txt}}, \text{att}_{\text{Img}}, \text{att}_{\text{Mix}}\} \in \mathbb{R}^{3 \times 1}$. The final aggregated feature is

$$m_{\text{Agg}} = \text{att}_{\text{Txt}} \cdot m_{\text{Txt}} + \text{att}_{\text{Img}} \cdot m_{\text{Img}} + \text{att}_{\text{Mix}} \cdot m_{\text{Mix}}$$

This allows the model to emphasize informative modalities while suppressing less relevant ones.

Final Classification Layer: m_{Agg} is passed through a two-layer fully connected (FC) network, where the first layer uses ReLU activation. The second layer outputs two unnormalized logits corresponding to the real and fake classes. These logits are passed through a softmax to obtain class probabilities, and the model is trained using standard cross-entropy (CE) loss. We designed this FC setup so that the final layer can be easily adapted in future work for more fine-grained misinformation labeling, such as the 3-way or 6-way labels used in the Fakeddit dataset, by simply changing the output dimension of the last layer.

5. Experimental Results

5.1. **Dataset.** We evaluate our model on the *Fakeddit* dataset, which contains Reddit posts with both textual content and associated images (provided as URLs), along with labels where 0 indicates fake news and 1 indicates real news. The dataset supports multiple classification settings; in this work, we focus on the binary fake news detection task.

5.2. **Data Preprocessing.** To ensure data quality, we first remove all posts containing null or missing values in either the text field, image field, or label. Since images in Fakeddit are provided as external links rather than stored locally, we crawl the web to download the corresponding images. After downloading, we perform an additional filtering step to remove posts with corrupted or unreadable image files. For computational efficiency, we subsample the dataset and construct fixed splits consisting of approximately 33,000 training posts, 5,000 validation posts, and 5,000 test posts. All splits are mutually exclusive and preserve the original label distribution as closely as possible.

5.3. **Model Results.** We note that CLIP: QKV-Attention refers to the ablation study described in Section 5.3.5.

Model	F1 Score
Text Unimodal (Pre-Trained BERT)	0.598
Text Unimodal (Fine-Tuned BERT)	0.819
Image Unimodal (Pre-Trained ResNet-101)	0.635
Image Unimodal (Fine-Tuned ResNet-101)	0.709
Multimodal Baseline (Pre-Trained)	0.714
Multimodal Baseline (Fine-Tuned)	0.810
CLIP: QKV-Attention	**0.846**
CLIP: Modality-Attention	0.843

Table 1. Unimodal and Multimodal model performance on the Fake News class.

We evaluate a series of unimodal and multimodal models to understand the relative contribution of textual and visual information for fake news detection, as well as the impact of task-specific fine-tuning and multimodal pre-training. Table 1 summarizes the F1 scores across all models on the Fake News class.

5.3.1. ***Text Unimodal Models.*** We begin with a text-only baseline using a pre-trained BERT encoder with frozen weights. While this model captures general linguistic patterns, it achieves a relatively low F1 score of 0.598, indicating that generic language representations alone are insufficient for this task. Fake news detection often relies on subtle stylistic cues, contextual framing, and domain-specific signals that are not explicitly captured during large-scale generic pre-training.

Fine-tuning BERT end-to-end on the fake news classification task leads to a substantial performance improvement, increasing the F1 score to 0.819. This highlights the importance of task-specific adaptation: fine-tuning enables the model to better leverage discriminative linguistic features relevant to misinformation, such as sensational phrasing or misleading narratives, which are underrepresented in general-purpose pre-training objectives.

5.3.2. ***Image Unimodal Models.*** Next, we evaluate image-only models to assess how much visual information alone can contribute to fake news detection. A pre-trained ResNet-101 encoder achieves an F1 score of 0.635, only slightly outperforming the pre-trained text-only model. This suggests that generic image representations alone are insufficient for this task. Fake news detection requires understanding subtle contextual and semantic cues in images that pre-trained models optimized for general visual tasks may fail to capture.

Fine-tuning ResNet-101 on the task improves performance to 0.709, confirming that visual features also benefit from task-specific supervision. However, after fine-tuning, image-only models consistently underperform text-based models. This aligns with the intuition that visual content alone is often ambiguous and insufficient to reliably identify fake news without accompanying textual context.

5.3.3. ***Multimodal Baselines.*** Motivated by the complementary strengths of text and image modalities, we next explore multimodal baselines that combine BERT and ResNet embeddings. Using frozen, pre-trained encoders yields an F1 score of 0.714, which improves over unimodal image models but still falls short of fine-tuned text-only performance. This suggests that the naïve fusion of generic embeddings is limited when the encoders lack task-specific alignment.

Allowing both encoders to be fine-tuned jointly results in a notable improvement to an F1 score of 0.810. This demonstrates that multimodal learning is effective when both modalities are adapted to the task, enabling the model to

exploit cross-modal correlations, such as mismatches between text claims and visual evidence, that are characteristic of fake news.

5.3.4. ***CLIP-Based Models.*** Finally, we evaluate CLIP-based architectures, which are pre-trained to align image and text representations in a shared embedding space. Both CLIP variants outperform all previous models, with the QKV-attention model achieving the best overall F1 score of 0.846 and the modality-attention variant achieving a comparable score of 0.843.

The strong and consistent performance across both attention mechanisms suggests that the primary advantage comes from CLIP's multimodal pre-training rather than the specific choice of attention fusion. By jointly modeling images and text during pre-training, CLIP captures semantically meaningful cross-modal relationships that transfer effectively to fake news detection. These results indicate that aligning modalities explicitly is more beneficial than the naive fusion of unimodal representations.

5.3.5. ***Attention Ablation Study***. As an ablation of the original modality-wise attention mechanism, we replace it with a QKV-based multi-head attention (MHA) module. Specifically, the text, image, and multimodal representations are treated as tokens within a single sequence and passed through the MHA layer. This yields three output token representations, one per modality, which are subsequently averaged to obtain a single fused representation before being fed into the final fully connected classification layer.

To compare the two attention mechanisms, we analyze and visualize the learned attention weights under each scheme. For the modality-wise attention model, Figure 2 shows that the distributions of attention weights across the test set are largely similar for samples predicted as Fake News and Real News. Moreover, the model assigns substantial attention to all three modalities (text, image, and multimodal), suggesting that it integrates information from multiple sources rather than relying on a single dominant modality. We observe a slight preference for textual features, which we hypothesize is due to the stronger baseline performance of the BERT encoder, indicating that the text modality provides relatively more discriminative information.

In contrast, Figure 3 illustrates the behavior of the QKV attention mechanism. For samples predicted as Real News, the model assigns slightly higher attention weights to the multimodal representations compared to samples predicted as Fake News. This pattern indicates that, for real-news predictions, the model relies slightly more on cross-modal text–image interactions, especially when unimodal cues are ambiguous.

Overall, both attention mechanisms demonstrate consistent reliance on all available modalities for decision-making, regardless of the predicted class. The modality-wise attention distributions remain nearly identical across Real and Fake News predictions, with all modalities receiving significant weights. The QKV attention weights are nearly uniform for both fake-news and real-news samples. These findings indicate that both models adopt a balanced, multi-modal integration strategy, analogous to human reasoning, wherein information from multiple modalities is jointly considered to arrive at a well-informed judgment.

Please refer to the appendix for more detailed information on how Figures 2 and 3 were computed.

6. Discussion

Overall, our experiments lead to three key conclusions. First, task-specific fine-tuning is essential for both text and image encoders, as generic pre-trained representations fail to capture the nuanced signals required for fake news detection. Second, when properly trained, multimodal models consistently outperform unimodal approaches, highlighting the importance of jointly reasoning over textual and visual information. Finally, our CLIP-guided multimodal framework achieves the strongest performance among all evaluated models, outperforming both unimodal approaches and conventional multimodal baselines. Our results indicate that performance gains stem from combining BERT and task-adapted ResNet representations with CLIP's semantically aligned image–text features to guide multimodal fusion. By explicitly measuring text–image alignment using the cosine similarity between CLIP text and image embeddings and learning to dynamically weight textual, visual, and multimodal signals via attention, the model is better able to suppress noisy or weakly aligned modalities, an important consideration in real-world social media data. The comparable performance observed across modality-wise and QKV attention further suggests that the benefits arise from principled multimodal interaction.

However, the current framework is limited to static images, while real-world misinformation increasingly appears in videos and GIFs, particularly in the form of deepfakes. A natural extension would be to replace the image encoder with a video backbone such as a 3D CNN or a vision transformer with temporal attention and integrate frame-level video–text embeddings within the CLIP-guided pipeline. Temporal aggregation modules could then capture inconsistencies across frames that are characteristic of manipulated media, albeit at the cost of increased computational and data requirements.

REFERENCES

[1] Y. Li, H. Su, X. Shen, W. Li, and Z. Cao. A Survey on Multimodal Fake News Detection, 2021.
[2] K. Nakamura, O. Levy, and W. Y. Wang. Fakeddit: A New Multimodal Dataset for Fine-Grained Fake News Detection, 2019.
[3] Y. Zhou, Q. Ying, Z. Qian, S. Li, and X. Zhang. Multimodal Fake News Detection via CLIP-Guided Learning, 2022.

7. APPENDIX

7.1. **Attention Ablation Study.**

7.1.1. *Modality-Wise Attention Analysis.* For the original modality-wise attention mechanism, the model outputs a scalar attention weight for each modality (text, image, multimodal) for each datum. Let $a_{\text{text}}^{(i)}, a_{\text{img}}^{(i)}, a_{\text{mult}}^{(i)}$ denote the text, image, and multimodal attention weights for the i-th datum in the test set. To analyze how the model distributes its focus across modalities depending on its prediction, we separate the test set into two groups: samples the model predicted as fake news and samples it predicted as real news. For each group, we collect the modality-wise attention weights across all samples.

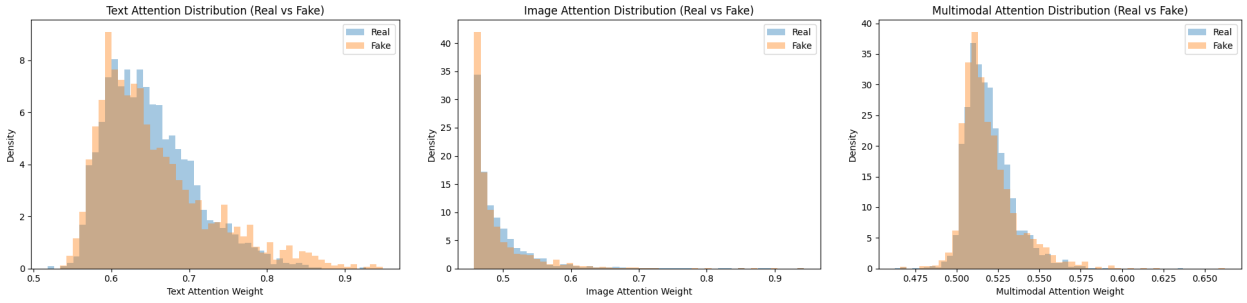
We then plot the empirical density of the attention weights for each modality (text, image, multimodal) separately for the fake-news predictions and the real-news predictions. Specifically, we use Python's `matplotlib.pyplot.hist` with `density=True` to normalize the counts, producing an estimate of the probability density function for each modality. Figure 2 shows these estimated density functions, clearly illustrating how the model distributes attention across text, image, and multimodal inputs for samples it predicted as fake versus samples it predicted as real.

7.1.2. *QKV Attention Analysis.* For the QKV multi-head attention ablation, we analyze the attention weights produced when treating the text, image, and multimodal representations as three tokens in a single attention sequence. For each test sample and each attention head, the model outputs a 3×3 attention matrix, where each entry represents the attention assigned from one modality token to another. To obtain a global view of the model's cross-modal behavior, we compute the expected attention weights separately over the test set for samples predicted as real and fake and by averaging over all attention heads:

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{test}}^{(c)},\, h}\left[\mathbf{A}^{(h)}(x)\right]$$

where $c \in \{\text{real}, \text{fake}\}$, $\mathcal{D}_{\text{test}}^{(c)}$ denotes the set of test samples predicted as class c, and $\mathbf{A}^{(h)}(x) \in \mathbb{R}^{3\times 3}$ is the attention matrix produced by head h for test datum x. This results in two 3×3 matrices summarizing the average cross-modal attention patterns for samples predicted as real and fake.

Figure 3 visualizes these expected attention matrices. Each heatmap entry indicates the average attention weight from a query modality to a key modality, providing insight into how the model distributes attention across modalities, under the QKV attention formulation, for samples predicted as real and fake respectively.

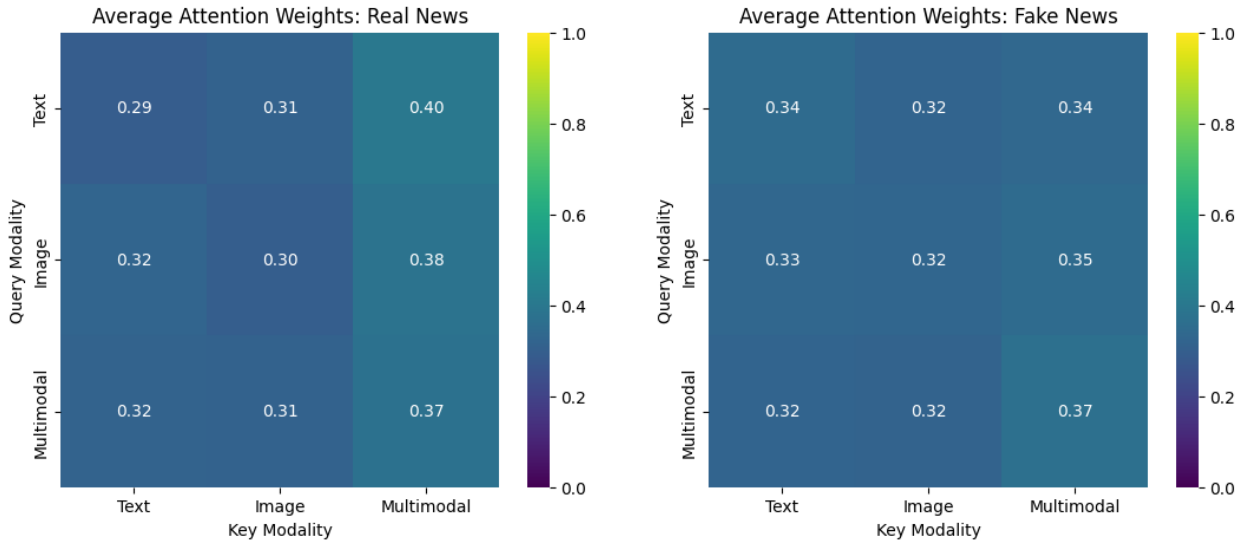


(A) Text Attention Weights (B) Image Attention Weights (C) Multimodal Attention Weights

FIGURE 2. Modality-Wise Attention: Distribution of Attention Weights across modalities.

7.2. **Cosine Similarities Between CLIP Embeddings.**

In this section, we analyze the cosine similarities between CLIP text and image embeddings on our training dataset. For each training sample, we took the text of the Reddit post and passed it through the CLIP text encoder to obtain $f_{\text{CLIP-T}}$, and similarly passed the associated image through the CLIP image encoder to obtain $f_{\text{CLIP-I}}$. We then computed the cosine similarity between these embeddings. Figure 4 shows the resulting distribution of these similarities across all training samples.



(A) QKV Attention: Average Attention Heatmap for Predicted Real News Data Samples

(B) QKV Attention: Average Attention Heatmap for Predicted Fake News Data Samples

FIGURE 3. QKV Attention: Average Attention Heatmaps

This analysis allows us to assess the degree of alignment between text and image modalities in the CLIP embedding space. Higher cosine similarity values indicate that the text and image are semantically aligned according to CLIP, while lower values suggest weaker alignment. Overall, this provides insight into the extent to which the model can leverage cross-modal information

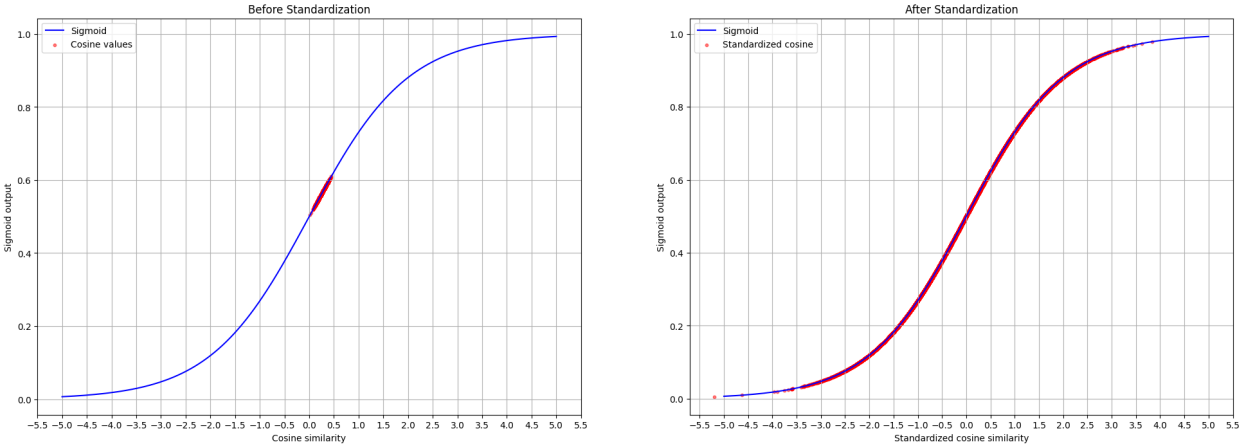


FIGURE 4. Effect of Standardization on Sigmoid-Transformed Cosine Similarities between CLIP Text and Image Embeddings.

Implementation Details: As previously discussed and illustrated in Figure 4, standardizing the cosine similarity values (s) in Equation (5) is essential for enabling the model to effectively distinguish between semantically aligned and unaligned text–image pairs.

During training, the cosine similarity values s within each mini-batch are standardized using the batch mean and variance computed over s, denoted by μ_s and σ_s^2, respectively. Simultaneously, we maintain running estimates of the mean and variance of s using running averages, analogous to Batch Normalization. At training step $t+1$, the running

statistics of s are updated as

$$\text{running\_mean}_s^{(t+1)} = (1-\rho)\cdot \text{running\_mean}_s^{(t)} + (\rho)\cdot \mu_s \tag{8}$$

$$\text{running\_var}_s^{(t+1)} = (1-\rho)\cdot \text{running\_var}_s^{(t)} + (\rho)\cdot \sigma_s^2 \tag{9}$$

where μ_s and σ_s^2 are computed from the current mini-batch of cosine similarity values s, and the momentum parameter is set to $\rho = 0.1$

During inference, each cosine similarity value s is standardized using the running mean and variance buffers

7.3. **CLIP-Based Models: Training Setup and Details.** All CLIP-based models were trained for 3 epochs using the `AdamW` optimizer with a learning rate of `5e-5` and a batch size of 32. To account for the slight class imbalance in our dataset (approximately 45%-55%), we use a **weighted cross-entropy loss**. Specifically, for a datum x_i with ground-truth label $y_i \in \{0, 1\}$, let $\hat{p}_{y_i}$ denote the model's predicted posterior probability assigned to the ground truth class y_i for the given datum. The **weighted cross-entropy loss** over a dataset of N samples is defined as

$$\mathcal{L}_{\text{dataset}} = -\frac{1}{\sum_{i=1}^{N} w_{y_i}} \sum_{i=1}^{N} w_{y_i} \log(\hat{p}_{y_i})$$

where w_y is the weight assigned to class y, set equal to the proportion of the *other* class:

$$w_0 \approx 0.55, \quad w_1 \approx 0.45$$

Since the class imbalance is slight, the weights are close to 0.5, so the weighted cross-entropy acts as a mild adjustment to the standard cross-entropy loss. This helps reduce bias toward the majority class while maintaining stable training.

Following the recommendations of [3] and due to computational constraints, during training we freeze the weights of both BERT (`bert-base-uncased`) and CLIP (`openai/clip-vit-base-patch32`), allowing only the ResNet backbone (`ResNet-101`) to be fine-tuned. The input images to ResNet are resized to 224×224. Projection heads for each modality consist of two fully connected layers with output dimensions 256 and 64, respectively, mapping the extracted features into a shared embedding space. Each linear layer is followed by a ReLU activation and batch normalization, with dropout applied at a rate of 0.2 for regularization.

In the Modality Wise Attention module, we pass $w_{\texttt{init}}$ into two FC layers with output dimension of 3. Each FC layer is followed by a `GELU` activation. At the end, a sigmoid normalization is applied to return the final modality-attention weights

The final classification head consists of two fully connected layers with dimensions 64 and 2, where the first layer is followed by a ReLU activation and the second layer produces the unnormalized logits, which are converted into predicted class probabilities via a softmax operation. This training strategy enables efficient optimization by focusing updates on the ResNet backbone while leveraging the rich, pre-trained representations from BERT and CLIP, keeping computational costs manageable.

Figures 5 and 6 denote the training and validation loss curves for the CLIP-based models with modality-wise attention and QKV attention respectively. To reduce noise and make the overall trend clearer, the training loss curves were smoothed using a moving average with a window size of 10. This smoothing helps to better visualize the convergence behavior during training. Both plots indicate that training converged, with no signs of overfitting.

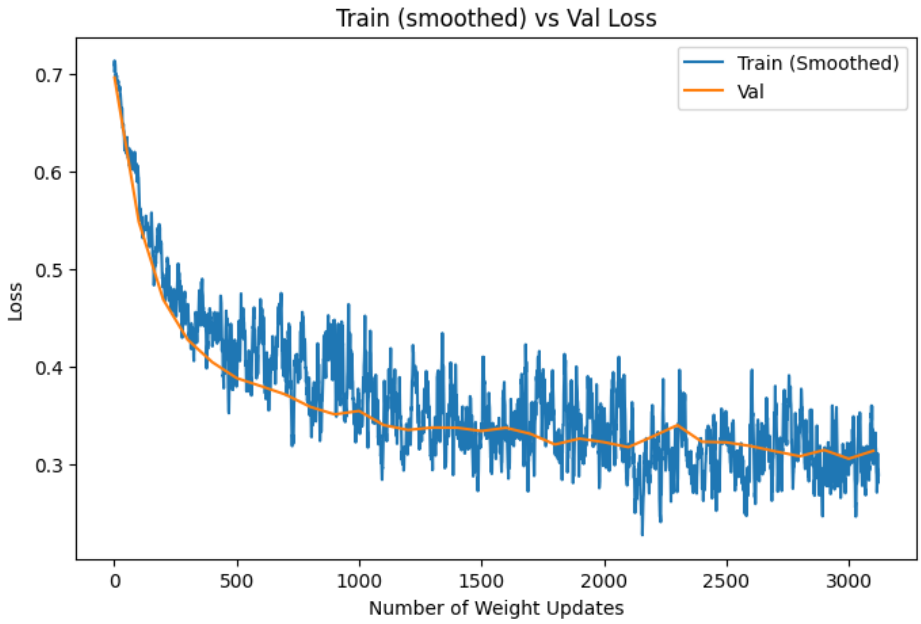


FIGURE 5. Training vs Validation Loss Curves for CLIP-Based Model with Modality-Wise Attention

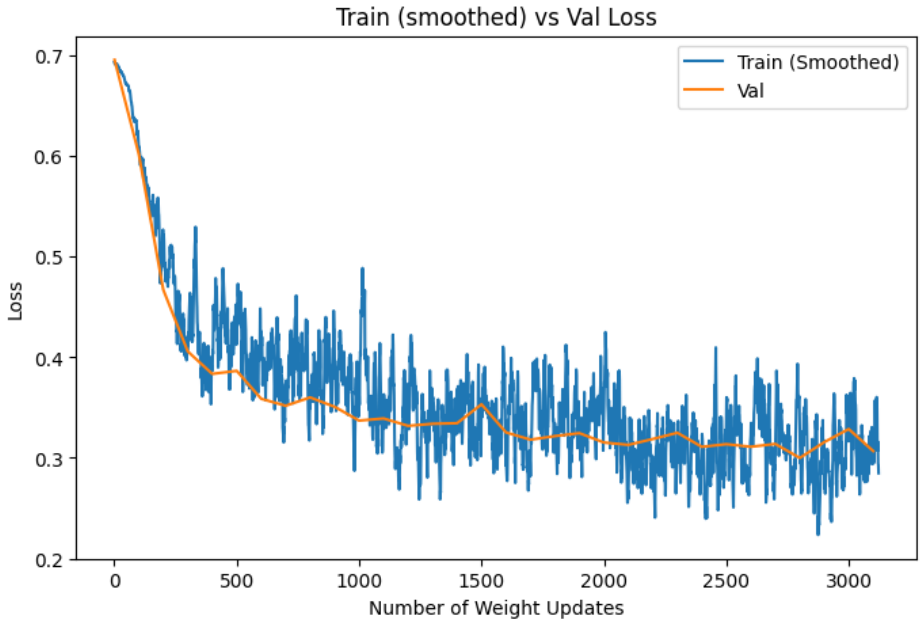


FIGURE 6. Training vs Validation Loss Curves for CLIP-Based Model with QKV Attention