Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.065872

ARTICLE

Optimizing Sentiment Integration in Image Captioning Using Transformer-Based Fusion Strategies

Komal Rani Narejo[1], Hongying Zan[1,*], Kheem Parkash Dharmani[2], Orken Mamyrbayev[3,*], Ainur Akhmediyarova[4], Zhibek Alibiyeva[4] and Janna Alimkulova[5]

[1]School of Computer and Artificial Intelligence, Zhengzhou University, Zhengzhou, 450001, China
[2]School of Computing, National University of Computer and Emerging Sciences, Islamabad, 04403, Pakistan
[3]Institute of Information and Computational Technologies, Almaty, 050010, Kazakhstan
[4]Institute of Automation and Information Technologies, Satbayev University, Almaty, 050013, Kazakhstan
[5]Turan University, Chaikina St 12a, Almaty, 050020, Kazakhstan

*Corresponding Authors: Hongying Zan. Email: iehyzan@zzu.ed.cn; Orken Mamyrbayev. Email: morkenj@mail.ru

Received: 24 March 2025; Accepted: 09 May 2025; Published: 03 July 2025

ABSTRACT: While automatic image captioning systems have made notable progress in the past few years, generating captions that fully convey sentiment remains a considerable challenge. Although existing models achieve strong performance in visual recognition and factual description, they often fail to account for the emotional context that is naturally present in human-generated captions. To address this gap, we propose the Sentiment-Driven Caption Generator (SDCG), which combines transformer-based visual and textual processing with multi-level fusion. RoBERTa is used for extracting sentiment from textual input, while visual features are handled by the Vision Transformer (ViT). These features are fused using several fusion approaches, including Concatenation, Attention, Visual-Sentiment Co-Attention (VSCA), and Cross-Attention. Our experiments demonstrate that SDCG significantly outperforms baseline models such as the Generalized Image Transformer (GIT), which achieves 82.01%, and Bootstrapping Language-Image Pre-training (BLIP), which achieves 83.07%, in sentiment accuracy. While SDCG achieves 94.52% sentiment accuracy and improves scores in BLEU and ROUGE-L, the model demonstrates clear advantages. More importantly, the captions are more natural, as they incorporate emotional cues and contextual awareness, making them resemble those written by a human.

KEYWORDS: Image-captioning; sentiment analysis; deep learning; fusion methods

1 Introduction

Image captioning stands out as a significant and fast-evolving research focus in computer vision [1], enabling the automatic generation of textual descriptions from images [2]. Its applications include aiding visually impaired individuals, improving content-based search engines, and organizing media databases [3,4]. As the volume of content in the digital environment increases, understanding not just the topic but also the emotions linked to it becomes more important. Despite breakthroughs in image captioning methods focusing on object recognition and scene depiction [5,6], often fail to capture sentiment, limiting interaction with visual content [7–9]. This limitation is especially significant in areas like social media, where personalized content drives user engagement, and mental health evaluation, where emotional context plays a key role in enhancing diagnostic tools [10–12]. The challenge of integrating sentiment into captions has been largely overlooked [13–16]. Current methods process both natural language and visual

data to generate captions expressing emotional content [17,18]. These models use neural network architectures such as CNNs for visual feature extraction [19,20] and attention mechanisms for generating textual descriptions [21]. Pretrained models like Inception, Xception, DenseNet, and EfficientNet help in building descriptive emotional captions [22–26]. For image captions to accurately convey sentiment, more advanced fusion techniques are needed to effectively integrate visual and sentiment embedding [27,28]. However, such techniques often struggle with imbalances in sentiment classes (e.g., positive, neutral, and negative), leading to suboptimal performance [29]. Furthermore, while multimodal fusion strategies using attention mechanisms have shown promise in improving sentiment alignment, fine-grained sentiment analysis still remains a challenge [30–34]. These challenges are addressed through in-depth experiments as part of our research, using the Emo-At-Cap dataset [35]. We obtain textual sentiment embeddings with the RoBERTa model [36] and extract visual features using ViT [37]. A custom transformer model is designed using these embeddings, employing fusion methods such as attention fusion, concatenation, cross-attention, VSCA, and our proposed model, SDCG, to seamlessly combine sentiment and visual embeddings. This research makes three key contributions: First, we introduce fusion strategies based on a custom transformer architecture that outperform traditional methods by more effectively integrating sentiment cues with visual information for image captioning. Second, we perform an extensive comparison of the proposed SDCG framework with two benchmark models, GIT [38] and BLIP [39], evaluating both the quality of generated captions and their impact on sentiment accuracy. Third, we combine Vision Transformer (ViT)-derived visual representations with RoBERTa textual embeddings, resulting in improved sentiment-aware caption accuracy. The remainder of the paper is structured as follows: Section 2 introduces prior research, Section 3 elaborates on our methodology, Section 4 describes the experimental setup and how evaluation was performed, Section 5 provides results with corresponding analysis and a discussion of limitations, and Section 6 offers concluding remarks and prospects for future study.

2 Related Work

2.1 Early Techniques in Image Captioning

Image captioning has progressed considerably over time, moving from simple neural network techniques to more complex models that incorporate sentiment analysis. One of the early models combined CNNs and RNNs to create captions and used an LSTM-based decoder setup to improve on the basic encoder–decoder design [40]. Similarly, reference [41] used convolutional neural networks to extract visual information from images, while recurrent networks with LSTM layers were responsible for forming the captions. These models helped show that when larger models are trained from start to finish, they tend to produce captions that are not only accurate but also better at reflecting the overall context of the image. They showed promising results, especially when tested on the COCO dataset, proving their effectiveness in generating coherent descriptions [42]. As deep learning technology advanced, more refined models built upon this encoder-decoder architecture began to emerge [43,44].

2.2 The Role of Sentiment in Image Descriptions

It became evident that simple image descriptions were insufficient, as they overlooked the emotional and personal nuances of language. To address this, reference [45] suggested concentrating on the relationships and positioning of objects in images. This approach improved the captions by considering the size and spatial connections of objects. Focusing on the need to grasp object interactions for richer captions, significant progress was made especially in evaluation metrics such as CIDEr-D and SPICE. A prime example is the Senti-Transformer model [46], whose multimodal transformer encoder links each description to the mood the image conveys, giving captions greater emotional weight. The model was trained in two steps.

Initially, it used cross-entropy loss to learn the basics, and then it was fine-tuned with full supervision to improve its overall performance. This setup helped the system become better at picking up on the emotional tone in images, leading to stronger results in both BLEU and CIDEr scores. It also performed reliably on benchmarks like MSR-VTT and COCO. Similarly, reference [47] introduced the In Senti-Cap architecture, which integrates CNN and LSTM modules to detect dominant sentiments in images. This approach, especially useful for visually impaired users and social media applications, coordinates sentiment regularization with traditional supervised learning, improving the alignment of captions with images conveying strong emotions.

2.3 Attention Mechanisms and Multimodal Fusion

Recent advances in sentiment-aware image captioning now use attention modules to more closely synchronize visual cues with their textual descriptions. Reference [48] proposed the ICAM model, which addresses challenges in multimodal sentiment analysis by using a Joint Attention mechanism for image captioning. To resolve issues like poor image-text alignment and adaptive attention, the model applies a Sentence-Image Cross Attention technique, allowing for dynamic adjustment of focus between the image and text. This technique integrates CATR and CBAM to improve feature refinement before caption generation. The combination of attention mechanisms and data has had a profound effect on both the linguistic and visual components of image captioning. In a similar vein, reference [49] applied Cross-on-Cross Attention (CoCA) and the Global Cross Encoder (GCE) to improve the model's focus on both visual and text features, especially on the MS-COCO dataset. They used layered attention to better align images with their descriptions. Additionally, they assessed a transformer architecture based on deep fusion (DFT) to evaluate how effectively it blends visual cues with semantic context during both encoding and decoding.

2.4 Task Adaptive Attention and Fine-Grained Sentiment Analysis

The Task Adaptive Attention (TAA) module combines non-visual and visual input through the application of task-specific vectors developed in transformer-based models. Reference [50] emphasized the importance of attention tuning in sentiment classification and caption accuracy, leading to significant improvements in BLEU, CIDEr, and ROUGE scores. Within few-shot learning, reference [51] introduced MATANet (Multi-scale Adaptive Task Attention Network) to overcome the shortcomings of earlier techniques. MATANet first produces feature maps at several spatial resolutions, then applies an adaptive task-attention module that zeroes in on the most informative local cues for a given task. This combination has proven highly effective when only a handful of training examples are available.

2.5 Fusion Methods in Multimodal Sentiment Analysis

Reference [52] introduced SENTI-ATTEND, an innovative framework that applies a dual attention mechanism to merge sentiment information at both the high-level and word levels. This model enhances caption generation by focusing on the layers of the image that carry the strongest emotional content, ensuring that the resulting captions are both emotionally accurate and factually correct. In a similar approach, reference [53] offered two ways to weave emotion into a captioning network. The first method, Direct Injection, adds a sentiment signal at every step of the RNN, while the sentiment flow method disperses sentiment across particular units of the LSTM, allowing for greater control over the affective tone and better coordination of sentiment in captions. Recent studies on sentiment-based image captioning have expanded to include complex emotions, such as sarcasm, by utilizing multimodal data.

2.6 Emerging Applications: Sarcasm and Complex Emotions

Reference [54] uses OpenAI's CLIP model for recognizing sentiment in images by continuously processing both visual and textual information. Two customized models, Emotion Clip and CIFAR100 Clip, demonstrated the adaptability of CLIP in both object classification and emotional analysis. Also, reference [55] proposed a different pipeline for recognizing sarcasm. Their system couples a residual CNN that adapts its spatial attention to pull out visual cues with a cross-lingual language model that interprets the accompanying text, while a Transformer generates descriptive captions to round out the context. By merging insights from all three streams, the model can more reliably pick up sarcastic intent in posts that mix images and words. Emerging image captioning models are focusing on enhancing contextual understanding and combining various modalities to improve the accuracy of the generated captions.

2.7 Advances in Hierarchical Attention Networks and Multimodal Fusion

With the goal of improving caption accuracy, reference [56] proposed HAN, a hierarchical attention-based model that extracts features from object detection systems, OCR-derived text, and patch-level visual data. The model adaptively adjusts feature importance through the pMRM, while a context gate ensures coherent interaction among these inputs. It also includes a histogram derived from the Multivariate Residual Module (MRM), which helps the model better understand the relationships between different elements in context. This type of design supports the main goals of SDCG by bringing together various types of data in a way that allows for accurate sentiment recognition without losing contextual meaning. To address the challenge of generating captions for images that contain embedded text, reference [57] introduced a model tested on the TextCaps dataset. By applying CLIP, the model captures robust features from both the image and OCR outputs, which are passed through multiple layers of attention. A decoder was implemented to select between fixed vocabulary tokens and detected text, demonstrating the model's strength in producing captions that account for the written content within images.

2.8 Challenges in Sentiment Prediction and Image Captioning

Accurate sentiment prediction and emotionally expressive captions depend to a large extent on effective use of visual attention [58]. Transformer-based models [59,60] use attention mechanisms to highlight emotionally relevant features in images, which can improve the quality of the caption, especially in emotionally nuanced scenarios. Despite this, matching the emotional tone of a caption with the degree of visual detail an application requires is not always straightforward. When captions lack this balance, the system's ability to recognize emotion can suffer, often because many models either overlook affective signals or fail to express them in a meaningful way [61,62].

3 Methodology

This section begins with an overview of the dataset, followed by an explanation of how we addressed the class imbalance problem using oversampling. Subsequently, we explain the extraction of textual and visual sentiment embeddings. The methodology also includes a custom transformer-based architecture for combining these embeddings with the help of advanced fusion techniques.

3.1 Dataset Overview

In this study, we utilized the Emo-At-Cap dataset [35], which includes 3840 images accompanied by human-generated captions. The sentiment of each image is classified into one of three categories: positive, negative, or neutral.

Fig. 1 illustrates how the language of a caption can impact its affective tone. Referring to laughter or euphoria, such as "The two laugh at a joke" or "Three friends leap with elation," evokes positivity. Referring to distress or violence, such as "The woman agonizes over her condition" or "A man tightly holds another," evokes negation. Captioning that simply states what is going on, such as 'Two men stand facing a third' or 'A man looks after a woman,' is neutral. The above examples shows that a captioning model can always add the underlying emotion to every description. The Emo-At-Cap dataset serves as an excellent choice because it contains a distinctive multimodal structure consisting of image-caption pairs along with sentiment annotations. The equal distribution of the dataset between image content and sentiment annotation labels made it perfect for conducting research into sentiment-based image captioning.

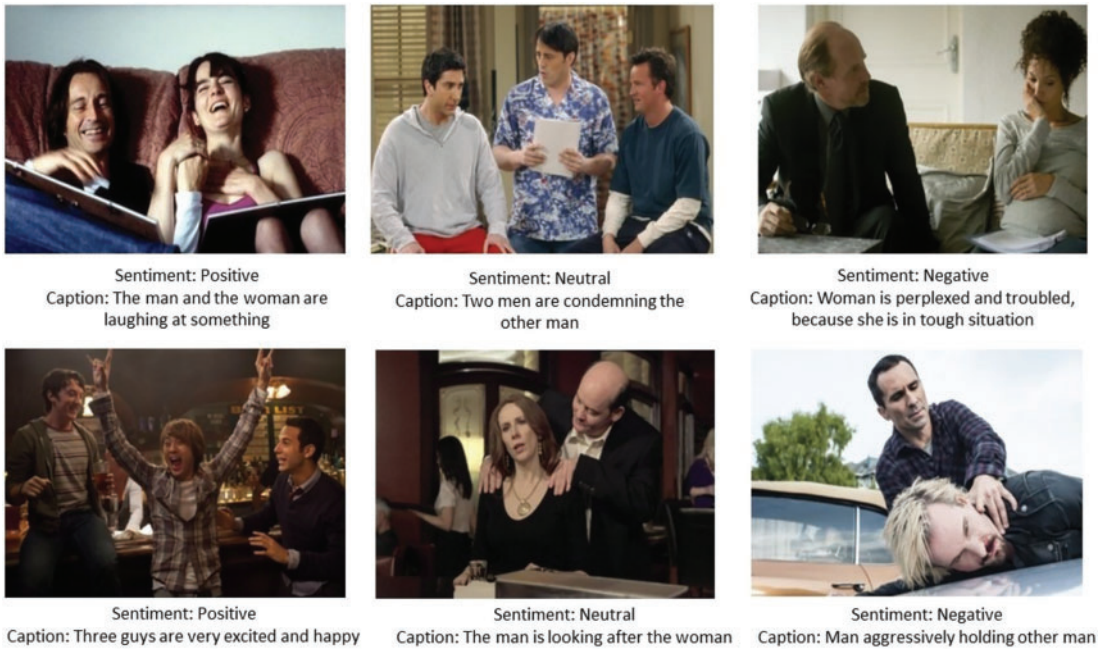

Figure 1: Sentiment-aware image captioning examples from the Emo-At-Cap dataset

Class Distribution and Addressing Imbalance

An unbalanced distribution between classes tends to create performance degradation because the model prefers majority category classifications, thus impairing caption generation accuracy for minority sentiments. The Random Over-Sampling technique [63] helps balance class distributions through its instantiation in the RandomOverSampler class of the imblearn library. This technique duplicates minority class samples multiple times to achieve a more appropriate class distribution. Oversampling was applied exclusively to the training data, increasing the number of samples per sentiment class from 3840 to 6111. We kept the validation and test data separate from the training process to reduce the risk of data leakage and ensure that the results reflect real-world performance.

Fig. 2 highlights the improved balance between sentiment classes. This adjustment reduced the risk of the model favoring one class over the another and helped it treat each sentiment category more equally during training.

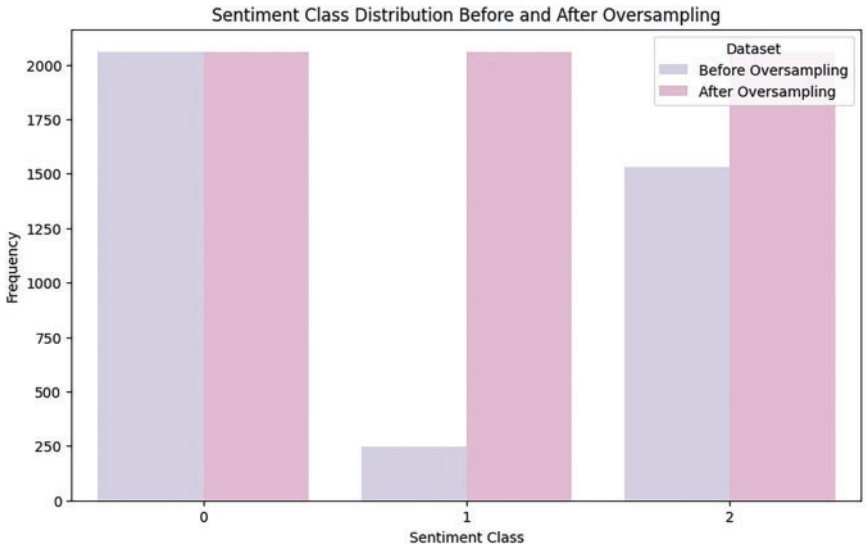

Figure 2: The sentiment class distribution before and after oversampling

3.2 Textual and Visual Sentiment Embeddings

Sentiment analysis often involves understanding input data across different modalities, such as text and images. RoBERTa and ViT both leverage transformer-based architectures to generate embeddings, but they process different types of input data: RoBERTa handles text, while ViT processes images. Despite sharing the transformer backbone, the encoding methods for text and images differ.

3.2.1 Textual Embedding Extraction (RoBERTa)

RoBERTa [36] processes the input text by using Byte Pair Encoding (BPE) for tokenization and includes special tokens to mark the boundaries of sequences. Each token is converted into an embedding vector and passed through several layers of a transformer model. The final hidden state of the [CLS] token is then extracted and used for sentiment classification. A weighted cross-entropy loss function is employed to handle class imbalance. The architecture of RoBERTa, shown in Fig. 3, is based on an encoder structure with multiple transformer layers. To assess the performance of RoBERTa's embeddings, t-SNE is employed to map the high-dimensional feature space to two dimensions. The chart in Fig. 4 reveals separate clusters for each sentiment, with neutral sentiment grouping closely together. While there is some overlap between the positive and negative sentiment representations, this is anticipated due to the subtle emotional distinctions in the text. This overlap, however, does not have a significant impact on the model's overall performance.

3.2.2 Visual Embedding Extraction (ViT)

In ViT [37], an image is first broken down into smaller patches. Each patch is then turned into a high-dimensional vector. These vectors pass through several transformer layers, where the [CLS] token gathers insights from all the patches. This combined information forms a complete image representation that helps with sentiment classification.

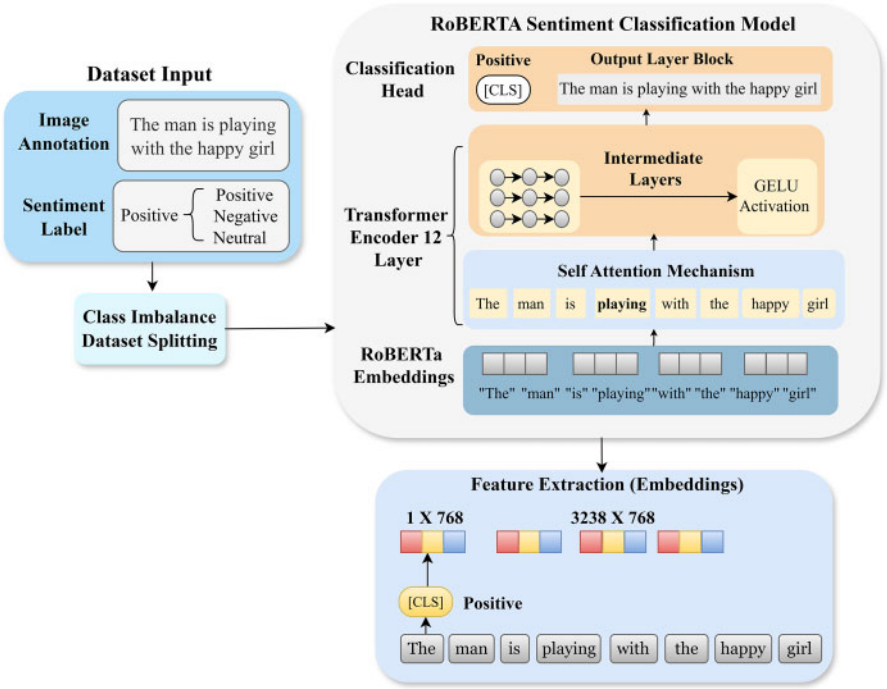

Figure 3: Sentiment-aware embedding extraction using RoBERTa architecture

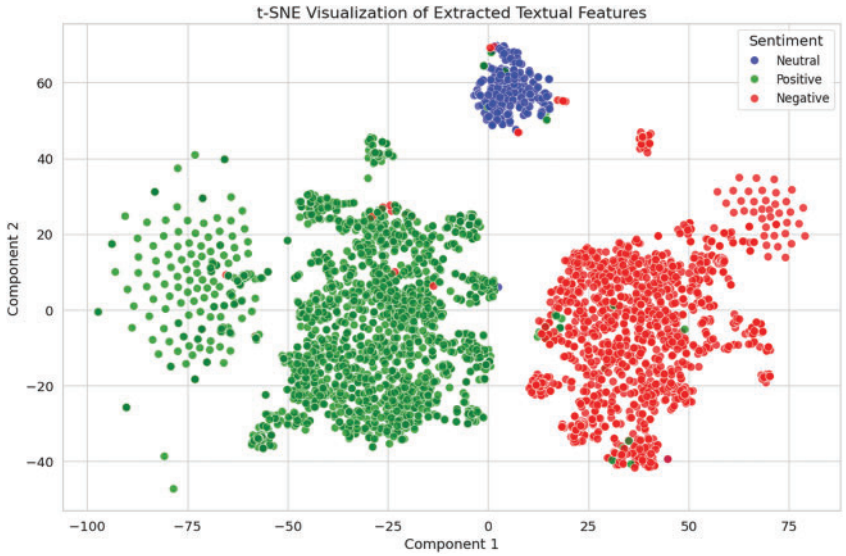

Figure 4: Visualization of textual embeddings clustered by sentiment categories using t-SNE

The diagram in Fig. 5 offers an in-depth view of the Vision Transformer architecture, demonstrating its process from image preprocessing to the final sentiment classification. The t-SNE visualization in Fig. 6 shows that neutral sentiment forms a compact cluster, indicating that the model effectively extracts sentiment features. While separating positive and negative sentiments remains somewhat tricky, the model performs well in recognizing and extracting sentiment details from visual content.

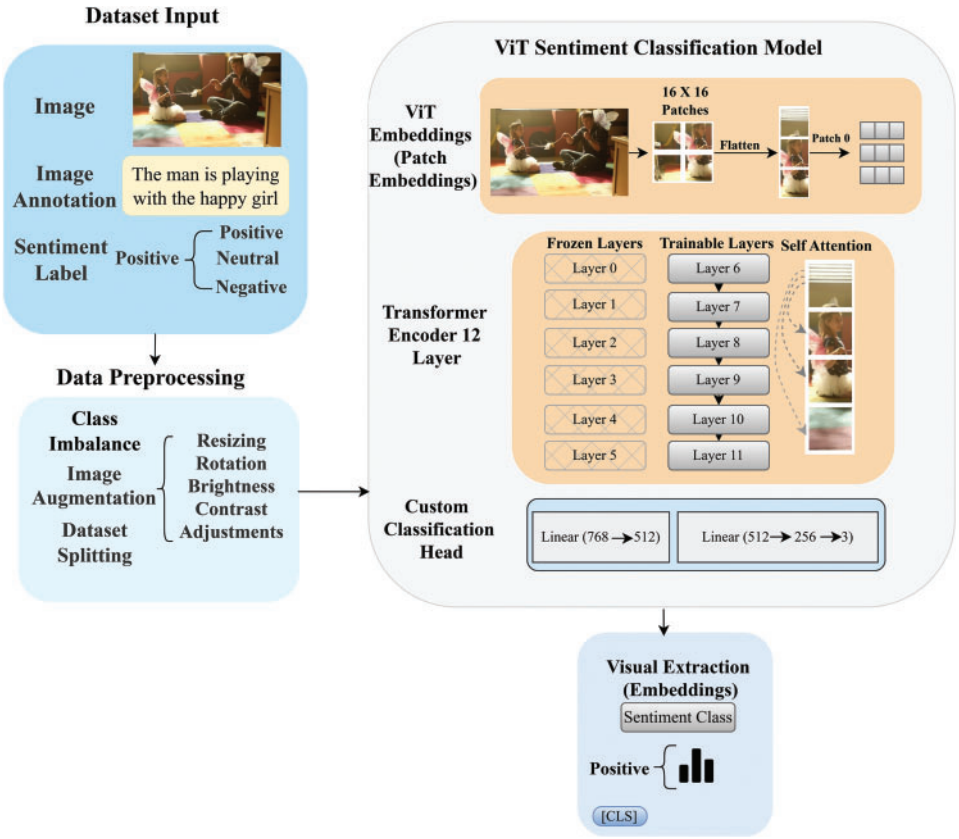

Figure 5: Visualization of the Vision Transformer (ViT) architecture

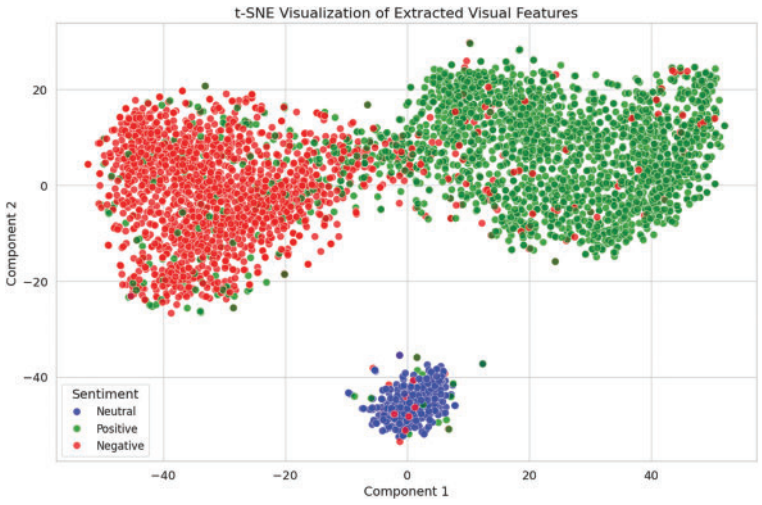

Figure 6: Visualization of visual embeddings clustered by sentiment categories using t-SNE

3.3 Fusion Techniques Leveraging Transformers

The proposed framework employs a Transformer-based decoder with novel fusion strategies to align and integrate textual and visual embeddings for sentiment-aware caption generation. These fusion

techniques enhance multimodal learning by capturing long dependencies and contextual details through self-attention mechanisms. Key fusion strategies include concatenation fusion, attention fusion, VSCA, cross-attention transformer and SDCG fusion. Each technique contributes uniquely to generating sentiment-aligned captions.

3.3.1 Concatenation Fusion

Concatenation fusion is computationally efficient and preserves the independence of low-level visual and textual features. The textual embedding h_{CLS} and the visual embedding v are projected into a shared D-dimensional space:

$$v' = W_v v + b_v, \quad t' = W_t h_{CLS} + b_t \tag{1}$$

where W_v and W_t are learnable projection matrices, and b_v and b_t are bias terms. The projected embeddings v' and t' are then concatenated into a single feature vector:

$$F_C = [t'; v'] \tag{2}$$

Transformers require this step because they fail to naturally grasp the order of input tokens. Each token receives positional encoding which contains information about its sequence position to allow the model to differentiate input tokens by their positions. Positional encoding refers to the method defined for the i-th token is defined as:

$$E(t_i) = t'_i + p_i \tag{3}$$

The symbol t'_i is the i-th projected textual embedding, and p_i provides its sequence position for the i-th token in the sequence. Through this addition, the model can determine word order relations in sequences during text processing. The fused embeddings become the input to the transformer encoder module before the attention mechanism operation. The computations for query Q, key K, and value V matrices are computed from the fused feature vector F_C using the following expressions:

$$Q = F_C W_Q, \quad K = F_C W_K, \quad V = F_C W_V \tag{4}$$

The calculation of attention mechanism follows this form:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{5}$$

The logits required for caption generation are calculated:

$$L = W_f F_C + b_f \tag{6}$$

The caption prediction depends on visual and textual embedding fusion which uses positional encoding to maintain textual sequence order from text input.

3.3.2 Attention Fusion

Attention fusion uses a dynamic attention mechanism to align textual and visual features. The textual embedding t' interacts with the visual embedding v' through text-to-visual and visual-to-textual attention

mechanisms:

$$\text{Attention}_{tv} = \text{Softmax}\left(\frac{Q_T K_V^T}{\sqrt{d_k}}\right) V_v \tag{7}$$

$$\text{Attention}_{vt} = \text{Softmax}\left(\frac{Q_v K_t^T}{\sqrt{d_k}}\right) V_t \tag{8}$$

The attention outputs get combined into a single consolidated representation:

$$F_A = [\text{Attention}_{tv}; \text{Attention}_{vt}] \tag{9}$$

Lastly, the combined attention output is mapped to the final latent space for generating captions.

$$L = W_f F_A + b_f \tag{10}$$

3.3.3 Visual-Sentiment Co-Attention Mechanism (VSCA)

The VSCA connects visual and textual embeddings through a two-way attention system, aiming to create captions that are more sensitive to sentiment.

$$\text{Attention}_{VS} = \text{Softmax}\left(\frac{Q_t K_v^T}{\sqrt{d_k}}\right) V_v \tag{11}$$

$$\text{Attention}_{SV} = \text{Softmax}\left(\frac{Q_v K_t^T}{\sqrt{d_k}}\right) V_t \tag{12}$$

The results from these bidirectional co-attention mechanisms are combined and normalized in the following manner:

$$F_{CA} = \text{LayerNorm}(\text{Attention}_{VS} + \text{Attention}_{SV}) \tag{13}$$

This stage is important for keeping calculations consistent and letting information pass smoothly both ways before it's merged into one final output. The final fused representation F_{CA} is used for sentiment-related caption generation.

$$L = W_f F_{CA} + b_f \tag{14}$$

3.3.4 Cross-Attention Transformers

The Cross-Attention Fusion mechanism integrates vision and sentiment embeddings to generate contextual captions. It utilizes cross-attention to dynamically align and merge features from multiple modalities. Text queries $Q_{t.}$ are applied to visual keys K_v, and the resulting attention weights highlight the most pertinent features.

$$\text{Attention}_{\text{Cross}} = \text{Softmax}\left(\frac{Q_{t.} K_v^T}{\sqrt{d_k}}\right) V_v \tag{15}$$

A unified representation is created by concatenating the output vectors from both attention systems.

$$F_{\text{CROSS}} = \text{LayerNorm}\left(\text{Attention}_{TV} + \text{Attention}_{VT}\right) \tag{16}$$

F_{CROSS} is used in the equation to blend the outputs from the (Attention$_{TV}$) and visual-to-textual attention (Attention$_{VT}$).

Once attention is fused, the resulting F_{CROSS} is projected into the latent space and used to derive logits.

$$L = W_f F_{\text{CROSS}} + b_f \tag{17}$$

In this phase, the learned weight matrix W_f and bias term b_f are applied to map the fused attention into the final feature space, facilitating the creation of the contextual caption.

3.3.5 Sentiment-Driven Caption Generator (SDCG)

In this research, we propose a transformer-based model called the SDCG fusion strategy, which combines visual and sentiment embeddings to generate captions that are both informative and in line with the sentiment of the image. The model uses a hierarchical fusion technique, along with multi-head attention and layer normalization, to improve the way visual and textual information work together. This framework helps the model create captions that are grammatically precise and that also pick up on the sentiment of the image. Fig. 7 presents the architecture of SDCG, describing how visual and textual data are integrated using hierarchical fusion and multi-head attention.

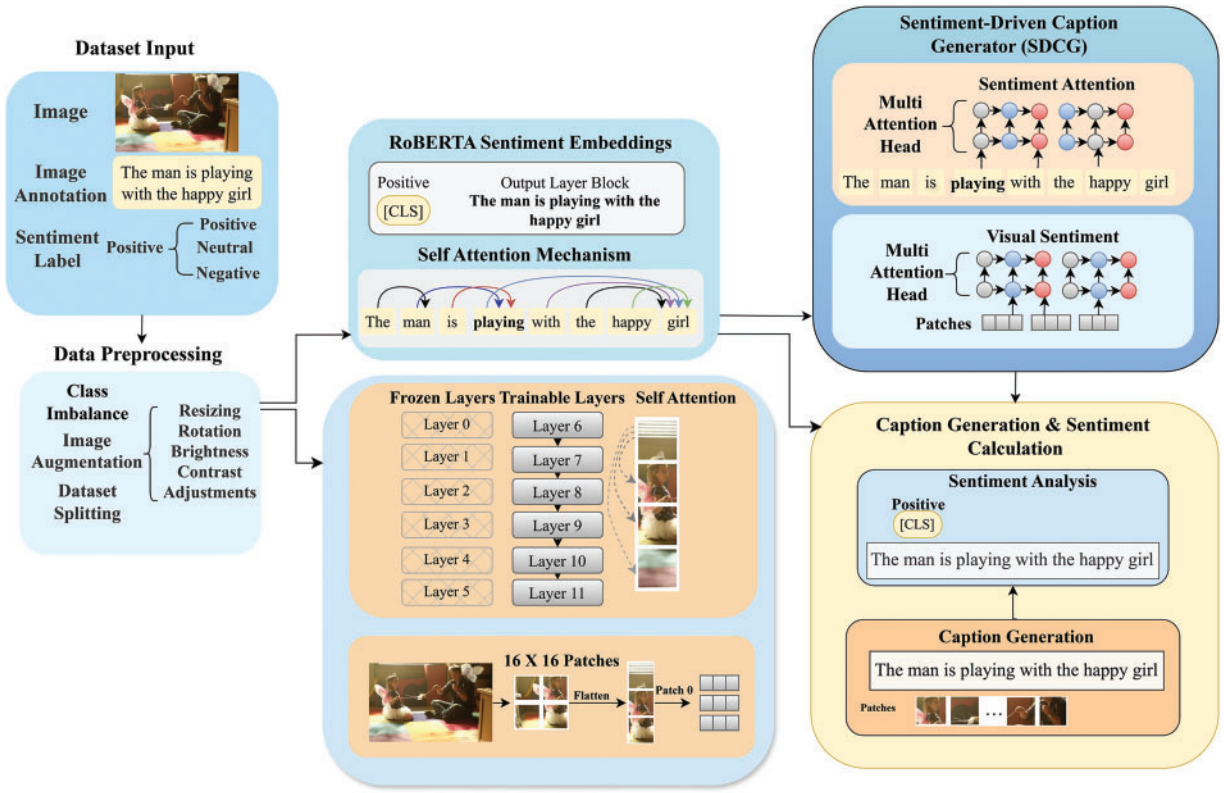

Figure 7: Sentiment-Driven Caption Generator (SDCG) architecture overview

The SDCG model starts by projecting textual embeddings (t') which undergo a projection process, allowing the model to set up attention between text and image. This step is important for understanding how sentiment and content align across the two. The model's design leans heavily on multi-head attention,

guiding textual embeddings to focus on the visual ones in the way described below. The SDCG model starts by projecting textual embeddings:

$$\text{Attention}_V = \text{MultiHead}(v', v', v') \tag{18}$$

The self-attention within multi-head attention results in the refined visual embeddings (v') performing an attention scan of themselves. Through the MultiHead function, the model develops the ability to focus simultaneously on multiple representation subspaces. The textual embeddings use the self-attention mechanism known as multi-head attention to process their elements.

Similarly, the textual embeddings attend to themselves through the multi-head attention mechanism:

$$\text{Attention}_T = \text{MultiHead}(t', t', t') \tag{19}$$

The multi-head attention mechanism works on the refined textual embeddings to yield Attention_T. The learned textual embeddings (t') use the same MultiHead function to self-attend during this operation. Results from visual and textual attention are combined. The results of the visual and textual attention operations are then concatenated:

$$F_{\text{SDCG}} = [\text{Attention}_V; \text{Attention}_T] \tag{20}$$

F_{SDCG} is the concatenation of the visual and textual attention outputs along a specific dimension. This fusion allows the model to integrate both modalities effectively.

The concatenated representation F_{SDCG} is then normalized using LayerNorm to eliminate numerical instability and ensure consistency:

$$F_{\text{SDCG}} = \text{LayerNorm}(F_{\text{SDCG}}) \tag{21}$$

Finally, the normalized representation is processed through a learnable weight matrix (W_f) and a bias vector (b_f) to produce the final logits:

$$L = W_f \cdot F_{\text{SDCG}} + b_f \tag{22}$$

L represents the output logits for caption generation, with W_f as a learnable weight matrix and b_f as a bias term that strengthens the bias vector on the normalized F_{SDCG}.

3.4 Baseline Models

To benchmark our fusion method, we compared it to GIT (Generative Image-to-Text Transformer) and BLIP (Bootstrapped Language-Image Pre-training) models known for top-quality captions. GIT excels at mapping visual content to natural text, and BLIP brings visual and linguistic data together seamlessly. We chose them for their reliability in producing coherent captions. The side-by-side results reveal both our model's advantages and its limitations.

3.4.1 Generative Image-to-Text Transformer (GIT)

We adapt the GIT [38] framework by integrating ViT for visual feature extraction and RoBERTa for sentiment-aware caption generation. In this architecture, the ViT encoder processes the input image, breaking it down into patches and encoding them into visual embeddings. The autoregressive decoder

generates captions token by token. At each step t, the probability of generating a token y_t is conditioned on the previously generated tokens $y_{1:t-1}$ and the visual embeddings v from ViT, as expressed by:

$$P(y_t \mid y_{1:t-1}, v) = \text{Decoder}(y_{1:t-1}, v) \tag{23}$$

After the caption is generated, RoBERTa is used to extract sentiment-aware embeddings from the [CLS] token. This enables sentiment alignment in the generated captions. We use this framework as a baseline because of its strong performance on benchmark datasets such as COCO, and its use of ViT (which is shared with our own architecture). The integration of RoBERTa allows for sentiment analysis directly in the generated captions, making it a relevant baseline for evaluating sentiment alignment.

3.4.2 Bootstrapping Language-Image Pre-Training (BLIP)

Similarly, we adapt the BLIP model [39], using ViT to extract visual features and RoBERTa to process the text for sentiment-aware analysis. The model uses ViT for visual feature extraction and a transformer-based encoder for text. In the cross-attention layer, visual queries Q_v attend to textual keys K_t and values V_t, as follows:

$$\text{Attention}_{\text{Cross}} = \text{Softmax}\left(\frac{Q_v K_t^T}{\sqrt{d_k}}\right) V_t \tag{24}$$

BLIP is trained with a hybrid loss function that combines captioning loss L_{caption} and contrastive loss $L_{\text{contrastive}}$:

$$L = L_{\text{caption}} + \lambda L_{\text{contrastive}} \tag{25}$$

This hybrid approach ensures strong alignment between the visual and textual modalities. We chose BLIP as a baseline because of its state-of-the-art performance in vision-language tasks and its ability to achieve robust cross-modal alignment.

4 Experimental Hyperparameter and Evaluation

This section of the paper discusses the experimental setup, while Table 1 presents the core hyperparameters for the different models. These configurations are essential for understanding the training process and have been carefully selected to optimize performance across various fusion models.

Table 1: Hyperparameter settings for model training

Parameter	Concatenation fusion	Attention fusion	Cross-attention fusion	VSCA	SDCG	BLIP	GIT
Optimizer	Adam	Adam	Adam	Adam	Adam	Adam	Adam
Step size	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001
Dropout rate	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Epochs	20	20	20	20	20	20	20
Batch size	32	32	32	32	32	16	16
Loss type	X-Entropy	X-Entropy	X-Entropy	X-Entropy	X-Entropy	X-Entropy	X-Entropy

(Continued)

Table 1 (continued)

Parameter	Concatenation fusion	Attention fusion	Cross-attention fusion	VSCA	SDCG	BLIP	GIT
Attention size	100	100	100	128	100	256	256
Text embedding	512	512	512	512	512	768	768
Visual embedding	768	768	768	768	768	768	768
Tokenizer	RoBERTa Tokenizer	RoBERTa Tokenizer	RoBERTa Tokenizer	RoBERTa Tokenizer	RoBERTa Tokenizer	BLIP Tokenizer	GIT Tokenizer
Captioning enabled	Yes	Yes	Yes	Yes	Yes	Yes	Yes

The key components of the model training and evaluation pipeline include batch size, learning rate, optimization strategy, and loss function, all of which are crucial for model convergence and generalization. In our evaluation, we benchmarked the captioning models with BLEU [64], ROUGE-L [65], and CIDEr [66]. BLEU gives us a quick count of overlapping n-grams with the reference, ROUGE-L measures recall by finding the longest shared subsequence, and CIDEr uses TF-IDF to reward those less common, more meaningful words. To make sure we weren't missing the emotional mark, we also calculated a sentiment-accuracy score, checking that each caption reliably hit its intended positive, negative, or neutral tone.

$$\text{BLEU}_n = \exp\left(\sum_{k=1}^{n} w_k \log p_k\right) \times \exp\left(\min\left(1 - \frac{c}{r}, 0\right)\right) \tag{26}$$

In this notation, n as the maximum n-gram order, w_k as the weight for each $\frac{1}{n}$, p_k refers to the modified n-gram precision, c is the generated caption's length, r indicates the effective size of the reference corpus, and other formulas for ROUGE-L and CIDEr follow similarly as shown in the initial setup.

$$\text{BLEU}_n = \exp\left(\sum_{k=1}^{n} w_k \log(p_k)\right) \times \exp\left(\min\left(1 - \frac{c}{r}, 0\right)\right) \tag{27}$$

Here, n stands for the top n-gram size, w_k is its usual weight $(1/n)$, p_k the adjusted n-gram precision, c indicates the limit of the generated caption, and and r denotes the effective corpus limit.

$$p_k = \frac{\sum_{c \in \text{candidate}} \min(\text{count}(c), \text{count}_{\text{ref}}(c))}{\sum_{c \in \text{candidate}} \text{count}(c)} \tag{28}$$

In this case, $\text{count}(c)$ measures how frequently an n-gram is found in the generated caption, and $\text{count}_{\text{ref}}(c)$ notes the highest number of times it appears in the reference captions.

$$\text{ROUGE-L} = F_\beta \times \frac{\text{LCS}(S, R)}{|R| + |S|} \tag{29}$$

For each caption, S be the generated caption and, R the true caption, $\mathrm{LCS}(S, R)$ finds how many elements they share in a row, and F_β gives a balanced mix of recall and precision.

$$F_\beta = \frac{(1+\beta^2) \times (\text{Precision} \times \text{Recall})}{\beta^2 \times \text{Precision} + \text{Recall}} \tag{30}$$

where β is typically set to 1.

$$\text{CIDEr}(S) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{|R|} \sum_{R \in R} \cos(\text{similarity}(S, R)) \tag{31}$$

In this metric, S is the produced caption, R the reference-caption collection, N the total produced captions, and $\cos(\text{similarity}(S, R))$ computes TF-IDF cosine similarity between S and R.

$$\text{TF-IDF}(t, D) = \text{TF}(t, D) \times \log\left(\frac{N}{\text{DF}(t)}\right) \tag{32}$$

Under this notation, t is the target term, D the document or caption, $\text{TF}(t, D)$ the count of occurrences of t in D, $\text{DF}(t)$ the number of reference captions where t appears, and N is the full set size of captions reference.

$$\text{Sentiment Accuracy} = \frac{\sum_{i=1}^{N} 1(y_i = \hat{y}_i)}{N} \tag{33}$$

Let N stands for the total samples number, y_i for the true label of the i-th sample, $\hat{y}_i$ for the predicted label, and i-th sample, and $1(\cdot)$ outputs 1 if the given condition is satisfied; otherwise, it returns 0.

5 Experimental Results

In this section, the results of the custom transformer model's performance in generating sentiment-aligned captions are presented. The results suggest that fusion techniques improve both the quality of the captions and the accuracy of the sentiment. Ablation studies comparing the proposed framework to baseline models demonstrate its superiority.

5.1 Performance of Models in Caption Generation

This research is built upon a transformer-based backbone. We evaluated five fusion techniques: concatenation fusion, attention fusion, cross-attention fusion, VSCA and SDCG. Evaluation metrics included BLEU, ROUGE-L, and CIDEr, as shown in Table 2. The visual sentiment co-attention model achieved a top CIDEr of 4.4273, while the sentiment-driven caption generator achieved 4.4211.

Table 2: Model performance on caption quality. Bold values indicate the best performance in each metric

Model	BLEU	ROUGE-L	CIDEr
Concatenation fusion	0.355	0.482	3.1025
Attention fusion	0.477	0.5101	3.6102
Cross-attention fusion	0.4912	0.5305	3.905
VSCA	0.5419	0.6045	4.4273
SDCG	**0.5382**	**0.6066**	**4.4211**

5.2 Sentiment Classification Results

Sentiment classification examines the performance of each model in matching the generated captions with the emotional tone of the image [67]. The highest sentiment accuracy was achieved by our proposed model, SDCG, at 94%, followed by the VSCA model at 93%, demonstrating their outstanding ability to capture sentiment nuances. The models that used fusion performed moderately: concatenation fusion was 85%, attention fusion was 86%, and cross-attention fusion was 87%, but they included a few misclassifications of positive, neutral, and negative sentiments. An exhaustive comparison of sentiment accuracy for all the models is provided in Table 3.

Table 3: Sentiment classification scores

Model	Accuracy rate
SDCG	**94%**
VSCA	93%
Cross-attention fusion	87%
Attention fusion	86%
Concatenation fusion	85%

In Fig. 8, each confusion matrix illustrates the performance of the models by showing correct predictions across the positive, neutral, and negative classes, as well as any misclassification errors. The matrices correspond to the following models: Concatenation Fusion 8(a), Attention Fusion 8(b), Cross-Attention Fusion 8(c), VSCA 8(d), and SDCG 8(e).

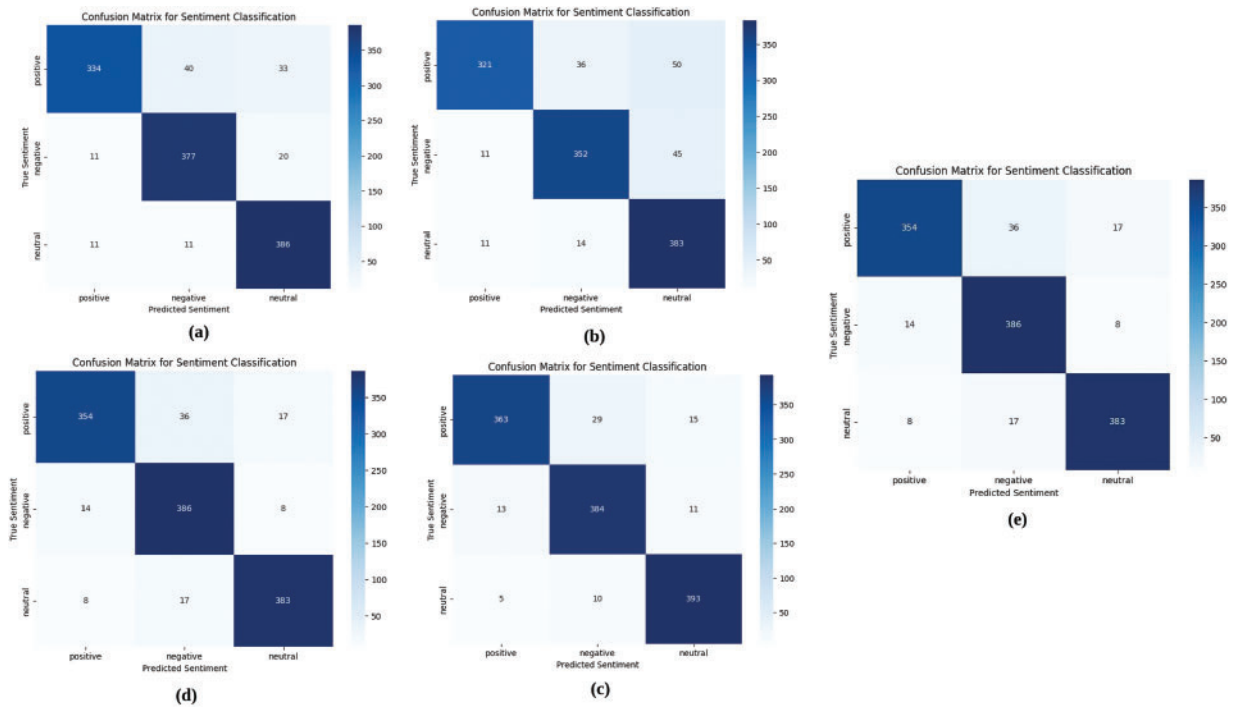

Figure 8: Sentiment classification evaluation using confusion matrix

5.3 Ablation Study

We evaluated the impact of fusion approaches on our custom transformer model and compared it with BLIP and GIT baseline models through an ablation study. The results are summarized in Table 4.

Table 4: Ablation study of baseline and enhanced transformer models with fusion techniques

Model	Fusion technique	CIDEr	BLEU	ROUGE-L
GIT (Baseline)	None	4.2461	0.5256	0.5536
BLIP (Baseline)	None	4.3458	0.5358	0.5654
Transformer variant	Concatenation fusion	3.5567	0.4892	0.5223
Transformer variant	Attention fusion	3.6482	0.4983	0.5350
Transformer variant	Cross-attention fusion	3.8546	0.5056	0.5425
Transformer variant	VSCA	4.0452	0.5154	0.5578
Transformer variant	SDCG	4.2467	0.5257	0.5658

5.4 Qualitative Observations

As can be seen in Fig. 9a, the caption generated, "Two people smile and clap happily," mirrors the scene's joyful tone and showcases the model's reliable detection of positivity in line with its strong classification metrics. as shown in Fig. 9b, for example "The family appears troubled and uncomfortable sitting together in the car," protrays a negative sentiment. Words such as "uncomfortable" and "troubled" match the expressions of the actors, highlighting the performance of the model to detect mood within complex environments. This accuracy is especially important for applications that depend on sentiment analysis, where handling negative sentiments is crucial, such as those that analyze people's mental states. Fig. 9c shows the caption 'Two cameramen capture an actor's performance on set.' It contains no negative language and provides a neutral description. This highlights the model's ability to capture subtle sentiment without exaggeration or distortion. The model shows effective performance when used with neutral situations that contain natural or unnoticed emotional expressions. However, it occasionally fails to identify subtle sentiment patterns, as demonstrated by its incorrect interpretation in Fig. 9d. In the original caption, 'The man is whispering something to the woman,' there is no specific emotional undertone. The model produces the caption, 'The man is whispering something romantic to the woman,' introducing excessive positive language that changes a neutral expression into something optimistic. This transformation from accurate to incorrect highlights both the strengths and weaknesses of the model, illustrating how sentiment alignment fails in this case. The artificial sentiment introduces romantic intentions even though the original staging provides no evidence for such romantic behavior. Fig. 9e further emphasizes contextual misinterpretation. The initial description indicates two people sitting/chatting peacefully, as shown in the picture. The model creates a flawed output that shows 'bored and dissatisfied' attitudes among both individuals, as indicated by the text, 'The two individuals are bored and dissatisfied.' The model misinterprets neutral visual cues as negative expressions, as it overrelies on such cues in this specific situation. This reveals limitations in the model's ability to recognize delicate feelings, resulting in incorrect sentiment readings. These instances demonstrate both the effective capabilities and key limitations of the SDCG model. The model shows strong results in sentiment identification but often misinterprets context when subtle emotional expressions occur.

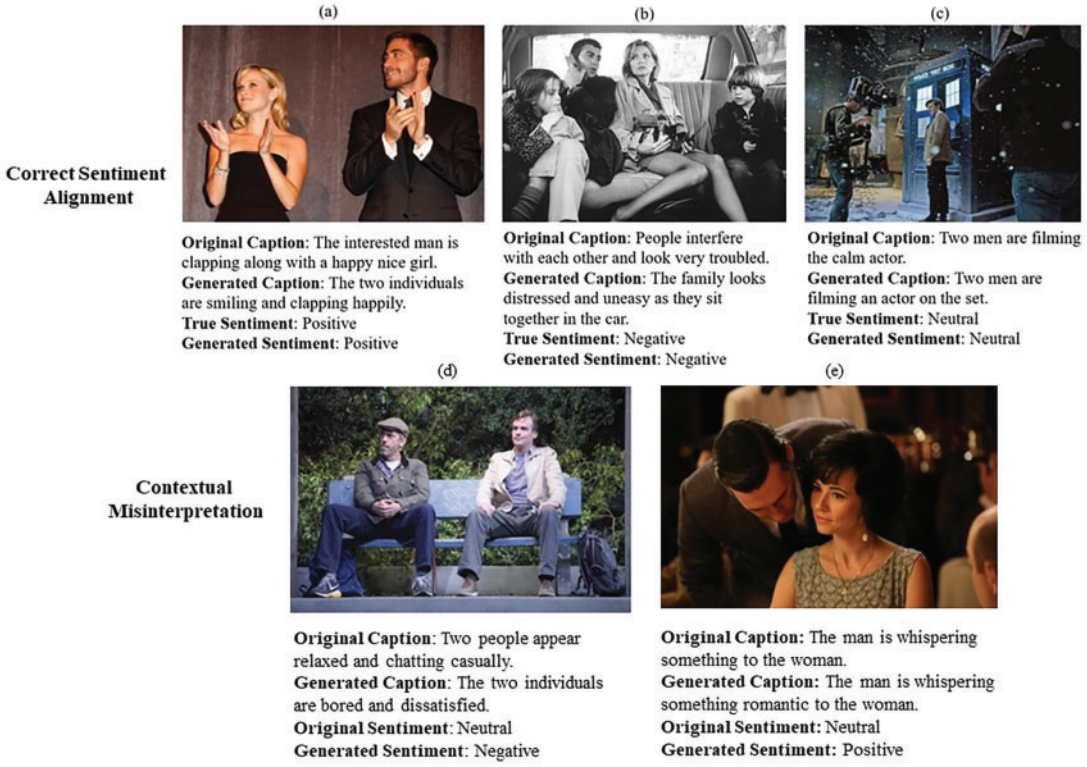

Figure 9: Generated caption illustrations: **(a)** Positive, **(b)** Negative, **(c)** Neutral, **(d)** Sentiment misalignment, **(e)** Sentiment misalignment

5.5 Training and Validation Loss

The SDCG model's performance, visualized in Fig. 10, shows a sharp initial drop in training loss from about 4.2 to below 2.0 by the fourth epoch. Afterward, it continues to decline steadily. Validation loss follows a similar trend, dropping from around 3.1 to just under 2.0 before leveling out at 1.5.

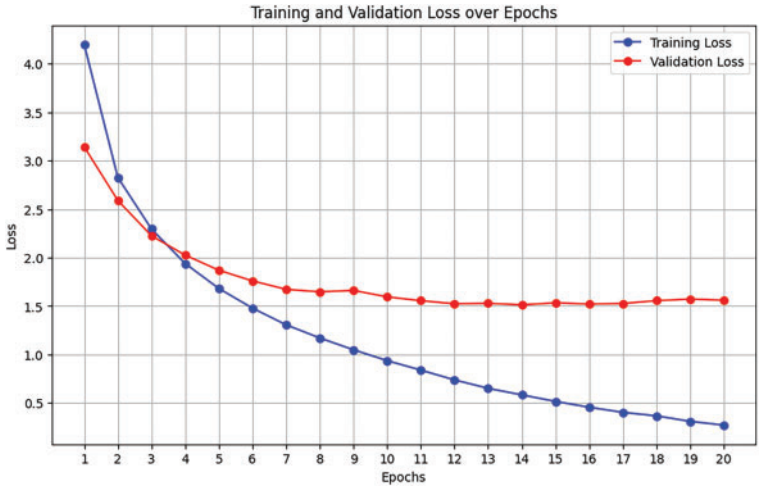

Figure 10: Performance of SDCG: training and validation loss

The similarity between both curves shows that the model maintains a healthy balance between fitting the training data and performing well on new data, which helps ensure reliable sentiment-based captions.

5.6 Limitations

A dual transformer architectural approach creates the main limitation because it leads to elevated computational complexity. The model's need for extensive computational power restricts its use on basic devices as well as its application on demand. Occasionally, our model produces incorrect image captions that do not correctly represent the emotional content. The sentiment integration functions well, although it sometimes misinterprets specific emotional signals. The model encounters major difficulties when analyzing emotions that adapt to contextual changes or when the emotional signals are ambiguous. The actual distribution of sentiments in real-world information demonstrates a significant problem because negative sentiments typically outnumber positive and neutral ones. Due to the unbalanced sentiment distributions that naturally occur in real scenarios, the model shows reduced generalization abilities, even though balancing methods were applied to the dataset. The sentiment model limits its analysis to basic positive, neutral, and negative categories, which makes the identification of advanced emotions like sarcasm, as well as nostalgic and excited feelings, impossible.

6 Conclusion & Future Work

The integration of sentiment into image captioning remains an underexplored area in computer vision and natural language processing. This paper introduced SDCG, a novel transformer-based model designed to generate sentiment-aware image captions with high accuracy. Using five transformer-based fusion strategies, along with RoBERTa for textual sentiment extraction and ViT for visual feature representation, the proposed approach effectively merged textual and visual sentiment embeddings. The experimental results demonstrated that SDCG outperformed existing fusion techniques and baseline models in terms of sentiment accuracy and caption quality, highlighting its effectiveness in generating captions that better align with both visual content and emotional context. For future work, we will focus on reducing computational overhead through model pruning and knowledge distillation, enabling a more efficient yet accurate implementation of SDCG. Additionally, the current model classifies sentiment into broad categories: positive, neutral, and negative, limiting its ability to capture fine-grained emotional nuances such as joy, nostalgia, or sarcasm. Expanding the sentiment classification framework to include more diverse emotional states could enhance the model's expressiveness and applicability. Furthermore, improving domain adaptation techniques will help enhance the generalization of SDCG across diverse datasets with varying linguistic styles, cultural contexts, and image content. Solving these problems may help researchers move closer to building image captioning models that are not only technically sound but also emotionally aware, which could benefit tasks such as analyzing social media, assisting visually impaired users, or tailoring digital content.

Acknowledgement: The authors wish to express their gratitude to the Ministry of Science and Higher Education of the Republic of Kazakhstan for their funding support.

Funding Statement: This research has been funded by the Committee of Science of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. BR24993166).

Author Contributions: The authors confirm their contributions to the paper as follows: conceptualization, methodology, data curation, and writing: Komal Rani Narejo; idea guidance, supervision, and review: Hongying Zan; formal analysis, validation, and visualization: Kheem Parkash Dharmani; funding acquisition and project administration:

Orken Mamyrbayev; resources and software: Janna Alimkulova; draft manuscript preparation: Ainur Akhmediyarova; resources and software: Zhibek Alibiyeva. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Voulodimos A, Doulamis N, Doulamis A, Protopapadakis E. Deep learning for computer vision: a brief review. Comput Intell Neurosci. 2018;2018:7068349. doi:10.1155/2018/7068349.
2. Sarto S, Cornia M, Baraldi L, Cucchiara R. BRIDGE: bridging gaps in image captioning evaluation with stronger visual cues. In: Proceedings of Computer Vision—ECCV 2024; 2024 Sep 29–Oct 4; Milan, Italy. p. 70–87.
3. Ganesan J, Azar AT, Alsenan S, Kamal NA, Qureshi B, Hassanien AE. Deep learning reader for visually impaired. Electronics. 2022;11(19):3335. doi:10.3390/electronics11193335.
4. Orynbay L, Razakhova B, Peer P, Meden B, Emeršič ž. Recent advances in synthesis and interaction of speech, text, and vision. Electronics. 2024;13(10):1726. doi:10.3390/electronics13101726.
5. Onuoha C, Flaherty J, Cong Thang T. Perceptual image quality prediction: are contrastive language–image pretraining (CLIP) visual features effective? Electronics. 2024;13(8):803. doi:10.3390/electronics13080803.
6. Farkh R, Oudinet G, Foued Y. Image captioning using multimodal deep learning approach. Comput Mater Contin. 2024;81:3951–68. doi:10.32604/cmc.2024.049855.
7. Yang J, Sun Y, Liang J, Ren B, Lai S-H. Image captioning by incorporating affective concepts learned from both visual and textual components. Neurocomputing. 2019;328(5):56–68. doi:10.1016/j.neucom.2018.07.086.
8. Ondeng O, Ouma H, Akuon P. A review of transformer-based approaches for image captioning. Appl Sci. 2023;13(19):11103. doi:10.3390/app131911103.
9. Shao M, Feng J, Wu J, Zhang H, Zheng Y. Fine-grained features for image captioning. Comput Mater Contin. 2023;75:4697–712. doi:10.32604/cmc.2023.055510.
10. Nezami OM, Dras M, Wan S, Paris C. Image captioning using facial expression and attention. J Artif Intell Res. 2020;68:661–89. doi:10.1613/jair.1.12025.
11. Al-Malla MA, Jafar A, Ghneim N. Image captioning model using attention and object features to mimic human image understanding. J Big Data. 2022;9:20. doi:10.1186/s40537-022-00562-z.
12. Geng Y, Mei H, Xue X, Zhang X. Image-caption model based on fusion feature. Appl Sci. 2022;12(19):9861.
13. Gherkar Y, Gujar P, Gaziyani A, Kadu S. Sentiment analysis of images using machine learning techniques. ITM Web Conf. 2022;44:03029. doi:10.1051/itmconf/20224403029.
14. Ortis A, Farinella GM, Battiato S. An overview on image sentiment analysis: methods, datasets and current challenges. In: Proceedings of the 16th International Joint Conference on e-Business and Telecommunications; 2019 Jul 26–28; Prague, Czech Republic. p. 290–300.
15. Singh H, Sharma A, Pant M. Pixels to prose: understanding the art of image captioning. arXiv:2408.15714. 2024.
16. Sun Q, Zhang J, Fang Z, Gao Y. Self-enhanced attention for image captioning. Neural Process Lett. 2024;56(2):131. doi:10.1007/s11063-024-11527-x.
17. Kavi Priya S, Pon Karthika K, Kaliappan J, Selvaraj S, Senthil Kumaran R, Nagalakshmi R, et al. Caption generation based on emotions using CSPDenseNet and BiLSTM with self-attention. Appl Comput Intell Soft Comput. 2022;2022:2756396. doi:10.1155/2022/2756396.
18. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2818–26.
19. Chollet F. Xception: deep learning with depthwise separable convolutions. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 1251–58.

20. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, et al. Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd International Conference on Machine Learning; 2015 Jul 6–11; Lille, France. p. 2048–57.
21. Zhu Y, Newsam S. Densenet for dense flow. In: IEEE International Conference on Image Processing (ICIP); 2017 Sep 17–20; Beijing, China. p. 790–4.
22. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition; 2017 Jul 21–26; Honolulu, HI, USA. p. 4700–8.
23. Tan M, Le Q. EfficientNet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning; 2019 Jun 9–15, Long Beach, CA, USA. p. 6105–114.
24. Koonce B, Koonce B. Efficientnet. In: Convolutional neural networks with swift for Tensorflow: image recognition and dataset categorization. Berkeley, CA, USA: Apress; 2021. p. 109–23.
25. Cordonnier J-B, Loukas A, Jaggi M. Multihead attention: collaborate instead of concatenate. arXiv:2006.16362. 2020.
26. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:6000–10. doi:10.5555/3295222.3295349.
27. Habib MB, Hafiz MFB, Khan NA, Hossain S. Multimodal sentiment analysis using deep learning fusion techniques and transformers. Int J Adv Comput Sci Appl. 2024;15(6):856–63. doi:10.14569/ijacsa.2024.0150686.
28. Kalimuthu M, Mogadala A, Mosbach M, Klakow D. Fusion models for improved image captioning. In: Proceedings of Pattern Recognition ICPR International Workshops and Challenges; 2021 Jan 10–15; Online. p. 381–95.
29. Mohamed Y, Khan FF, Haydarov K, Elhoseiny M. It is okay to not be okay: overcoming emotional bias in affective image captioning by contrastive data collection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2022 Jun 18–24; New Orleans, LA, USA. p. 21231–40.
30. Ren S, Li S. Image emotion analysis combining attention mechanism and multi-level correlation. Front Comput Intell Syst. 2023;6(1):60–5. doi:10.54097/fcis.v6i1.12.
31. Chen S, Jin Q, Wang P, Wu Q. Say as you wish: fine-grained control of image caption generation with abstract scene graphs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2020 Jun 13–19; Seattle, WA, USA. p. 9959–68.
32. Yang C, Wang Y, Han L, Jia X, Sun H. Fine-grained image emotion captioning based on generative adversarial networks. Multimed Tools Appl. 2024;83(34):81857–75. doi:10.1007/s11042-024-18680-4.
33. Aziz A, Chowdhury NK, Kabir MA, Chy AN, Siddique MJ. MMTF-DES: a fusion of multimodal transformer models for desire, emotion, and sentiment analysis of social media data. arXiv:2310.14143. 2023.
34. Zhu L, Zhu Z, Zhang C, Xu Y, Kong X. Multimodal sentiment analysis based on fusion methods: a survey. Inf Fusion. 2023;95(3):306–25. doi:10.1016/j.inffus.2023.02.028.
35. Kovenko V, Abdullaiev O, Maliovanyi D, Tarasovskyi D, Bogach I, Bisikalo O. EmoAtCap: emotional attitude captioning dataset. Mendeley Data. 2021. doi: 10.17632/dym6p2pvbt.1.
36. Liu Y. Roberta: a robustly optimized BERT pretraining approach. arXiv:1907.11692. 2019.
37. Alexey D. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
38. Wang J, Yang Z, Hu X, Li L, Lin K, Gan Z, et al. GIT: a generative image-to-text transformer for vision and language. arXiv:2205.14100. 2022.
39. Li J, Li D, Xiong C, Hoi SCH. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. Proc Mach Learn Res. 2022;162:12888–900.
40. He S, Lu Y. A modularized architecture of multi-branch convolutional neural network for image captioning. Electronics. 2019;8(12):1417. doi:10.3390/electronics8121417.
41. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA. p. 3156–64.
42. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: common objects in context. In: Proceedings of Computer Vision–ECCV 2014; 2014 Sep 6–12; Zurich, Switzerland. p. 13–6.

43. Shao J, Yang R. Controllable image caption with an encoder-decoder optimization structure. Appl Intell. 2022;52(10):11382–93. doi:10.1007/s10489-021-02988-x.
44. Ye Z, Khan R, Naqvi N, Islam MS. A novel automatic image caption generation using bidirectional long-short term memory framework. Multimed Tools Appl. 2021;80(17):25557–82. doi:10.1007/s11042-021-10632-6.
45. Herdade S, Kappeler A, Boakye K, Soares J. Image captioning: transforming objects into words. Adv Neural Inf Process Syst. 2019;32:11135–45.
46. Wu X, Li T. Sentimental visual captioning using multimodal transformer. Int J Comput Vis. 2023;131(4):1073–90. doi:10.1007/s11263-023-01752-7.
47. Li T, Hu Y, Wu X. Image captioning with inherent sentiment. In: Proceedings of 2021 IEEE International Conference on Multimedia and Expo (ICME); 2021 Jul 5–9; Shenzhen, China. p. 1–6.
48. Sun Y, Jin G, Zhao Y, Cui R. Multimodal sentiment analysis based on image captioning and attention mechanism. In: Proceedings of the 2023 IEEE 5th International Conference on Civil Aviation Safety and Information Technology (ICCASIT); 2023 Oct 11–13; Dali, China. p. 296–301.
49. Zhang J, Xie Y, Ding W, Wang Z. Cross on cross attention: deep fusion transformer for image captioning. IEEE Trans Circuits Syst Video Technol. 2023;33(8):4257–68. doi:10.1109/tcsvt.2023.3243725.
50. Yan C, Hao Y, Li L, Yin J, Liu A, Mao Z, et al. Task-adaptive attention for image captioning. IEEE Trans Circuits Syst Video Technol. 2021;32(1):43–51. doi:10.1109/tcsvt.2021.3067449.
51. Chen H, Li H, Li Y, Chen C. Multi-scale adaptive task attention network for few-shot learning. In: Proceedings of the 2022 26th International Conference on Pattern Recognition (ICPR); 2022 Aug 21–25; Montreal, QC, Canada. p. 4765–71.
52. Nezami OM, Dras M, Wan S, Paris C. Senti-attend: image captioning using sentiment and attention. arXiv:1811.09789. 2018.
53. You Q, Jin H, Luo J. Image captioning at will: a versatile scheme for effectively injecting sentiments into image descriptions. arXiv:1801.10121. 2018.
54. Bondielli A, Passaro LC. Leveraging clip for image emotion recognition. In: Proceedings of CEUR Workshop Proceedings, CEUR-WS; 2021 Jun 23–26; St. Petersburg, Russia.
55. Aggarwal S, Pandey A, Vishwakarma DK. Modelling visual semantics via image captioning to extract enhanced multi-level cross-modal semantic incongruity representation with attention for multimodal sarcasm detection. arXiv:2408.02595. 2024.
56. Wang W, Chen Z, Hu H. Hierarchical attention network for image captioning. In: Proceedings of AAAI Conference on Artificial Intelligence; 2019 Jan 27–Feb 1; Honolulu, HI, USA. p. 8957–64.
57. Ueda A, Yang W, Sugiura K. Switching text-based image encoders for captioning images with text. IEEE Access. 2023;11:55706–15. doi:10.1109/access.2023.3282444.
58. Wu Z, Meng M, Wu J. Visual sentiment prediction with attribute augmentation and multi-attention mechanism. Neural Process Lett. 2020;51(3):2403–16. doi:10.1007/s11063-020-10201-2.
59. Xin Q, Zhang Y, Tan B. Image captioning with vision/text transformers. In: Proceedings of International Conference on Machine Learning; 2021 Jul 18–24; Virtual.
60. Jamil A, Mahmood K, Villar MG, Prola T, Diez IDLT, Samad MA, et al. Deep learning approaches for image captioning: opportunities, challenges and future potential. IEEE Access. 2024. doi:10.1109/access.2024.3365528.
61. Staniūtė R, Šešok D. A systematic literature review on image captioning. Appl Sci. 2019;9(10):2024. doi:10.3390/app9102024.
62. Li B, Zhou Y, Ren H. Image emotion caption based on visual attention mechanisms. In: Proceedings of 2020 IEEE 6th International Conference on Computer and Communications (ICCC); 2020 Dec 11–14; Chengdu, China. p. 1456–60.
63. Hayaty M, Muthmainah S, Ghufran SM. Random and synthetic over-sampling approach to resolve data imbalance in classification. Int J Artif Intell Res. 2020;4(2):86–94. doi:10.29099/ijair.v4i2.152.

64. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics; 2002 Jul 7–12; Philadelphia, PA, USA. p. 311–8.
65. Lin C-Y. Rouge: a package for automatic evaluation of summaries. In: Proceedings of Text Summarization Branches Out; 2004; Barcelona, Spain. p. 74–81.
66. Vedantam R, Lawrence Zitnick C, Parikh D. Cider: consensus-based image description evaluation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition; 2015 Jun 7–12; Boston, MA, USA. p. 4566–75.
67. Krotov A, Tebo A, Picart DK, Algave AD. Evaluating authenticity and quality of image captions via sentiment and semantic analyses. arXiv:2409.09560. 2024.