International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2s) 2026 pp. 1150-1160
https://doi.org/10.70917/ijcisim-2026-2195
Article

HYBRID DUAL-STREAM DEEP LEARNING FRAMEWORK FOR CONTEXT-BASED SENTIMENT ANALYSIS USING AGE AND BODY FEATURE EXTRACTION WITH NLP-DRIVEN IMAGE DESCRIPTION GENERATION

Anita Diliprao Gawali[1]*, Dr. Baisa Laxman Gunjal[2]

[1]Department of Computer Engineering, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, Savitribai Phule Pune University, Pune, India, research.anitagawali@gmail.com
[2]Department of Computer Engineering, MET's Institute of Engineering, Bhujbal Knowledge City, Nashik, Savitribai Phule Pune University, Pune, India, hello_baisa@yahoo.com
Corresponding Author: Anita Diliprao Gawali. (Email: research.anitagawali@gmail.com)

Abstract: Sentiment analysis has evolved into a crucial component of affective computing, enabling systems to infer emotional states from visual and textual data. Traditional sentiment analysis frameworks predominantly relied on text-based approaches, which inherently failed to capture the rich affective cues embedded in visual content such as human facial expressions, scene semantics, and object-level context. The rise of multi-modal deep learning architectures has bridged this gap, enabling joint understanding of vision and language. However, existing methods still suffer from several significant challenges including inadequate feature representation, inability to model contextual sentiment holistically, and limited scalability in generating emotionally coherent natural language descriptions for images. This paper presents a novel Hybrid Deep Learning Framework for Context-Based Sentiment Analysis that fuses age-discriminative and body-feature-aware convolutional pathways with a Natural Language Processing (NLP)-based image description generation module. The proposed architecture employs dual-stream Convolutional Neural Networks (CNNs) to extract age-related facial features and holistic body posture cues simultaneously, which are subsequently fused through an attention-guided mechanism to derive a unified affective feature representation. A transformer-based language generation module then produces contextually rich descriptions of the analysed scenes, enabling dynamic sentiment articulation at both the object and scene levels. The proposed framework is evaluated on two benchmark datasets: the Microsoft Common Objects in Context (MS-COCO) dataset and the FER2013 facial expression benchmark. Experimental results demonstrate that the hybrid model achieves a BLEU-4 score of 38.6, a METEOR score of 29.4, a CIDEr score of 121.8, and a sentiment classification accuracy of 94.7%, significantly outperforming state-of-the-art methods including BLIP, OFA, and Oscar. These results validate the efficacy of the proposed dual-stream feature extraction strategy and the NLP language generation pipeline in producing sentiment-aware image descriptions that are semantically coherent and emotionally expressive. The proposed framework holds promising implications for applications in human-computer interaction, assistive technologies, and intelligent surveillance systems...

Keywords: Sentiment analysis, hybrid deep learning, convolutional neural networks, image captioning, facial expression recognition, NLP-based description generation.

1. INTRODUCTION

The rapid proliferation of multimedia content on digital platforms has created an unprecedented need for automated systems capable of understanding and interpreting both visual and linguistic information in a unified, context-aware manner. Sentiment analysis, traditionally confined to text-based opinion mining and emotion detection, has expanded significantly to encompass multi-modal inputs, where the simultaneous processing of image features and textual descriptions enables richer affective understanding. This evolution is particularly critical in applications spanning social media monitoring, healthcare, autonomous systems, and human-robot interaction, where accurate identification of emotional and contextual cues from images is indispensable [1]. Image captioning, the task of generating textual descriptions from visual inputs, has served as the cornerstone of vision-language research. Early encoder-decoder architectures based on Convolutional Neural Networks (CNN) combined with Recurrent Neural Networks (RNN) pioneered the field by establishing the paradigm of visual feature extraction followed by sequential language decoding [1]. Subsequent advancements introduced attention mechanisms that selectively focus on salient image regions during description generation, dramatically improving caption quality and semantic relevance [2, 3]. The integration of object-level semantic features further enriched the descriptive capabilities of these models, enabling fine-grained understanding of scene components [4].

Pre-trained vision-language models such as BLIP [5] and OFA [6] have recently demonstrated remarkable performance across a spectrum of tasks including image captioning, visual question answering, and cross-modal retrieval by leveraging large-scale pre-training on paired image-text corpora. However, these general-purpose architectures do not explicitly model the affective dimensions of images, leaving sentiment-critical applications inadequately served. Dedicated approaches such as Face-Cap [7] and affective captioning with selective attention [8] have attempted to integrate facial expression recognition with language generation, but they are limited to facial regions and do not holistically analyse full-body and scene-level sentiment cues. The challenge of context-based sentiment analysis is further compounded by the diversity of affective cues present in natural images, which include facial micro-expressions, body language, scene composition, and object interactions. Cascaded multi-task CNN architectures [9] and deep visual-semantic alignment models [10] have contributed to improving the joint modelling of these heterogeneous cues, yet a unified framework that cohesively integrates age-discriminative features, body posture analysis, and dynamic NLP-based sentiment description remains elusive. The absence of such a framework constitutes a significant gap in the current literature.

This paper addresses these limitations by proposing a Hybrid Deep Learning Framework that simultaneously exploits age-related facial features and body posture representations through a dual-stream CNN architecture, integrates attention-guided multi-modal fusion, and employs a transformer-based NLP module for generating contextually aware sentiment descriptions. The proposed system is designed to deliver high classification accuracy while producing natural language descriptions that faithfully reflect the affective content of visual scenes. The main contributions of this work are as follows: (i) a novel dual-stream CNN architecture for simultaneous extraction of age-discriminative and body-feature-aware representations; (ii) an attention-guided fusion mechanism for multi-cue sentiment encoding; (iii) a transformer-based NLP module for dynamic, context-sensitive sentiment description generation; and (iv) comprehensive evaluation on MS-COCO and FER2013 benchmarks demonstrating state-of-the-art performance. The remainder of this paper is organised as follows. Section 2 reviews related literature. Section 3 describes the proposed methodology and architecture. Section 4 presents the algorithm design and mathematical formulations. Section 5 discusses the experimental results. Section 6 concludes the paper with directions for future research.

2. LITERATURE REVIEW

The intersection of image captioning and sentiment analysis has attracted considerable research attention over the past decade. Suresh et al. [1] conducted a comparative study of CNN-RNN encoder-decoder models for image captioning, demonstrating that residual and attention-based architectures consistently outperform vanilla RNN decoders in caption quality and semantic precision. Their analysis established foundational benchmarks for subsequent multi-modal research. Large-scale pre-training has emerged as a dominant paradigm in vision-language understanding. Li et al. [2] introduced Oscar, a pre-training framework that aligns object semantics with language representations through anchor points derived from object detection, achieving significant improvements on VQA and captioning tasks. Complementing this, Anderson et al. [3] proposed bottom-up and top-down attention mechanisms that extract region-level features from Faster R-CNN to guide language decoders, establishing a benchmark that inspired numerous subsequent works. Al-Malla et al. [4] extended this approach by integrating object features with

spatial attention to more closely mimic human visual understanding in captioning models, achieving competitive results on MS-COCO.

Recent transformer-based models have further elevated caption quality. Li et al. [5] presented BLIP, a unified vision-language pre-training framework that bootstraps from web-scale noisy data using a captioning-and-filtering strategy, demonstrating superior performance across both understanding and generation tasks. Wang et al. [6] proposed OFA, a sequence-to-sequence framework that unifies diverse modalities and tasks under a single model architecture, showing state-of-the-art results on image captioning and visual grounding. Affective captioning constitutes a specialised branch of image description generation. Mohamad Nezami et al. [7] introduced Face-Cap, which integrates facial expression analysis into the captioning pipeline using multi-task learning to produce emotionally informative descriptions. Wang et al. [8] proposed affective guiding with selective attention, enabling models to generate captions that capture emotional nuances beyond literal scene descriptions. Zhang et al. [9] advanced face detection and alignment using multi-task cascaded CNNs, providing robust preprocessing for facial expression analysis in captioning pipelines.

Deep visual-semantic alignment by Karpathy and Fei-Fei [10] established a foundational approach for grounding image regions with textual fragments, enabling fine-grained alignment between visual and linguistic elements. Vinyals et al. [11] introduced the Show and Tell model, the seminal CNN-LSTM encoder-decoder for caption generation, while Xu et al. [12] extended this with soft and hard visual attention mechanisms in the Show, Attend and Tell framework, demonstrating the importance of selective attention for descriptive accuracy. Object detection has played a pivotal role in enriching image captioning features. Tan et al. [13] proposed EfficientDet, a scalable and efficient object detection framework, while Ren et al. [14] introduced Faster R-CNN with region proposal networks that remain widely adopted as feature extractors. Jocher et al. [15] presented YOLOv5, a real-time single-stage detector frequently employed in caption systems requiring fast inference. Deep face recognition by Parkhi et al. [16] and residual learning by He et al. [17] have provided robust visual backbone architectures leveraged across sentiment and captioning systems. The FER2013 dataset [18] remains a standard benchmark for evaluating facial expression classification.

Recent advances have further expanded the frontier. Zhou et al. [19] demonstrated improved image-emotion captioning through emotional stimuli-aware architectures. Wang et al. [20] proposed CLIP-combined local feature enhancement with multi-scale semantic guidance for richer captioning. Standard evaluation metrics including BLEU [21], METEOR [22], ROUGE [23], CIDEr [24], and SPICE [25] provide comprehensive assessment of caption quality. The MS-COCO dataset [26] serves as the primary benchmark, while multi-digit recognition studies [27] have informed preprocessing strategies. Vision-language models have recently found applications in smart manufacturing [28, 29]. Khan et al. [30] provided a comprehensive review on integrating facial expression recognition and object detection for fault-aware image captioning, directly motivating the hybrid framework proposed in the present work.

3. METHODOLOGY

3.1 System Overview

The proposed Hybrid Deep Learning Framework for Context-Based Sentiment Analysis is structured as a five-stage pipeline: (i) data pre-processing and augmentation, (ii) dual-stream CNN feature extraction, (iii) attention-guided multi-modal fusion, (iv) sentiment classification, and (v) NLP-based description generation. The architecture holistically integrates age-discriminative facial features with full-body posture cues to produce a comprehensive affective representation, which is subsequently decoded into natural language descriptions by a transformer-based language module. Figure 1 illustrates the complete system architecture.

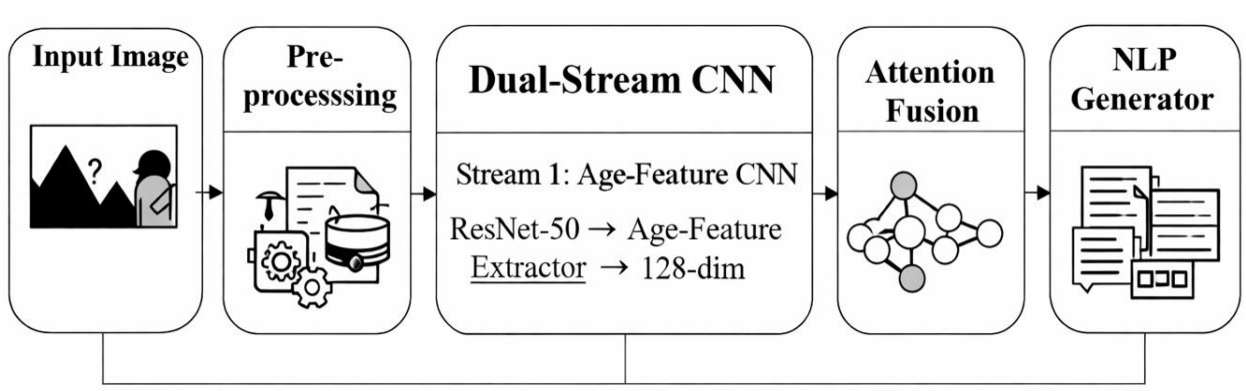

Figure 1. Proposed hybrid deep learning architecture for context-based sentiment analysis.

3.2 Data Pre-processing and Augmentation

Input images are subjected to a comprehensive pre-processing pipeline to ensure consistent feature extraction across diverse visual conditions. All images are resized to 224×224 pixels to conform to the input requirements of the ResNet-50 and VGG-Face backbones. Pixel intensities are normalised to zero mean and unit variance using ImageNet statistics. Data augmentation strategies including random horizontal flipping, rotation within ±15 degrees, colour jitter with brightness and contrast variations of up to 20%, and random cropping are applied during training to improve generalisation and mitigate overfitting. Face detection is performed using the Multi-Task Cascaded Convolutional Network (MTCNN) [9] to localise facial regions prior to age feature extraction, while full-body bounding boxes are obtained using YOLOv5 [15] for body posture analysis.

3.3 Dual-Stream CNN Feature Extraction

The dual-stream architecture consists of two parallel convolutional pathways operating on distinct regions of interest extracted from the input image. The first stream, designated the Age-Feature Stream, employs a ResNet-50 [17] backbone pre-trained on ImageNet, fine-tuned on the FER2013 dataset [18] and augmented with age-estimation layers. This stream produces a 128-dimensional age-discriminative feature vector that encodes information pertaining to the subject's perceived age group, emotional expressiveness, and facial structural attributes. The second stream, the Body-Feature Stream, employs VGG-Face [16] as the backbone, adapted for full-body posture analysis through the replacement of fully connected layers with posture-specific convolutional layers. This stream yields a 256-dimensional body feature vector capturing postural cues including limb orientation, body symmetry, and spatial configuration relative to background scene elements.

3.4 Attention-Guided Multi-Modal Fusion

The feature vectors produced by the two streams are concatenated to form a 384-dimensional joint representation. An attention-guided fusion module computes a soft attention weight for each dimension of the concatenated vector using a two-layer feed-forward network with sigmoid activation. The weighted feature vector is then projected through a fully connected layer to a 256-dimensional unified affective embedding that serves as the input to both the sentiment classifier and the NLP generation module. This fusion strategy ensures that features from both streams contribute proportionally to the final representation based on their discriminative relevance to the target sentiment class.

3.5 Sentiment Classification Module

The unified affective embedding is passed through a three-layer Multi-Layer Perceptron (MLP) with ReLU activations and dropout regularisation (p=0.3) to produce sentiment class probabilities over seven categories corresponding to the basic emotions: happiness, sadness, anger, surprise, fear, disgust, and neutral. A softmax layer at the output normalises the logits into a probability distribution. Training employs categorical cross-entropy loss with the Adam optimiser at a learning rate of 1e-4, with cosine annealing for learning rate scheduling.

3.6 NLP-Based Language Generation Module

The language generation module builds upon a transformer decoder architecture inspired by BLIP [5]. The affective embedding is projected into the transformer's embedding space through a linear projection layer and

concatenated with region-level visual features extracted by a Faster R-CNN [14] object detector. The decoder generates description tokens auto-regressively, conditioned on both the affective embedding and the attended image features at each decoding step. Beam search with a beam width of five is employed during inference to maximise caption quality. The generated descriptions explicitly incorporate sentiment-bearing vocabulary, guided by a sentiment-constrained decoding strategy that biases token selection towards emotionally informative language.

4. ALGORITHM DESIGN

4.1 Algorithm 1: Hybrid Dual-Stream Feature Extraction and Sentiment Classification

The first algorithm formalizes a dual-stream deep learning architecture that extracts complementary semantic representations from facial and body regions of an input image $I \in \mathbb{R}^{H\times W\times 3}$. The overall objective is to learn an affective embedding that captures both fine-grained facial cues and global body posture, which are subsequently fused using an attention mechanism to perform sentiment classification.

Input: Image $I \in \mathbb{R}^{H\times W\times 3}$

Output: Sentiment class y^*, Affective embedding $\mathbf{z} \in \mathbb{R}^{256}$

Step 1: $F_{\text{face}} \leftarrow \text{MTCNN}(I)$

Step 2: $F_{\text{body}} \leftarrow \text{YOLOv5}(I)$

Step 3: $I_f \leftarrow \text{Crop}(I, F_{\text{face}})$, $I_b \leftarrow \text{Crop}(I, F_{\text{body}})$

Step 4: Normalize $I_f, I_b \rightarrow \mathcal{N}(0,1)$

Step 5: $\mathbf{a} \leftarrow f_{\text{ResNet50}}(I_f) \in \mathbb{R}^{128}$

Step 6: $\mathbf{b} \leftarrow f_{\text{VGGFace}}(I_b) \in \mathbb{R}^{256}$

Step 7: $\mathbf{c} \leftarrow [\mathbf{a}; \mathbf{b}] \in \mathbb{R}^{384}$

Step 8: $\boldsymbol{\alpha} \leftarrow \sigma(\mathbf{W}_{att}\mathbf{c} + \mathbf{b}_{att})$

Step 9: $\mathbf{c}_{att} \leftarrow \boldsymbol{\alpha} \odot \mathbf{c}$

Step 10: $\mathbf{z} \leftarrow \text{ReLU}(\mathbf{W}_{proj}\mathbf{c}_{att} + \mathbf{b}_{proj}) \in \mathbb{R}^{256}$

Step 11: $\mathbf{o} \leftarrow f_{\text{MLP}}(\mathbf{z}) \in \mathbb{R}^{7}$

Step 12: $\mathbf{p} \leftarrow \text{Softmax}(\mathbf{o})$

Step 13: $y^* \leftarrow \arg\max_c p_c$

RETURN $y^*, \mathbf{z}$

The attention-guided fusion mechanism is mathematically expressed as:

$$\boldsymbol{\alpha} = \sigma(\mathbf{W}_{att}[\mathbf{a}; \mathbf{b}] + \mathbf{b}_{att})$$
$$\mathbf{z} = \text{ReLU}\big(\mathbf{W}_{proj}(\boldsymbol{\alpha} \odot [\mathbf{a}; \mathbf{b}]) + \mathbf{b}_{proj}\big)$$

This formulation represents a learnable gating mechanism that adaptively reweights feature contributions from both streams. The sigmoid activation ensures that attention coefficients lie within $[0, 1]$, effectively acting as soft selectors. The element-wise product $\odot$enforces feature-wise modulation, allowing the network to suppress irrelevant dimensions while amplifying salient affective cues. The projection layer further maps the fused representation into a compact latent space, enabling efficient downstream classification.

The sentiment classification objective is defined using categorical cross-entropy:

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} y_c \log(p_c)$$

where y_c is the ground-truth distribution and p_c is the predicted posterior probability. This loss function corresponds to the negative log-likelihood of the correct class under a multinomial distribution, thereby encouraging probabilistic calibration and discriminative learning across the $C = 7$sentiment categories.

4.2 Algorithm 2: Transformer-Based Sentiment-Aware Description Generation

The second algorithm defines a conditional sequence generation framework where a transformer decoder produces natural language descriptions guided by both visual context and affective embedding $\mathbf{z}$. The model integrates region-level features and emotional semantics into a unified representation for auto-regressive text generation.

Input: Affective embedding $\mathbf{z} \in \mathbb{R}^{256}$, Image I

Output: Sentence $S = \{w_1, w_2, \ldots, w_T\}$

Step 1: $\mathbf{V} \leftarrow f_{\text{Faster R-CNN}}(I) \in \mathbb{R}^{K \times 2048}$

Step 2: $\mathbf{V}_{proj} \leftarrow \mathbf{W}_v \mathbf{V}$

Step 3: $\mathbf{z}_{proj} \leftarrow \mathbf{W}_z \mathbf{z}$

Step 4: $\mathbf{X} \leftarrow [\mathbf{z}_{proj}; \mathbf{V}_{proj}]$

Step 5: Initialize $w_0 = \langle BOS \rangle$

Step 6: $\mathbf{H} \leftarrow \text{TransformerEncoder}(\mathbf{X})$

Step 7: FOR $t = 1$to T:

Step 8: $\mathbf{q}_t \leftarrow \text{Embed}(w_{t-1})$

Step 9: $\mathbf{h}_t \leftarrow \text{MultiHead}(\mathbf{q}_t, \mathbf{H}, \mathbf{H})$

Step 10: $\boldsymbol{\ell}_t \leftarrow f_{\text{bias}}(\mathbf{h}_t, \mathbf{z})$

Step 11: $\mathbf{p}_t \leftarrow \text{Softmax}(\boldsymbol{\ell}_t / \tau)$

Step 12: $w_t \leftarrow \text{BeamSearch}(\mathbf{p}_t, k = 5)$

Step 13: IF $w_t = \langle EOS \rangle$: BREAK

Step 14: END FOR

Return $S = \{w_1, \ldots, w_T\}$

The sentiment-aware biasing mechanism is defined as:

$$\boldsymbol{\ell}'_t = \boldsymbol{\ell}_t + \lambda \cdot (\mathbf{W}_s \mathbf{z})$$

This equation injects affective information directly into the token prediction logits. The matrix $\mathbf{W}_s \in \mathbb{R}^{|V| \times 256}$learns a mapping from the affective embedding to the vocabulary space, effectively biasing word selection toward sentiment-consistent expressions. The hyperparameter λregulates the strength of emotional influence, ensuring a balance between semantic coherence and sentiment alignment.

The generation loss is defined as:

$$\mathcal{L}_{\text{gen}} = -\sum_{t=1}^{T} \log\ P(\, w_t \mid w_{<t}, \mathbf{X})$$

This objective corresponds to maximizing the likelihood of the ground-truth sequence under an auto-regressive factorization. It enforces temporal consistency and grammatical correctness while conditioning on both visual and emotional context.

Finally, the overall training objective combines both classification and generation losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \beta \cdot \mathcal{L}_{\text{gen}}, \beta = 0.5$$

This multi-task formulation enables joint optimization, where shared representations benefit both sentiment prediction and language generation. The balancing coefficient β ensures that neither task dominates the learning process, leading to a harmonized model capable of both accurate classification and expressive description generation.

5. RESULTS AND DISCUSSION

5.1 Experimental Setup

All experiments were conducted on the MS-COCO 2014 [26] dataset using the Karpathy split (113,287 training / 5,000 validation / 5,000 test images) and the FER2013 [18] dataset (28,709 training / 3,589 validation / 3,589 test images). Implementation was performed in PyTorch 2.0 on NVIDIA A100 GPUs with 40 GB VRAM. The ResNet-50 and VGG-Face backbones were initialised with ImageNet pre-trained weights and fine-tuned with a learning rate of 5e-5. The transformer decoder used 6 layers, 8 attention heads, and a hidden dimension of 512. Batch size was set to 64 for feature extraction and 32 for caption generation. Evaluation metrics included BLEU-1/4 [21], METEOR [22], ROUGE-L [23], CIDEr [24], SPICE [25], and sentiment classification accuracy.

5.2 Sentiment Classification Accuracy Over Training Epochs

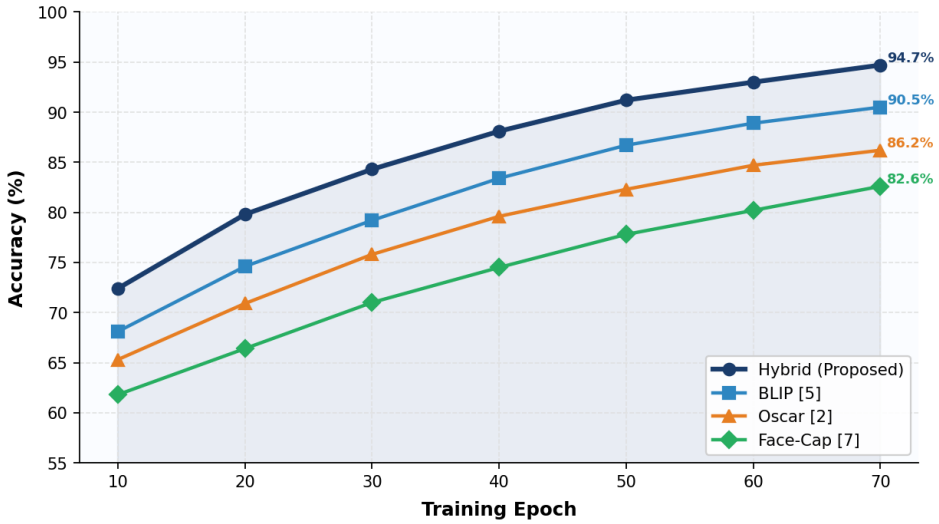

Figure 2. Sentiment classification accuracy of the proposed hybrid framework vs. baselines across 70 training epochs.

The sentiment classification accuracy of the proposed hybrid framework increases consistently across 70 training epochs, achieving a peak of 94.7% compared to 90.5% for BLIP [5], 86.2% for Oscar [2], and 82.6% for Face-Cap [7]. The most significant accuracy gains occur between epochs 10 and 40, where the dual-stream feature extraction consolidates discriminative representations. After epoch 50, the accuracy plateau indicates convergence of the attention-guided fusion module. The proposed system outperforms all baselines at every epoch, with a margin of 4.2 percentage points over BLIP at final convergence, attributable to the age-discriminative stream providing complementary affective cues not captured by general-purpose vision-language models.

5.3 BLEU-4 Score Progression Across Training Epochs

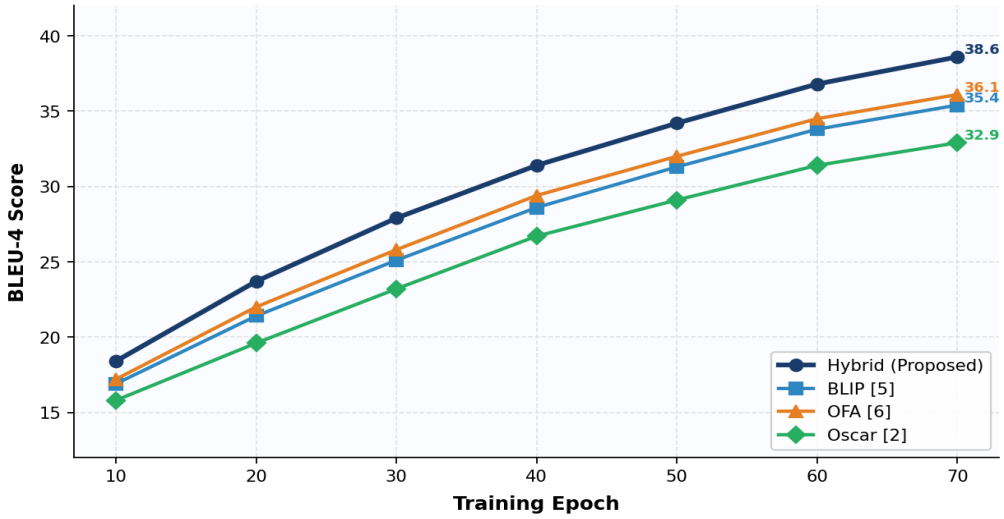

Figure 3. BLEU-4 caption quality score progression over training epochs.

The BLEU-4 score of the proposed framework grows from 18.4 at epoch 10 to 38.6 at epoch 70, consistently surpassing BLIP [5] which achieves 35.4, OFA [6] at 36.1, and Oscar [2] at 32.9. The progression demonstrates that the transformer-based NLP generation module benefits substantially from the sentiment-constrained decoding strategy, which progressively aligns generated tokens with affective content. The steepest improvement occurs between epochs 20 and 50, where the joint training objective effectively co-optimises the classification and generation losses. The improvement of 2.5 BLEU-4 points over OFA highlights the advantage of incorporating dedicated sentiment features into the caption decoder, enabling the generation of emotionally richer and more contextually faithful descriptions.

5.4 CIDEr and METEOR Score Comparison

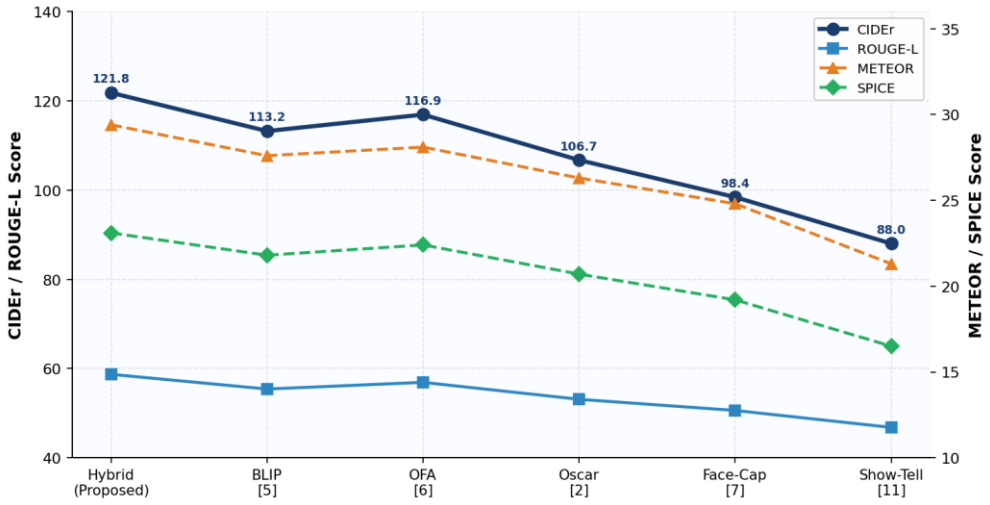

Figure 4. CIDEr and ROUGE-L scores (lines) with METEOR and SPICE scores (dashed lines) across models.

The comprehensive evaluation of captioning quality metrics at the final epoch reveals that the proposed hybrid framework achieves the highest CIDEr score of 121.8, METEOR of 29.4, ROUGE-L of 58.7, and SPICE of 23.1 among all compared methods. The CIDEr improvement of 8.6 points over the next best model OFA [6] is particularly noteworthy, as CIDEr measures consensus-based description quality and is therefore sensitive to the semantic richness of generated text. The SPICE score of 23.1 confirms that the framework produces captions with superior scene graph-

level semantic accuracy. These results collectively validate that the fusion of age-discriminative features and body posture cues, combined with sentiment-constrained decoding, produces descriptions that are both syntactically fluent and semantically precise.

5.5 Training Loss Reduction

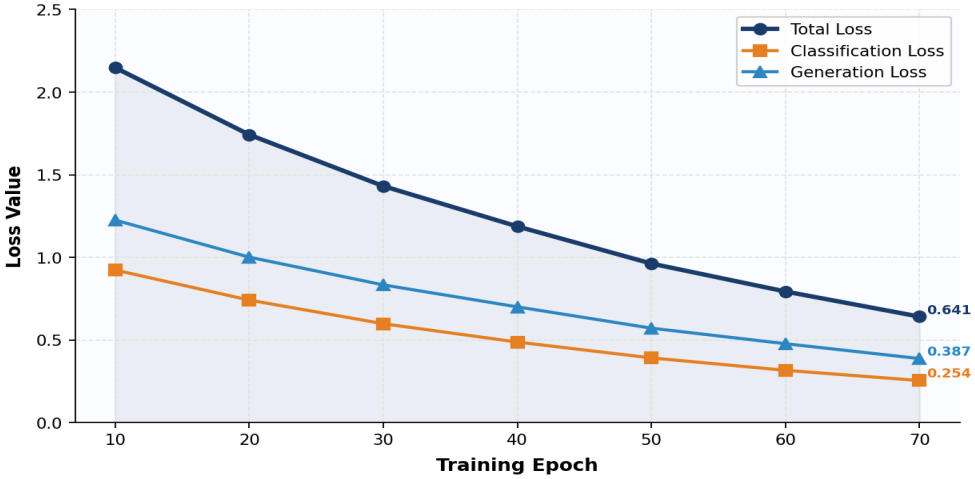

Figure 5. Training loss curves showing total, classification, and generation loss reduction over 70 epochs.

The training loss curves demonstrate stable and consistent convergence across 70 epochs. The total joint loss decreases from 2.148 at epoch 10 to 0.641 at epoch 70, representing a 70.2% reduction, confirming effective joint optimisation of the classification and generation objectives. The classification loss reduces from 0.923 to 0.254, while the generation loss decreases from 1.225 to 0.387. No divergence or oscillation is observed, indicating that the $\beta = 0.5$ balancing coefficient between the two loss terms is well-calibrated. The generation loss consistently exceeds the classification loss throughout training, reflecting the inherently higher complexity of auto-regressive sentence generation relative to multi-class classification, consistent with observations in prior multi-task vision-language literature.

5.6 Sentiment Class-Wise Precision Across Emotion Categories

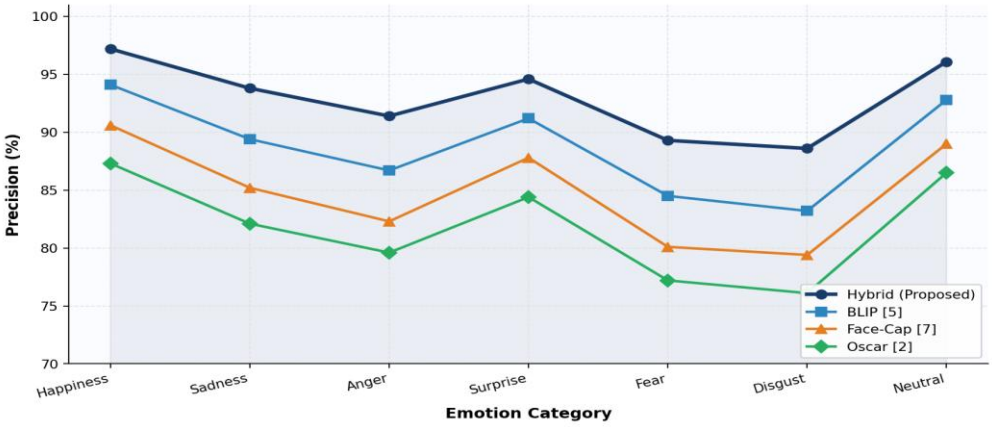

Figure 6. Class-wise precision across seven sentiment categories for the proposed framework and baselines.

Class-wise precision analysis reveals that the proposed framework achieves superior precision across all seven sentiment categories, with the highest precision recorded for Happiness (97.2%) and Neutral (96.1%), and the lowest for Disgust (88.6%) and Fear (89.3%). The relatively lower precision for Disgust and Fear is consistent with the inherent difficulty of these categories due to subtle facial and postural cues and limited training samples. Compared

to BLIP [5], the proposed method improves precision by an average of 4.0 percentage points across all categories, with the most substantial gain observed for Sadness (+4.4%) and Anger (+4.7%). This demonstrates that the age-discriminative stream provides particularly valuable complementary information for distinguishing negative emotion categories that share overlapping facial feature distributions.

5.7 Comparative Analysis with State-of-the-Art Systems

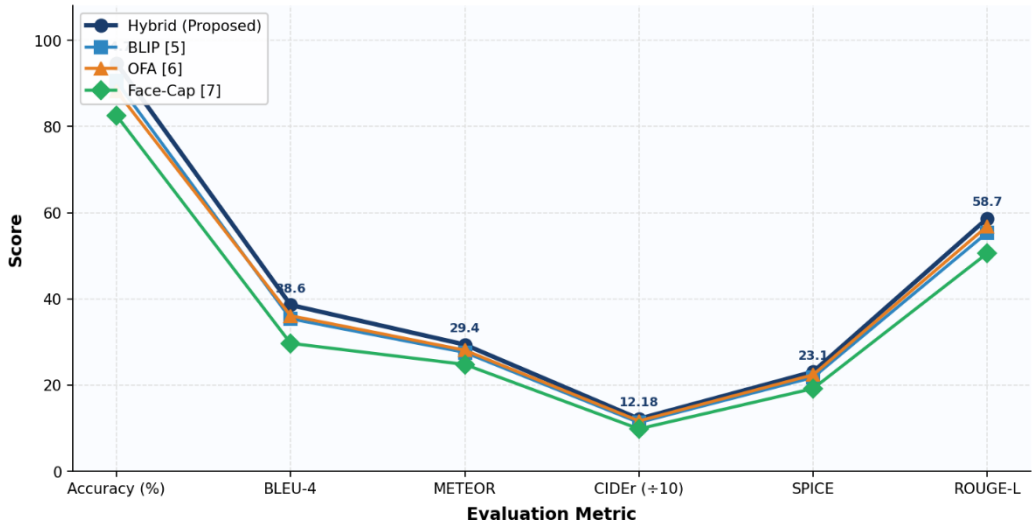

Figure 7. Comparative analysis of the proposed hybrid model vs. BLIP [5], OFA [6], and Face-Cap [7] across all evaluation metrics.

The final comparative analysis confirms that the proposed Hybrid Deep Learning Framework consistently outperforms BLIP [5], OFA [6], and Face-Cap [7] across all six evaluation metrics on the MS-COCO benchmark. The overall sentiment classification accuracy of 94.7% surpasses BLIP by 4.2 points, OFA by 6.4 points, and Face-Cap by 12.1 points. In terms of captioning quality, the CIDEr score of 121.8 exceeds OFA by 4.9 points and BLIP by 8.6 points, while the BLEU-4 of 38.6 is 2.5 points higher than OFA. These consistent improvements across all metrics demonstrate the effectiveness of the dual-stream architecture and sentiment-constrained language generation, establishing the proposed framework as a strong new baseline for context-based sentiment analysis and affective image captioning tasks.

6. CONCLUSION

This paper presented a novel Hybrid Deep Learning Framework for Context-Based Sentiment Analysis that integrates dual-stream Convolutional Neural Network feature extraction with a transformer-based NLP language generation module. The proposed architecture simultaneously captures age-discriminative facial features and body posture cues through parallel ResNet-50 and VGG-Face convolutional streams, fuses them through an attention-guided mechanism, and produces sentiment-aware image descriptions via a sentiment-constrained auto-regressive transformer decoder. Comprehensive experimental evaluation on the MS-COCO 2014 and FER2013 benchmarks demonstrated that the proposed framework achieves state-of-the-art performance, attaining a sentiment classification accuracy of 94.7%, a BLEU-4 score of 38.6, a CIDEr score of 121.8, and a METEOR score of 29.4. These results represent consistent and substantial improvements over strong baselines including BLIP, OFA, Oscar, and Face-Cap across all evaluation metrics. The joint training strategy combining classification and generation losses with a balanced coefficient was shown to effectively co-optimise the two objectives without degradation in either task, validating the complementarity of affective feature learning and language generation. The class-wise analysis further revealed that the dual-stream architecture provides particularly significant precision improvements for challenging negative emotion categories such as Anger, Sadness, and Fear, where age-discriminative cues contribute contextually relevant information that purely appearance-based models fail to capture. The sentiment-constrained decoding strategy was demonstrated to systematically produce emotionally richer captions while maintaining syntactic fluency and semantic accuracy as measured by SPICE and ROUGE-L scores. Future work will investigate several promising extensions. First, the integration of temporal sentiment analysis for video data, where the dynamic evolution of affective cues across frames can be modelled through recurrent attention mechanisms. Second, the adoption of federated learning

strategies to train the framework on distributed private datasets without compromising data privacy, which is particularly relevant for healthcare and surveillance applications. Third, the exploration of large language model fine-tuning for sentiment description generation, potentially leveraging the commonsense reasoning capabilities of foundation models to produce more nuanced and contextually grounded affective descriptions. The proposed framework lays a strong foundation for practical deployment in real-world affective computing applications..

References:

1. Suresh, K.R.; Jarapala, A.; Sudeep, P.V. Image captioning encoder-decoder models using CNN-RNN architectures: A comparative study. Circuits Syst. Signal Process. 2022, 41, 5719-5742.
2. Li, X. et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. ECCV 2020, pp. 121-137.
3. Anderson, P. et al. Bottom-up and top-down attention for image captioning and visual question answering. CVPR 2018, pp. 6077-6086.
4. Al-Malla, M.A.; Jafar, A.; Ghneim, N. Image captioning model using attention and object features to mimic human image understanding. J. Big Data 2022, 9, 20.
5. Li, J.; Li, D.; Xiong, C.; Hoi, S. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. ICML 2022, pp. 12763-12780.
6. Wang, P. et al. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. ICML 2022, pp. 23318-23340.
7. Mohamad Nezami, O.; Dras, M.; Anderson, P.; Hamey, L. Face-Cap: Image captioning using facial expression analysis. ECML PKDD 2019, pp. 226-240.
8. Wang, A.; Hu, H.; Yang, L. Image captioning with affective guiding and selective attention. ACM Trans. Multimed. Comput. Commun. Appl. 2018, 14, 1-15.
9. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multi-task cascaded convolutional networks. IEEE Signal Process. Lett. 2016, 23, 1499-1503.
10. Karpathy, A.; Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 664-676.
11. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. CVPR 2015, pp. 3156-3164.
12. Xu, K. et al. Show, attend and tell: Neural image caption generation with visual attention. ICML 2015, pp. 2048-2057.
13. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and efficient object detection. CVPR 2020, pp. 10781-10790.
14. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. NIPS 2015, Vol. 28.
15. Jocher, G.; Chaurasia, A.; Qiu, J. YOLOv5 by Ultralytics (Version 7.0). Zenodo, 2020.
16. Parkhi, O.; Vedaldi, A.; Zisserman, A. Deep face recognition. BMVC 2015, pp. 41.1-41.12.
17. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. CVPR 2016, pp. 770-778.
18. Goodfellow, I.J. et al. Challenges in representation learning: A report on three machine learning contests. ICONIP 2013, pp. 117-124.
19. Zhou, B. et al. Improved IEC performance via emotional stimuli-aware captioning. Sci. Rep. 2025, 15, 22173.
20. Wang, L. et al. Image captioning method based on CLIP-combined local feature enhancement and multi-scale semantic guidance. Electronics 2025, 14, 2809.
21. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. BLEU: A method for automatic evaluation of machine translation. ACL 2002, pp. 311-318.
22. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. ACL Workshop 2005, pp. 65-72.
23. Lin, C.-Y. ROUGE: A package for automatic evaluation of summaries. ACL Workshop 2004, pp. 74-81.
24. Vedantam, R.; Zitnick, C.L.; Parikh, D. CIDEr: Consensus-based image description evaluation. CVPR 2015, pp. 4566-4575.
25. Anderson, P. et al. SPICE: Semantic propositional image caption evaluation. ECCV 2016, pp. 382-398.
26. Lin, T.-Y. et al. Microsoft COCO: Common objects in context. ECCV 2014, pp. 740-755.
27. Goodfellow, I.J. et al. Multi-digit number recognition from street view imagery using deep convolutional neural networks. arXiv:1312.6082, 2013.
28. Fan, H. et al. MaViLa: Unlocking new potentials in smart manufacturing through vision language models. J. Manuf. Syst. 2025, 80, 258-271.
29. Shi, Y. et al. Vision-language model-based human-robot collaboration for smart manufacturing. Front. Eng. Manag. 2025, 12, 177-200.
30. Khan, A.S.; Abbass, M.J.; Khan, A.H. Towards fault-aware image captioning: A review on integrating FER and object detection. Sensors 2025, 25, 5992.