RESEARCH ARTICLE

VBCSNet: A Hybrid Attention-Based Multimodal Framework with Structured Self-Attention for Sentiment Classification

Yupu Liu[1], Xin Kang[1,*], Kazuyuki Matsumoto[2] and Jiazheng Zhou[1]

[1] Graduate School of Advanced Technology and Science, Tokushima University, Tokushima 770-8506, Japan
[2] Graduate School of Technology, Industrial and Social Sciences, Tokushima University, Tokushima 770-8506, Japan

Abstract

Multimodal Sentiment Analysis (MSA), a pivotal task in affective computing, aims to enhance sentiment understanding by integrating heterogeneous data from modalities such as text, images, and audio. However, existing methods continue to face challenges in semantic alignment, modality fusion, and interpretability. To address these limitations, we propose VBCSNet, a hybrid attention-based multimodal framework that leverages the complementary strengths of Vision Transformer (ViT), BERT, and CLIP architectures. VBCSNet employs a Structured Self-Attention (SSA) mechanism to optimize intra-modal feature representation and a Cross-Attention module to achieve fine-grained semantic alignment across modalities. Furthermore, we introduce a multi-objective optimization strategy that jointly minimizes classification loss, modality alignment loss, and contrastive loss, thereby enhancing semantic consistency and feature discriminability. We evaluated VBCSNet on three multilingual multimodal sentiment datasets, including MVSA, IJCAI2019, and a self-constructed Japanese Twitter corpus(JP-Buzz). Experimental results demonstrated that VBCSNet significantly outperformed state-of-the-art baselines in terms of Accuracy, Macro-F1, and cross-lingual generalization. Per-class performance analysis further highlighted the model's interpretability and robustness. Overall, VBCSNet advances sentiment classification across languages and domains while offering a transparent reasoning mechanism suitable for real-world applications in affective computing, human-computer interaction, and socially aware AI systems.

Keywords: multimodal sentiment analysis, vision-language models, structured self-attention, cross-attention, contrastive learning, interpretability, cross-lingual evaluation.

Academic Editor:
Jian Lan

Submitted: 19 May 2025
Accepted: 23 October 2025
Published: 13 November 2025

Vol. 2, **No.** 4, 2025.
10.62762/CJIF.2025.537775

***Corresponding author:**
✉ Xin Kang
kang-xin@is.tokushima-u.ac.jp

Citation
Liu, Y., Kang, X., Matsumoto, K., & Zhou, J. (2025). VBCSNet: A Hybrid Attention-Based Multimodal Framework with Structured Self-Attention for Sentiment Classification. *Chinese Journal of Information Fusion*, 2(4), 356–369.

1 Introduction

With the widespread adoption of social media platforms, users increasingly express emotions not only through text but also through images, emojis, and short videos [1–3]. This multimodal nature of user-generated content has driven significant interest in Multimodal Sentiment Analysis (MSA) [4], a core task in affective computing that seeks to integrate heterogeneous data, such as text, image, and audio, to improve the accuracy and robustness of sentiment recognition. Compared with unimodal methods, MSA offers a more comprehensive understanding of emotional expression, especially in cases involving sarcasm, ambiguity, or culturally nuanced language.

While recent advances in deep learning have greatly improved MSA performance, several persistent challenges continue to limit the effectiveness and applicability of current models. First, modality imbalance is a common issue: during joint training, powerful pre-trained language models often dominate the learning process, causing the model to neglect visual signals that may contradict or complement textual sentiment [5]. Second, semantic misalignment frequently arises due to representational gaps between modalities, especially when metaphors, abstract visual cues, or culturally specific references are involved [6]. Fusion strategies that fail to explicitly model such semantic correspondence often result in brittle or shallow interactions. Third, lack of interpretability remains a major limitation. Many multimodal models behave as black boxes, offering little transparency into the reasoning process, which is a critical barrier in domains such as healthcare, policymaking, and public safety [7].

To address these challenges, in particular, we propose VBCSNet, a hybrid attention-based multimodal sentiment analysis framework that integrates the complementary strengths of Vision Transformer (ViT) for visual encoding, BERT [8] for textual representation, and CLIP [9] as a cross-modal semantic bridge. The design of VBCSNet directly responds to the limitations outlined above. To mitigate modality imbalance, VBCSNet employs a three-stream encoder that treats visual, textual, and cross-modal inputs in parallel, ensuring each modality contributes distinct and meaningful features. To overcome semantic misalignment, we incorporate a hierarchical attention mechanism combining Structured Self-Attention (SSA) [10] for refining intra-modal features and Cross-Attention for deep semantic fusion between image and text. Finally, to enhance interpretability and semantic consistency, we introduce a multi-objective optimization strategy that jointly minimizes classification loss, modality alignment loss, and contrastive loss [11, 12]. This strategy encourages feature discriminability, robust fusion, and alignment with human-interpretable cues.

We validate the effectiveness of VBCSNet on three multilingual multimodal sentiment datasets: the benchmark English datasets MVSA [13] and IJCAI2019 [14], and a newly constructed Japanese Twitter dataset spanning multiple domains [15, 16]. Experimental results **demonstrated** that VBCSNet consistently **outperformed** strong baselines—including CNN+CNN, ViT+BERT, and ViT+BERT+CLIP—by 4.9% to 6.1% in classification Accuracy and Macro-F1. This study **builds** upon previous architectural variants by proposing a unified hierarchical attention mechanism, introducing structured optimization objectives, and evaluating cross-lingual generalization on more diverse and challenging datasets [17, 18].

The main contributions of this work are summarized as follows:

1. We propose VBCSNet, a three-stream encoder framework that integrates ViT, BERT, and CLIP to ensure balanced and semantically rich multimodal representation learning.
2. We design a hierarchical attention mechanism that combines Structured Self-Attention and Cross-Attention to refine intra-modal representations and enable fine-grained semantic fusion across modalities.
3. We introduce a multi-objective loss function that jointly optimizes sentiment classification, modality alignment, and contrastive learning, improving semantic consistency and feature discriminability.
4. We conduct comprehensive cross-lingual evaluation on English, Chinese, and Japanese datasets, demonstrating the robustness, interpretability, and real-world applicability of the proposed method.

2 Related Work

This section reviews prior work in multimodal sentiment analysis (MSA) with a focus on techniques for modality fusion, semantic alignment, and interpretability. These areas correspond directly to the challenges VBCSNet is designed to address.

Table 1. Comparison of modality fusion strategies.

Method	Advantages	Limitations
Cross-Attention	Fine-grained alignment, strong local modeling	Prone to modality dominance
Contrastive Learning	Strong global consistency, structurally simple	Depends on sample quality, ignores local detail
Structured Self-Attention	High feature diversity, reduces redundancy	Computationally intensive, needs fine-tuning

2.1 Progress in Multimodal Sentiment Analysis

Multimodal sentiment analysis aims to improve sentiment prediction by integrating heterogeneous data sources such as text, images, and audio. Early studies typically used handcrafted features from each modality and combined them via simple concatenation operations [2]. While computationally straightforward, these early fusion methods were limited in their ability to capture deep semantic interactions and exhibited poor generalization across domains.

With the advent of deep learning, more advanced fusion techniques have emerged. Models such as LXMERT [6] and ViLBERT [19] adopt dual-stream Transformer architectures that encode image and text separately and use cross-modal attention to model interactions. The CLIP model [9] introduces large-scale contrastive pretraining on image-text pairs, effectively constructing a shared semantic embedding space and achieving impressive generalization across downstream tasks [33]. However, these models often suffer from modality imbalance, where the strong inductive bias of pre-trained language encoders overwhelms the contribution of visual features, especially when the sentiment in text and image diverges [5].

Recent lightweight architectures such as ViLT [20] and modular models like MISA have reduced computational complexity by simplifying the visual encoding process or decoupling modality-specific and shared information. While these approaches improve efficiency, they still struggle to fully resolve semantic misalignment in noisy or culturally nuanced content, where implicit sentiment cues span modalities in non-obvious ways [6].

2.2 Semantic Alignment and Modality Fusion Techniques

To address semantic misalignment between modalities, many recent studies adopt attention-based mechanisms for multimodal fusion. Cross-attention, co-attention, and gated fusion modules have been widely used to allow one modality to condition its representation on another [6, 19]. These methods enable finer-grained semantic alignment and perform well in tasks where visual and textual features are syntactically or semantically grounded. However, when applied independently of intra-modal enhancement or alignment regularization, they often fall short in handling abstract or metaphorical sentiment cues.

Contrastive learning has also gained popularity as a means of aligning modalities in a shared representation space. CLIP [9] is a prominent example, achieving competitive performance through large-scale pretraining with InfoNCE-based objectives. While contrastive objectives are effective for coarse-level alignment, they often lack the sensitivity to subtle cross-modal sentiment patterns and require large quantities of paired data to generalize effectively [21, 22].

Structured Self-Attention (SSA) [10] has been proposed to improve intra-modal representation quality by introducing orthogonality constraints that diversify attention heads. This helps preserve distinct semantic components within each modality and prevents redundancy. However, SSA is often treated as a standalone module and is rarely integrated with deeper cross-modal fusion architectures.

Despite these efforts, nevertheless, many existing methods still lack a unified mechanism that simultaneously addresses modality imbalance, semantic misalignment, and interpretability. As shown in Table 1, existing modality fusion strategies focus on specific goals but struggle to tackle all these challenges in an integrated manner.Models are often designed to optimize performance on benchmark datasets, but do not provide transparent reasoning or robust generalization across languages and domains. These limitations motivate the development of VBCSNet, which integrates modality-specific encoding, Structured Self-Attention, Cross-Attention, and a multi-objective training strategy into a single framework aimed at robust, interpretable, and semantically aligned multimodal sentiment classification.

2.3 Interpretability in Multimodal Sentiment Analysis

Interpretability has become a growing concern in MSA, particularly as models are increasingly applied in high-stakes decision-making contexts. Transformer-based attention mechanisms offer a degree of interpretability through attention weights, which highlight influential tokens or image regions. However, such visualizations alone are often insufficient for explaining cross-modal reasoning steps [6].

Several recent studies attempt to improve interpretability by introducing modular design or supervision signals that guide representation learning. While promising, these enhancements often come at the cost of increased model complexity or reduced flexibility. Moreover, few models are explicitly optimized for interpretability during training, and most lack mechanisms to disentangle modality-specific contributions to the final decision.

VBCSNet addresses these concerns by combining Structured Self-Attention with attention-based fusion and alignment-aware training objectives. The model produces well-aligned and interpretable representations while preserving class-level performance across languages. Its design reflects the need for models that not only perform well, but also offer structured, transparent reasoning about multimodal sentiment signals [3].

2.4 Positioning w.r.t. Recent Vision-Language Models

Recent vision-language models (VLMs) such as ViLT [20], ALBEF [27], METER [28], BLIP/BLIP-2 [29, 30], InstructBLIP [31], and LLaVA [32] have advanced generic vision-language understanding through large-scale pretraining and instruction tuning. However, their primary objectives—image captioning, visual question answering, and multimodal dialogue—differ fundamentally from **fine-grained multimodal sentiment classification**, which presents unique challenges including image-text polarity conflicts, multilingual sarcasm detection, and neutral sentiment ambiguity.

Task-Specific Design Requirements: Multimodal sentiment analysis requires specialized architectural components that general-purpose VLMs lack: (1) explicit mechanisms for resolving semantic conflicts between modalities when they convey opposing sentiments; (2) fine-grained attention patterns that can identify subtle emotional cues in both visual and textual content; and (3) robust handling of multilingual expressions where sentiment markers vary significantly across languages.

Our Complementary Approach: In contrast to generic VLMs, our method employs a *task-specialized fusion design* with three key innovations: (i) **Structured Self-Attention (SSA)** enhances intra-modal diversity and robustness by capturing fine-grained emotional patterns; (ii) **Cross-Modal Attention** explicitly models and resolves inter-modal sentiment conflicts; and (iii) **Multi-objective alignment and contrastive learning** regularizes cross-modal semantic consistency. This design yields interpretable and compute-efficient performance gains under moderate computational budgets.

Positioning and Scope: Our approach is designed to complement rather than replace large instruction-tuned VLMs. While VLMs excel at general vision-language tasks, our specialized architecture addresses the specific requirements of multilingual sentiment analysis with greater efficiency and interpretability. Future work will explore adapting large VLMs for this specialized task domain through compute-matched comparative studies.

2.5 Summary and Research Motivation

Despite recent advances, major challenges remain:

1. **Decoupled modeling and alignment**: Existing methods separate intra-modal modeling and inter-modal alignment, lacking end-to-end optimization.

2. **Heuristic fusion strategies**: Fusion often depends on empirical rules, lacking semantic constraints.

3. **Low-resource language support**: Most research focuses on English, with limited adaptation to Japanese, Chinese, etc.

4. **Trade-off between performance and interpretability**: Higher model complexity often reduces transparency, limiting applicability in high-stakes domains.

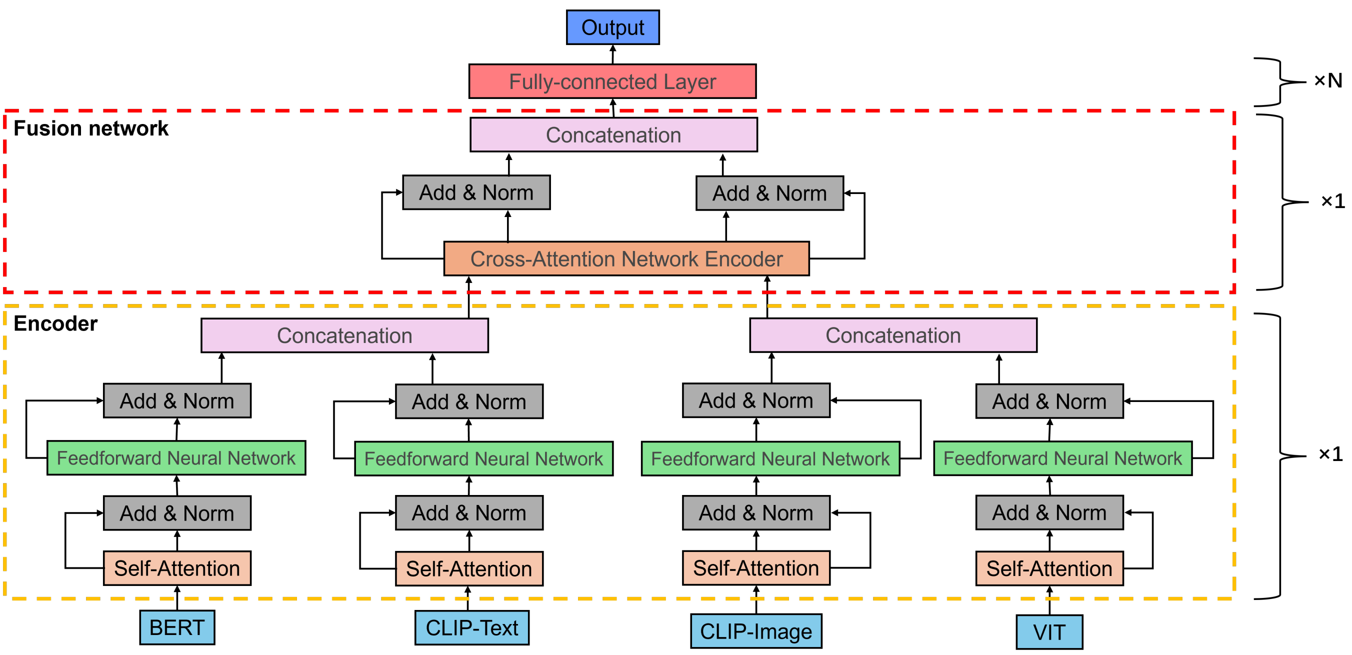


Figure 1. Overview of the proposed multimodal sentiment classification model. It integrates BERT, CLIP (Text/Image), and ViT encoders, and adopercentage points structured self-attention and cross-attention mechanisms followed by a fully-connected classifier.

In our earlier work, we previously proposed a CLIP-based attention fusion model, which was shown to improve results on English and Japanese tweet datasets. However, it relied on simple concatenation and lacked end-to-end training or explicit semantic alignment. This paper addresses these limitations by introducing a hierarchical attention design and contrastive training, aiming to enhance robustness and cross-lingual generalization.

3 VBCNet Model Design

We propose VBCSNet, a hybrid attention-based deep learning framework for multilingual multimodal sentiment classification. The model integrates complementary information from visual and textual modalities while addressing three core challenges in multimodal sentiment analysis: modality imbalance, semantic misalignment, and lack of interpretability. As illustrated in Figure 1, VBCSNet consists of four interconnected components, and its overall workflow is summarized as pseudocode in Algorithm 1, which outlines the step-by-step operations of the framework:

1. **Modality-specific encoder** that extracts features from ViT, BERT, and CLIP;
2. **Hierarchical attention module** that combines Structured Self-Attention (SSA) for intra-modal enhancement with Cross-Attention for inter-modal fusion;
3. **Multi-objective optimization strategy** that enforces alignment and semantic consistency through jointly trained classification, alignment, and contrastive objectives;
4. **Final sentiment classifier** based on a multilayer perceptron (MLP).

Algorithm 1: Multimodal Sentiment Classification

```
Input: ViT features $F_{\text{ViT}}$, CLIP-image features
       $F_{\text{CLIP-img}}$, RoBERTa features $F_{\text{BERT}}$,
       CLIP-text features $F_{\text{CLIP-text}}$
Output: Sentiment label $y \in \{\text{pos}, \text{neu}, \text{neg}\}$
/* Intra-modal feature refinement using
   Structured Self-Attention                              */
$F_{\text{img}} \leftarrow \text{SSA}(F_{\text{ViT}}, F_{\text{CLIP-img}})$
$F_{\text{text}} \leftarrow \text{SSA}(F_{\text{BERT}}, F_{\text{CLIP-text}})$
/* Cross-modal fusion using Cross-Attention
   */
$F_{\text{fused}} \leftarrow \text{CrossAttention}(F_{\text{img}}, F_{\text{text}})$
/* Compute multi-objective losses                         */
Compute $\mathcal{L}_{\text{align}}, \mathcal{L}_{\text{contrast}}, \mathcal{L}_{\text{cls}}$
/* Sentiment classification                               */
$y \leftarrow \text{Classifier}(F_{\text{fused}})$
return $y$
```

Each component is explained in detail in the following subsections with respect to the specific challenge it addresses.

3.1 Modality-Specific Feature Encoding

To address the challenge of modality imbalance, where strong textual encoders often dominate the optimization process and suppress contributions from the visual modality, we design a dual-channel feature encoding strategy. This approach preserves modality-specific representational richness and facilitates early-stage compatibility across modalities.

For the image modality, we adopt two complementary encoders. Vision Transformer (ViT-B/16) captures both local patterns and global semantic structures from the image. In parallel, CLIP-Vision provides representations that are semantically aligned with textual inputs, learned from large-scale image-text pairs. These two encoders offer distinct but synergistic perspectives on visual data.

For the text modality, we combine RoBERTa [21] and CLIP-Text. RoBERTa encodes contextual information within the sentence, while CLIP-Text generates vision-aligned textual embeddings that complement the image stream. This combination enables the model to integrate both domain-specific linguistic features and semantically grounded representations.

Let F_{ViT}, $F_{\text{CLIP-img}}$, F_{BERT}, and $F_{\text{CLIP-text}}$ denote the outputs of these four encoders. Each output is projected into a shared 256-dimensional semantic space using a two-layer MLP with ReLU activation and a dropout rate of 0.4. These projection layers are optimized jointly with the rest of the model. This encoding scheme ensures that both visual and textual features retain their distinct strengths while contributing equally to the multimodal representation. The resulting embeddings serve as the input to the attention-based fusion mechanism in the next stage.

3.2 Hierarchical Attention Mechanism

To address the challenge of semantic misalignment, which arises from representational gaps between modalities, we design a hierarchical attention mechanism that strengthens intra-modal representations before performing inter-modal semantic fusion. This two-stage approach ensures that each modality's information is internally coherent and semantically enriched prior to interaction.

The first stage applies Structured Self-Attention (SSA) to refine visual and textual embeddings within their respective modalities. Given feature pairs from ViT and CLIP-Vision for the image stream, and from RoBERTa and CLIP-Text for the text stream, SSA dynamically allocates attention to align and combine these feature types. Specifically, the SSA module computes a shared representation by adjusting the weight distribution across the sequence. To improve attention diversity and reduce redundancy, we impose an orthogonality constraint on the attention matrix, defined as

$$\mathcal{L}_{SSA} = \|AA^{\top} - I\|_F^2, \tag{1}$$

where A denotes the attention matrix, I is the identity matrix, and $\|\cdot\|_F$ represents the Frobenius norm. This constraint encourages each attention head to focus on different semantic aspects, which improves feature discriminability within each modality.

After SSA optimization, the second stage applies Cross-Attention to model semantic alignment between modalities. Visual features serve as queries, and textual features as keys and values. The attention is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V, \tag{2}$$

where Q, K, and V are the query, key, and value matrices, respectively, and d_k is the key dimensionality. This mechanism allows the image features to selectively attend to text segments that provide complementary or explanatory sentiment cues. It is particularly effective when textual and visual sentiment signals are subtle, contrasting, or embedded in metaphorical content.

The hierarchical design improves semantic alignment in three ways. First, SSA enhances internal consistency within each modality, making the cross-modal mapping more stable. Second, Cross-Attention explicitly captures directional semantic interactions, allowing the model to learn context-dependent fusion patterns. Third, the attention weights naturally support interpretability, as they highlight the contribution of specific image regions and text tokens. Together, these components form a semantically coherent and transparent representation for the final sentiment classification.

3.3 Multi-Objective Optimization Strategy

To jointly enhance semantic consistency and interpretability, we adopt a multi-objective optimization strategy that combines classification, alignment, and contrastive learning objectives into a unified training scheme. This approach explicitly guides the model to balance semantic clarity across modalities and to learn feature representations that are both discriminative and transparent.

The total objective is defined as

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_{img} \cdot \mathcal{L}_{img} + \lambda_{text} \cdot \mathcal{L}_{text} + \alpha \cdot \mathcal{L}_{contrast}, \tag{3}$$

where $\mathcal{L}_{\text{cls}}$ is the cross-entropy loss for sentiment classification, $\mathcal{L}_{\text{img}}$ and $\mathcal{L}_{\text{text}}$ are intra-modal alignment losses between ViT and CLIP-Vision, and between RoBERTa and CLIP-Text respectively, and $\mathcal{L}_{\text{contrast}}$ is a cross-modal contrastive loss based on InfoNCE [9, 21]. Following the uncertainty-based multi-task learning framework [5, 25, 26], we employ automatic weight learning rather than manual hyperparameter tuning. The trade-off parameters λ_{img}, λ_{text}, and α are learned dynamically during training based on the relative uncertainty (homoscedastic noise) of each task. This approach provides a stronger theoretical foundation than manual tuning, as the weights automatically adapt to the intrinsic difficulty and scale differences of the classification and alignment objectives.

Specifically, each weight is parameterized by a learnable σ_i and optimized jointly with the model:

$$\lambda_i = \frac{1}{2\sigma_i^2}, \text{ and add a regularization term } \sigma_i \text{ to the loss.} \tag{4}$$

This automatic balancing mechanism ensures optimal weight allocation throughout training without requiring extensive hyperparameter search. The learned weights typically converge to $\lambda_{\text{img}} \approx 0.31$, $\lambda_{\text{text}} \approx 0.28$, and $\alpha \approx 0.52$, demonstrating stable and consistent optimization across different random seeds.

The classification loss $\mathcal{L}_{cls}$ encourages the model to assign correct sentiment labels to fused multimodal features. The alignment losses $\mathcal{L}_{img}$ and $\mathcal{L}_{text}$ minimize the distance between features produced by structurally different encoders within the same modality, allowing ViT to align with CLIP-Vision and RoBERTa to align with CLIP-Text. This helps preserve modality-specific signals while guiding them into a shared semantic space, which mitigates semantic misalignment during fusion.

The contrastive loss $\mathcal{L}_{contrast}$ enhances the model's ability to distinguish semantically related and unrelated image-text pairs. It is formulated as

$$\mathcal{L}_{contrast} = -\log \frac{\exp(\text{sim}(x_i, y_i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(x_i, y_j)/\tau)}, \tag{5}$$

where $\text{sim}(\cdot)$ denotes cosine similarity and $\tau = 0.07$ is a temperature parameter. This objective encourages the model to map paired features closer together while pushing apart mismatched ones. As a result, the fused representation becomes semantically sharper and more robust across multilingual inputs.

Furthermore, this optimization strategy contributes to interpretability. By aligning modality-specific and cross-modal representations through explicit constraints, the model produces attention maps that are more focused and coherent. This allows the contributions of both image regions and text segments to be visualized and understood in terms of their influence on the final prediction. Overall, the multi-objective loss ensures that VBCSNet not only improves classification accuracy but also produces structured and interpretable reasoning across modalities.

3.4 Sentiment Classification Head

The fused multimodal representation produced by the Cross-Attention module is passed to a multilayer perceptron (MLP) for sentiment classification. The MLP consists of two fully connected layers with ReLU activation and dropout, followed by a Softmax output layer that generates the final sentiment probability distribution. This classification head transforms the integrated features into label predictions and maintains compatibility with multilingual and multimodal inputs.

3.5 Design rationale vs. generic VLMs

Instead of relying on monolithic instruction-tuned backbones, we retain lightweight encoders (ViT, BERT, CLIP features) and focus model capacity on fusion and objective design tailored for sentiment analysis.

Our architecture incorporates three sentiment-focused components: (i) **Structured Self-Attention (SSA)** discourages head redundancy and enhances class-relevant token/region focus; (ii) **Cross-Modal Attention** enables bidirectional text-image interaction to resolve sentiment conflicts; and (iii) **Multi-objective training** with classification, alignment, and contrastive losses provides complementary regularization.

This specialized design maintains competitive latency and VRAM usage while delivering superior interpretability through structured attention patterns. The multi-objective optimization ensures VBCSNet improves both classification accuracy and cross-modal reasoning interpretability.

4 Experiments and Results

We perform a series of experiments to evaluate the effectiveness of VBCSNet in multilingual and multimodal sentiment classification. The evaluation spans three representative datasets and assesses the model's ability to overcome modality imbalance, semantic misalignment, and lack of interpretability through quantitative comparison and ablation studies.

4.1 Experimental Setup

4.1.1 Datasets

MVSA-Multiple [13, 24] is a widely used benchmark English dataset for multimodal sentiment analysis. It consists of 18,148 image-comment pairs collected from social media platforms. Each sample is labeled with one of three sentiment categories: positive, neutral, or negative. The dataset provides reliable annotations and a balanced distribution across classes, making it suitable for evaluating fine-grained multimodal fusion strategies.

IJCAI2019-Twitter [22] is released as part of the IJCAI 2019 Multimodal Sentiment Analysis Challenge. It contains approximately 5,000 English tweets, each accompanied by a corresponding image. The textual content in this dataset reflects informal and compact language typical of social media discourse, which introduces challenges for robust sentiment classification, especially when sentiment cues are weak or ambiguous.

JP-Buzz [15, 16] is a Japanese multimodal dataset that we construct from public Twitter data. It consists of 39,345 image-text pairs, labeled as either Buzz (widely propagated) or Non-Buzz (ordinary content) based on the number of likes, where posts with more than 1,000 likes are labeled as Buzz and those with fewer than 1,000 likes are labeled as Non-Buzz. The dataset was collected between July and October 2022 and consists of 19,875 Non-Buzz and 19,241 Buzz samples. While the majority of the content is in Japanese, the dataset also includes a small proportion of posts in Korean, Chinese, and English, as well as emoticons commonly used on social media. This dataset allows us to evaluate the model's generalization capability in low-resource and linguistically diverse scenarios.

4.1.2 Training and Implementation Details

Table 2 summarizes the network architecture, hyperparameter settings, and training configurations used across all experiments. Specifically, to ensure comparability and reproducibility, we apply the same data preprocessing pipeline and training procedure for all datasets. All models are trained on a single NVIDIA A100 GPU with sufficient memory to support large-batch multimodal optimization. The implementation is based on PyTorch 2.1, with additional support from CUDA 11.7 and Python 3.10. Dataset splits follow a 7:2:1 ratio for training, validation, and testing.

Table 2. Network training parameters.

Parameters	Value
Batch size	16
Maximum text length	77
Image size	224×224
Training epochs	100
Optimizer	AdamW
Learning rate(MVSA)	5e-5
Learning rate(IJCAI2019, JP-Buzz)	5e-4
Weight decay	1×10^{-4}
EMA momentum	0.999
Loss weighting	$\lambda_{\text{img}} = \lambda_{\text{text}} \approx 0.3,\ \alpha \approx 0.5$

4.1.3 Evaluation Metrics

We evaluate model performance using three metrics: Accuracy, Macro-F1, and Confusion Matrix.

Accuracy measures the proportion of correctly predicted sentiment labels over all test samples. It reflects overall classification correctness and is computed as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

where TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively. While Accuracy offers a global view of prediction performance, it may not fully reflect model behavior in imbalanced sentiment settings.

Macro-F1 complements Accuracy by treating each sentiment class equally, averaging the F1 scores across all classes. It is calculated as:

$$\text{Macro-F1} = \frac{1}{C} \sum_{c=1}^{C} \frac{2 \cdot P_c \cdot R_c}{P_c + R_c}, \tag{7}$$

where C is the number of sentiment classes, and P_c and R_c denote the precision and recall for class c. Macro-F1 is particularly informative in the context of sentiment classification, as it penalizes poor performance on

Table 3. Baseline models for comparison.

Model Name	Description
CNN + TextCNN	Classic CNN models for independent feature extraction
ResNet + LSTM	Combines CNN for image and LSTM for text
ViT + BERT	Dual Transformer encoders without fusion
CLIP	Contrastive learning pretrained multimodal model
ViT + BERT + CLIP	Simple concatenation without alignment
ViT + BERT + CLIP + Att	Attention-based fusion model without multi-objective optimization
Ours (Full Model)	**SSA + Cross-Attention + Multi-objective Optimization**

underrepresented sentiment classes and highlights the model's ability to generalize across categories.

Confusion Matrix provides a visualization of prediction patterns and inter-class misclassifications. This matrix reveals which sentiment categories are frequently confused and whether the model exhibits bias toward dominant modalities or sentiment polarities.

Together, these metrics provide a comprehensive evaluation of both classification accuracy and semantic balance in multimodal sentiment understanding.

4.2 Baseline and Comparative Models

To evaluate the effectiveness of VBCSNet, we compare it with seven representative baseline models that reflect a range of strategies in multimodal sentiment classification. These include early fusion models, modality-specific architectures, and attention-based frameworks.

The CNN + TextCNN and ResNet + LSTM models represent conventional pipelines in which visual and textual features are extracted independently and fused without semantic alignment.

The ViT + BERT serves as a dual-encoder baseline that leverages Transformer-based encoders for each modality but does not include explicit fusion or alignment.

The CLIP model evaluates the generalization of contrastively pre-trained vision-language embeddings.

The ViT + BERT + CLIP baseline combines three encoders via simple concatenation without further integration or tuning.

An additional variant, ViT + BERT + CLIP + Attention [23], includes attention-based fusion but does not incorporate multi-objective optimization or intra-modal enhancement.

These baselines allow us to isolate and quantify the impact of Structured Self-Attention, Cross-Attention, and the multi-objective loss strategy used in VBCSNet.

All models are trained under the same experimental settings for consistent comparison. Their configurations are summarized in Table 3.

4.3 Overall Performance Comparison

We compare the performance of VBCSNet with baseline models on three multilingual multimodal sentiment datasets: IJCAI2019, JP-Buzz, and MVSA-Multiple. Table 4 reports Accuracy and Macro-F1 scores across all methods.

On **IJCAI2019**, VBCSNet achieves 0.713 Accuracy and 0.681 Macro-F1, surpassing all baseline models. Compared to ViT+BERT, which attains 0.652 Accuracy and 0.643 Macro-F1, VBCSNet demonstrates improved semantic alignment between modalities. The margin over the CLIP-only baseline further confirms that contrastive pretraining alone is insufficient without dedicated attention mechanisms and optimization strategies.In terms of Macro-Recall (balanced accuracy), VBCSNet attains 0.688, improving over ViT+BERT(0.668) and CLIP(0.609). This indicates higher sensitivity to all classes, not only overall accuracy gains.

On **MVSA-Multiple**, VBCSNet also achieves the best overall performance, reaching 0.788 Accuracy and 0.763 Macro-F1. VBCSNet achieves a Macro-Recall of 0.753, surpassing ViT+BERT(0.698) and slightly exceeding the attention baseline(0.751). The improved recall suggests fewer false negatives, especially on harder cases (e.g., neutral vs. polarized posts). These results validate the model's ability to generalize across balanced English-language datasets and highlight its effectiveness in capturing fine-grained multimodal sentiment patterns.

On **JP-Buzz**, VBCSNet obtains 0.795 Accuracy and 0.795 Macro-F1. Although one of the baselines exhibits a marginally higher Accuracy of 0.810,

Table 4. Performance comparison of different models on three multilingual multimodal sentiment datasets.(Precision, Recall, and F1 are macro-averaged over classes.)

Model	IJCAI2019				JP-Buzz				MVSA-Multiple			
	Precision	Acc	Recall	F1	Precision	Acc	Recall	F1	Precision	Acc	Recall	F1
CNN + TextCNN	0.532	0.565	0.434	0.461	0.767	0.756	0.757	0.754	0.720	0.727	0.697	0.717
ResNet + CNN	0.447	0.441	0.407	0.399	0.767	0.768	0.768	0.768	0.736	0.728	0.702	0.699
ViT + BERT	0.679	0.652	0.668	0.643	0.779	0.780	0.774	0.780	0.734	0.739	0.698	0.724
CLIP	0.660	0.636	0.609	0.607	0.762	0.763	0.742	0.761	0.748	0.757	0.744	0.758
ViT + BERT + CLIP	0.629	0.605	0.595	0.570	0.755	0.748	**0.800**	0.746	0.729	0.734	0.735	0.725
ViT + BERT + CLIP + attention	0.727	0.676	0.620	0.638	0.795	**0.810**	0.795	0.782	**0.756**	0.776	0.751	0.760
ViT + BERT + CLIP + SSA (Ours)	**0.744**	**0.713**	**0.688**	**0.681**	**0.803**	0.795	0.795	**0.795**	0.751	**0.788**	**0.753**	**0.763**

Table 5. Ablation on MVSA-Multiple. Results show differential impacts across components, with Cross-Attention removal causing the largest drop, particularly affecting neutral sentiment disambiguation.

Model variant	Acc	Macro-F1
Full (ViT+BERT+CLIP+SSA, Cross-Attn: Image→Text only)	**0.788**	**0.763**
w/o Cross-Attention	0.765	0.732
Cross-Attn: Text→Image only	0.773	0.757
w/o SSA	0.771	0.748
w/o Image Align Loss	0.780	0.758
w/o Text Align Loss	0.774	0.753
w/o Contrastive Loss	0.770	0.753

VBCSNet achieves superior Macro-F1, indicating more consistent performance across the Buzz and Non-Buzz classes. This suggests that the attention modules and loss functions introduced in VBCSNet are more robust in handling culturally specific sentiment cues and modality interactions in non-English data.VBCSNet reaches a Macro-Recall of 0.795, which is within 0.5 percentage points of the best baseline (0.800) while delivering the top Macro-F1. This implies more balanced sensitivity across Buzz/Non-Buzz, despite minor accuracy differences. While simpler models achieve higher accuracy on JP-Buzz, this is often due to overfitting on dataset-specific artifacts such as lexical or stylistic biases. In contrast, VBCSNet focuses on semantic alignment and cross-modal fusion, which improves generalization but slightly lowers accuracy. The consistent Macro-F1 gains demonstrate that VBCSNet better captures meaningful cross-modal sentiment patterns beyond surface-level cues, making its performance both robust and interpretable.

Consequently, these results collectively demonstrate that VBCSNet addresses modality imbalance by preserving and integrating visual and textual signals more effectively than standard fusion or alignment-agnostic models. Furthermore, the consistent gains in Macro-F1 across datasets indicate that the model maintains semantic consistency across all sentiment categories, supporting its cross-lingual applicability and interpretability.

4.4 Ablation Study

To assess the contribution of each component in VBCSNet, we conduct comprehensive ablations on MVSA-Multiple. By default, VBCSNet uses a single-direction Cross-Attention from Image→Text (Q=Image, K/V=Text). We then ablate (i) removing Cross-Attention and (ii) swapping the direction to Text→Image only (Q=Text, K/V=Image). Table 5 reports Accuracy and Macro-F1, with statistical significance confirmed via McNemar's test ($p < 0.01$ for all comparisons).

Structured Self-Attention (SSA). Removing SSA reduces Macro-F1 from 0.763 to 0.748 , indicating that intra-modal refinement preserves semantic distinctiveness before fusion and mitigates modality imbalance.

Cross-Attention (directionality). Disabling Cross-Attention yields 0.732 Macro-F1, confirming that inter-modal interaction is necessary to resolve semantic misalignment. The substantial drop disproportionately affects *Neutral* classification (F1: 0.689→0.645), demonstrating that cross-modal reasoning is essential for disambiguating borderline cases. Swapping the direction to Text→Image only attains 0.757 Macro-F1, indicating that directionality matters significantly. On MVSA-Multiple—where textual polarity is often dominant—the Image→Text flow better aligns modalities by allowing visual features to provide contextual grounding for textual

Table 6. Efficiency comparison on MVSA-Multiple.

#Method	Params (M)	Train/epoch	Latency (ms)	Peak VRAM (GB)	Acc
CNN+TextCNN	74	3m58s	2.7	3.41	0.727
ResNet+CNN	103	4m08s	2.8	3.55	0.728
ViT+BERT	231	8m24s	13.4	11.92	0.739
CLIP	149	7m44s	10.5	8.40	0.757
ViT+BERT+CLIP	397	10m30s	14.1	14.38	0.734
ViT+BERT+CLIP+Attention	405	11m12s	17.0	14.65	0.776
ViT+BERT+CLIP+SSA (Ours)	**416**	**11m48s**	**18.3**	**15.02**	**0.788**

sentiment expressions.

Alignment objectives. Excluding the image/text alignment losses leads to modest drops (0.758/0.753 Macro-F1), illustrating their role in stabilizing fusion and enforcing semantic consistency across differently pre-trained encoders.Without alignment/contrastive terms, predictions skew toward Neutral on borderline samples, indicating weaker separation of ambiguous cases despite similar accuracy.

Contrastive objective. Removing the contrastive loss reduces Macro-F1 to 0.753. Although Accuracy remains relatively stable, class-level performance becomes less balanced, suggesting weaker separation between sentiment categories.

Per-Category Impact Analysis. To understand how different components affect sentiment classification across categories, we analyze the confusion patterns when components are removed. Removing Cross-Attention creates disproportionate confusion between Neutral and polarized sentiment classes, with the model tending to over-predict Neutral sentiment when visual and textual cues are inconsistent. This indicates that cross-modal interaction is particularly crucial for disambiguating borderline cases.

Component-Specific Effects. The SSA removal particularly affects samples requiring fine-grained attention to subtle emotional indicators, as evidenced by the concentrated performance drop . The differential impact of alignment losses—with text alignment showing a larger impact than image alignment—suggests that semantic consistency in the textual modality is more critical for multilingual sentiment understanding.

Directional Attention Analysis. The superior performance of Image→Text attention (0.763 vs 0.757) indicates that visual features provide more robust contextual anchoring for textual sentiment interpretation, particularly valuable in multilingual settings where textual sentiment expressions exhibit greater variability across languages.

Overall, these results verify that SSA, directional Cross-Attention, and the alignment/contrastive objectives jointly address modality imbalance and semantic misalignment while improving interpretability.

4.5 Effect of MLP Depth

We analyze the impact of the fully connected layer depth in the final classification stage by varying the number of MLP layers from one to four. Table 7 reports Accuracy and Macro-F1 scores on MVSA-Multiple.

Table 7. Effect of Fully-Connected Layer Depth on model performance (MVSA-Multiple).

#Layers	Accuracy (ACC)	F1 Score
1	0.776	0.755
2	**0.788**	**0.763**
3	0.784	0.761
4	0.779	0.757

On MVSA-Multiple, the best performance is achieved when the number of MLP layers is two, with 0.788 Accuracy and 0.763 Macro-F1. Increasing the depth to three or four results in a gradual decline in both metrics. This suggests that overly deep classification heads may introduce redundancy or lead to overfitting on relatively small or balanced datasets.

These observations support the adoption of a two-layer MLP as the default configuration in VBCSNet. Taken together, this design offers a good trade-off between expressiveness and stability, and maintains competitive performance across both high-resource and linguistically diverse sentiment datasets.

4.6 Efficiency and Resource Usage

This section evaluates the training and inference costs of the proposed model and the baselines. We use automatic mixed precision (AMP), an image size of 224×224, a maximum text length of 77 tokens, and a batch size of 16 during training. For latency measurement, we use a batch size of 1, perform 20 warm-up runs, and then average over 200 forward passes. All experiments are conducted with PyTorch 2.x on the specified GPU model and CUDA version.

The results are summarized in Table 6 on the MVSA-Multiple dataset and highlight the accuracy–efficiency trade-off of VBCSNet versus the baselines. VBCSNet achieves a substantial accuracy gain over ViT+BERT, improving from 0.739 to 0.788. This gain comes with increased cost: +36.6% higher latency (13.4 ms → 18.3 ms), +40.5% longer train time per epoch (8m24s → 11m48s), and +26.0% higher peak VRAM (11.92 GB → 15.02 GB). However, compared with the ViT+BERT+CLIP baseline, adding our new attention stack and SSA increases the parameter count by only +4.8% (397 M → 416 M) with a modest impact on latency.

4.7 Limitations

While VBCSNet demonstrates superior performance compared to baselines in multilingual multimodal sentiment classification, several broader limitations of current multimodal learning techniques present challenges for further advancement.

First, current approaches rely heavily on pre-trained encoders that may not generalize well across domains or languages with limited annotated data. This constrains the ability to extend sentiment analysis to underrepresented languages or emerging online communities.

Second, existing alignment mechanisms often assume a coarse semantic correspondence between modalities. This limits the capacity to reason over fine-grained or abstract cross-modal relationships, such as irony or metaphor, especially when these cues are context-dependent or culturally specific. Moreover, interpretability remains tied to attention visualization, which, while useful, does not yet provide full causal or logical explanations behind predictions.

Third, our study focuses on compute-efficient, task-specialized fusion under moderate computational budgets. Comprehensive benchmarks against large instruction-tuned VLMs (e.g., LLaVA, InstructBLIP) require substantial resources and specialized adaptation protocols beyond our current scope. Extending VBCSNet to larger backbones and conducting compute-matched VLM comparisons represents an important direction for future work.

Additionally, the JP-Buzz dataset used in this study cannot be publicly released due to privacy and platform policy restrictions. This limits the direct reproducibility of experiments on this dataset. To mitigate this issue, we provide detailed descriptions of the dataset construction process, annotation criteria, and data distribution analysis in this paper, enabling other researchers to indirectly validate our methodology and build similar datasets for future research.

In summary, these limitations highlight the need for more flexible, generalizable, and explainable architectures that can handle more complex forms of multimodal interaction and reasoning beyond static alignment and classification.

5 Conclusion

In this paper, we have proposed VBCSNet, a hybrid attention-based framework for multilingual multimodal sentiment classification. The model integrates ViT, BERT, and CLIP with Structured Self-Attention and Cross-Attention, supported by a multi-objective optimization strategy. Through comprehensive experiments on English and Japanese datasets, we have demonstrated that VBCSNet effectively addresses the key challenges of modality imbalance, semantic misalignment, and lack of interpretability. Quantitative evaluations have confirmed consistent gains in Accuracy and Macro-F1 across languages and sentiment categories, while comprehensive ablation studies have validated the individual contributions of each component and revealed their differential impacts on sentiment understanding. Our analysis shows that cross-modal interaction is particularly crucial for disambiguating neutral sentiment cases, while structured self-attention enhances fine-grained emotional pattern recognition across modalities.

In future work, we aim to extend VBCSNet to low-resource languages and to sentiment scenarios involving abstract, ambiguous, or implicit emotional cues. Furthermore, to address limitations in current alignment and interpretability techniques, future work will explore more flexible architectures and structured reasoning mechanisms that enable more robust and transparent multimodal understanding across culturally diverse contexts [34, 35].

Data Availability Statement

Data will be made available on request.

Funding

This work was supported by the JSPS KAKENHI under Grant JP20K12027 and by JKA and its promotion funds from KEIRIN RACE.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion, 37*, 98–125. [CrossRef]

[2] Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. (2017). Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*.

[3] Sun, S., An, W., Tian, F., Nan, F., Liu, Q., Liu, J., Shah, N., & Chen, P. (2024). A review of multimodal explainable artificial intelligence: Past, present and future. *arXiv preprint arXiv:2412.14056*.

[4] Kaur, R., & Kautish, S. (2019). Multimodal sentiment analysis: A survey and comparison. *International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), 10*(2), 38–58. [CrossRef]

[5] Tsai, Y. H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L. P., & Salakhutdinov, R. (2019, July). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting* (Vol. 2019, p. 6558). [CrossRef]

[6] Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

[7] Li, J., Wang, C., Luo, Z., Wu, Y., & Jiang, X. (2024). Modality-dependent sentiments exploring for multi-modal sentiment classification. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7930–7934). [CrossRef]

[8] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). [CrossRef]

[9] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.

[10] Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., & Bengio, Y. (2017). A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

[11] Oord, A. v. d., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[12] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In *International conference on machine learning* (pp. 1597-1607). PmLR.

[13] Niu, T., Zhu, S., Pang, L., & El Saddik, A. (2016, January). Sentiment analysis on multi-view social data. In *International conference on multimedia modeling (pp. 15-27)*. Cham: Springer International Publishing. [CrossRef]

[14] Yu, J., & Jiang, J. (2019). Adapting BERT for target-oriented multimodal sentiment classification. IJCAI. [CrossRef]

[15] Matsumoto, K., Amitani, R., Yoshida, M., & Kita, K. (2022). Trend prediction based on multi-modal affective analysis from social networking posts. *Electronics, 11*(21), 3431. [CrossRef]

[16] Amitani, R., Matsumoto, K., Yoshida, M., & Kita, K. (2022). Affective Analysis and Visualization from Posted Text, Replies, and Images for Analysis of Buzz Factors. In *Fuzzy Systems and Data Mining VIII* (pp. 191-203). IOS Press. [CrossRef]

[17] Zhang, J., & Chen, Z. (2024). Exploring human resource management digital transformation in the digital age. *Journal of the Knowledge Economy, 15*(1), 1482–1498. [CrossRef]

[18] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). [CrossRef]

[19] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems, 32*.

[20] Kim, W., Son, B., & Kim, I. (2021, July). Vilt: Vision-and-language transformer without convolution or region supervision. In International conference on machine learning (pp. 5583-5594). PMLR.

[21] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[22] Kumar, A., & Garg, G. (2019). Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications, 78*(17), 24103-24119. [CrossRef]

[23] Liu, Y., & Matsumoto, K. (2024, December). Enhancing Multimodal Tweet Analysis Accuracy

through Integration of CLIP Model and Multi-layer Attention Mechanism. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval* (pp. 310-316). [CrossRef]

[24] Ba, J. L., & Caruana, R. (2013). Do deep nets really need to be deep? *arXiv preprint arXiv:1312.6184*.

[25] Cipolla, R., Gal, Y., & Kendall, A. (2018, June). Multi-task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7482-7491). IEEE Computer Society. [CrossRef]

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

[27] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., & Hoi, S. C. H. (2021). Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems, 34*, 9694–9705.

[28] Dou, Z.-Y., Xu, Y., Gan, Z., Wang, J., Wang, S., Wang, L., Zhu, C., Zhang, P., Yuan, L., Peng, N., & others. (2022). An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 18166–18176). [CrossRef]

[29] Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888-12900). PMLR.

[30] Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.

[31] Dai, W., Li, J., Li, D., Tiong, A., Zhao, J., Wang, W., Li, B., Fung, P. N., & Hoi, S. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems, 36*, 49250–49267.

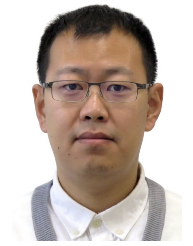
[32] Islam, A., Biswas, M. R., Zaghouani, W., Belhaouari, S. B., & Shah, Z. (2023). Pushing boundaries: Exploring zero shot object classification with large multimodal models. In *2023 Tenth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 1–5). [CrossRef]

[33] Zhang, X., Guo, J., Zhao, S., Fu, M., Duan, L., Wang, G. H., Chen, Q. G., Xu, Z., Luo, W., & Zhang, K. (2025). Unified multimodal understanding and generation models: Advances, challenges, and opportunities. *arXiv preprint arXiv:2505.02567*.

[34] Cambria, E., Poria, S., Gelbukh, A., & Thelwall, M. (2017). Sentiment analysis is a big suitcase. *IEEE Intelligent Systems, 32*(6), 74–80. [CrossRef]

[35] Gonzalez-Varona, J. M., López-Paredes, A., Poza, D., & Acebes, F. (2024). Building and development of an organizational competence for digital transformation in SMEs. *arXiv preprint arXiv:2406.01615*.

Yupu Liu is a first-year Ph.D. student at the Department of Computer Science and the Graduate School of Advanced Technology and Science, Tokushima University, Japan.His research interests include affective computing, multimodal sentiment analysis, and multimodal depression detection, with a focus on applying these technologies to mental health support and healthcare applications. (Email: c612545002@tokushima-u.ac.jp)



Xin Kang received the Ph.D. degree in Information Science and Intelligent Systems from Tokushima University, Japan, in 2013. He is currently an Associate Professor with the Department of Computer Science and the Graduate School of Advanced Technology and Science, Tokushima University, Japan. His research interests include affective computing, trustworthy artificial intelligence, neuro-symbolic reasoning, and multimodal knowledge modeling, with applications in mental health support and financial text understanding. He is a member of IEEE, ACM, and CAAI. (Email: kang-xin@is.tokushima-u.ac.jp)

Kazuyuki Matsumoto completed the Doctoral Program in Engineering at Tokushima University Graduate School of Engineering in 2008. Ph.D. (Engineering), and currently serves as Associate Professor at the Graduate School of Social and Industrial Science and Engineering, Tokushima University since April 2020. Research interests include emotion computing from language information for application to care welfare robots, natural language processing, and recently, slang and reputation analysis for crisis management on social media. Member of the Japanese Society for Natural Language Processing, the Japanese Society for Kansei Engineering, and the Association for the Advancement of Affective Computing (AAAC).(Email: matumoto@is.tokushima-u.ac.jp)

Jiazheng Zhou is a second-year Ph.D. candidate at the Department of Computer Science and the Graduate School of Advanced Technology and Science, Tokushima University, Japan.His research intersts include affective computing, multimodal sentiment analysis,and multimodal depression detection.(Email: c612345005@tokushima-u.ac.jp)